Atomwise

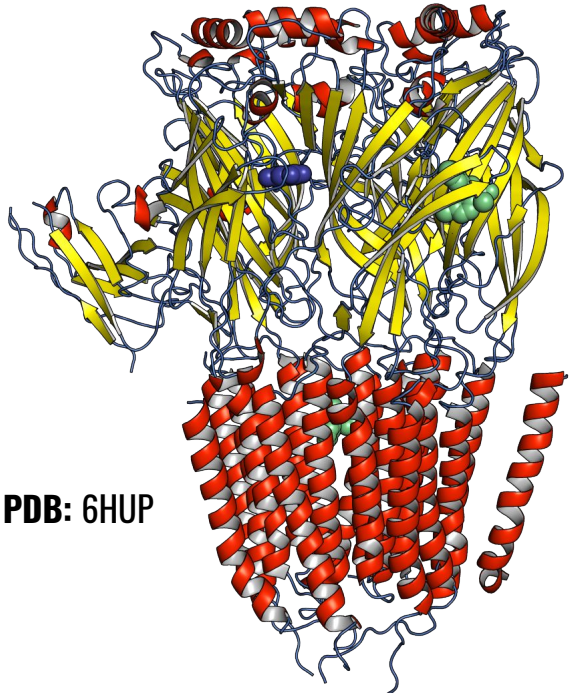
Where do my compounds bind?

Automated assignment of active compounds to non-primary sites helps deep-learning uncover allosteric modulators.

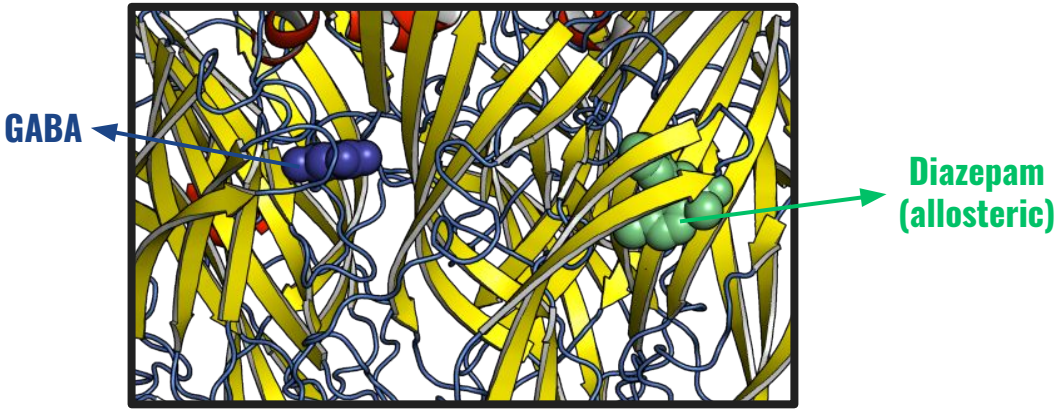
Saulo de Oliveira, DPhil
Cheminformatics Scientist
Atomwise Inc.

What is allostery?

Binding outside of active/catalytic site capable of modulating a protein's function



PDB: 6HUP



Human full-length α1ß3γ2L $GABA_A$ receptor in complex with diazepam (Valium).

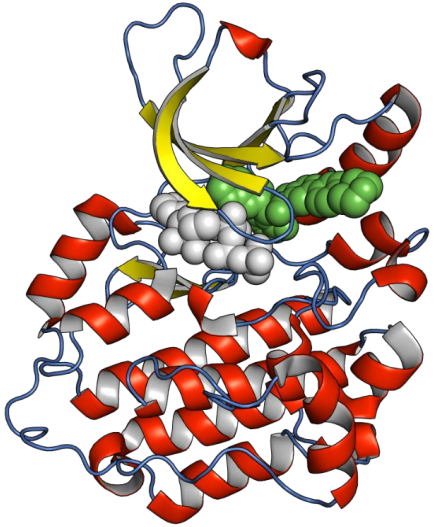
[*] Cully, M., 2019. GABA. sub. A receptor structures solved. *Nature Reviews Drug Discovery*, *18*(2), pp.98-99.

The allure of allostery to Pharma research

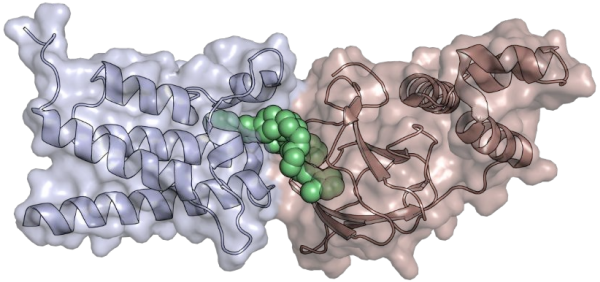
Non-orthosteric molecules offer a path to tackle common problems in the pipeline

SELECTIVITY

Allosteric sites are less conserved and can be exploited to attain selectivity (EGFR - PDB: 6DUK).



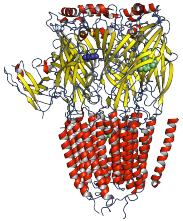
DEGRADERS



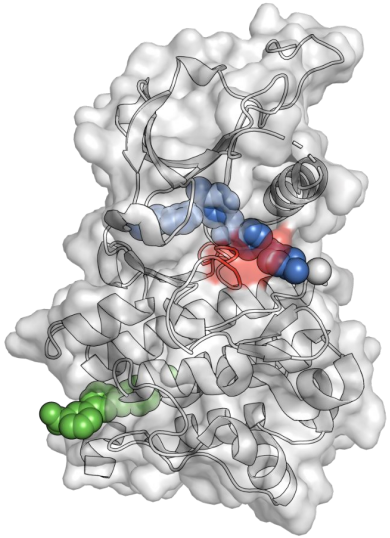
Crystal structure of PROTAC 1 in complex with the bromodomain of human SMARCA2 and pVHL (PDB: 6HAY).

MODULATION

Non-orthosteric binders can act as activators instead of inhibitors ($GABA_A$ receptor - PDB: 6HUP).



DRUG RESISTANCE

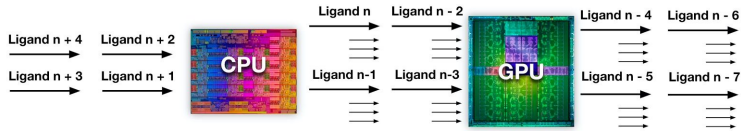


Allosteric binders can provide additive inhibitory activity against T315I mutant human Bcr-Abl (PDB: 3K5V).

In silico identification of (allosteric) binders

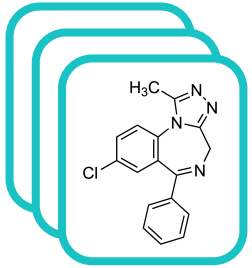
Step 1: molecular docking

https://blog.atomwise.com/efficient-gpu-implementation-of-autodock-vina



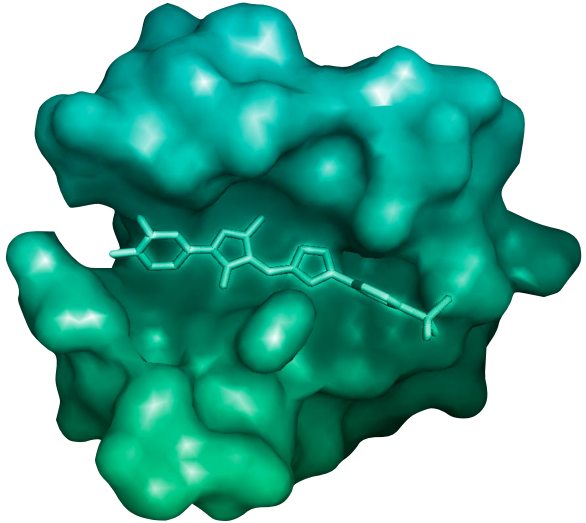
Software Design: CUina

ChEMBL



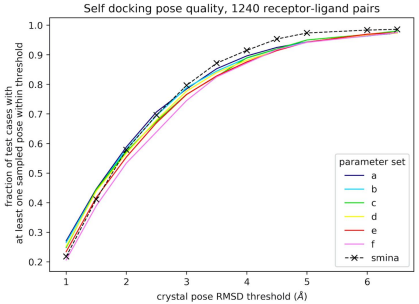
Compound DB

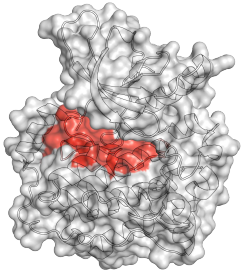
CUina



Docked complex (pose)

CUina Pose Quality





RCSB PDB PROTEIN DATA BANK

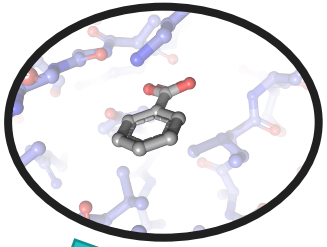
Protein structures

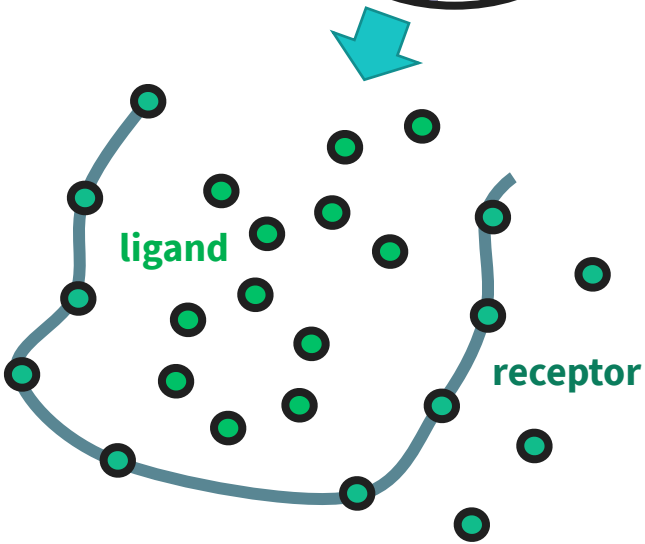
SCAN ME

In silico identification of (allosteric) binders

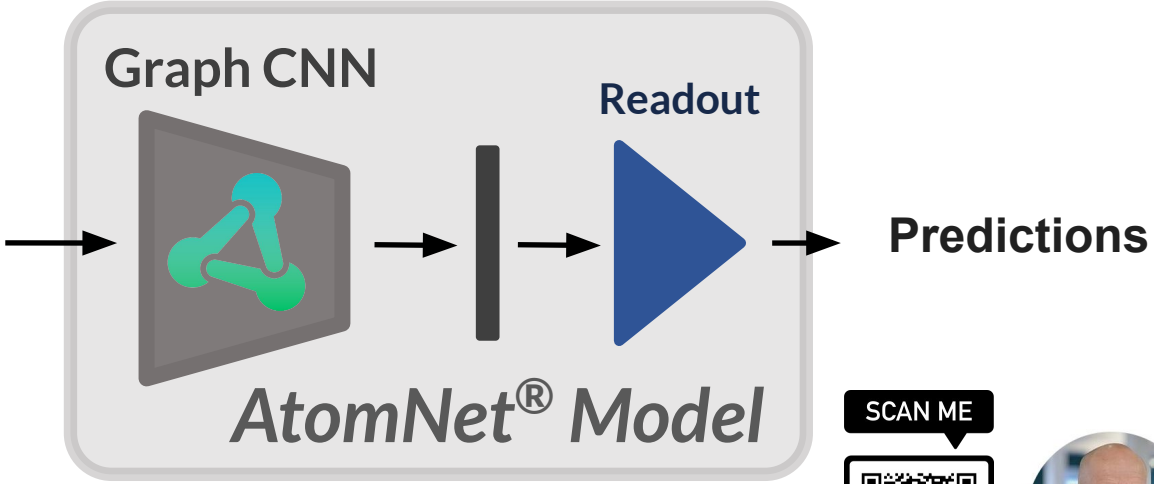
Step 2: feed the docked poses into a neural network to train/perform predictions

AtomNet®
GRAPHite





https://blog.atomwise.com/revealing-hidden-determinants-of-molecular-recognition-with-graph-convolutional-attention-mechanisms



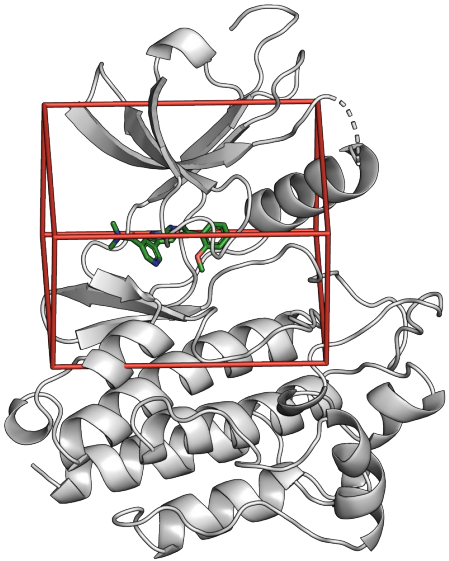
SCAN ME

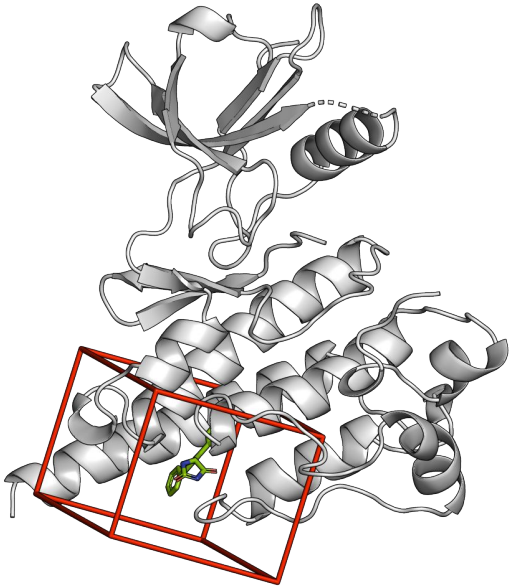
Where do my active compounds bind?

Most activity measurements cannot be reliably mapped to a specific binding site

ChEMBL

Molecule ChEMBL ID	Compound Key	Standard Type	Standard Relation	Standard Value	Standard Units
CHEMBL125331	7r	IC50	=	1790.0	nM
CHEMBL325949	10m	IC50	=	46000.0	nM
CHEMBL91250	4e	IC50	>	50000.0	nM





When training a machine learning model, we need a **reliable** label (ground truth). Which bounding box do we choose for each case?

Identifying known multi-site proteins

This can be performed using publicly available data (PDB + PubMed)

Text mining

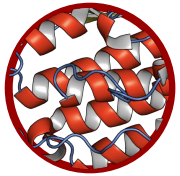
PubMed

ALLOSTERY

ALLOSTERIC

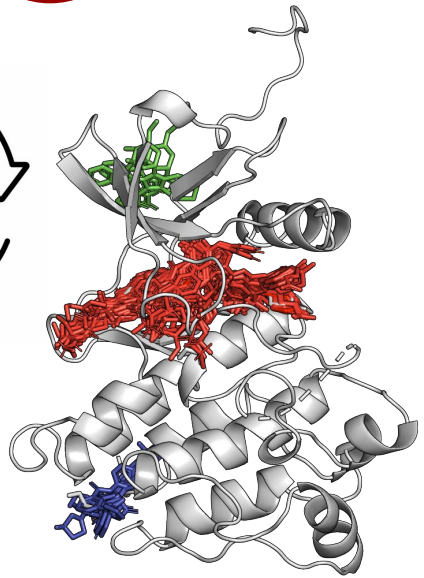
NON-COMPETITIVE

ACTIVITY MEASUREMENTS



Structural analysis

RCSB PDB PROTEIN DATA BANK



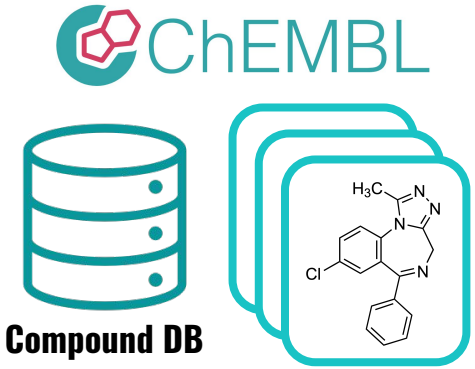
Crystal structure of Abl1

PDB: 5HU9

Mapping known actives to detected sites

Many ways of doing it… Here, we describe a method based on t-SNE projection.

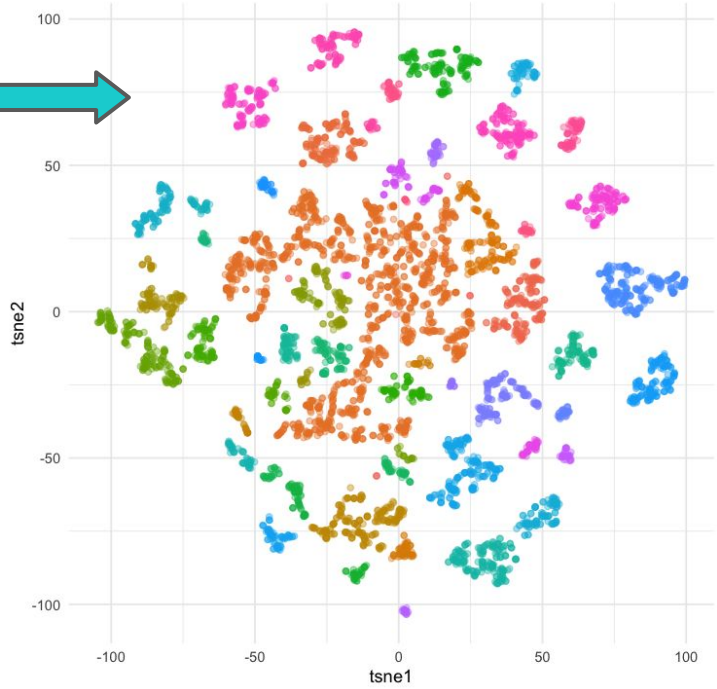
1. Take all measured compounds for a given target from a compound database (e.g. ChEMBLdb).



2. Represent each compound via molecular fingerprints (e.g. ECFP4).

3. Project into lower dimensional space (e.g. using t-SNE).

4. Cluster compounds (e.g. agglomerative clust.).



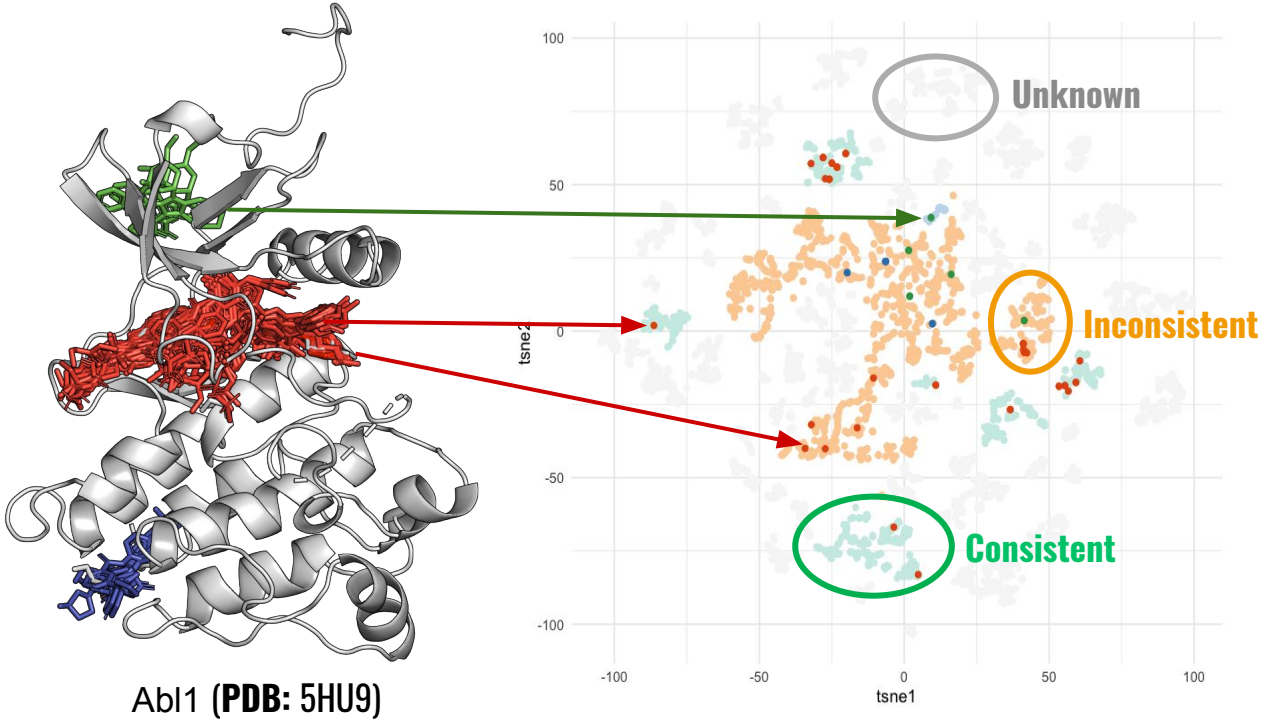
Mapping known actives to detected sites

We can use the compound clusters + structural data to map more compounds

1. Map the structurally resolved compounds to their clusters in the projection.

2. Identify clusters that are consistent: all structurally resolved compounds in the cluster bind to the same site.

3. **An educated guess:** other compounds in consistent cluster also bind to that site.



Abl1 (**PDB:** 5HU9)

Our site-specific data set is highly enabling

We compiled a unique set of activity measurements mapped to binding sites

>500

MULTI-SITE PROTEINS

>0.5 M

COMPOUNDS

18% of compounds

MAPPED TO NON-PRIMARY BINDING SITE*

[*] Data reported for a highly-curated subset of 103 known allosteric proteins.

Pose quality prediction using deep learning

In this context, classifying docked compound poses into good/bad

Model construction

Protein-ligand complex structures

↓

Cross-docking

↓

Deep convolutional neural network

Model use

Target ensemble | Docked molecules

↓

Pose rank model

↓

RMSD rank prediction

↓

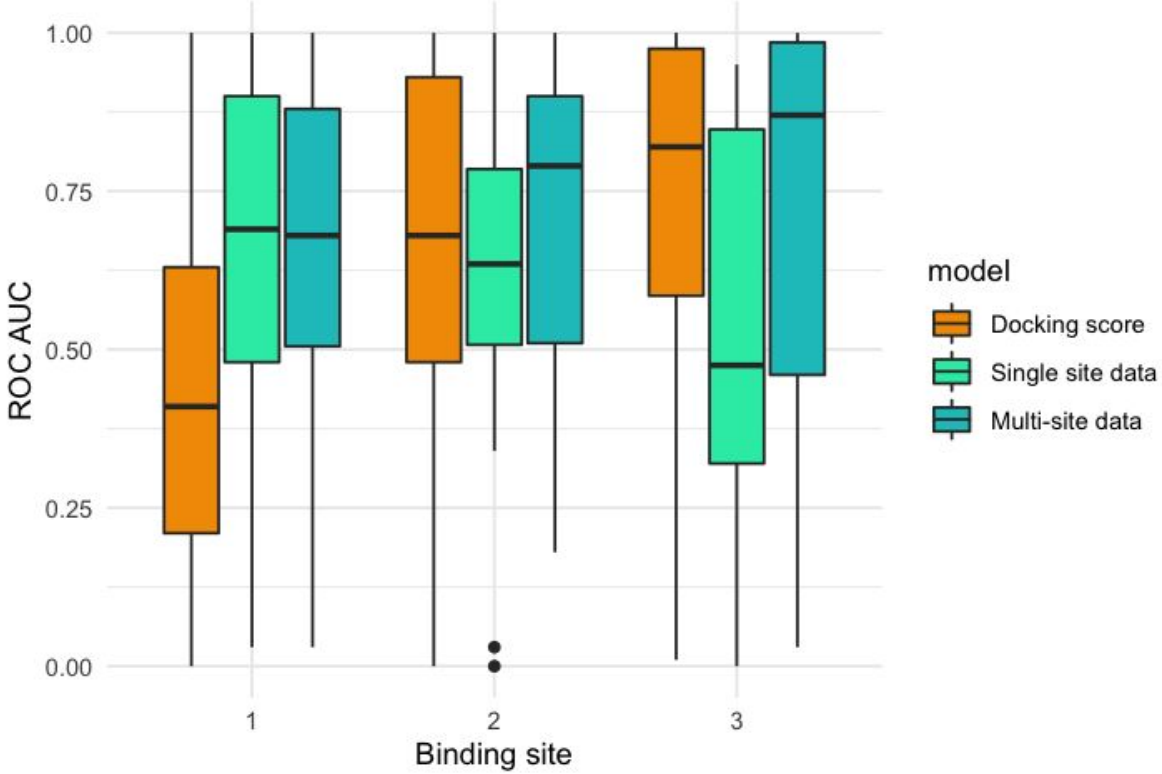
Results

SCAN ME

https://blog.atomwise.com/ligand-pose-ensembles-improves-affinity-prediction-in-structure-based-virtual-screening

Multi-site data leads to better pose prediction

Two exploratory models illustrating the importance of better site annotation

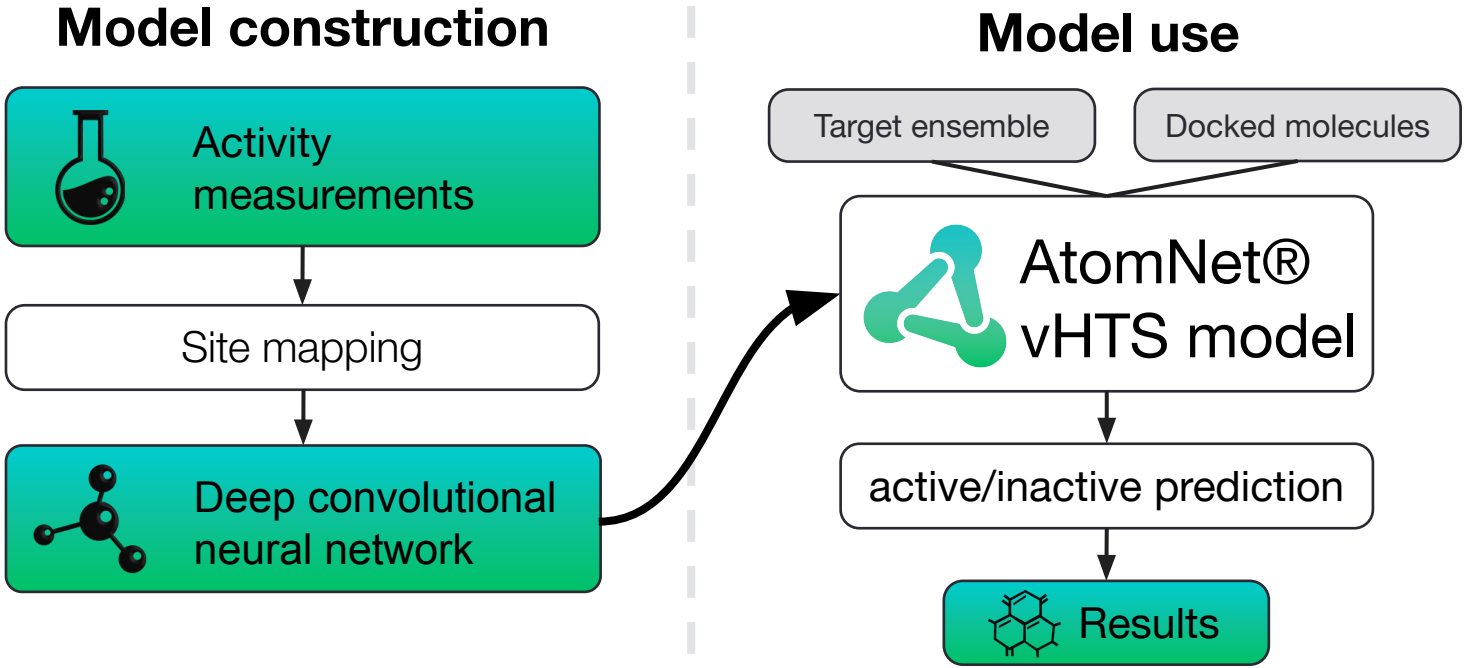


PoseNet training:

- Cross-docking performed using Smina, with the vina scoring function.
- Good poses (positives) defined as < 2.5 Å.
- Bad poses (negatives) defined as > 4.0 Å.
- Models trained as a classifier using a binary cross-entropy loss.

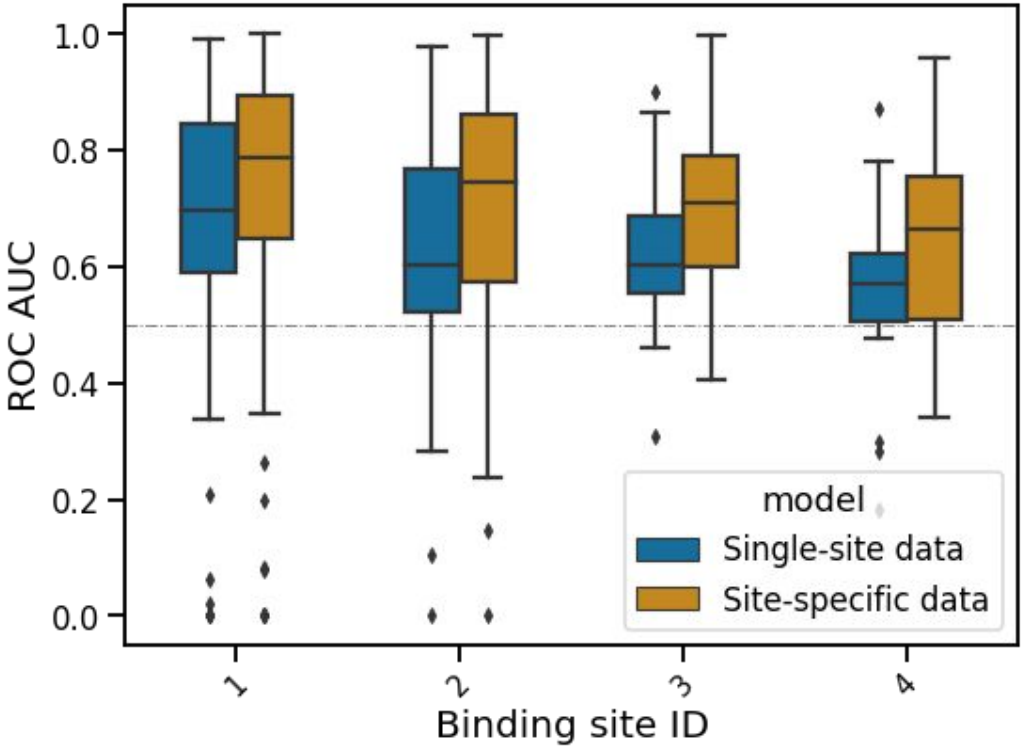
Virtual high throughput screening (vHTS)

A classification task predicting active compounds from inactive compounds



Multi-site data improves vHTS performance

Performance improvement is also observed for primary site



- These exploratory models were trained specifically for ACS using data from public databases such as ChEMBLdb.
- No binding sites for the proteins in our test set were used during training (70% sequence similarity split).
- Improvement in performance for primary site highlights the impact of incorrect data labeling.

Conclusions

1. **Missing a piece of the puzzle:** our current activity data paradigm, for the most part, lacks binding site annotation.

2. **These are not rare events:** for known allosteric proteins, 1 out of 5 compounds is mapped to a non-primary site.

3. **Why should we care?** Site-specific data increase our odds of finding novel allosteric modulators *in silico*.

Acknowledgments

In collaboration with Paweł Gniewek, Kate Stafford, Henry van den Bedem and Atomwise Team!



Our Other Talks & Posters

https://info.atomwise.com/acs_fall2021

Join Us!

https://www.atomwise.com/careers/