Selvita

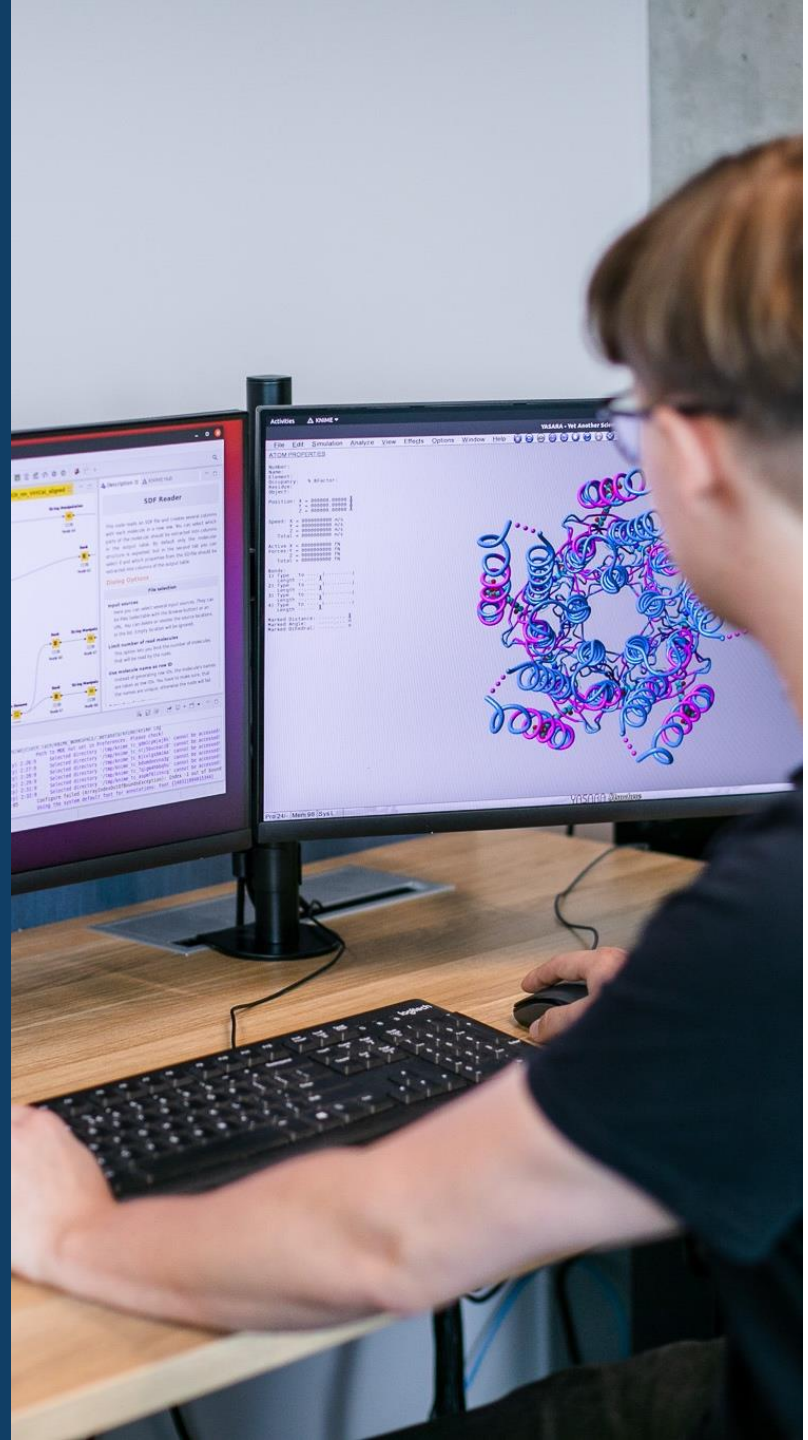
Target-Aware Drug Activity Model:
A deep learning approach to virtual HTS

Szymon Czaplak, Fabrizio Ambrogi

ICANN conference 2024

Agenda

1. **Virtual screening with TADAM**
   - Challenges
   - Solutions
2. **Case studies**
   - Retrospective comparison with docking pipeline
   - Experimental pipeline description
   - Results & conclusion
   - Comparison with other ML methods
3. **Other applications of TADAM**
4. **Summary**



Selvita

Virtual Screening with TADAM

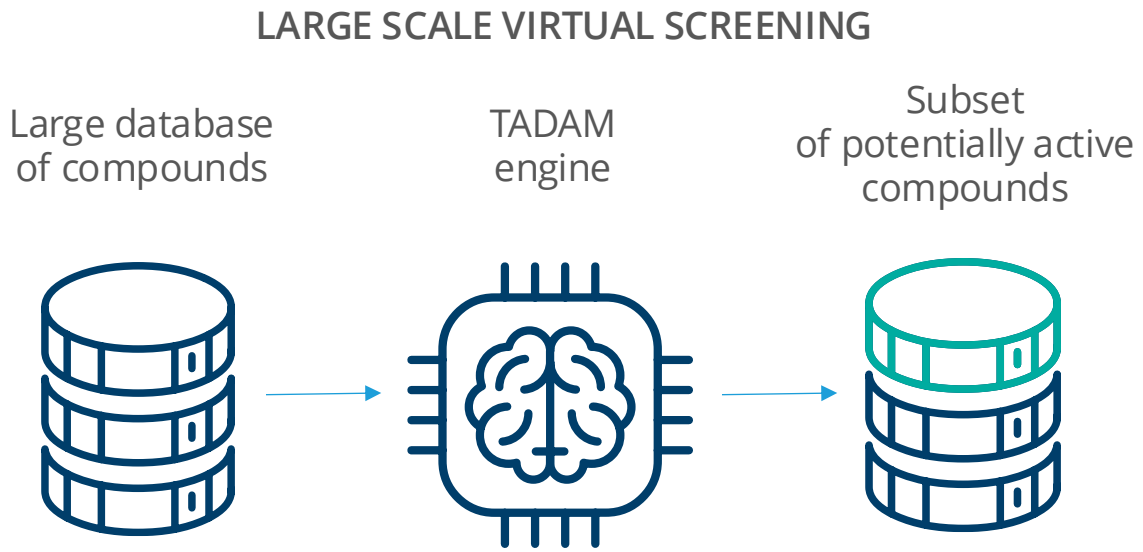
Challenges

- Very large databases of chemical compounds, too many for traditional methods
- Protein targets without any available activity data, hard to use homology models
- A limited set of potential ligands available, that we want to diversify

Solution

We have trained a large proprietary deep learning model, that is able to predict a compound's **activity towards any protein target's pocket.**

It can **screen very large libraries** of compounds (like MolPort or Enamine) in hours, and **identify which pharmacophores are the most influential** in the prediction.



We called our model TADAM, which stands for **T**arget **A**ware **D**rug **A**ffinity **M**odel**.**

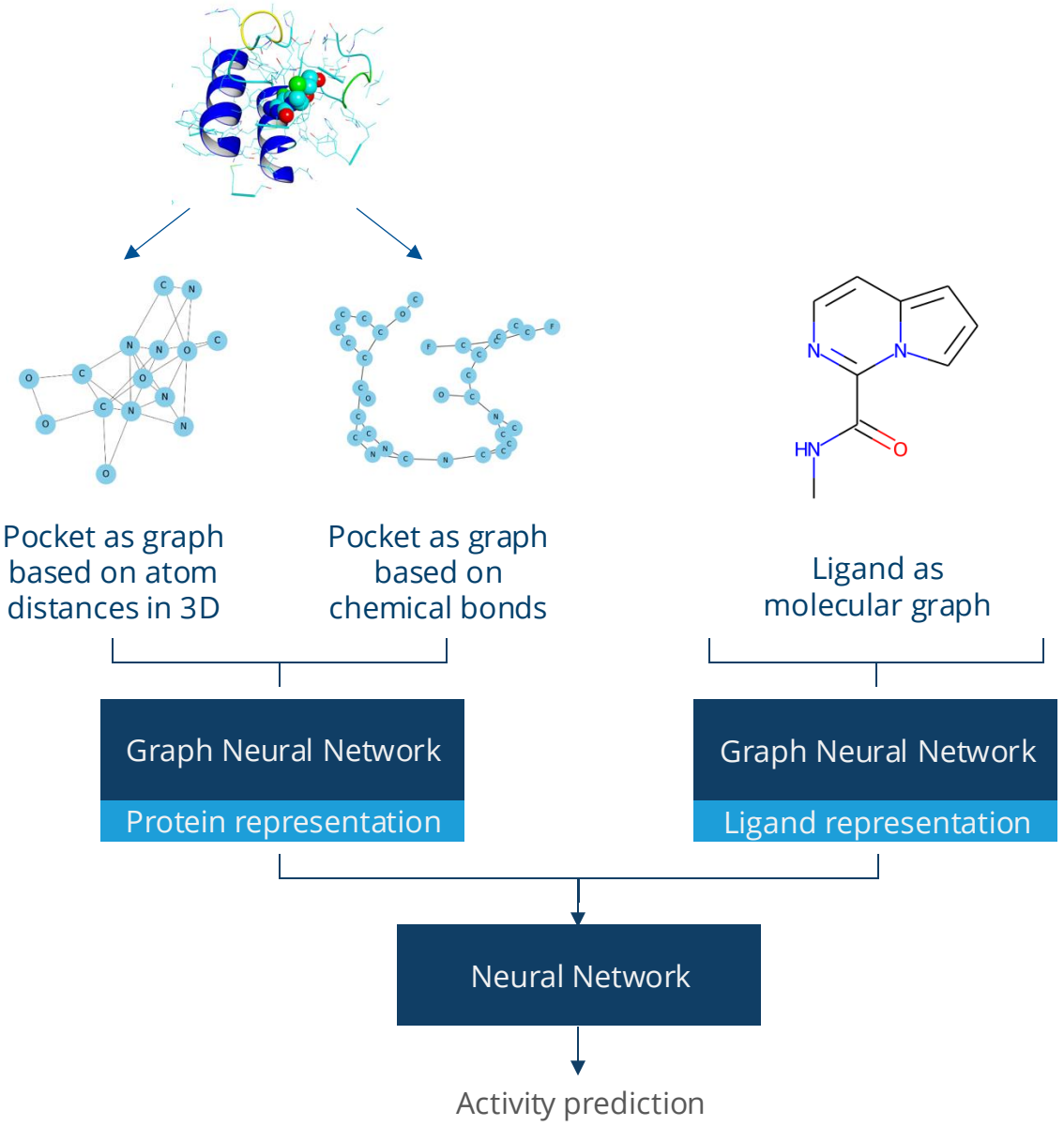
TADAM Model Description

Data representation

- Pocket definition: residues in the 10A sphere around ligand
- Pocket representation
  - We are using innovative representation of proteins' pockets that utilizes information from **both atom connectivity and spatial distances** in 3D space
- Ligand representation: molecular graph with atoms as nodes and chemical bonds as edges

Model

- Model was trained to predict activity of any given compounds towards any target
- Trained on carefully tailored dataset from collected data from ChEMBL and PDB
- It utilizes 3d information about protein's binding pocket as well as connectivity between atoms
- It is **considerably faster** than standard docking approaches



TADAM Compared with Traditional Docking

Dataset

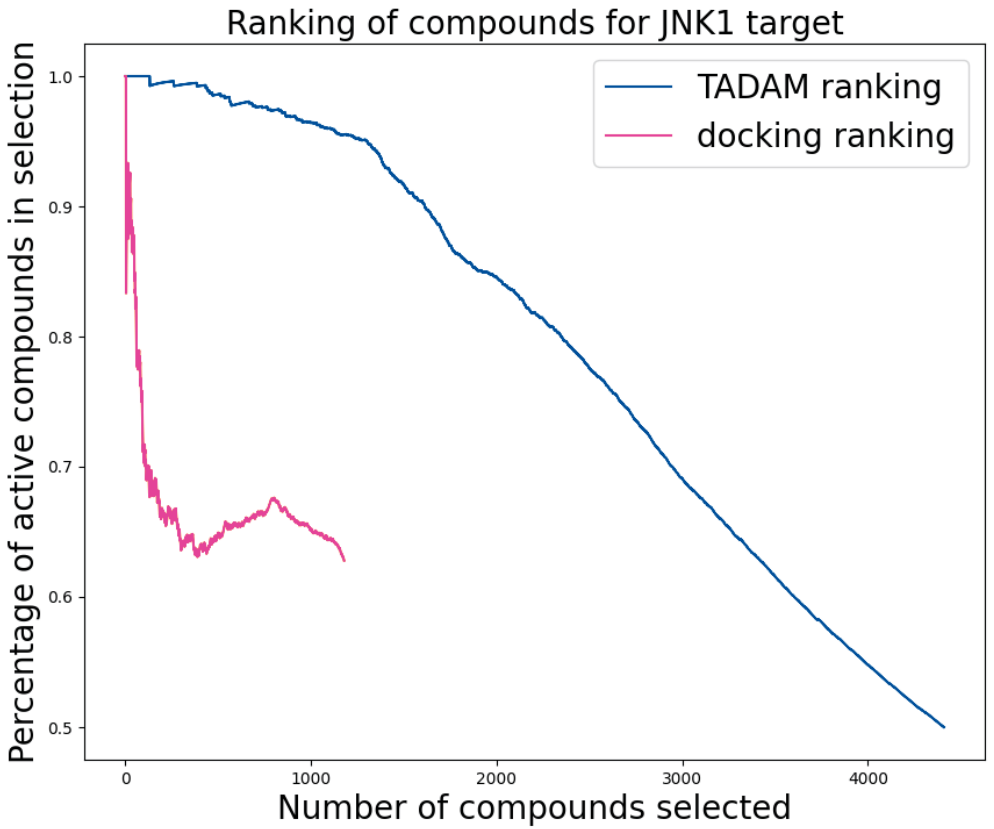
- JNK1, target protein excluded from TADAM training
- Data collected from ChEMBL
- Decoys generated to achieve equal class distribution:
  - dissimilar by Tanimoto distance
  - similar by phys-chem properties

Docking procedure

- The docking was done with FlexX and Molegro software
- Scoring of poses was done with MOE (GBVI/WSA dG score)
- Docking was successful for only ~27% of all compounds

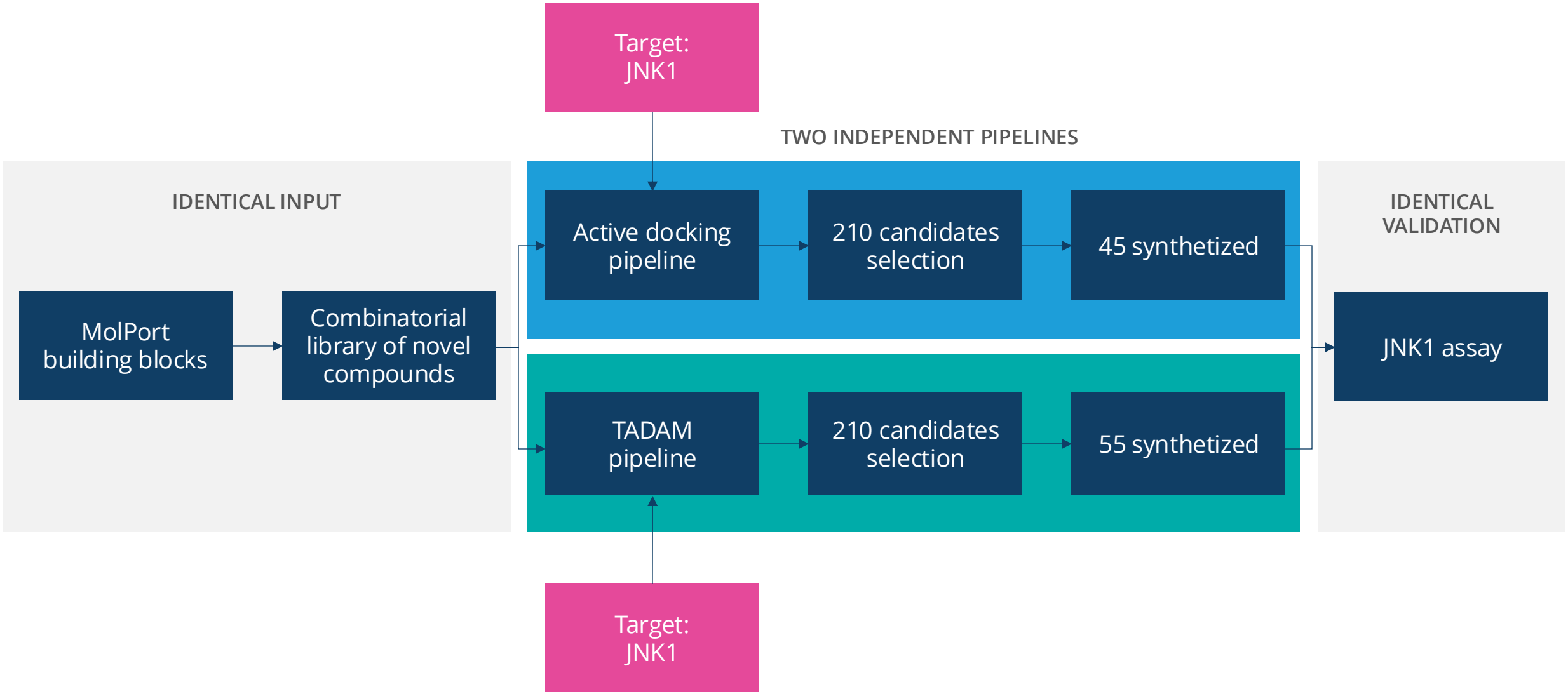
Results

- Rankings are compared by precision@K, over all possible Ks
- Our model vastly outperforms the traditional method
- Not all docked compounds were active and, more importantly, many ligands would have been discarded by the software.



In summary, TADAM is a good alternative to classical docking, that can find ligand that would have been missed in standard approach.

Case Study – Process Overview



Final Results

ASSAY PIPELINE

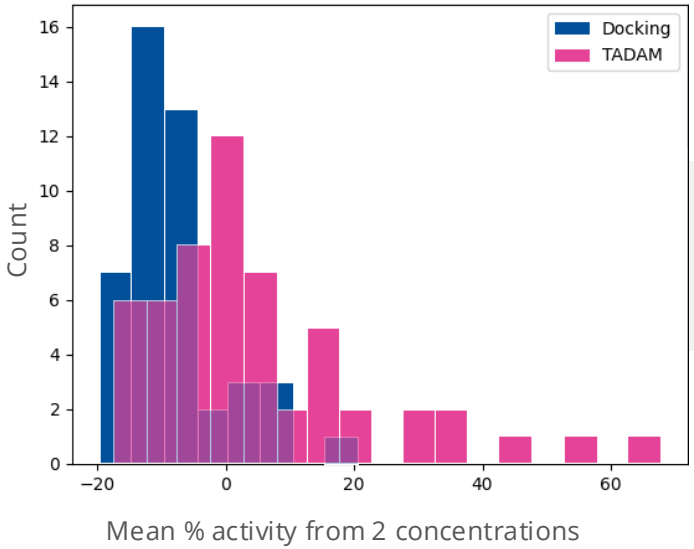
All compounds (100)

Quick HTS-like run in two concentrations

Active compounds and uncertain results

Detailed full dose response assay

Results from HTS-like screening



MEASURED ACTIVITY FOR TADAM AND DOCKING

Results from detailed full dose response assay

- 7 confirmed ligands coming from TADAM
  5 of which < 5 µM IC50, the others < 12
- Only one candidate from docking got into the top 10 most active compounds, with an IC50 around 69 µM

All confirmed hits were selected by TADAM!

TADAM Comparison with Other ML Methods

Dataset

- PDB's reported after 2022
- 2359 protein-ligand complexes (1931 unique ligands, 303 unique proteins)

Decoys

- Compounds taken from MolPort
- Kept compounds within 1-sigma of Phys-Chem properties among known ligands
- Selection of decoys done by sampling representatives of 50 clusters
- 50 diverse decoys matched with each unique protein, totaling 15k+ negative pairs

Performance comparison between TADAM, DiffDock and HyperPCM on the complexes and decoys from PDB 2023

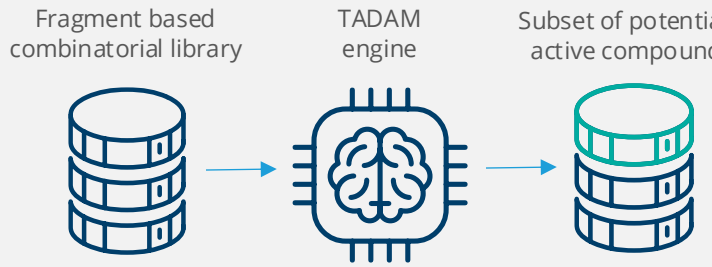
Model	Recall	Specificity	AUROC	Model Size
DiffDock	19.5%	79.8%	0.5	4M (docking) + 4M (scoring)
HyperPCM	18.1%	**99.1%**	**0.59**	220M
TADAM	**24.6%**	85%	0.57	**880k**

Recall: True positive rate.
Specificity: True negative rate.
AUROCC: Theoretical discriminatory power of a model.

Recall is key for high throughput screening, as it represents the power of the model in avoiding false negatives

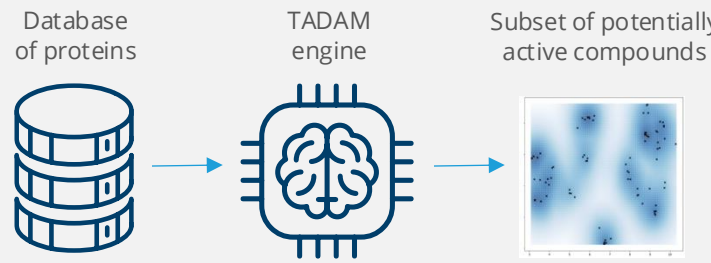
Other Applications

FRAGMENT GROWING IDEAS



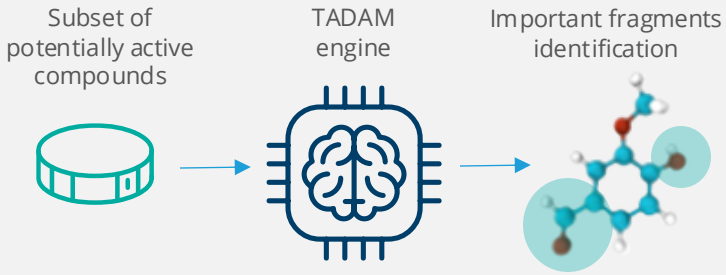
- Prioritization of fragment growing ideas
- **Compounds prioritization and selection** based on **predicted activity** towards a given target

PROTEINS' POCKETS SIMILARITY SEARCH



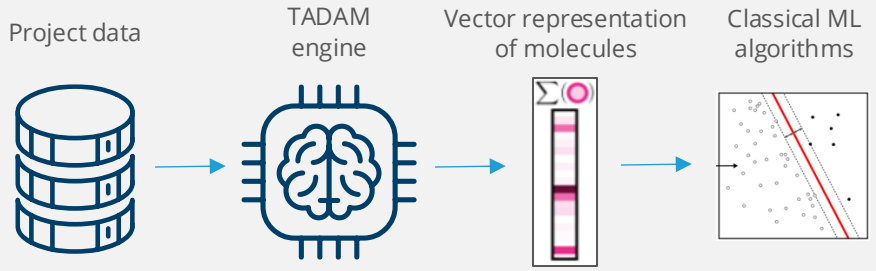
- **Analyze and visualize** database of proteins
- **Off-target search**
- **Search for targets with similar pockets** for reference

ACTIVITY INTERPRETATION



- Help in understanding **which parts of the compound are most important** for activity
- Guidance in ligand optimization processes

REPRESENTATION LEARNING

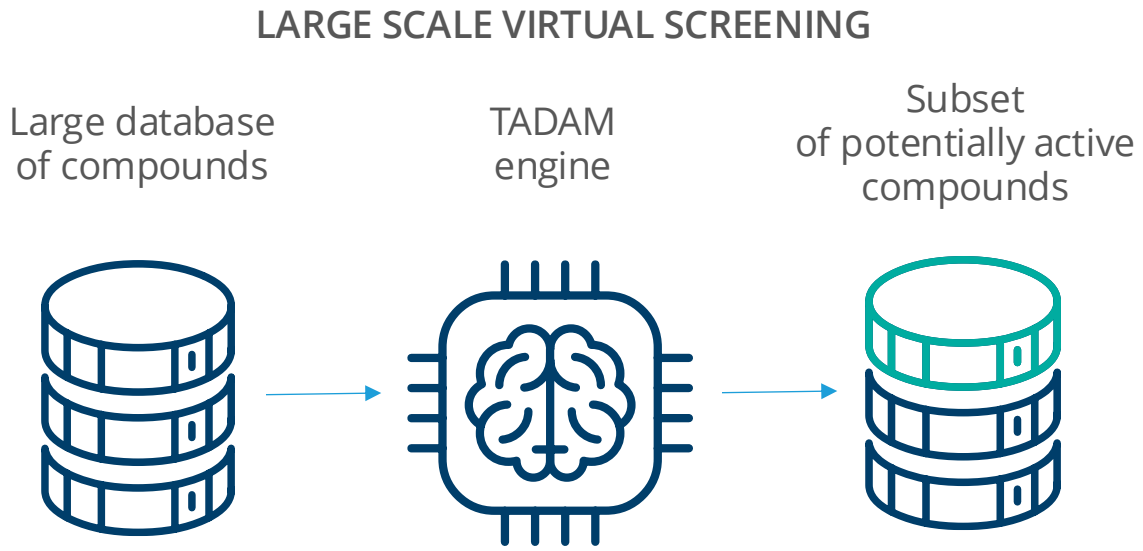


- **Extraction of generic molecular representation conditioned on activity**
- Project data augmented with knowledge extracted from large databases

Summary

Key Takeaways

- TADAM can rapidly screen **very large** databases of compounds
- Trained to predict activity between any protein and small molecule
- It utilizes a graph representation based on both chemical bonds and 3D confirmation of protein
- The model outcomes can be used in many other applications
- The in vitro evaluations mark it way above docking in detecting real ligands
- It surpasses other SotA ML methods in our retrospective analysis, in both recall and speed of screening.



Acknowledgements

Fundusze Europejskie
Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

Selvita

Thank you for your attention!

Szymon Czaplak, Senior Machine Learning Specialist

szymon.czaplak@selvita.com

Fabrizio Ambrogi, Senior Machine Learning Specialist

fabrizio.ambrogi@selvita.com

SelvitaKrakow

/company/selvita/

/Biotechnology-Company/Selvita-SA