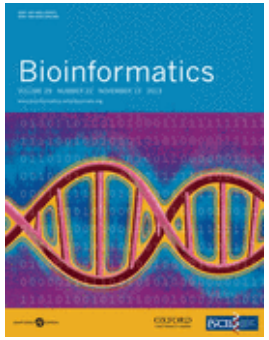
MolTrans: Molecular Interaction Transformer for Drug Target Interaction Prediction



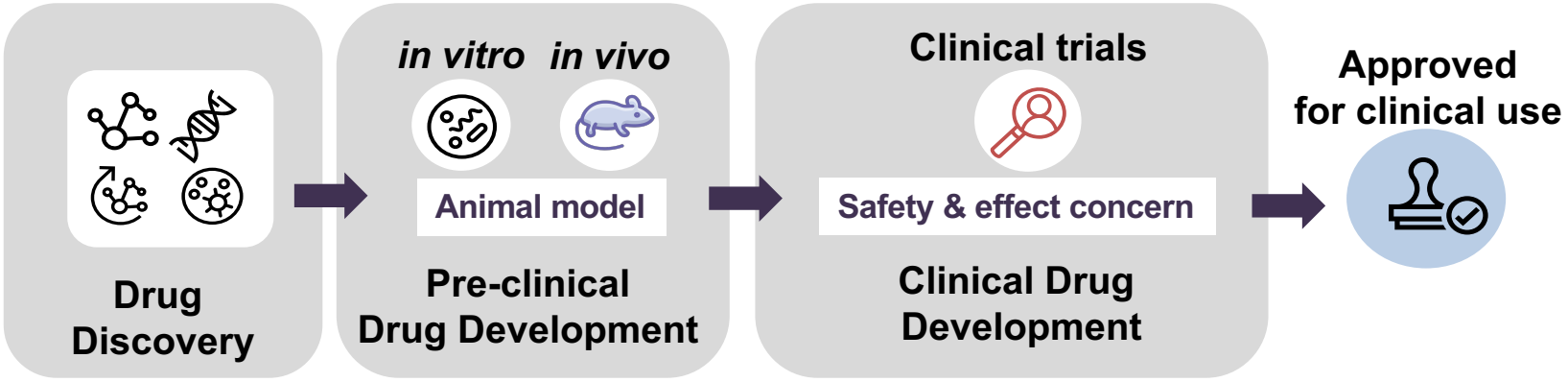
Kexin Huang
Harvard University
kexinhuang@hsph.harvard.edu

Cao Xiao
IQVIA
cao.xiao@iqvia.com

Lucas Glass
IQVIA
lucas.glass@iqvia.com

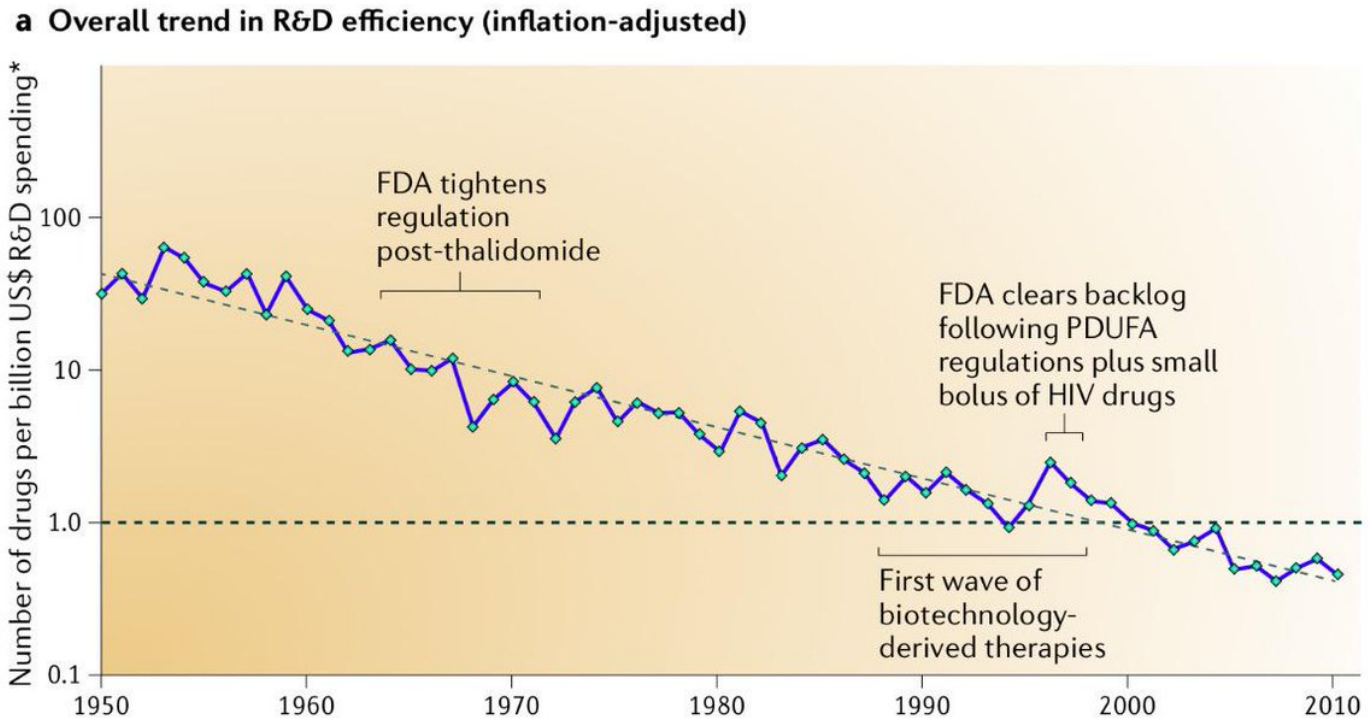
Jimeng Sun
UIUC
jimeng@illinois.edu

Traditional Drug Discovery & Development Process



	Drug discovery	Pre-clinical	Phase 1	Phase 2	Phase 3
Time spent	4-5 years	1-2 years	1-2 years	1-2 years	2-3 years
$ spent	$550M	$125M	$225M	$250M	$250M
Output	5,000 - 10,000 compounds	10-20 candidates	5-10 candidates	2-5 candidates	1-2 candidates

Eroom's Law


a Overall trend in R&D efficiency (inflation-adjusted)
Number of drugs per billion US$ R&D spending*
100
10
1.0
0.1
1950
1960
1970
1980
1990
2000
2010
FDA tightens regulation post-thalidomide
FDA clears backlog following PDUFA regulations plus small bolus of HIV drugs
First wave of biotechnology-derived therapies

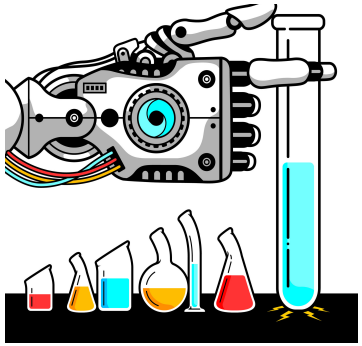
ML Accelerates Drug Discovery

MERCK
Be well

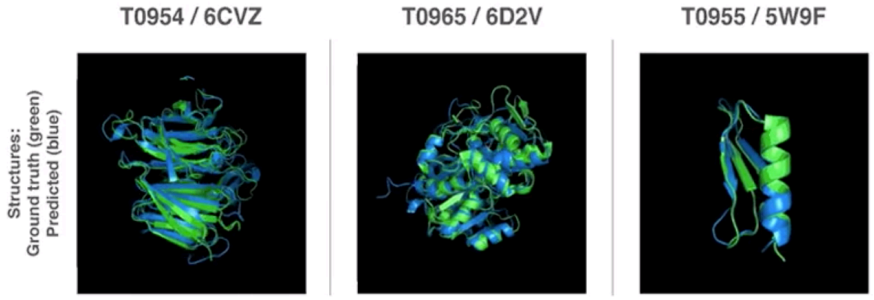
Merck Molecular Activity Challenge

Help develop safe and effective medicines by predicting molecular activity.

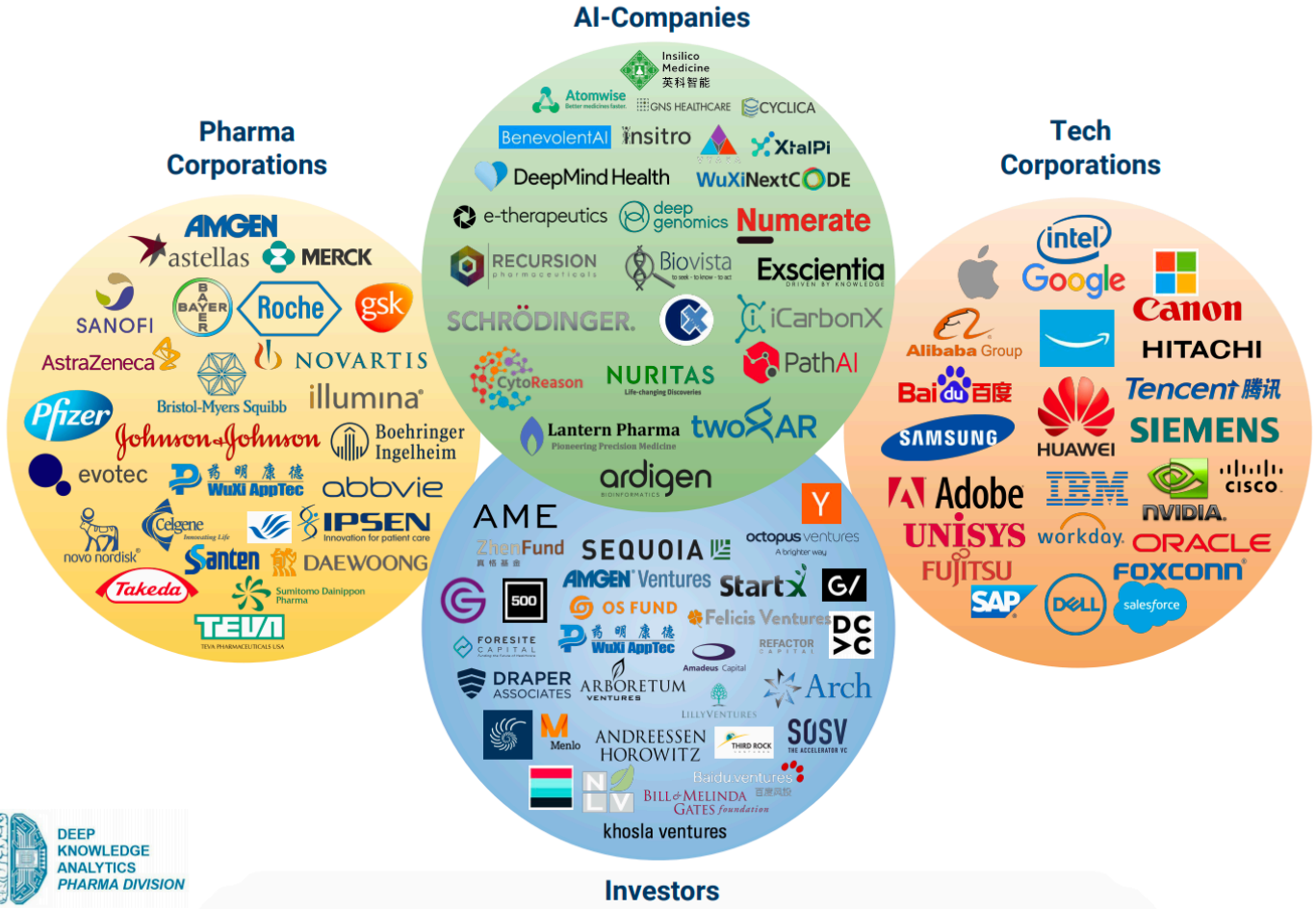
$40,000 · 236 teams · 7 years ago



DeepMind's AI will accelerate drug discovery by predicting how proteins fold



Leading Companies - Advanced AI in Healthcare and Drug Discovery / 2019 Q1



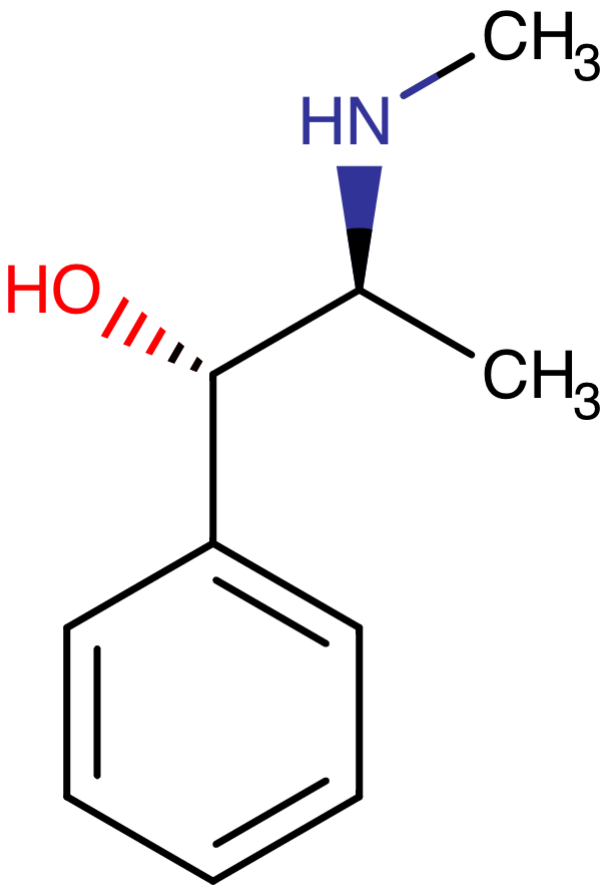
DEEP KNOWLEDGE ANALYTICS *PHARMA DIVISION*

What's a compound?

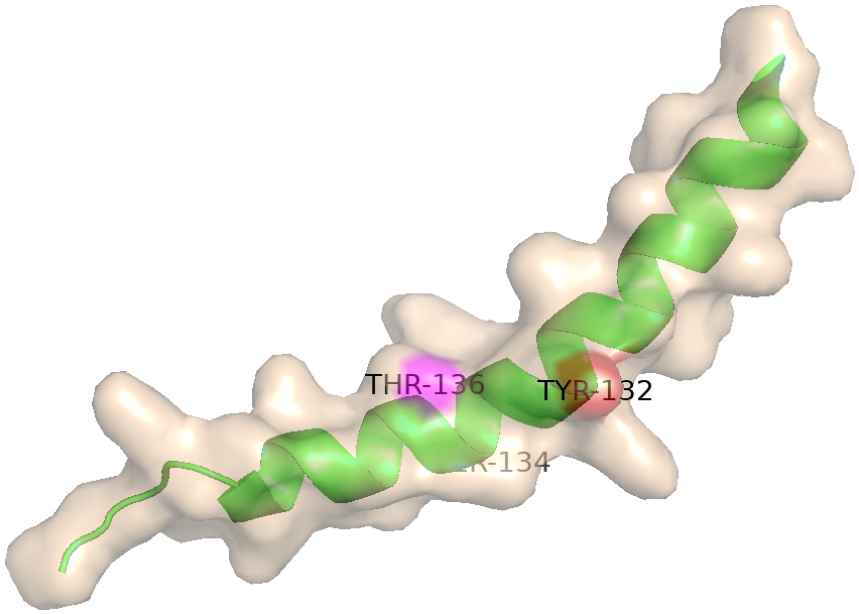
Pseudoephedrine

SMILES
Simplified molecular-input line-entry system

CC(C(C1=CC=CC=C1)O)NC



What's a protein?

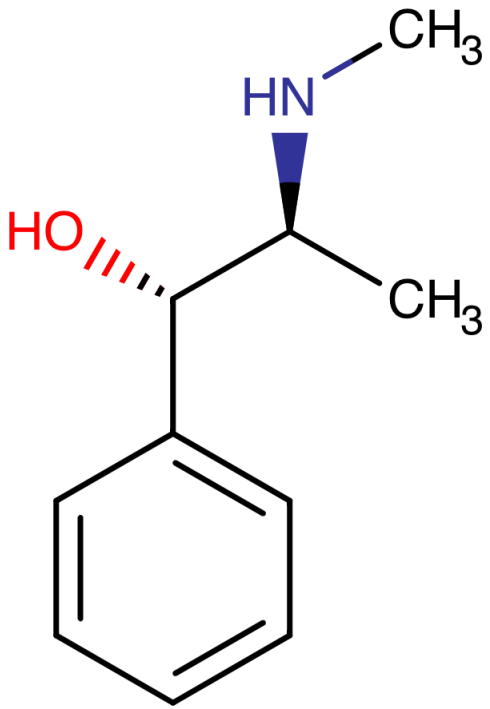


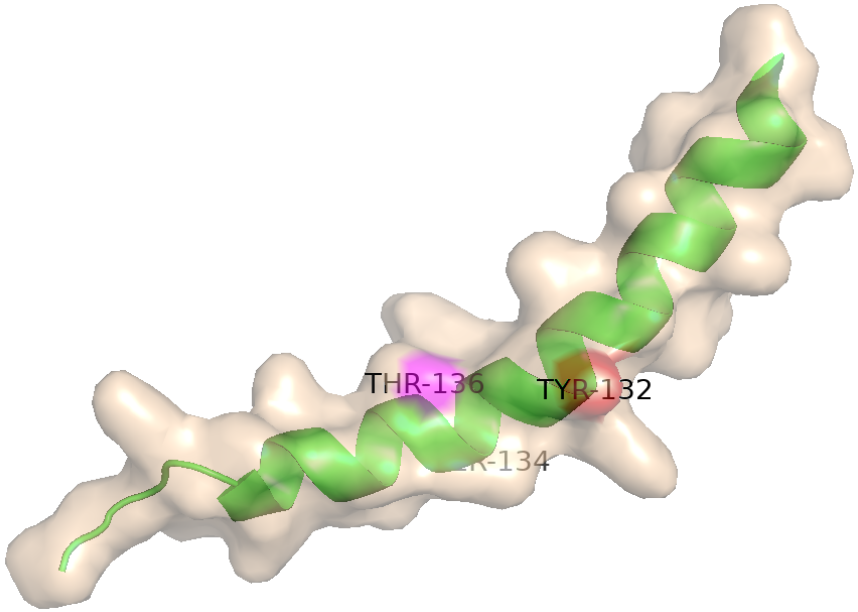
Alpha-2A receptor

Amino Acid Sequence

MFRQEQPLAEGSFAPMGSLQPDAGNASWNGTEA
PGGGARATPYSLQVTLTLVCLAGLLMLLTVFGNVL
VIIAVFTSRALKAPQNLFLVSLASADILVATLVIPFSL
ANEVMGYWYFGKAWCEIYLALDVLFCTSSIVHLCA
ISLDRYWSITQAIEYNLKRTPRRIKAIIITVWVISAVIS
FPPLISIEKKGGGGGPQPAEPRCEINDQKWYVISSC
IGSFFAPCLIMILVYVRIYQIAKRRTRVPPSRRGPDA
VAAPPGGTERRPNGLGPERSAGPGGAEAEPLPTQ
LNGAPGEPAPAGPRDTDALDLEESSSSDHAERPP
GPRRPERGPRGKGKARASQVKPGDSLPRRGPGA
TGIGTPAAGPGEERVGAAKASRWRGRQNREKRFT
FVLAVVIGVFVVCWFPFFFTYTLTAVGCSVPRTLFK
FFFWFGYCNSSLNPVIYTIFNHDFRRAFKKILCRGD
RKRIV

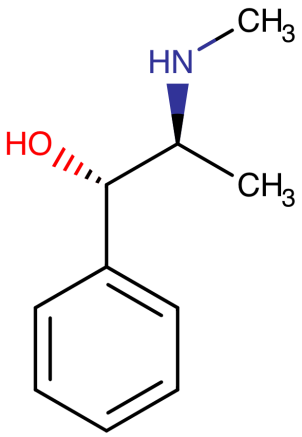
What's DTI?

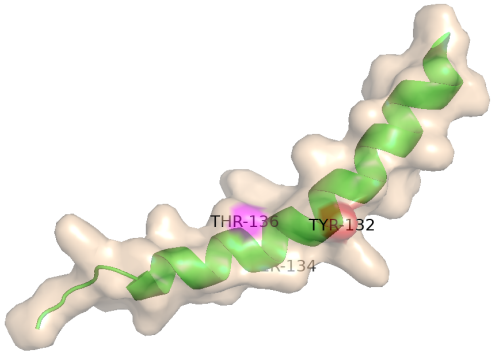




Q: Will they bind?

What's DTI?

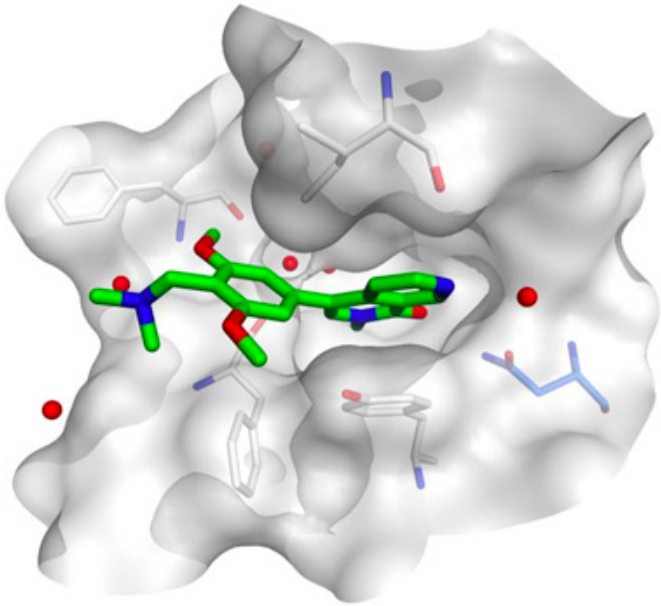




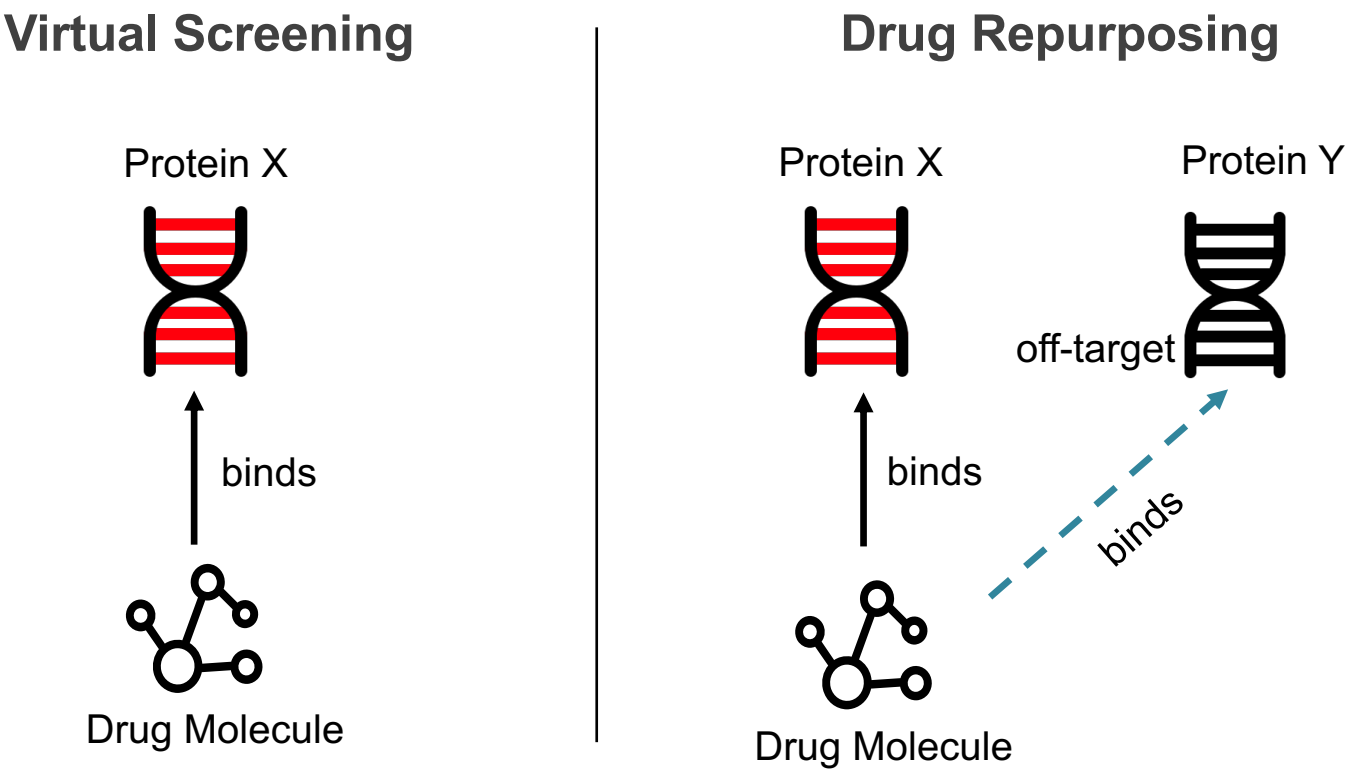
A machine learning question:

Given drug SMILES, target amino acid sequence, what is their predicted binding affinity score?

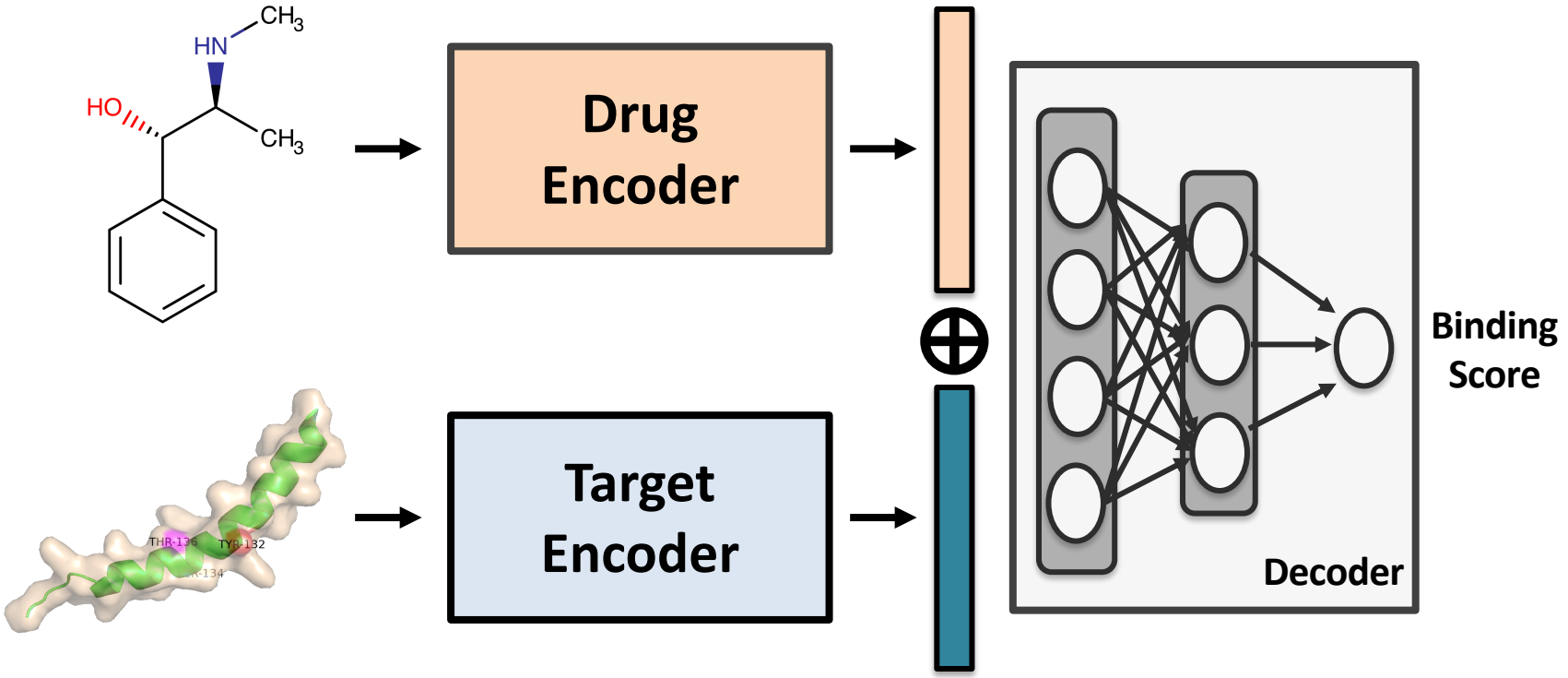
DTI Mechanism



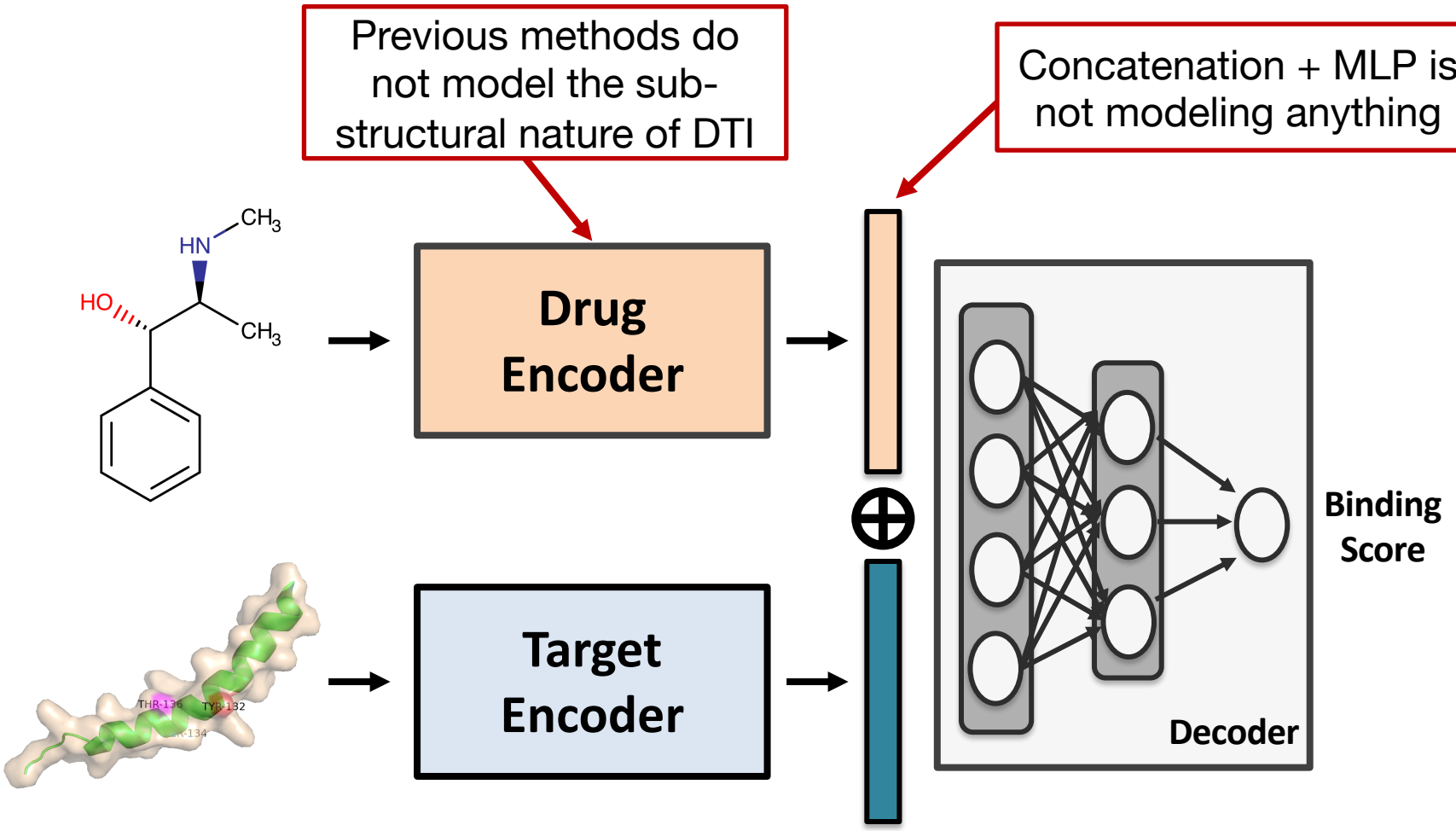
DTI Application



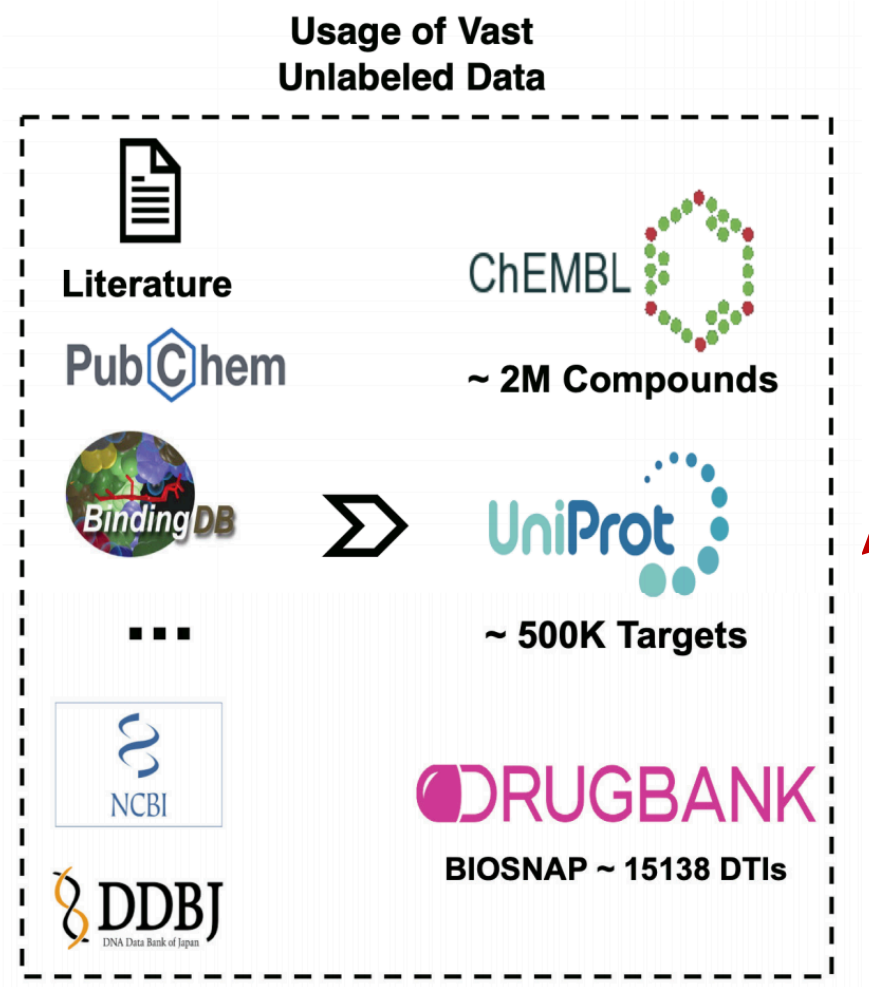
A Typical Deep Learning DTI Framework



Challenge 1: Inadequate modeling of interaction mechanism

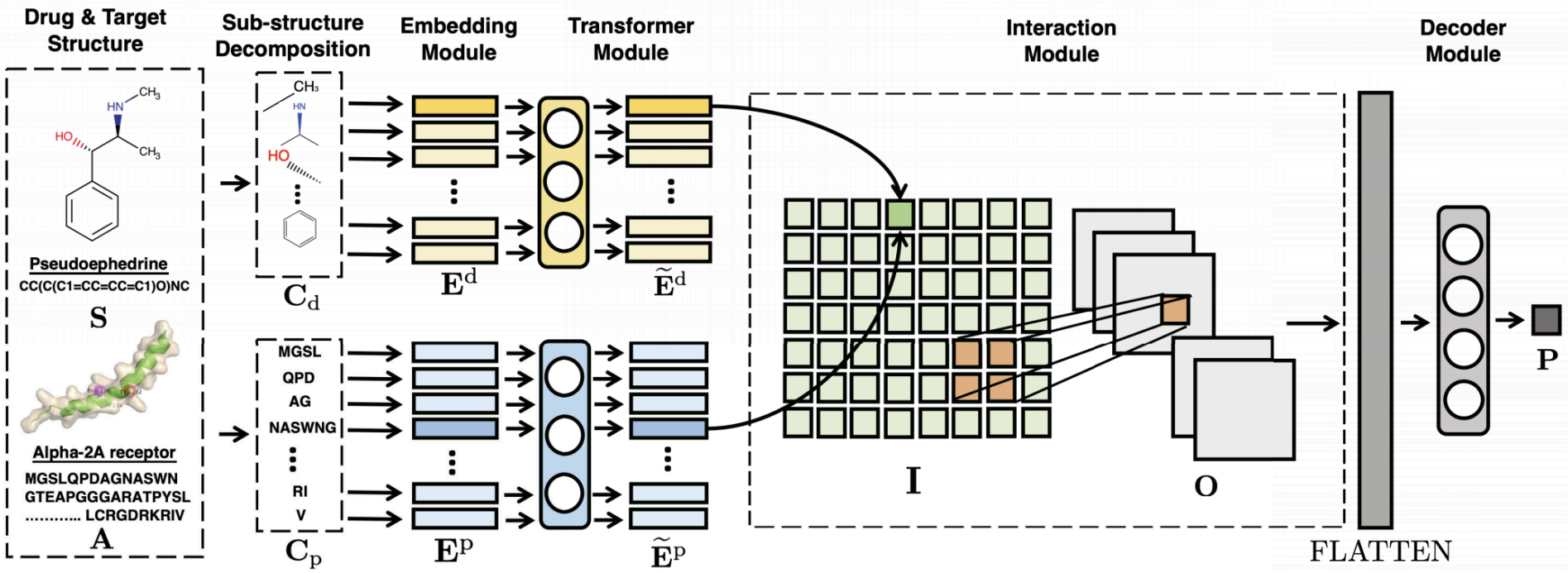


Challenge 2: Restricted to limited labeled data



The model architectures in previous works (DeepDTA, DeepDTI, GraphDTA) are not designed to enable the integration of massive dataset.

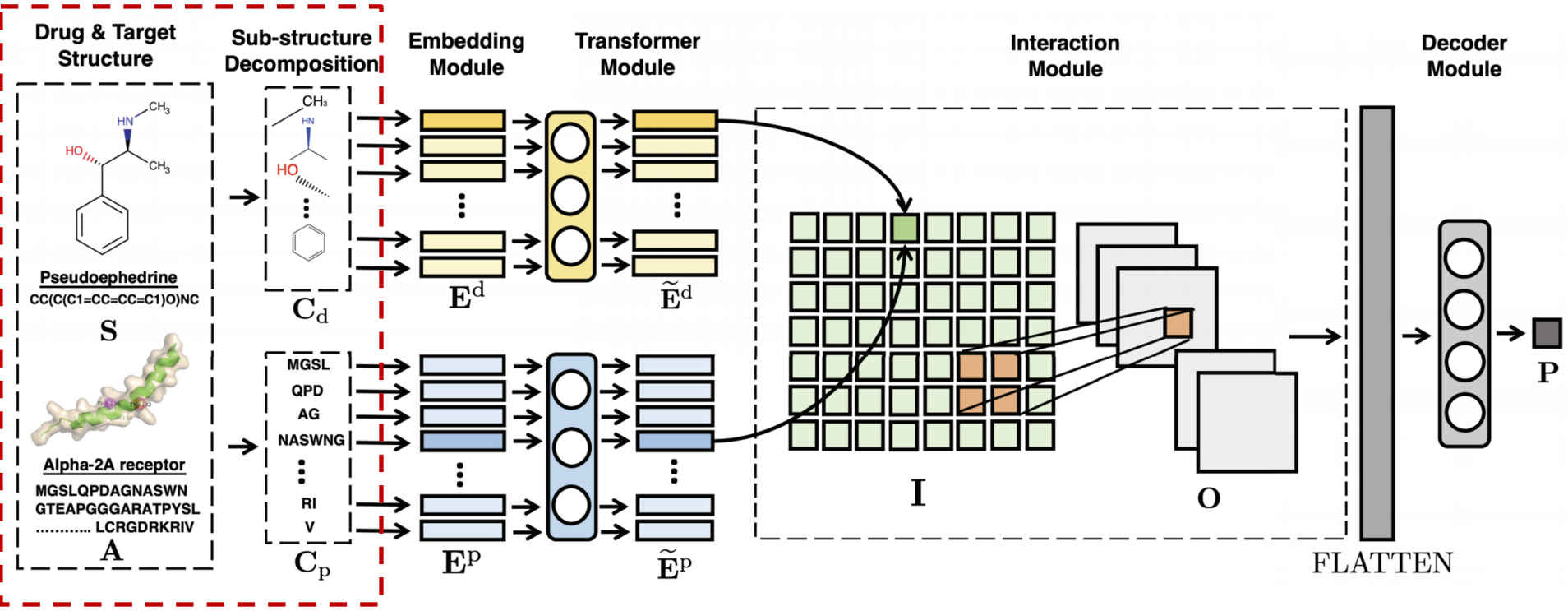
MolTrans



A new data-driven substructural fingerprint that uses vast data sources.

A new model block motivated by and simulated DTI mechanism.

Substructure Decomposition



In order to model the DTI substructural interaction, we have to identify substructure in the input sequence.

Substructure Decomposition

A New Data-Driven Algorithm!

Algorithm 1: Frequent Consecutive Sub-sequence Mining

Input: $\mathbb{V}$ as the set of all initial amino acids/SMILES tokens; $\mathbb{W}$ as the set of tokenized proteins/drugs; θ as the specified frequency threshold; ℓ as the maximum size of $\mathbb{V}$.

Output: $\mathbb{W}$, the updated tokenized proteins/drugs; $\mathbb{V}$, the updated token vocabulary set.

for $t = 1 \ldots \ell$ **do**
- (A, B), freq $\leftarrow$ scan $\mathbb{W}$
- // (A, B) is the frequentest consecutive tokens.
- **if** freq $< \theta$ **then**
  - break // (A, B) 's frequency lower than threshold
- $\mathbb{W} \leftarrow$ find(A, B) $\in \mathbb{W}$, replace with (AB)
- // update $\mathbb{W}$ with the combined token (AB)
- $\mathbb{V} \leftarrow \mathbb{V} \cup$ (AB) // add (AB) to the token vocabulary set $\mathbb{V}$

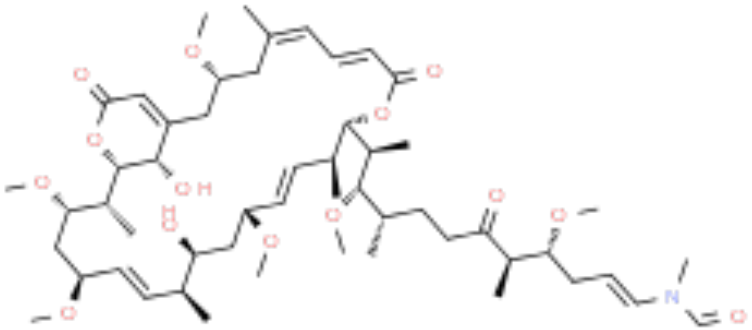
Partition each SMILES/Amino Acid Sequence into reasonable-sized high-quality substructures.

Mine through vast ChEMBL and UniProt Database!

~2M SMILES Strings

~500K target sequence

Substructure Decomposition Example



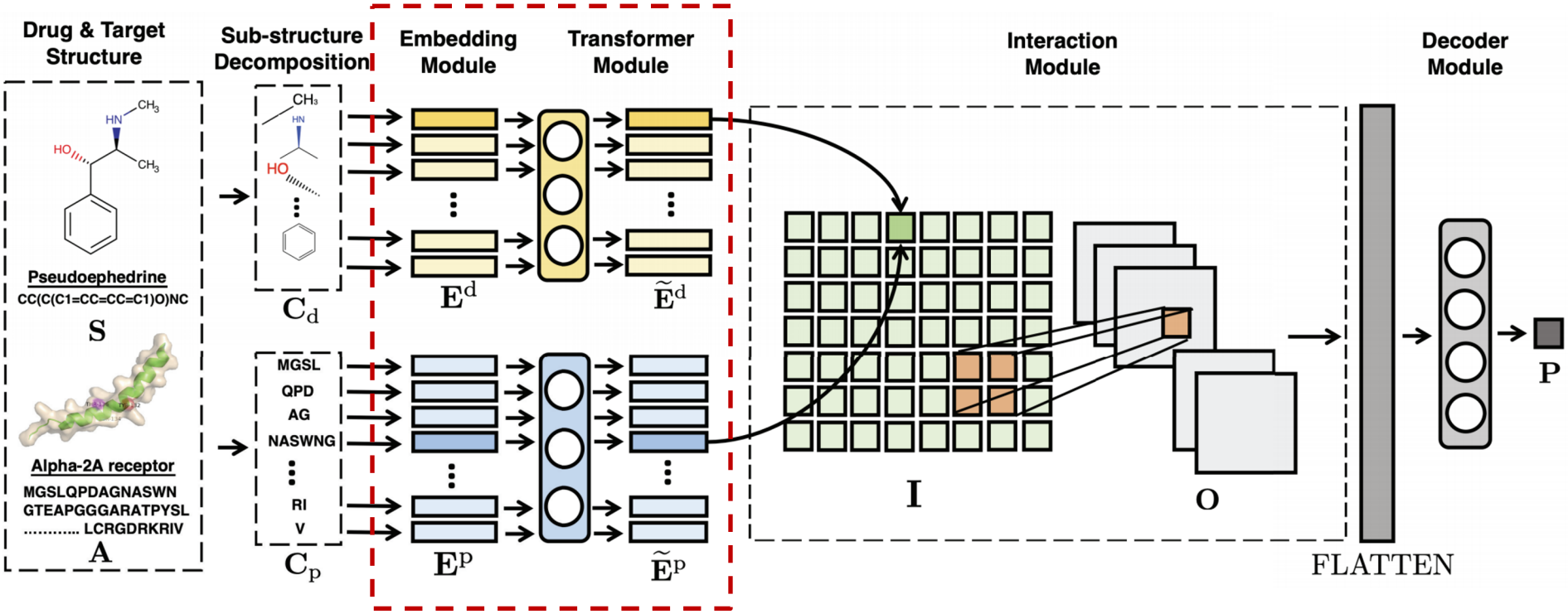
Formamide

CO[C@@H]1[C@H](O)[C@@H](C)O[C@@H](OC[C@@H]2[C@@H](C)OC(=O)\\C=C\\[C@H](C)[C@H](CC[C@@H](C)C(=O)\\C=C\\[C@H]3O[C@@H]23)O[C@@H]4O[C@H](C)C[C@H](O)[C@H]4O)[C@@H]1OC

CO[C@@H]1
[C@H](O)[C@@H](C)
O[C@@H](OC[C@@H]2
[C@@H](C)
OC(=O)\\C=C\\
[C@H](C)[C@H](CC
[C@@H](C)C(=O)
\\C=C\\
[C@H]3
O[C@@H]23)
O[C@@H]4
O[C@H](C)C
[C@H](O)
[C@H]4O)
[C@@H]1
OC

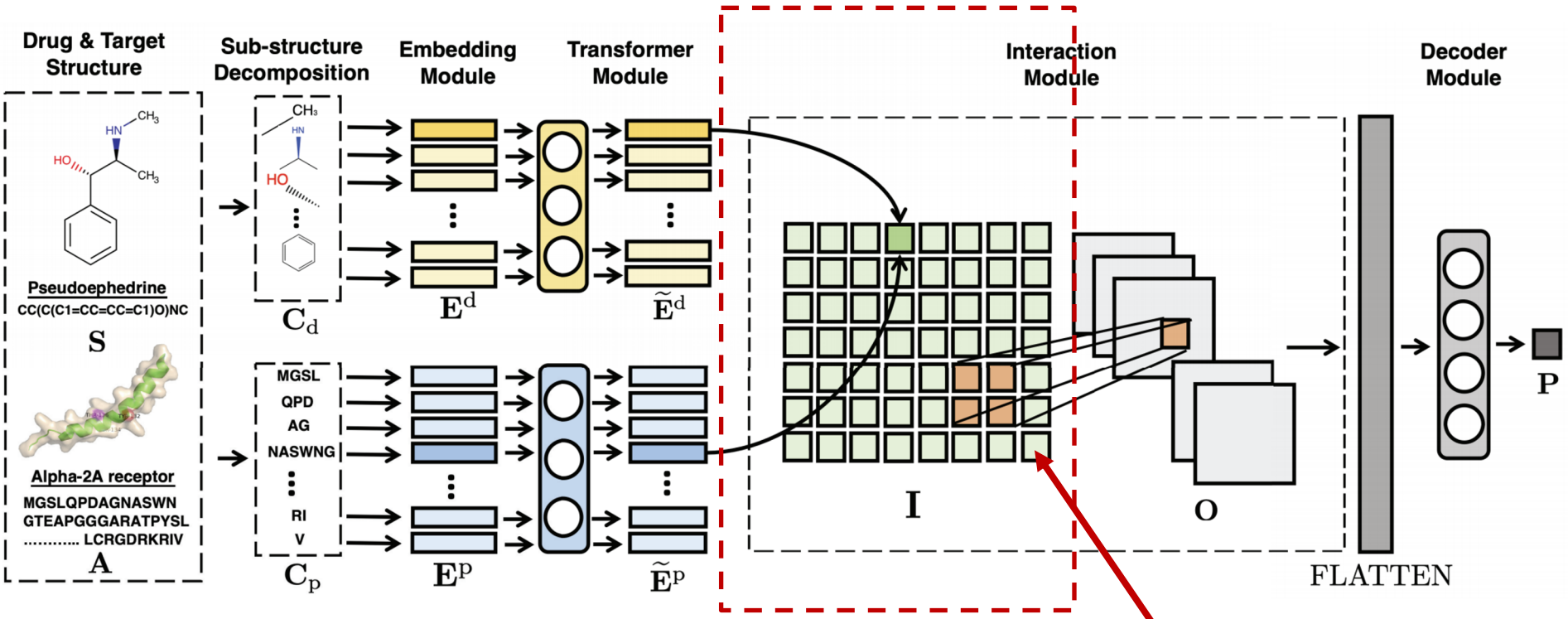
moderate-sized partition with each partition associated with sub-structures

Augmented Embedding



To capture relations among each substructure in the input, we leverage Transformer's self-attention mechanism!

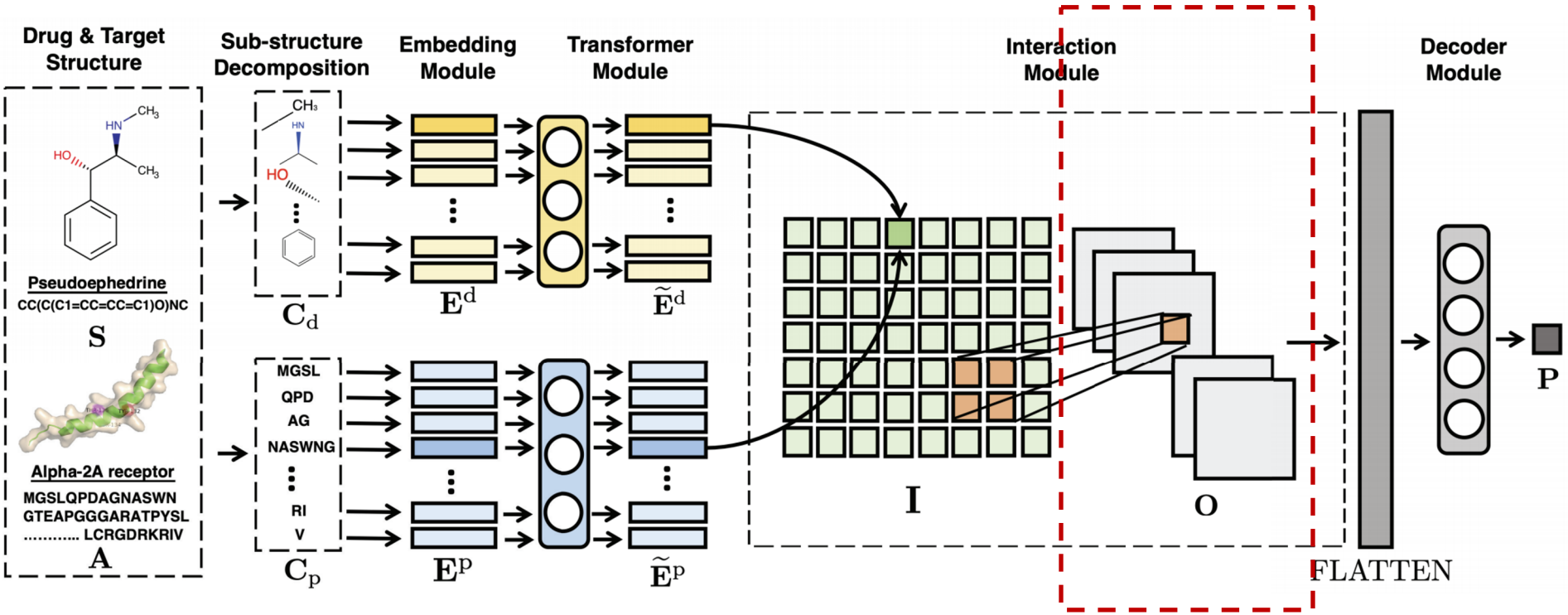
Modeling Sub-structural Interaction



Motivated by the fact that DTI happens in sub-structural level, the interaction module pair each drug-target substructure fingerprint and generate a scalar that measures their interaction,

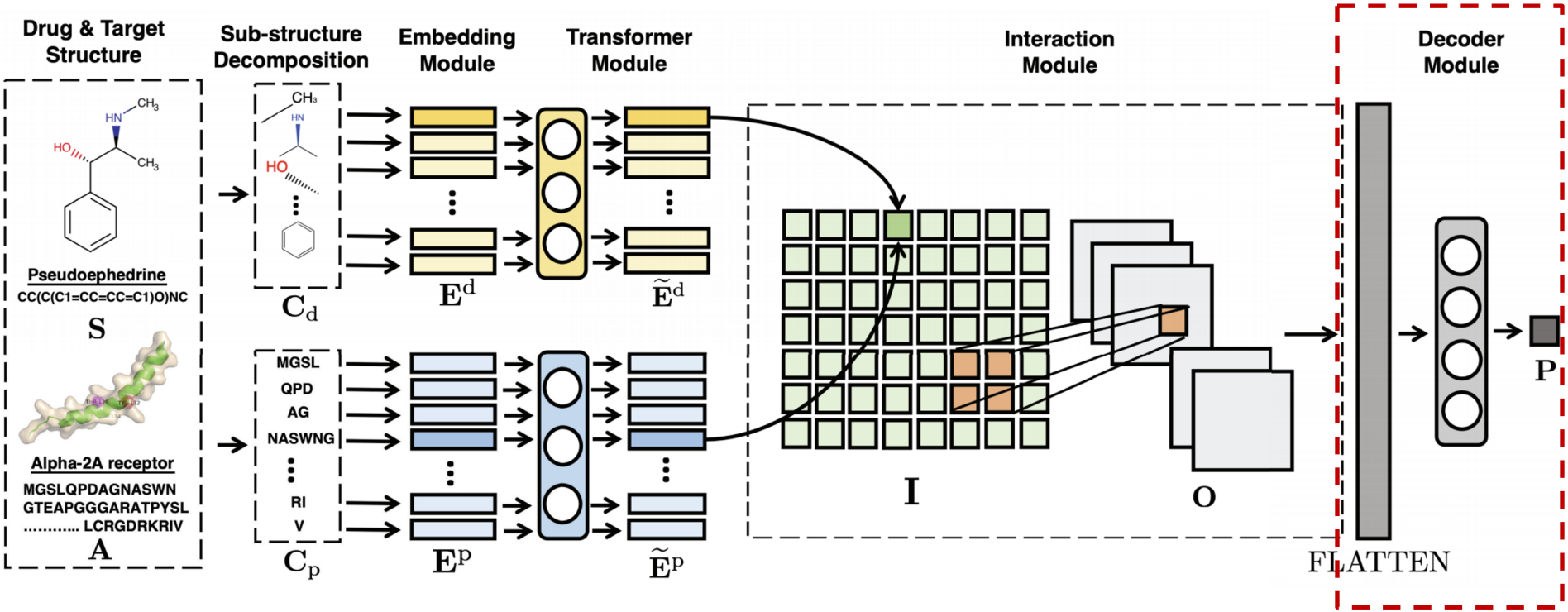
Each position corresponds to an interaction intensity between a drug and a target substructure!

Capturing Higher Order Interaction



Nearby sub-structure of proteins and drugs also influence each other in triggering the interactions. We include a convolutional neural network to model these higher-order interactions.

Prediction



By flattening the CNN output, we generate an embedding for the DTI pair. The embedding is fed into a decoder for prediction.

MolTrans Achieves Superior Predictive Performance

Dataset 1: BIOSNAP

Method	ROC-AUC	PR-AUC	Sensitivity	Specificity	Threshold
LR	0.846 ± 0.004	0.850 ± 0.011	0.755 ± 0.039	0.800 ± 0.018	0.434
DNN	0.849 ± 0.003	0.855 ± 0.010	0.776 ± 0.040	0.838 ± 0.024	0.499
GNN-CPI	0.879 ± 0.007	0.890 ± 0.004	0.780 ± 0.014	0.819 ± 0.012	0.349
DeepDTI	0.876 ± 0.005	0.876 ± 0.006	$\mathbf{0.789 \pm 0.027}$	0.845 ± 0.017	0.347
DeepDTA	0.876 ± 0.005	0.883 ± 0.006	0.781 ± 0.015	0.824 ± 0.012	0.466
DeepConv-DTI	0.883 ± 0.002	0.889 ± 0.005	0.770 ± 0.023	0.832 ± 0.016	0.441
MolTrans	$\mathbf{0.895 \pm 0.002}$	$\mathbf{0.901 \pm 0.004}$	0.775 ± 0.032	$\mathbf{0.851 \pm 0.014}$	0.431

Dataset 2: DAVIS

Method	ROC-AUC	PR-AUC	Sensitivity	Specificity	Threshold
LR	0.835 ± 0.010	0.232 ± 0.023	0.699 ± 0.051	0.842 ± 0.033	0.399
DNN	0.864 ± 0.009	0.258 ± 0.024	0.764 ± 0.045	0.860 ± 0.038	0.489
GNN-CPI	0.840 ± 0.012	0.269 ± 0.020	0.696 ± 0.047	0.842 ± 0.039	0.487
DeepDTI	0.861 ± 0.002	0.231 ± 0.006	0.751 ± 0.015	0.853 ± 0.012	0.387
DeepDTA	0.880 ± 0.007	0.302 ± 0.044	0.764 ± 0.045	0.865 ± 0.020	0.482
DeepConv-DTI	0.884 ± 0.008	0.299 ± 0.039	0.754 ± 0.040	$\mathbf{0.880 \pm 0.024}$	0.438
MolTrans	$\mathbf{0.907 \pm 0.002}$	$\mathbf{0.404 \pm 0.016}$	$\mathbf{0.800 \pm 0.022}$	0.876 ± 0.013	0.447

Dataset 3: BindingDB

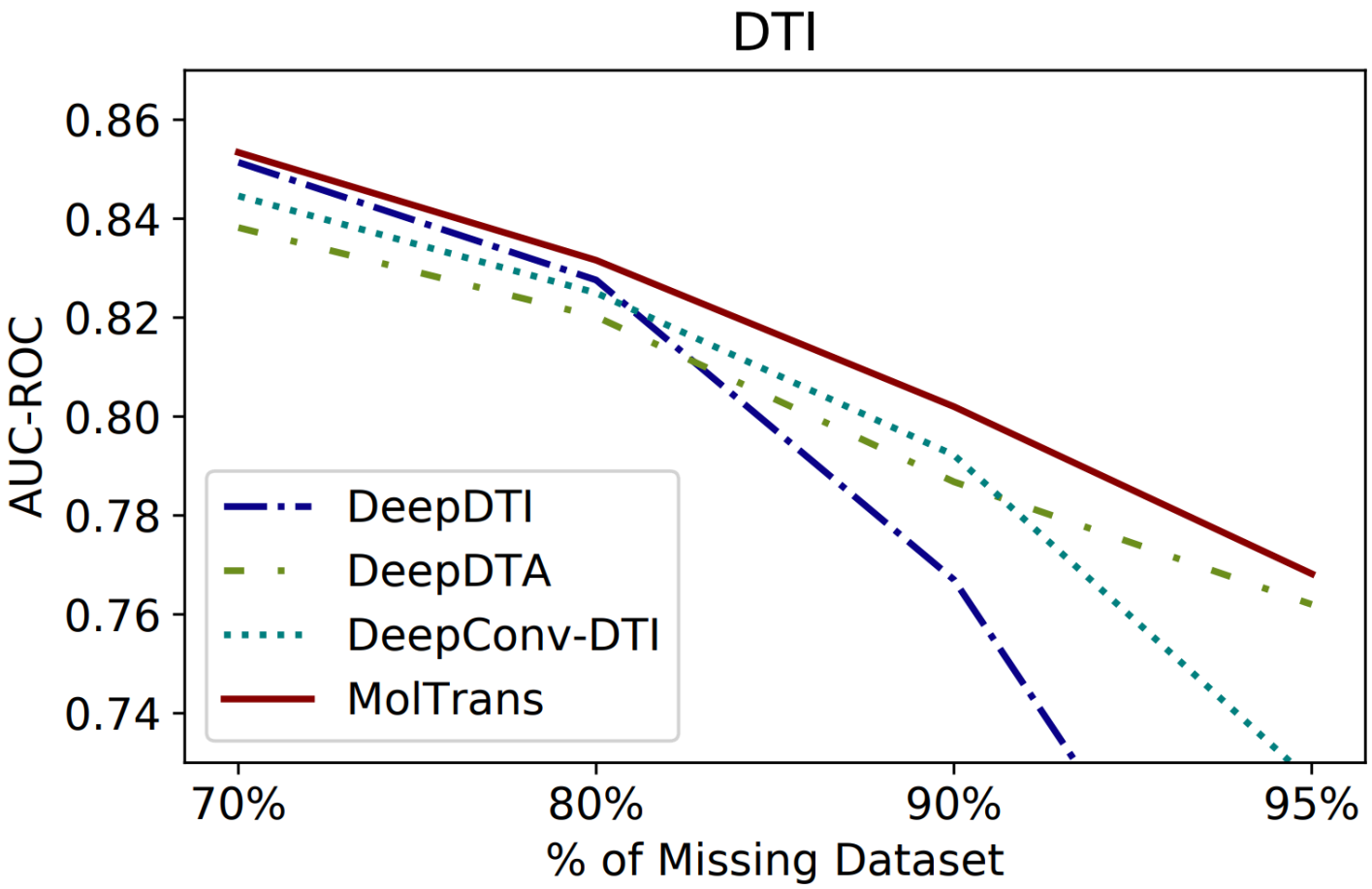
Method	ROC-AUC	PR-AUC	Sensitivity	Specificity	Threshold
LR	0.887 ± 0.002	0.557 ± 0.015	0.741 ± 0.013	0.896 ± 0.011	0.394
DNN	0.908 ± 0.003	0.613 ± 0.015	0.769 ± 0.028	0.914 ± 0.021	0.371
GNN-CPI	0.900 ± 0.004	0.578 ± 0.015	0.754 ± 0.015	0.903 ± 0.011	0.406
DeepDTI	0.844 ± 0.002	0.429 ± 0.005	0.651 ± 0.024	0.895 ± 0.023	0.060
DeepDTA	0.913 ± 0.003	0.622 ± 0.012	0.780 ± 0.035	$\mathbf{0.915 \pm 0.016}$	0.305
DeepConv-DTI	0.908 ± 0.004	0.611 ± 0.015	0.781 ± 0.015	0.905 ± 0.013	0.318
MolTrans	$\mathbf{0.914 \pm 0.001}$	$\mathbf{0.622 \pm 0.007}$	$\mathbf{0.797 \pm 0.005}$	0.896 ± 0.007	0.355

MolTrans has up to 25% increase over best performing baseline!

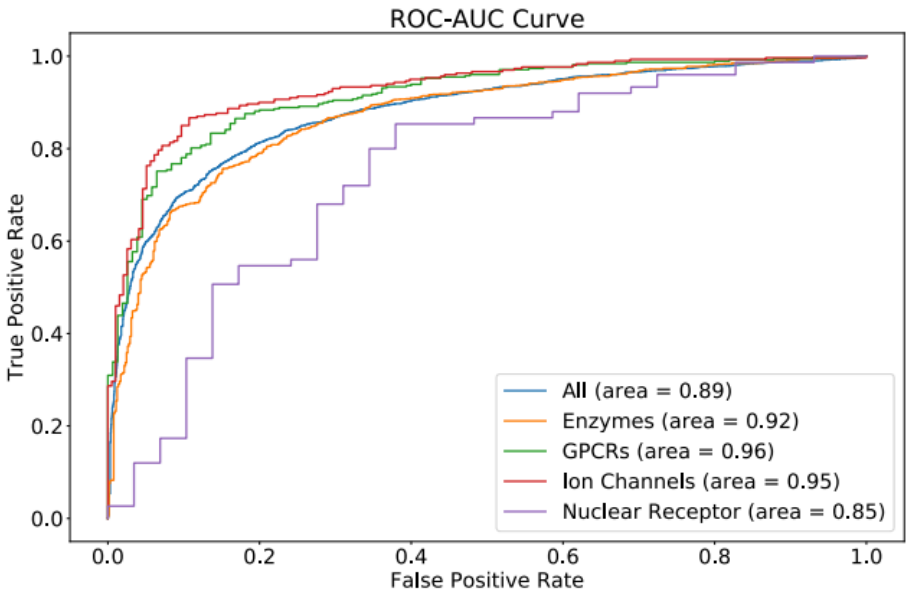
MolTrans has competitive performance in unseen drug and target setting

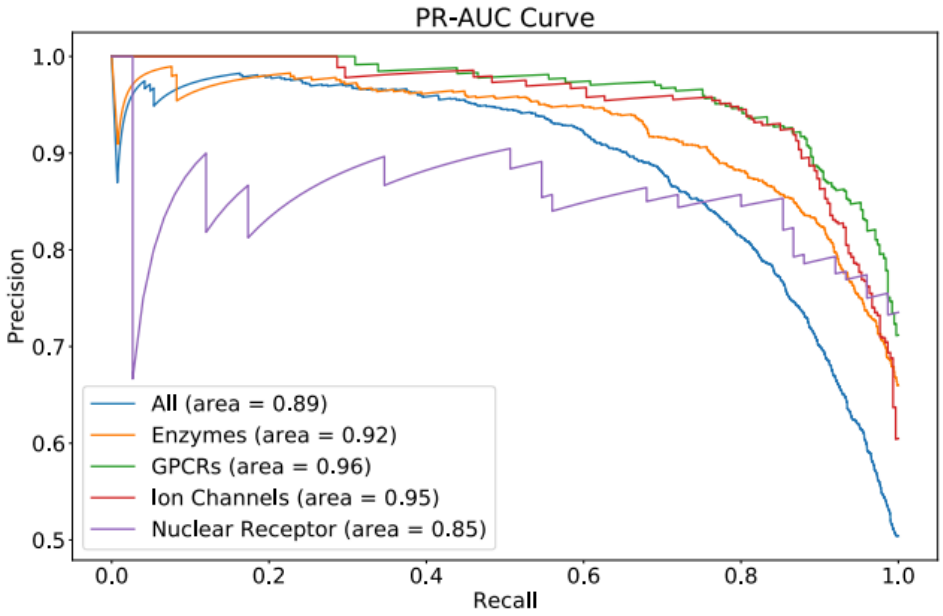
Settings	DeepDTI	DeepDTA	DeepConv-DTI	MolTrans
Unseen Drugs	0.843 ± 0.003	0.849 ± 0.007	0.847 ± 0.009	$\mathbf{0.853 \pm 0.011}$
Unseen Proteins	0.759 ± 0.029	0.767 ± 0.022	0.766 ± 0.022	$\mathbf{0.770 \pm 0.029}$

MolTrans performs best with scarce data



MolTrans is robust in various protein families



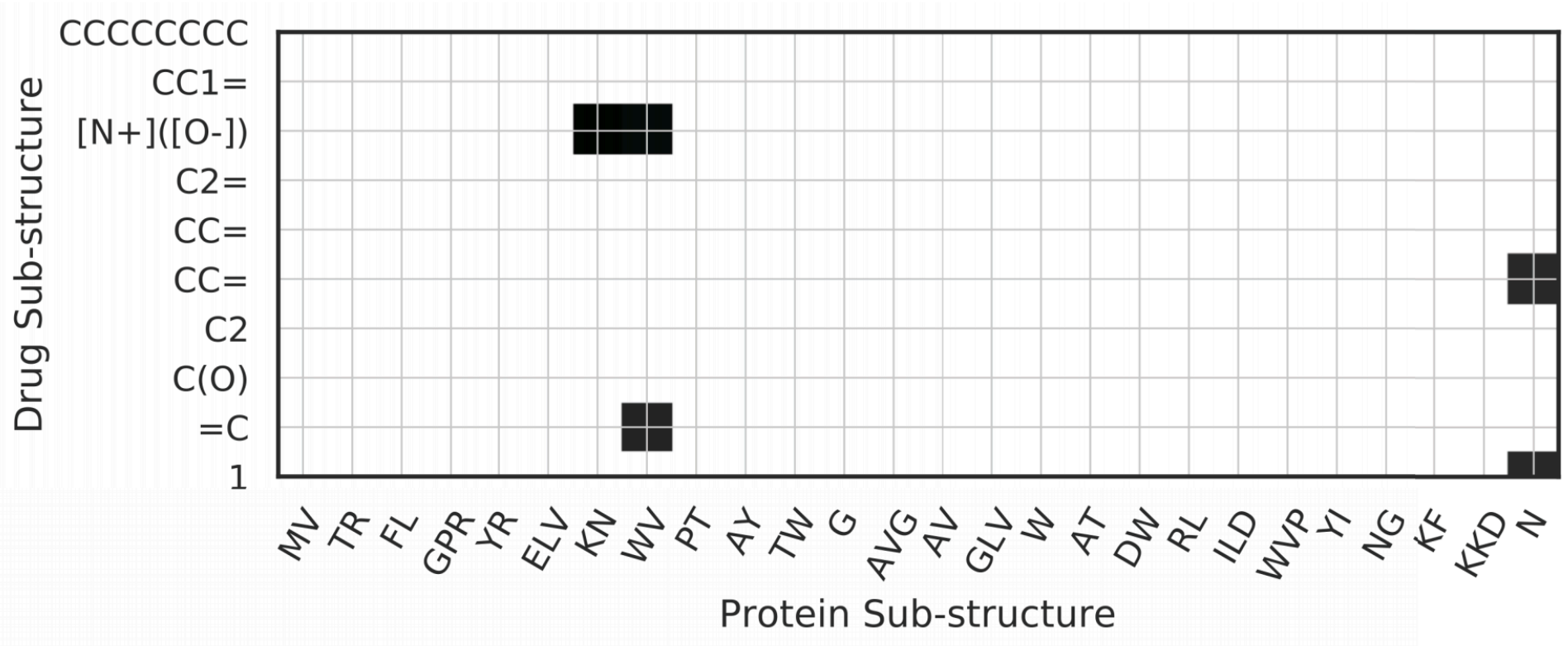


MolTrans allows model understanding

Drug: 2-nonyl n-oxide

Protein: Cytochrome bc1 complex unit 1

[N+]([O-]) and KNWV has the highest interaction coefficient, matching with the previous study (Lightbown and Jackson, 1956) who showed that nitrogen oxide group is essential for cytochrome inhibition activity.



Ablation Study

Setup	ROC-AUC	PR-AUC
MolTrans	$\mathbf{0.895 \pm 0.002}$	$\mathbf{0.901 \pm 0.004}$
-CNN	0.876 ± 0.003	0.883 ± 0.006
-AugEmbed	0.876 ± 0.004	0.870 ± 0.004
-Interaction	0.847 ± 0.003	0.859 ± 0.005
Small	0.888 ± 0.001	0.888 ± 0.007
-FCS	0.887 ± 0.004	0.887 ± 0.004

Code

Paper

Thank you!