Rapid Traversal of Vast Chemical Space Using ML-Guided Docking Screens to Accelerate Drug Discovery

Israel Cabeza de Vaca Lopez

Jens Carlsson Lab

SciLifeLab AI Seminar Series

Nov 2025

Outline

1. Introduction
2. Conformal predictor
3. Benchmarking of conformal predictors
4. Optimized workflow for ultralarge chemical libraries
5. Prospective virtual screen of a multi-billion-scale library
6. Machine learning-guided design of polypharmacology
7. Conclusions

Introduction: The Problem

A 'Vast' Chemical Space

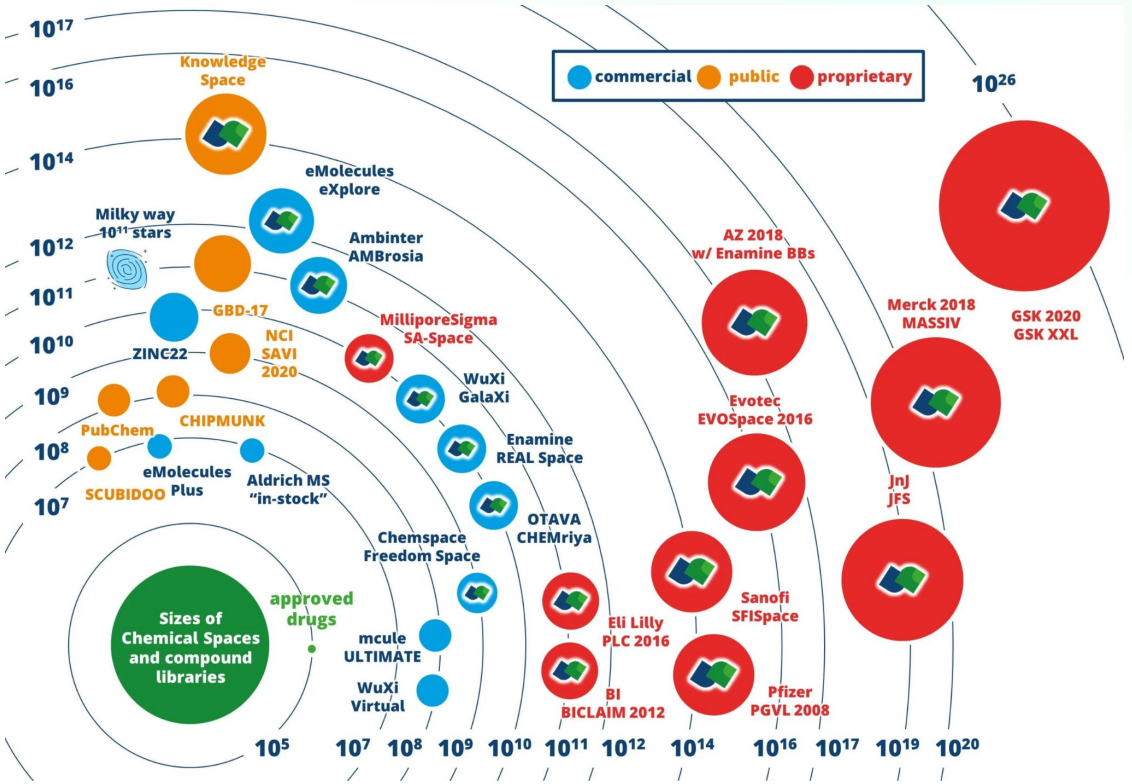
The scale of potential drug-like molecules is staggering, estimated at over 10^{60}

Commercially available *make-on-demand* libraries have exploded, now containing **over 70 billion compounds**

In contrast, *in-stock* libraries contain only ~13 million compounds. We are barely scratching the surface of what's possible

Chemical libraries contain little overlap providing a great opportunity to search a massive diverse space

BioSolveIT



https://www.biosolveit.de/chemical-spaces/

Introduction: The Problem

A 'Vast' Chemical Space

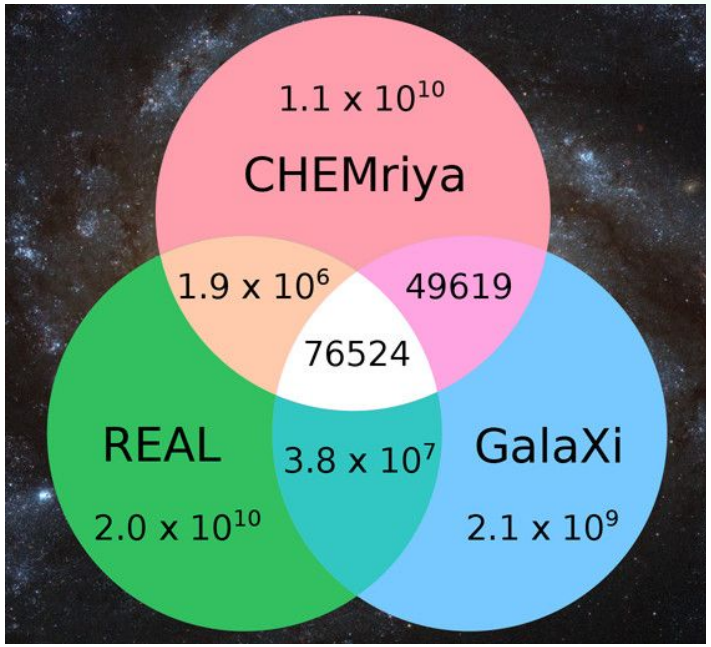
The scale of potential drug-like molecules is staggering, estimated at over 10^{60}

Commercially available 'make-on-demand' libraries have exploded, now containing over 70 billion compounds

In contrast, 'in-stock' libraries contain only ~13 million compounds. We are barely scratching the surface of what's possible

Chemical libraries contain **little overlap** providing a great opportunity to search a massive diverse space

Otava CHEMriya



Enamine REAL Space

WuXi GalaXi Space

(https://doi.org/10.1021/acs.jcim.1c01378)

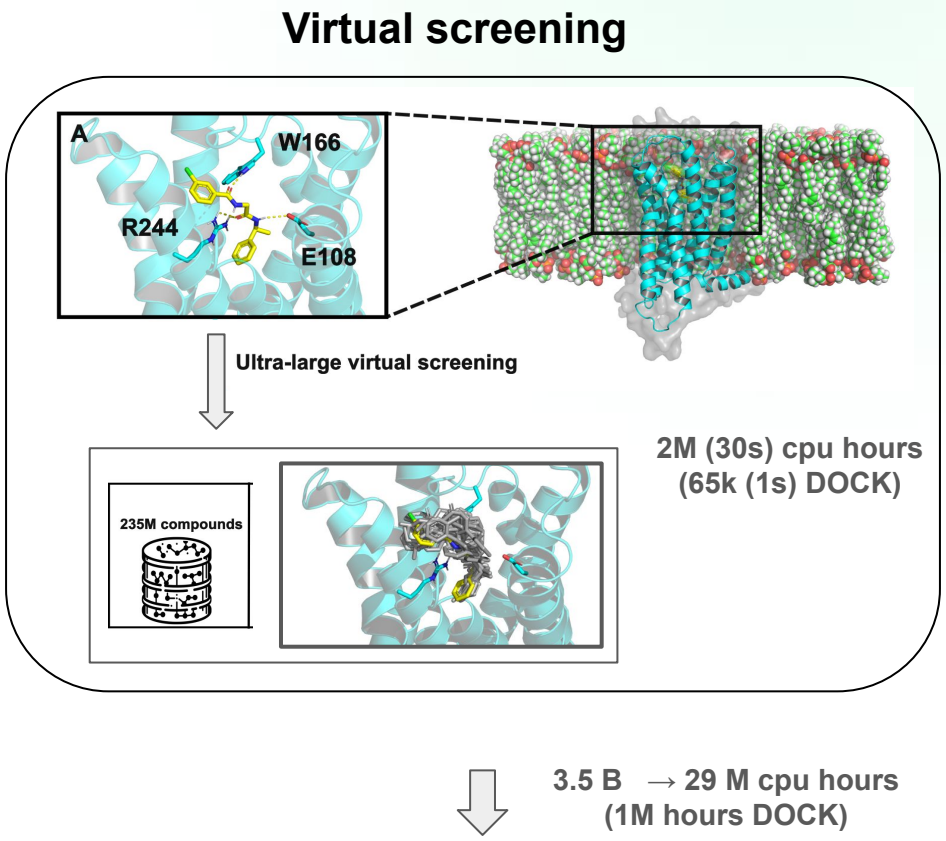
Introduction: The Problem

The 'Docking' Bottleneck

Structure-Based Virtual Screening is a powerful tool for finding ligands. However, it's computationally slow

Screening 70 billion (or future **trillion**-scale) compounds is unfeasible or prohibitively expensive, requiring millions of CPU-hours

How can we find the best needles in this enormous haystack?



Impossible to perform virtual screenings over the full library

A Hybrid ML-Docking Strategy as a solution

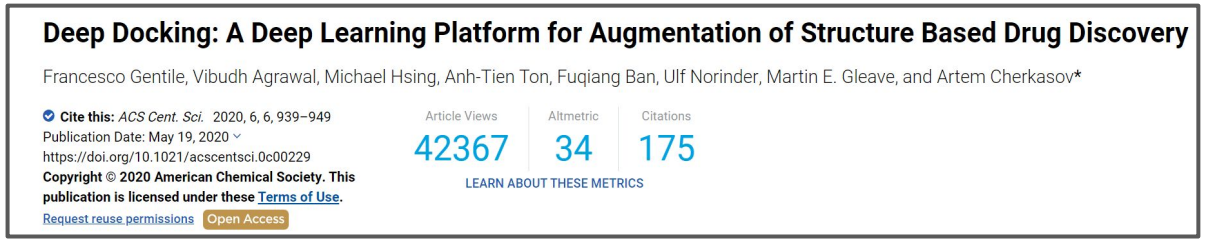
The Proposed Solution

A workflow that does **NOT** dock all compounds. Instead, it uses Machine Learning (ML) as a fast, intelligent "filter."

- **Training Dataset:** Perform expensive docking on a **small random sample.**
- **Train:** Train an ML model on these results to learn what a top molecule looks like.
- **Filter Fast:** Use the fast ML model to classify the **entire** multi-billion compound library.
- **Dock Smart:** Only perform the expensive docking on the small, **high-potential set selected** by the ML.

Studies using this approach

GCNN
(Graph-Convolutional Neural Networks)



DNN

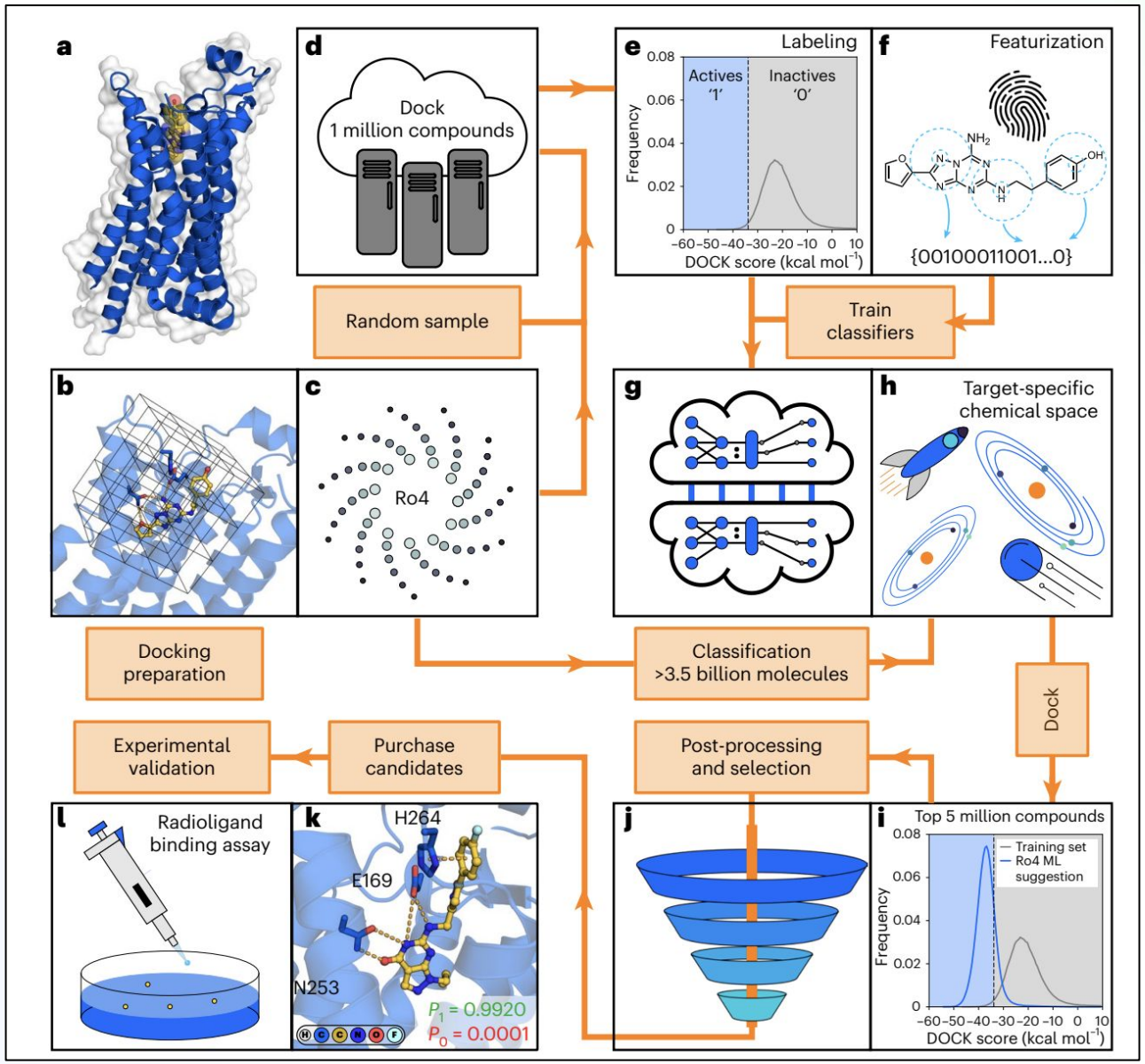
A Hybrid ML-Docking Strategy as a solution

1. Sample & Dock

Dock a random 1-million-compound "training set" from the vast library. Label the top 1% as '**Virtual Actives**' (Class 1) and the rest as '**Inactives**' (Class 0)

5. Validate

The final, top-scoring compounds from the docking run are purchased, synthesized, and experimentally tested in binding assays



2. Train Model

Train a machine learning classifier on the 1M labeled compounds to distinguish 'Actives' from 'Inactives' based on their molecular features

3. Predict Library

Use the fast, trained ML model to make predictions for every single compound in the entire **billion sized** compound library

4. Prioritize & Dock

The ML model selects a much smaller set of predicted high-scorers. **Only** this reduced set is run through the expensive docking

Outline

1. Introduction
2. **Conformal predictor**
3. Benchmarking of conformal predictors
4. Optimized workflow for ultralarge chemical libraries
5. Prospective virtual screen of a multi-billion-scale library
6. Machine learning-guided design of polypharmacology
7. Conclusions

Methodology: Controlling the Error Rate

Conformal Predictor (CP)

A machine learning framework for making predictions with a built-in **measure of certainty or confidence**.

Main Components:

- **Nonconformity Measures:** Quantifies how "different" a new example is from a set of training examples.
- **Confidence Level:** Determined by a user-chosen significance level (ε).

CP is crucial for estimating the **reliability** of predictions, especially for imbalanced datasets where 'Actives' (top 1%) are rare.

Regression

Estimate continuous values (like a dock score) with a statistically valid confidence interval.

Classification

Estimate the probability of the correct class (Active vs. Inactive) while controlling the error rate.

Methodology: Controlling the Error Rate

Conformal Predictor (CP)

The workflow doesn't just give a "yes/no" answer. It uses **Conformal Prediction** to provide a **confidence level** for each prediction.

This is crucial because it allows researchers to **control the error rate**. They can tune the **significance level (ε)** to balance a trade-off:

- **Strict ε (such as 0.01):** Fewer compounds to dock, high confidence, but might miss some hits.
- **Loose ε (such as 0.20):** More compounds to dock, lower confidence, but finds more potential hits.

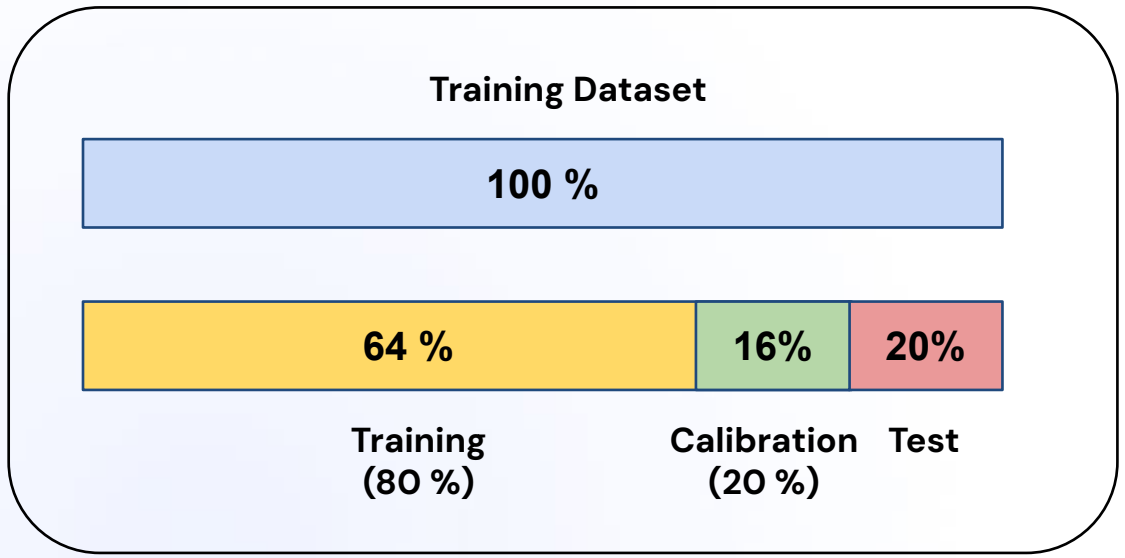
Our Goal

Classify molecules as 'Active' or 'Inactive' while **knowing and controlling the error** tolerated in our estimations.

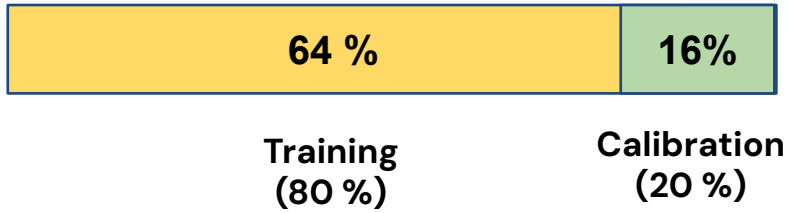
P-Values in CP

CP requires the **p-values** of each **predicted probability** of the ML prediction

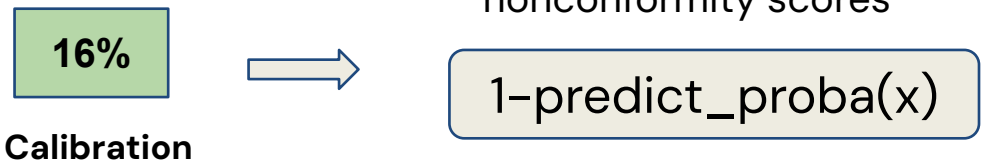
The **p-values** quantify the **statistical evidence** that a test molecule **belongs to a given class** by measuring the rank of its **nonconformity score** against the empirical distribution of scores established by the **calibration set**



Train ML classifier



Generate the ML predicted probabilities distribution



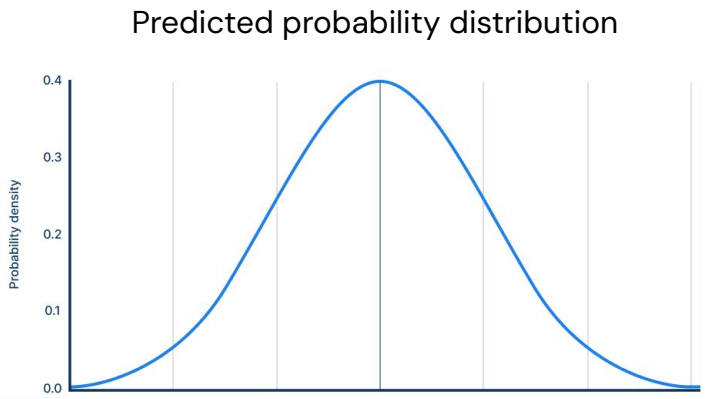
Ex. Probability 0.9 means an nonconf score of 0.1

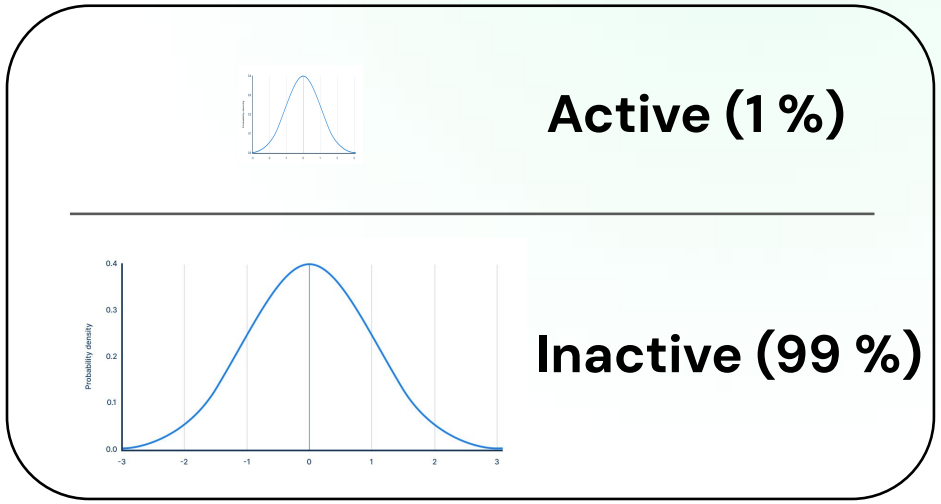
P-Values in CP

Generate the ML predicted probabilities distribution

16%

Calibration

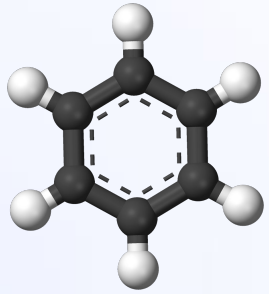




Required exchangeability between the training set and the objective set !!!

20%

Test

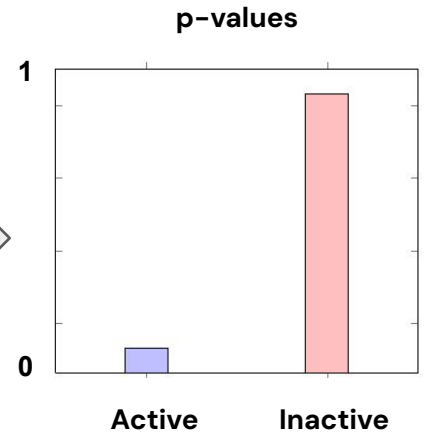


ML classifier

Predicted probabilities of active/inactive

Calibration distributions

P-values of active/inactive



The P-value is then derived by comparing this score to the scores of the Calibration Set. It tells us what fraction of calibration examples were 'stranger' than our current molecule

P-Values in CP

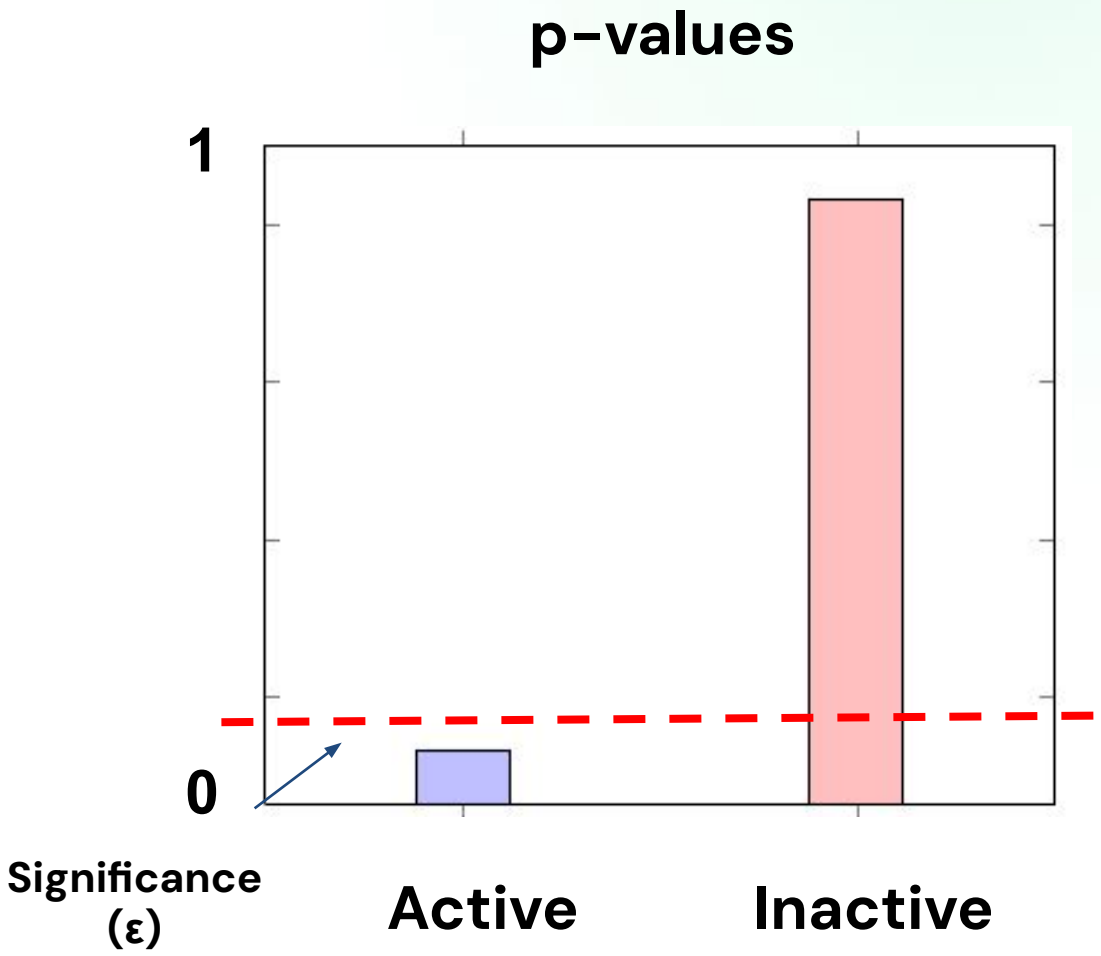
P-Values for Confidence

CP uses p-values to measure confidence. It doesn't just ask "Is this active?" It asks "**How well does this molecule conform to the pattern of known Actives?**"

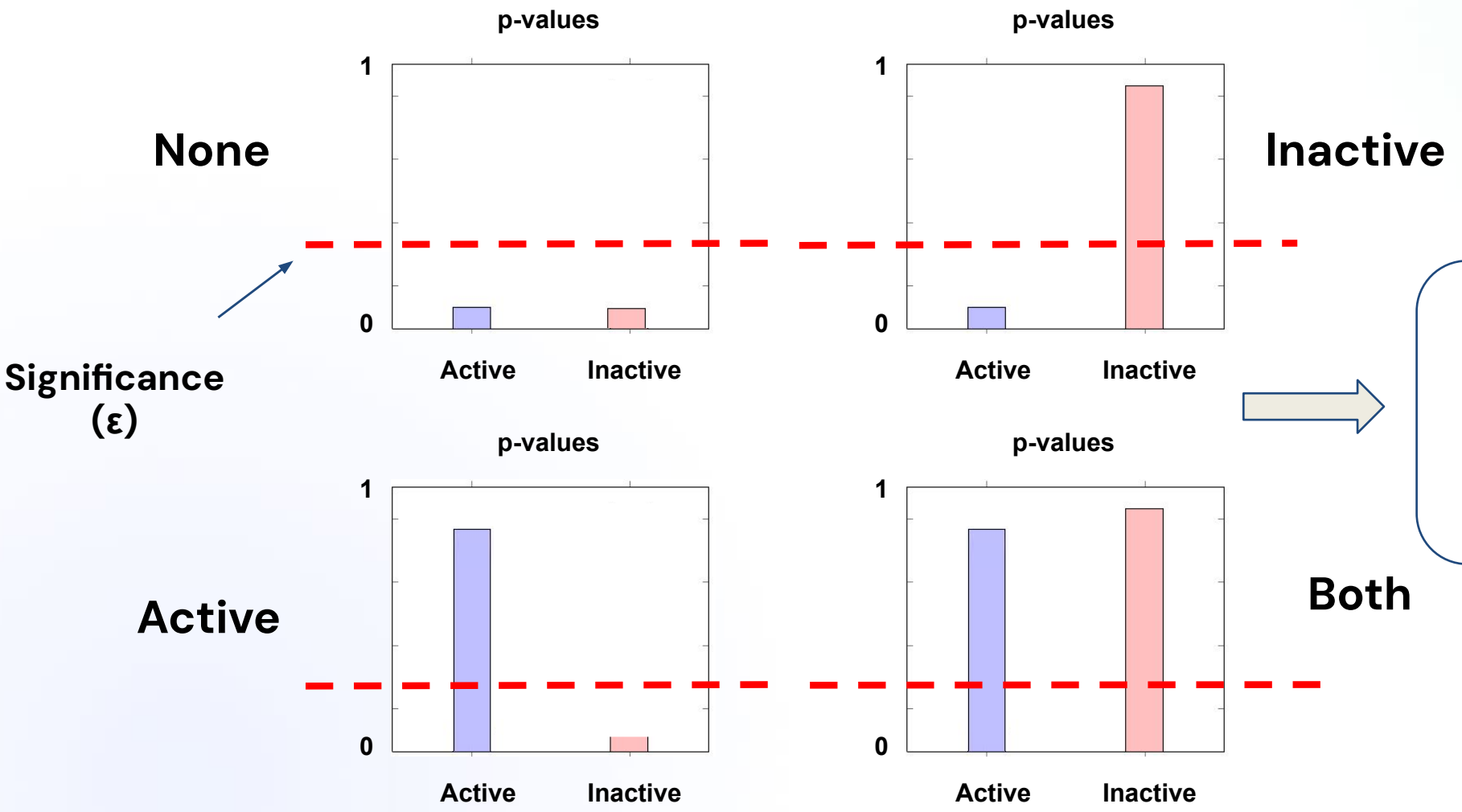
- **High p-value (p > ε):** It fits the profile well **(Keep label)**
- **Low p-value (p < ε):** It is too different/weird **(Reject label)**

We generate two p-values for each molecule:

- **p(Active):** A measure of how compatible the molecule is with the 'Active' class.
- **p(Inactive):** A measure of how compatible the molecule is with the 'Inactive' class



P-Values in CP



How to select the significance?

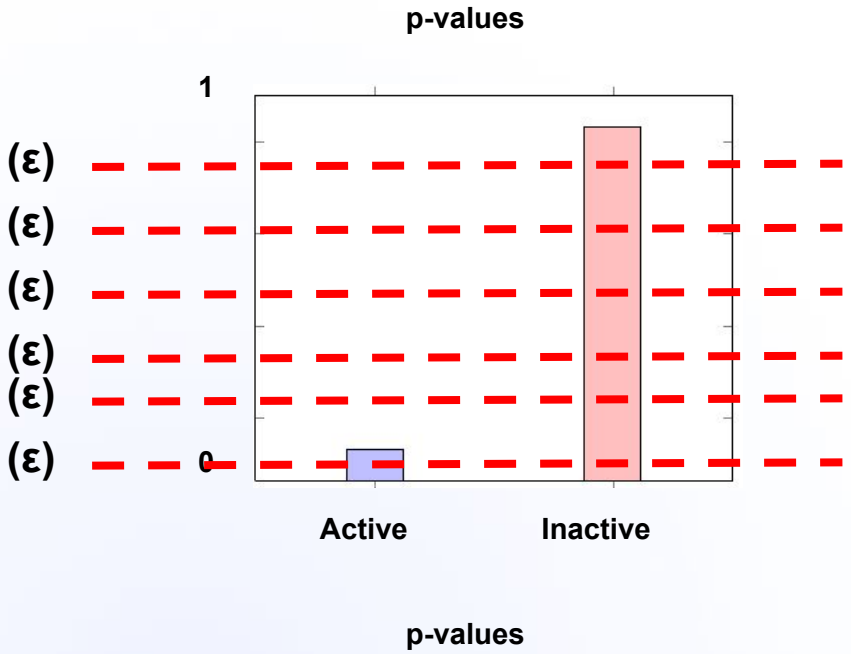
"Looking for the **optimal** significance"

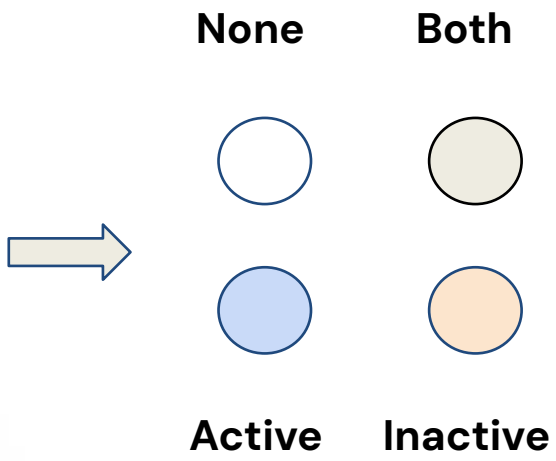
Rank predictions

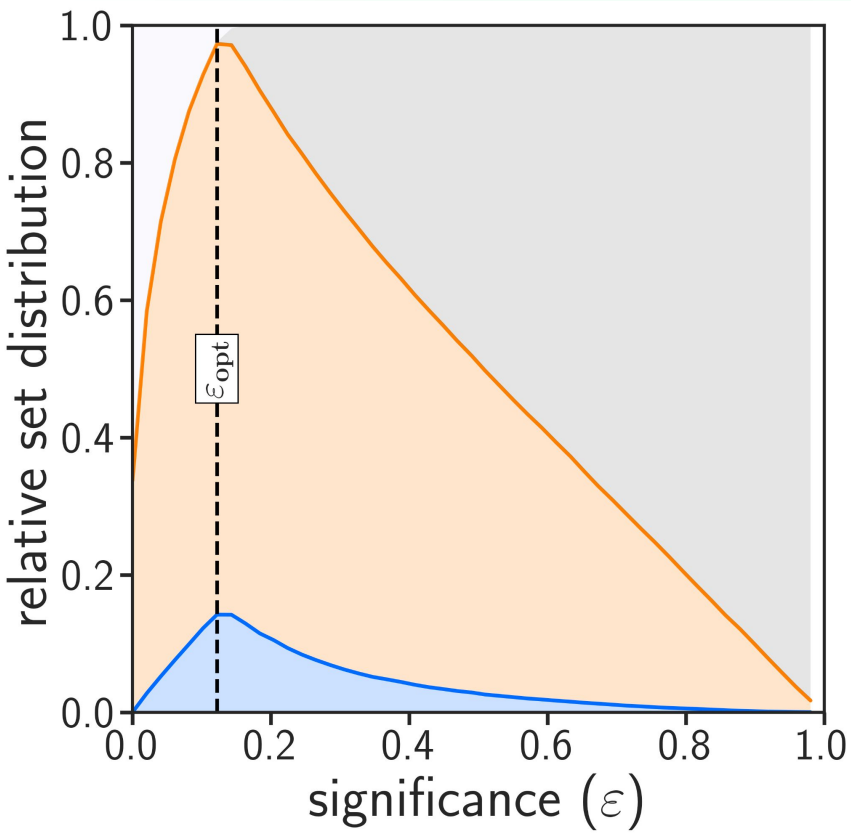
Quality of information
$\Delta P = P_1 - P_0$

P-Values in CP

Finding Optimal significance (ε_{opt})

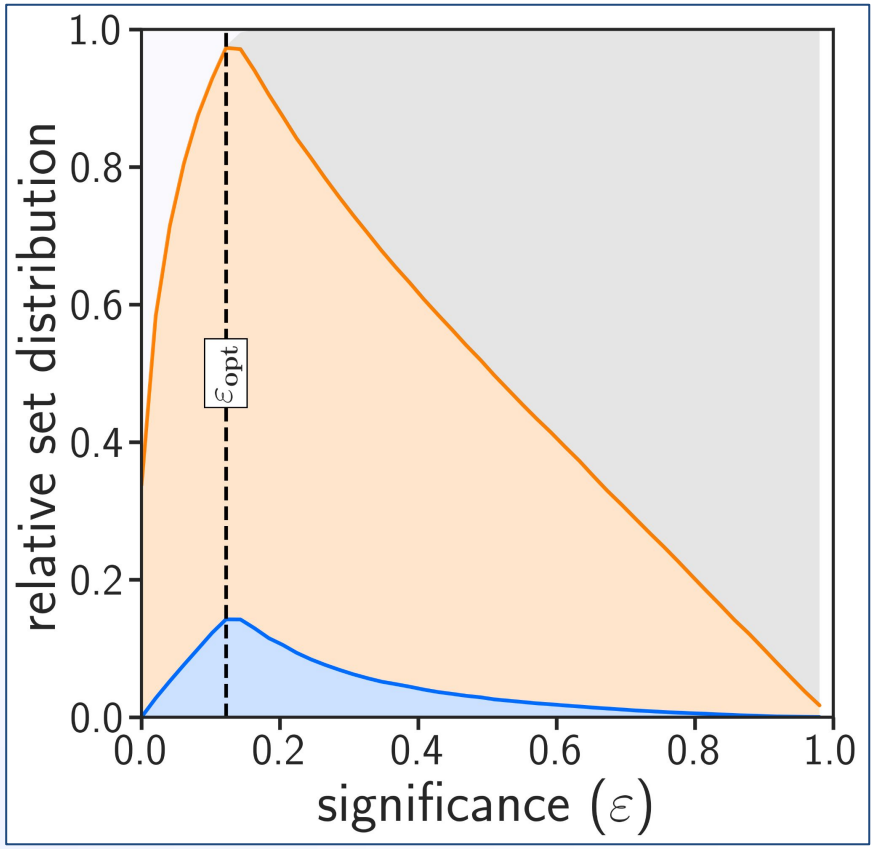


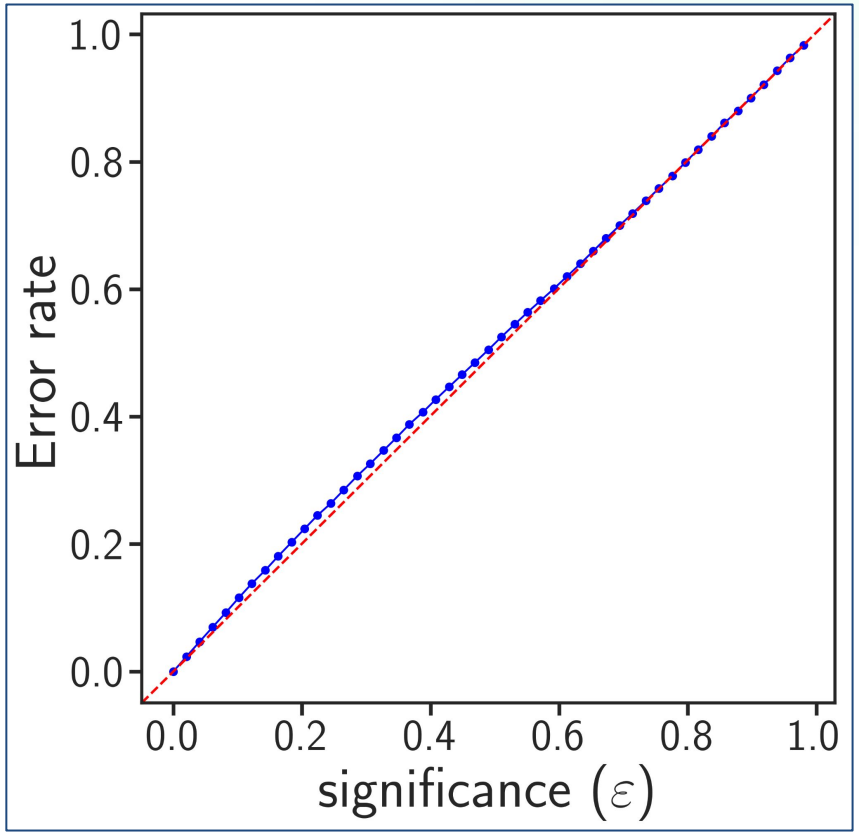




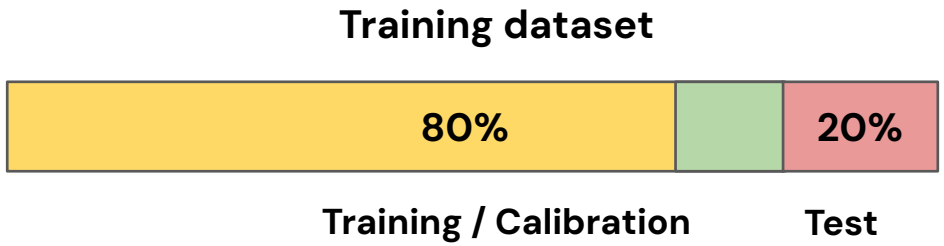
P-Values in CP

Finding Optimal significance (ε_{opt})

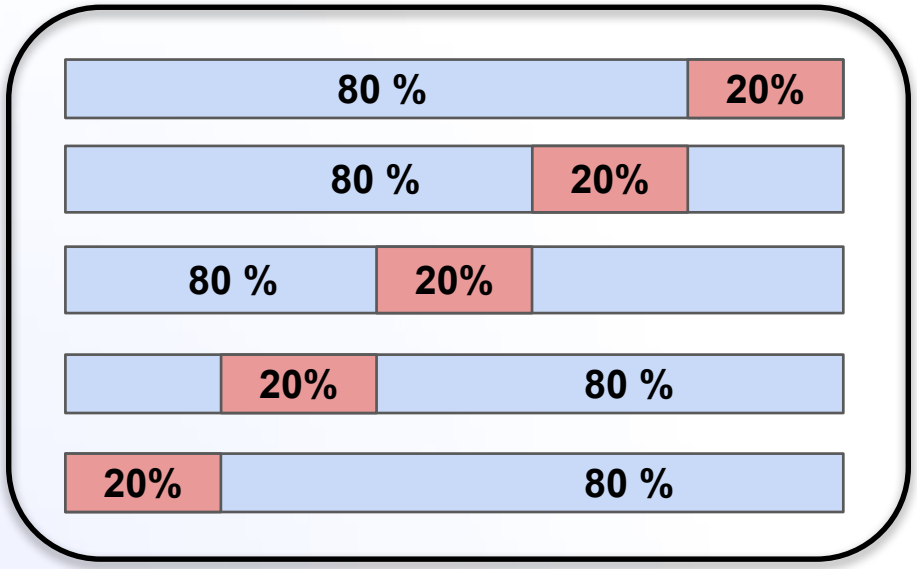




Aggregated Mondrian CP



Splitting is performed keeping the same distribution active/inactives



$P_{0,0}$ $P_{1,0}$
$P_{0,1}$ $P_{1,1}$
$P_{0,2}$ $P_{1,2}$
$P_{0,3}$ $P_{1,3}$
$P_{0,4}$ $P_{1,4}$

(Aggregation)
Median

P_0 P_1

Outline

1. Introduction
2. Conformal predictor
3. **Benchmarking of conformal predictors**
4. **Optimized workflow for ultralarge chemical libraries**
5. **Prospective virtual screen of a multi-billion-scale library**
6. Machine learning-guided design of polypharmacology
7. Conclusions

Outline

3

Benchmarking of conformal predictors

Goal: Identify the optimal workflow using a diverse set of 8 protein targets and **11 M** random compounds

4

Optimized workflow for ultralarge chemical libraries

Goal: Check the workflow in two protein targets using **235 M** compounds docked data

5

Prospective virtual screen of a multi-billion-scale library

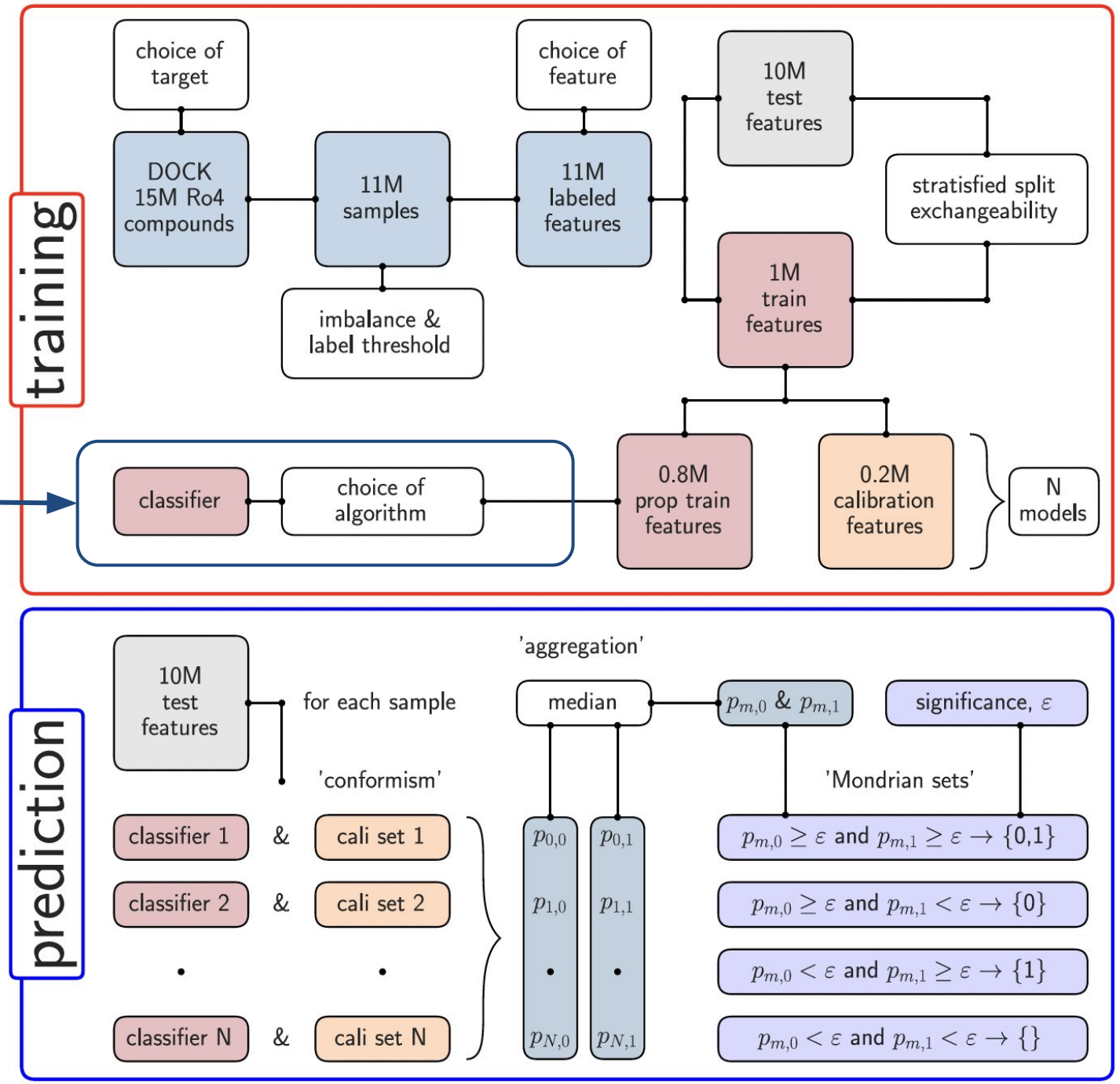
Goal: Apply prospectively the optimal workflow to one protein target using **3.5 B** compounds

Conformal prediction workflow to identify optimal parameters

Find the best ML model

Ro4:

- MW < 400 Da
- clogP < 4



Methodology: Finding the Best ML Model

Benchmarking Classifiers

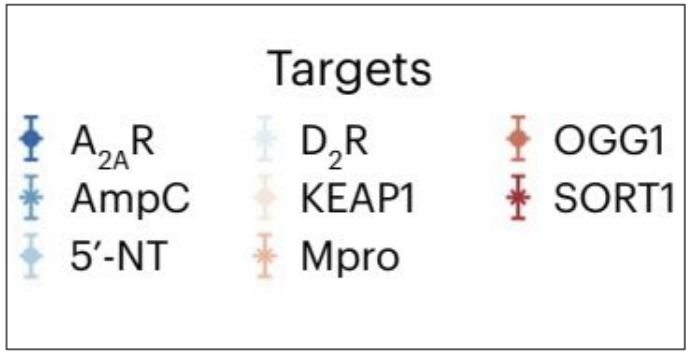
We benchmarked models on 8 different protein targets to find the best-performing combination.

ML Algorithms Tested:

- CatBoost (a gradient boosting method)
- Deep Neural Networks (DNN)
- RoBERTa (a transformer-based model)

Eight diverse protein targets

Different types of protein folds, binding sites, protein-ligand interactions, and ligand chemotypes



Methodology: Finding the Best ML Model

Benchmarking Classifiers

We benchmarked models on 8 different protein targets to find the best-performing combination.

ML Algorithms Tested:

- CatBoost (a gradient boosting method)
- Deep Neural Networks (DNN)
- RoBERTa (a transformer-based model)

Benchmarking Features

Molecular Features (Descriptors) Tested:

- Morgan2 Fingerprints
- Continuous Data-Driven Descriptors (CDDD)
- Transformer-based Descriptors

The Winner: CatBoost + Morgan2 Fingerprints

This combination achieved the optimal balance of precision, sensitivity, and, critically, **computational speed and storage of molecular descriptors**

Benchmarking of conformal predictors

$A_{2A}R$ validation

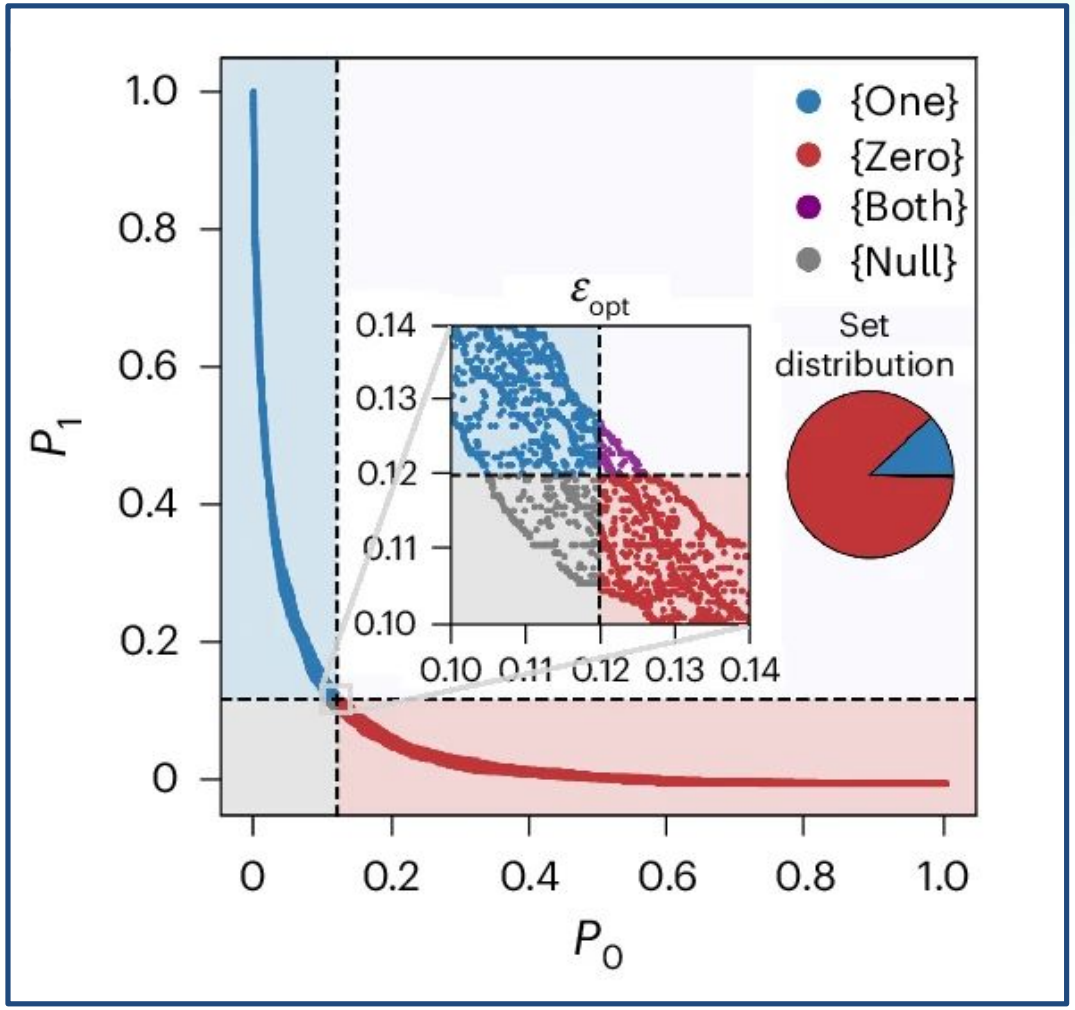
Virtual actives (blue, 1 class)

Virtual inactives (red, 0 class)

Both (purple, 1 or 0 class)

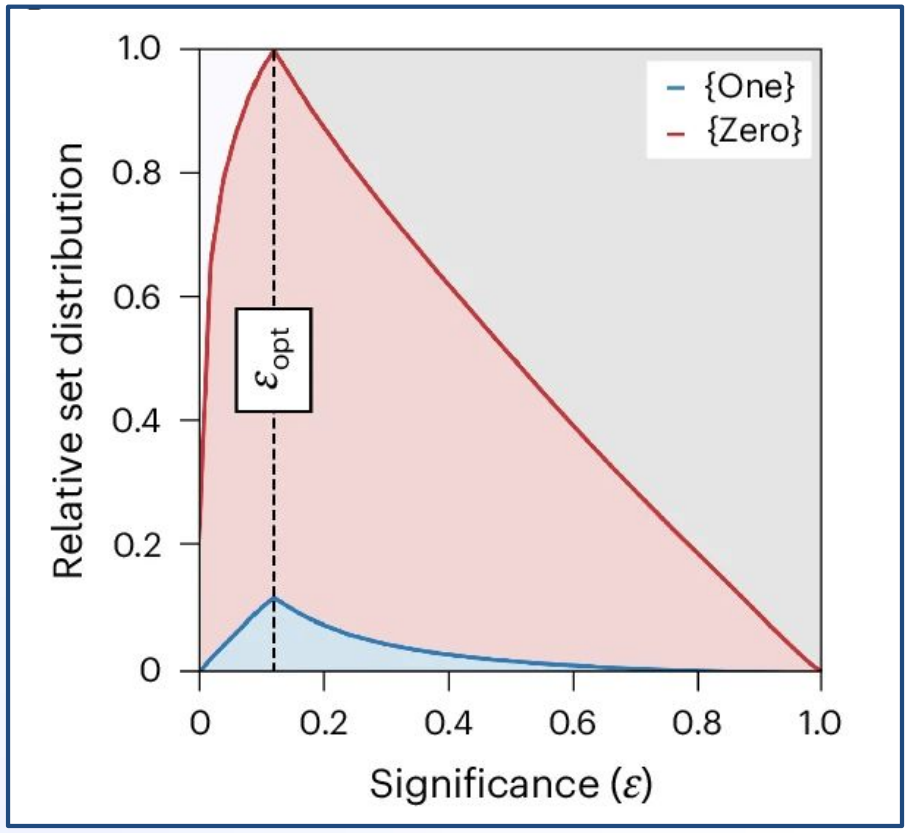
Null (gray, no class assignment)

Relative fractions of each set are represented by a pie chart

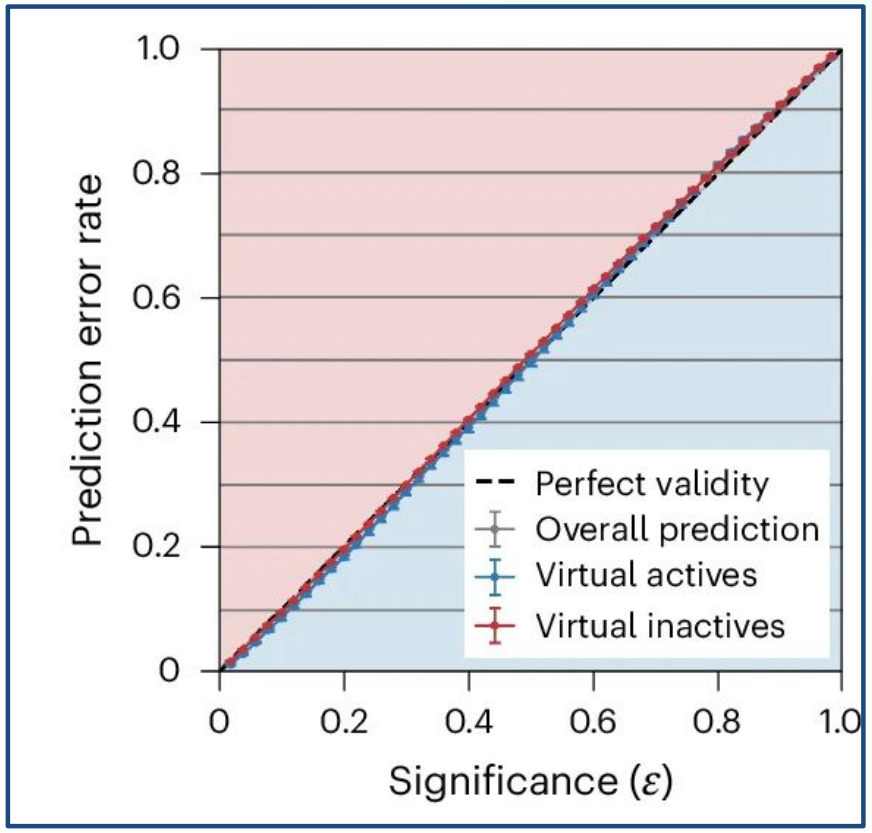


Benchmarking of conformal predictors

$A_{2A}R$ validation



The $A_{2A}R$ test set divided into four prediction sets depending on the significance level.

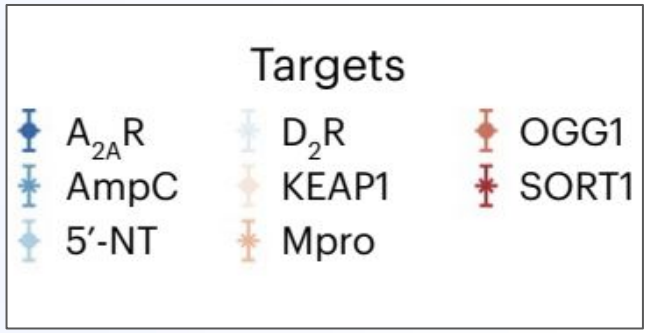


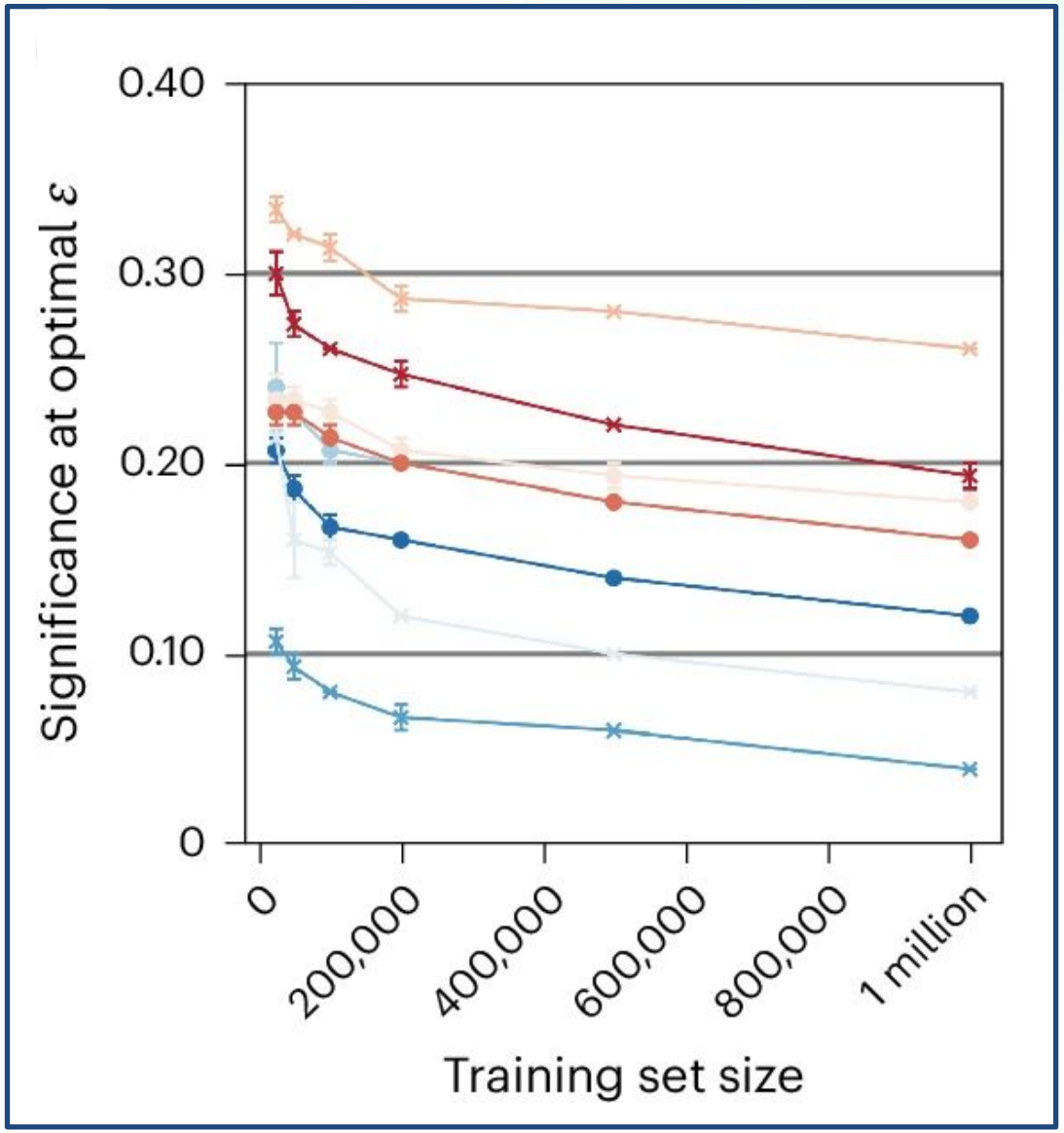
There was a close agreement between the significance value and the prediction error rate.

Benchmarking of conformal predictors

Training Set Size Matters

- The team explored training set sizes from 25,000 to 1 million compounds
- **Key Finding:** Error at optimal significance (ε) improved asymptotically as the training set size increased
- The models' performance **stabilized at 1 million** compounds
- This size was therefore established as the standard for all future screens



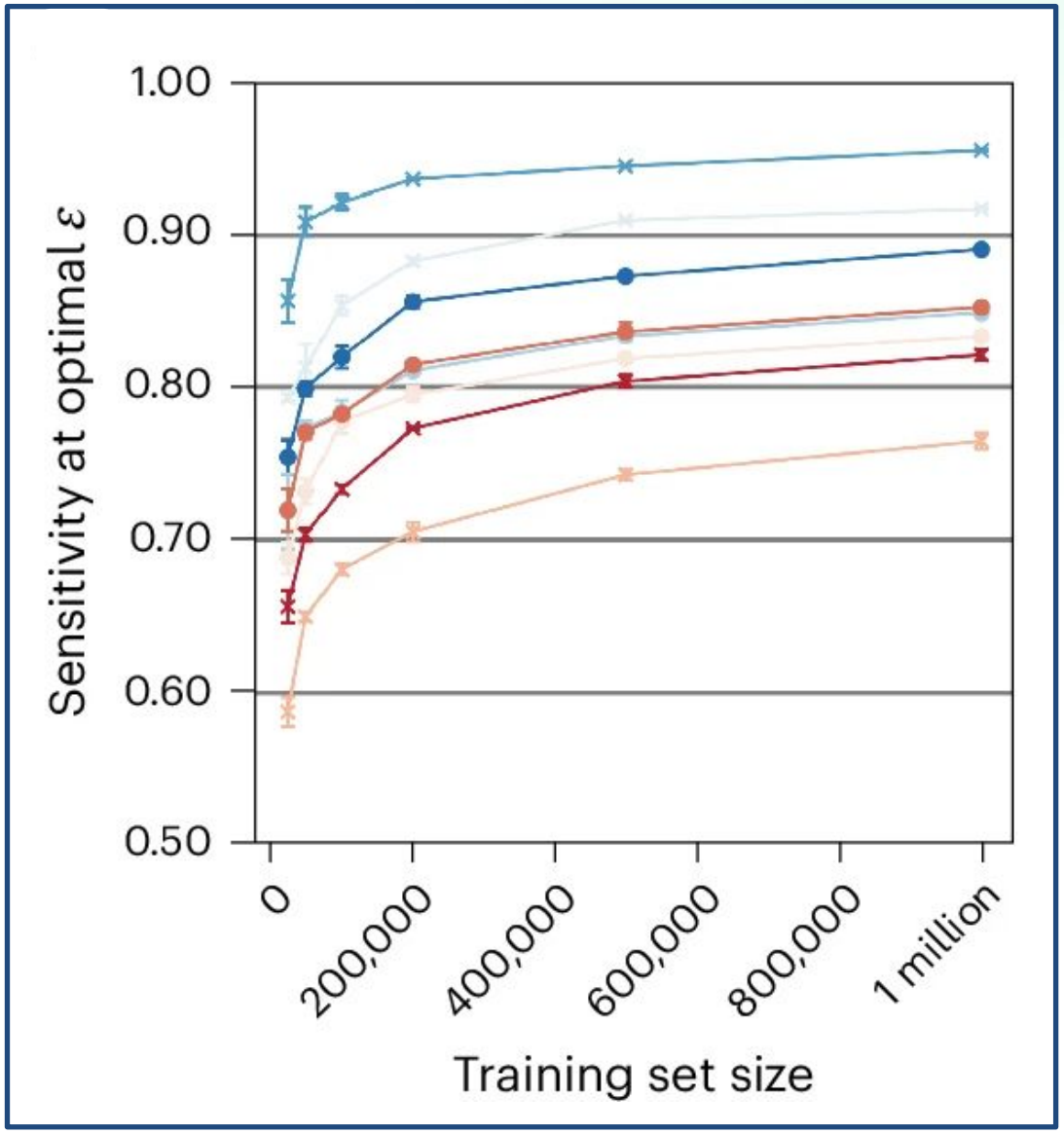


Benchmarking of conformal predictors

Training Set Size Matters

- The team explored training set sizes from 25,000 to 1 million compounds
- **Key Finding:** Performance (**Sensitivity** & Precision) improved significantly as the training set size increased
- The models' performance **stabilized at 1 million** compounds
- This size was therefore established as the standard for all future screens

$$\text{Sensitivity} = \frac{\text{TP}}{\text{AP}}$$

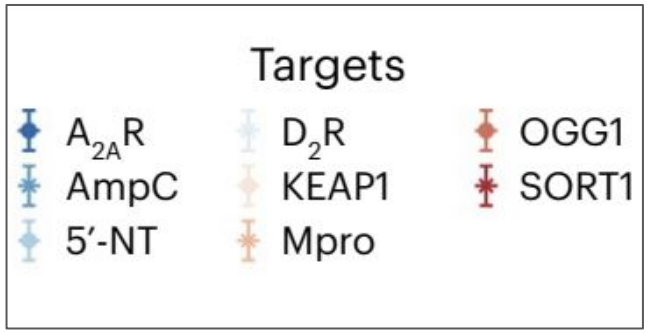


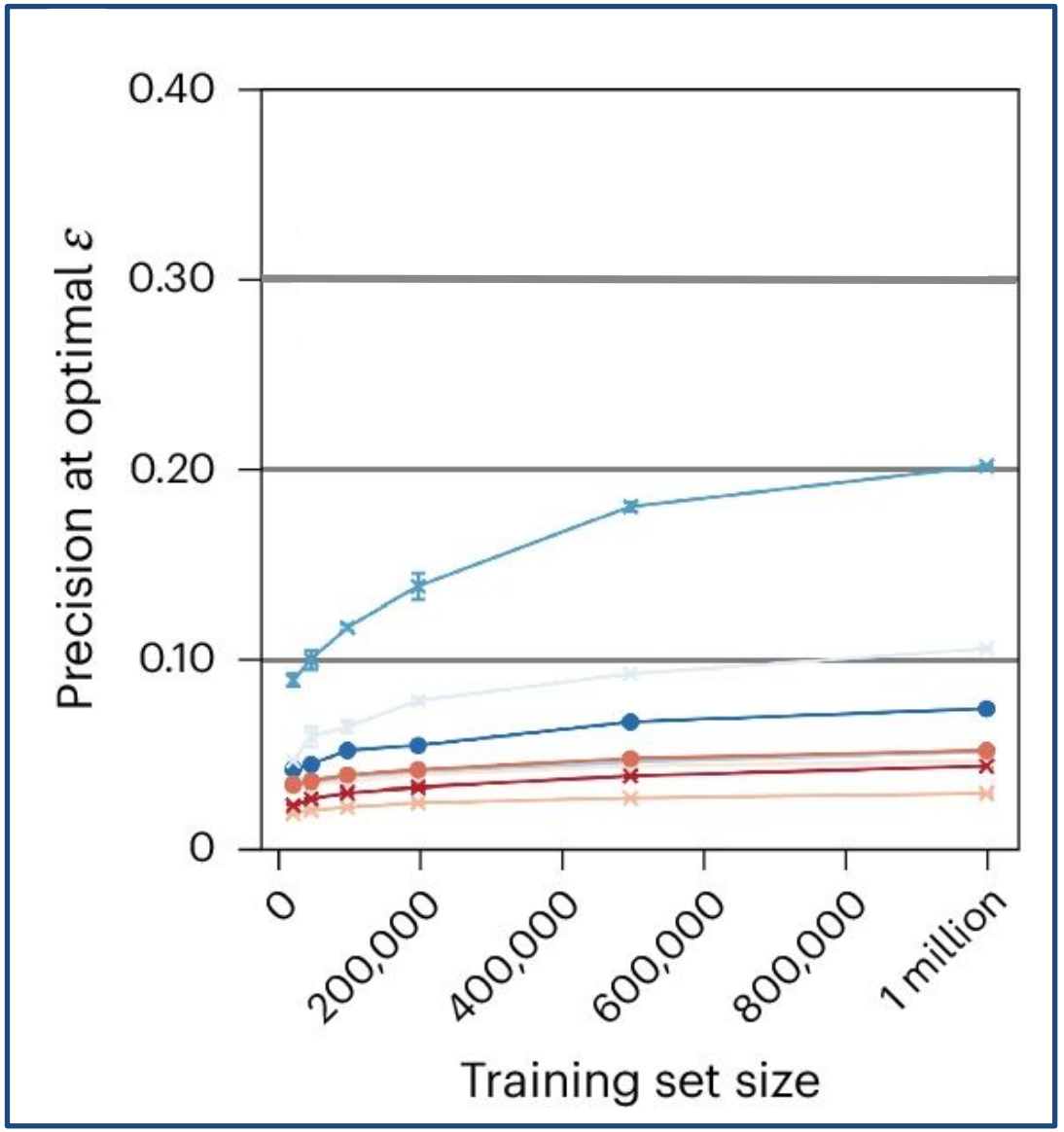
Benchmarking of conformal predictors

Training Set Size Matters

- The team explored training set sizes from 25,000 to 1 million compounds
- **Key Finding:** Performance (Sensitivity & **Precision**) improved significantly as the training set size increased
- The models' performance **stabilized at 1 million** compounds
- This size was therefore established as the standard for all future screens

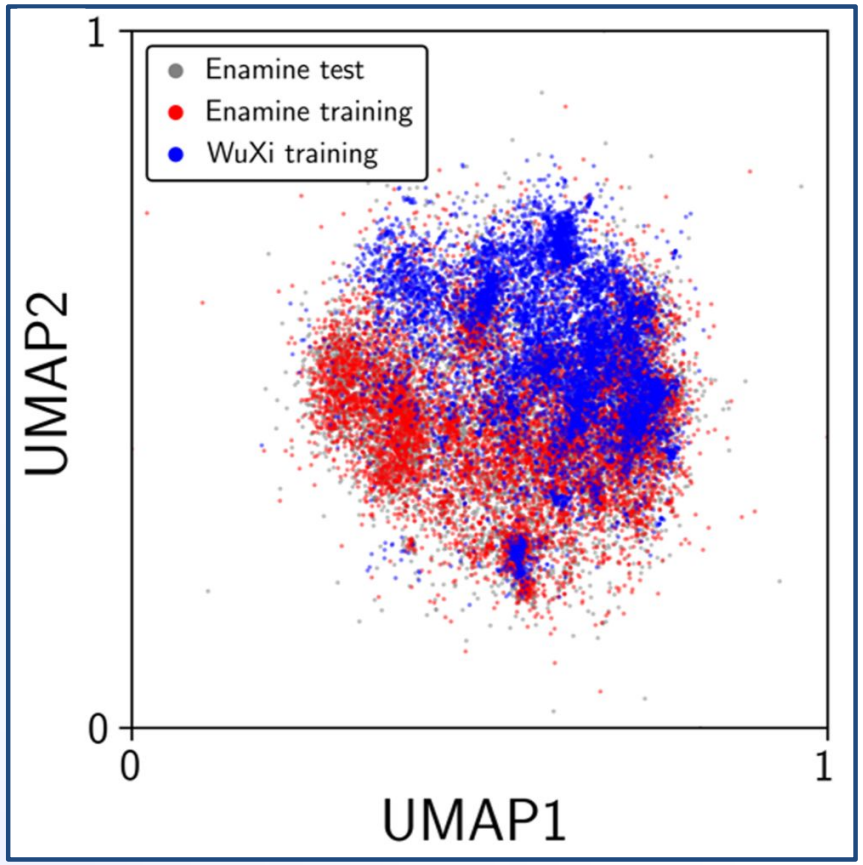
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$



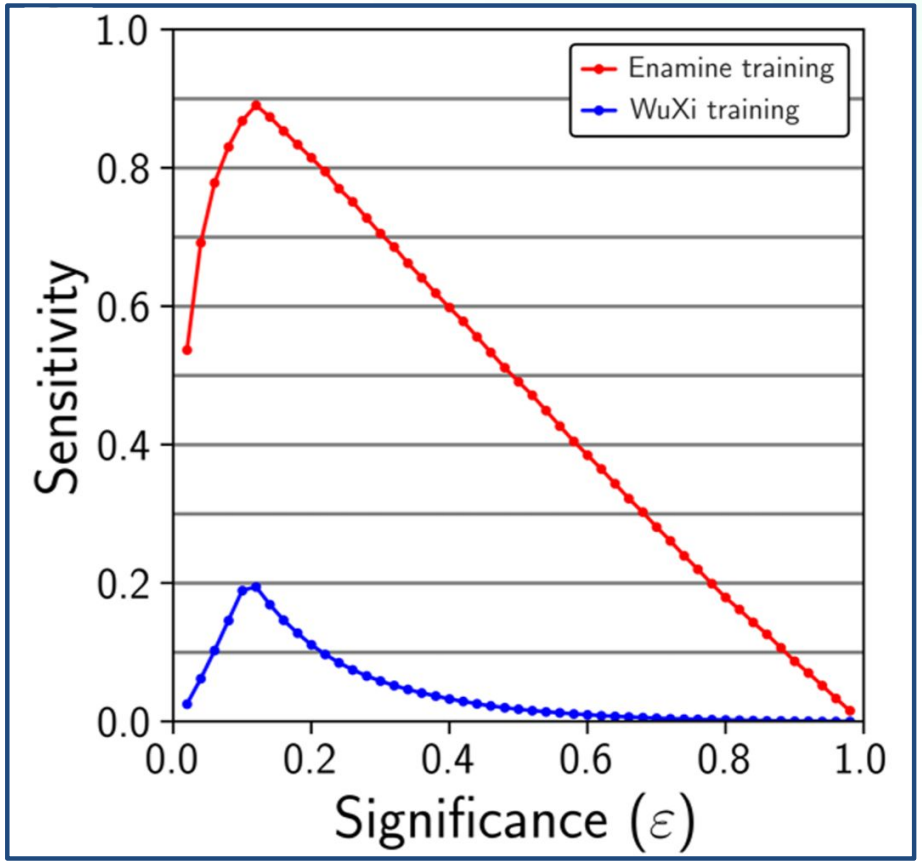


Analysis: Non-exchangeable datasets

Structural similarity between non-exchangeable datasets and conformal predictor performance



Similarity: Two-dimensional unsupervised UMAP projection illustrates the chemical relationships in high-dimensional feature space



Confidence: Difference in sensitivity values obtained from conformal predictors trained on one million exchangeable (red) and one million non-exchangeable (blue) molecules as a function of the significance value (ε)

Outline

3

Benchmarking of conformal predictors

Goal: Identify the optimal workflow using 11 M random compounds from Enamine REAL chemical space.

4

Optimized workflow for ultralarge chemical libraries

Goal: Check the workflow in two targets using 235 M compounds docked data.

5

Prospective virtual screen of a multi-billion-scale library

Goal: Apply prospectively the optimal workflow using 3.5 B compounds from the Enamine REAL chemical space.

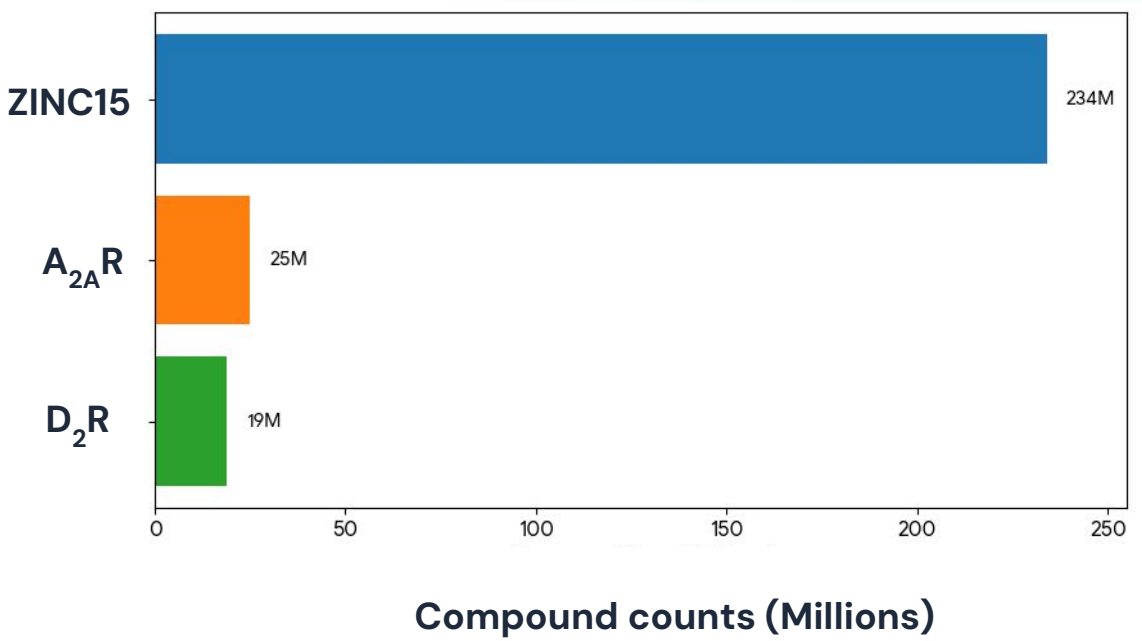
The New Problem: Scaling to Ultralarge Libraries

Initial Workflow (on 235M)

- **D_2R**: An **homology model** of the active **D_2R** was constructed using a cryo-EM structure of the **D_3** dopamine receptor
- **$A_{2A}R$:** An antagonist-bound **crystal structure**

The standard CP workflow (at ε_{opt}) on the **ZINC15** library was a success, but still resulted in a large number of hits:

- **$A_{2A}R$:** Reduced 234M compounds to **25 million**.
- **D_2R:** Reduced 234M compounds to **19 million**.



This is viable for 235M, but not for billion sized libraries

Docking even 10% of a 3.5B library (350M) is computationally demanding

Solution 1: Tuning Significance (ε)

Stricter Threshold, Better Scores

We hypothesized that a stricter significance level (a smaller ε) would reduce the set size and enrich for higher-confidence hits.

Result: This worked. As ε decreased, the docking score distribution shifted dramatically toward better energies.

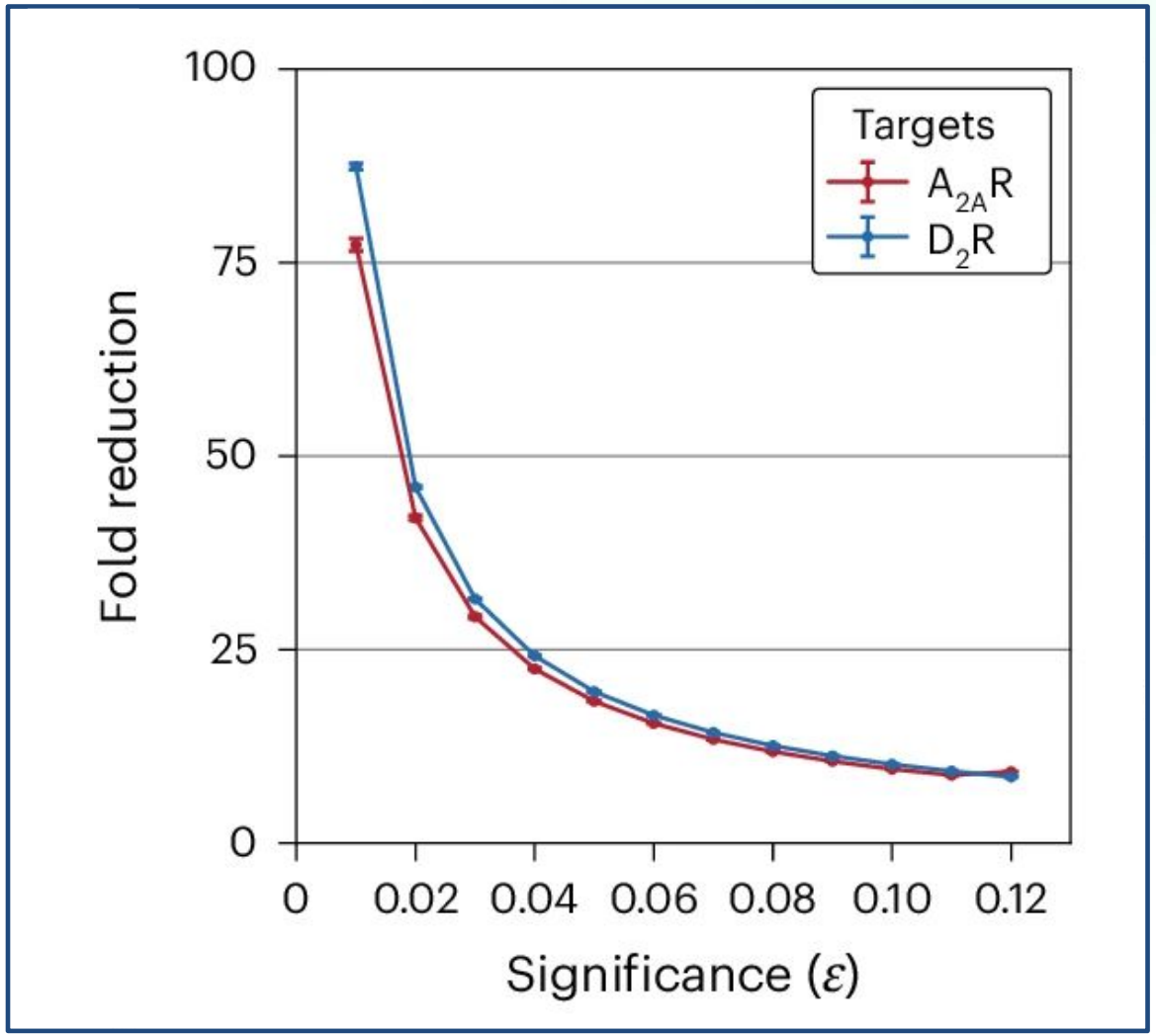
- **D_2R Example:**

  Training Set (Avg): -23.8 kcal/mol

  At ε = 0.08 ($ε_{opt}$): -47.7 kcal/mol

  At ε = 0.01 (Strictest): -50.9 kcal/mol

At ε=0.01, the library was reduced to ~3M compounds, still capturing 64-80% of the top 10,000 hits.



Solution 1: Tuning Significance (ε)

Stricter Threshold, Better Scores

We hypothesized that a stricter significance level (a smaller ε) would reduce the set size and enrich for higher-confidence hits.

Result: This worked. As ε decreased, the docking score distribution shifted dramatically toward better energies.

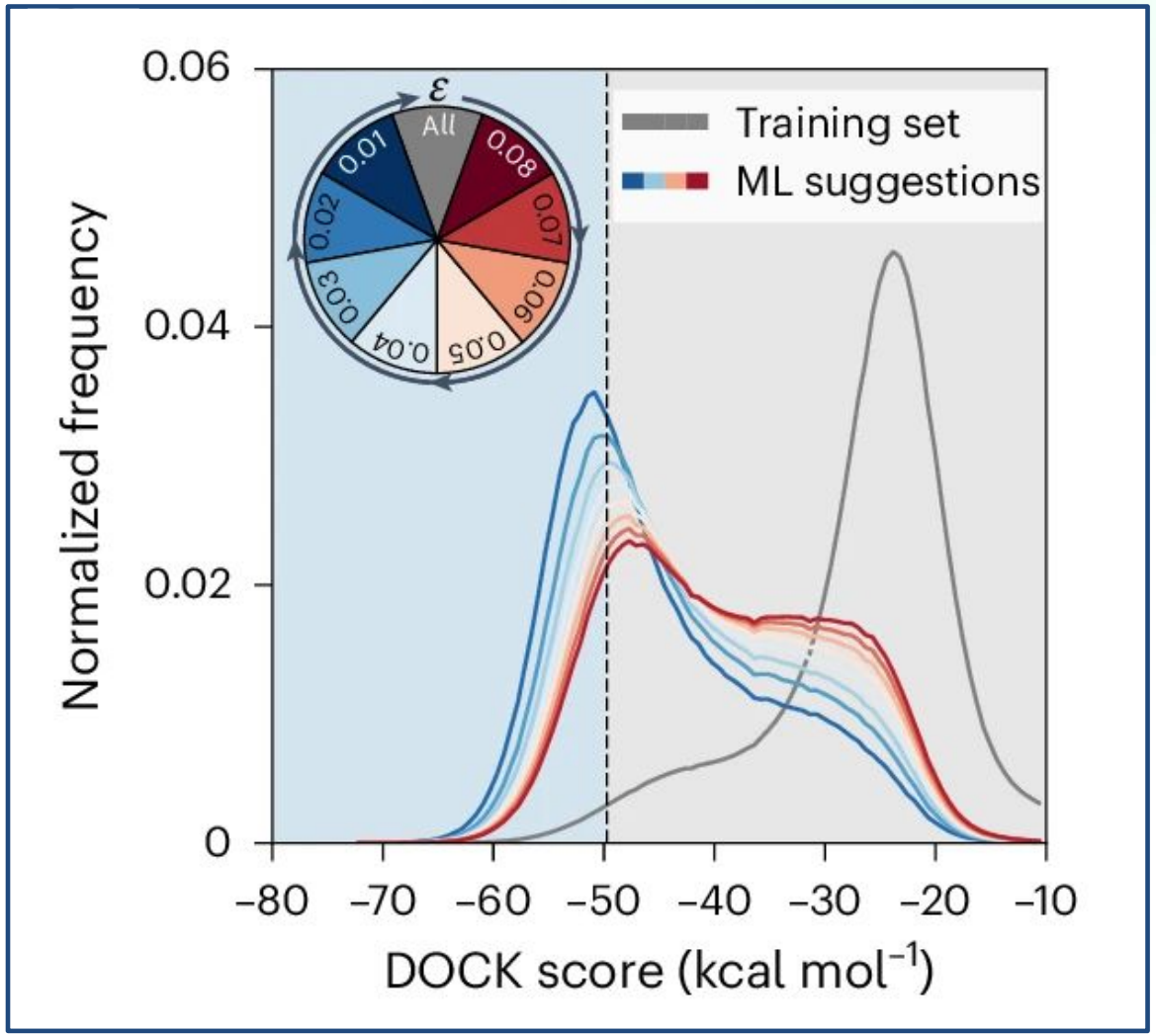
- **D_2R Example:**

  Training Set (Avg): -23.8 kcal/mol

  At ε = 0.08 ($ε_{opt}$): -47.7 kcal/mol

  At ε = 0.01 (Strictest): -50.9 kcal/mol

At ε=0.01, the library was reduced to **~3M** compounds, still **capturing 64-80%** of the top 10,000 hits.



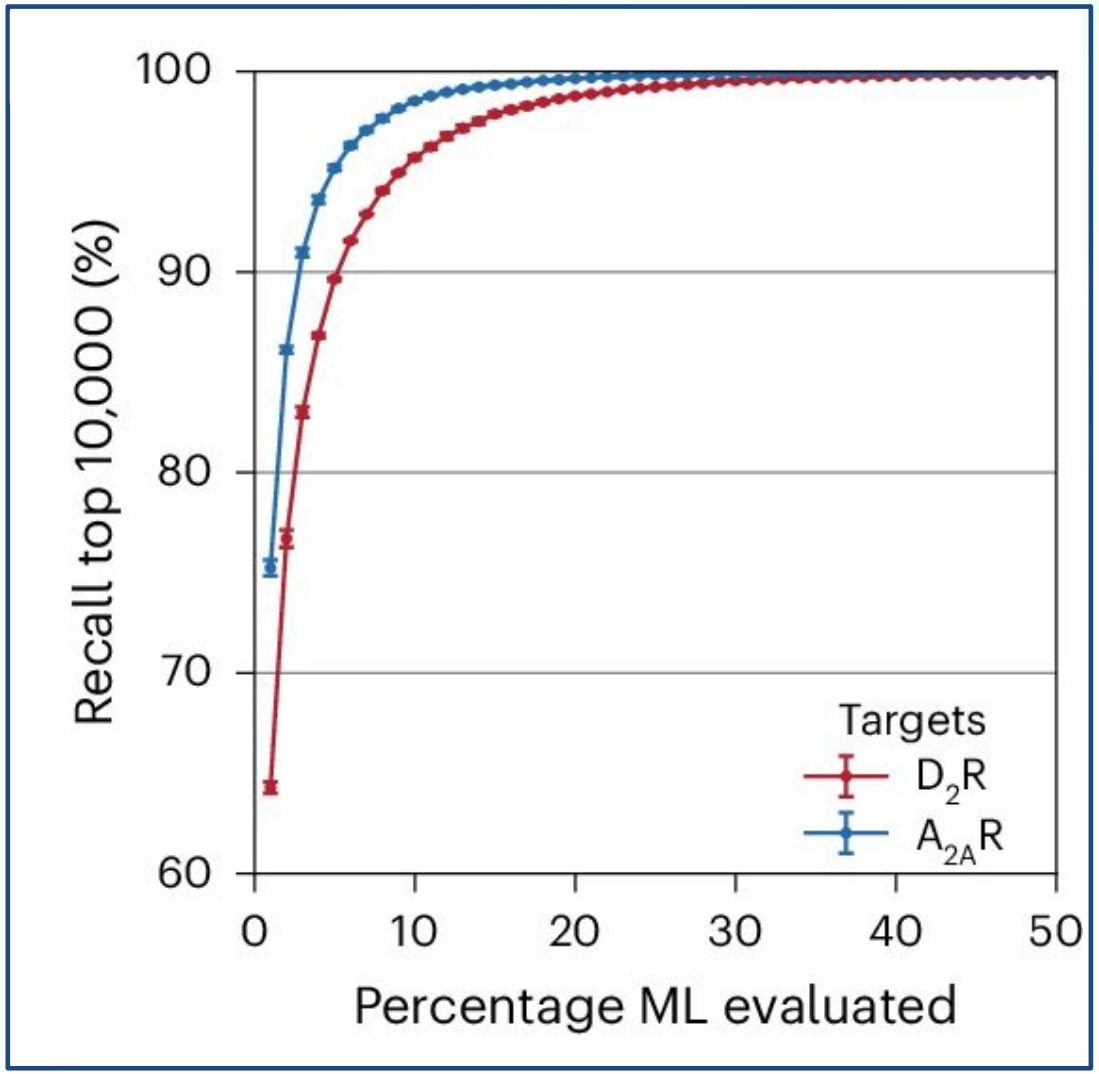
Solution 2: Prioritizing Candidates

Quality of Information

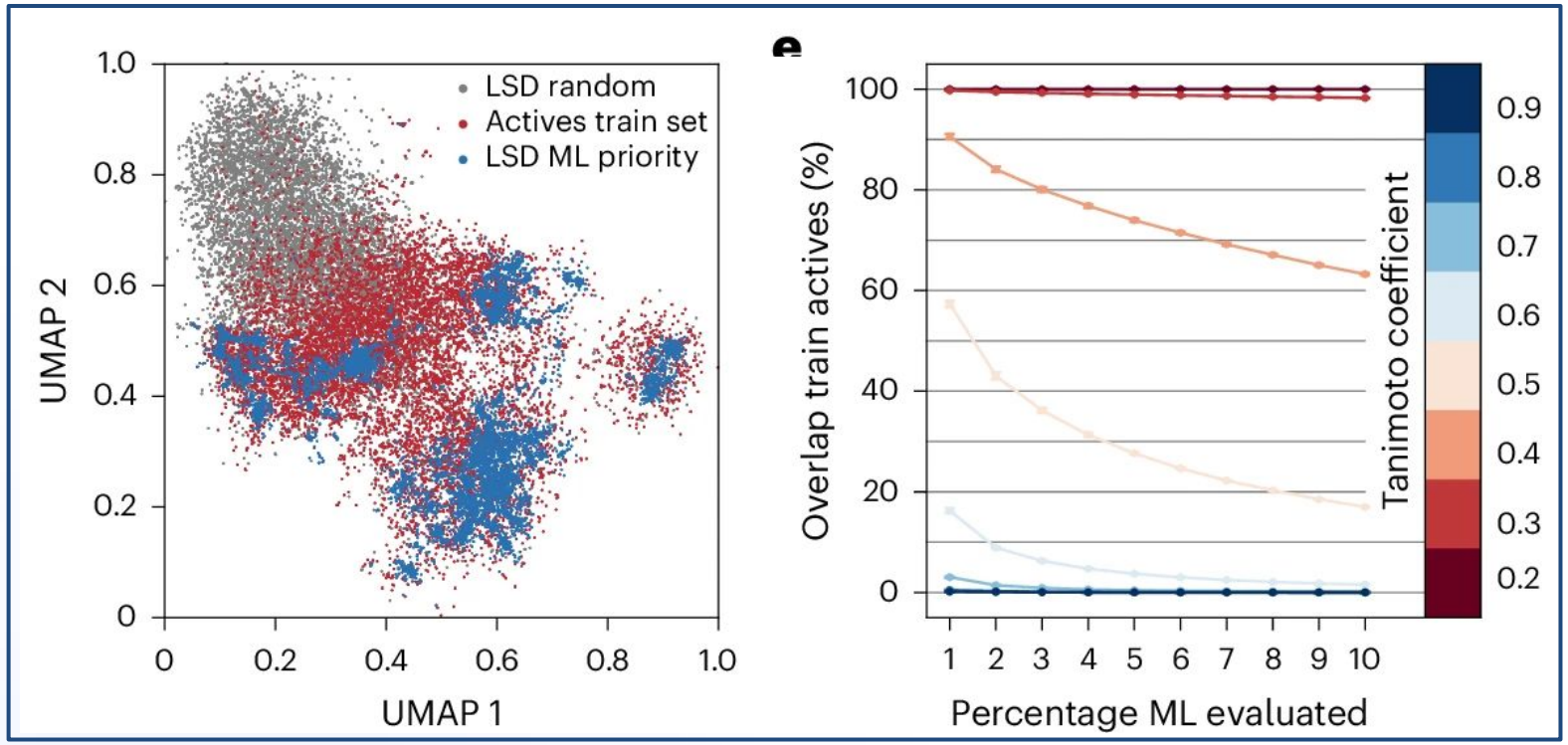
An alternative is to rank by "quality of information"

$$\Delta P = P_1 - P_0$$

This identifies **>90%** of the top 10k hits by evaluating only the **top 3–5%** of the ML's predictions.



Analysis: Similarity vs. Diversity



Similarity: Two-dimensional unsupervised UMAP projection. The molecules in which the predictor had higher confidence generally showed greater structural similarity to actives from the training set

Confidence: Molecules with higher confidence (top 1-10% evaluated) show much greater Tanimoto similarity to the training set actives.

What about Diversity?

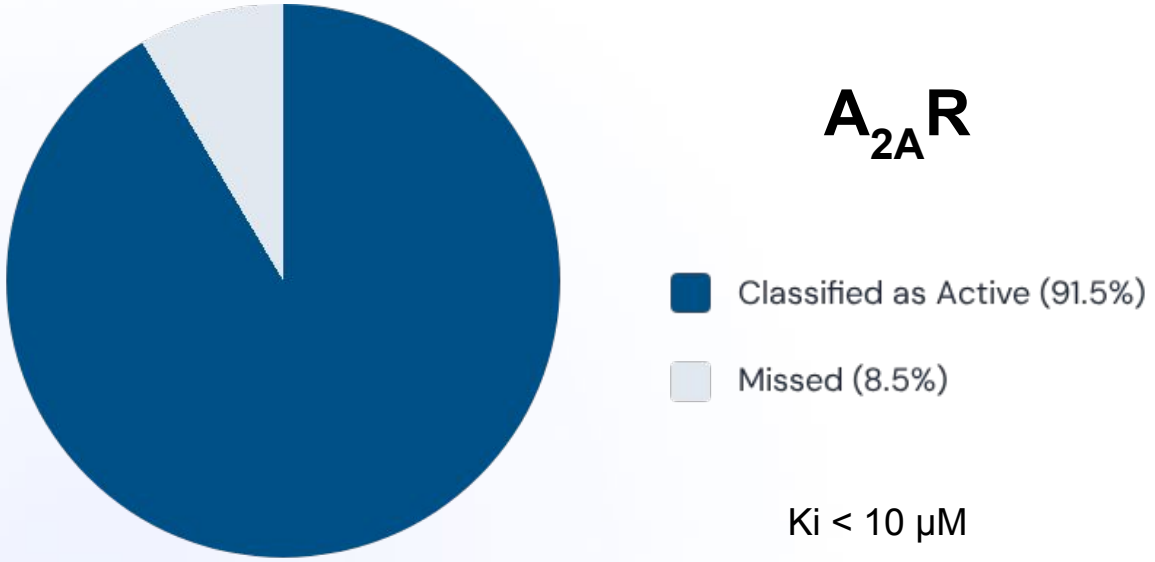
A key concern is "scaffold collapse" finding only minor variations of the same hits.

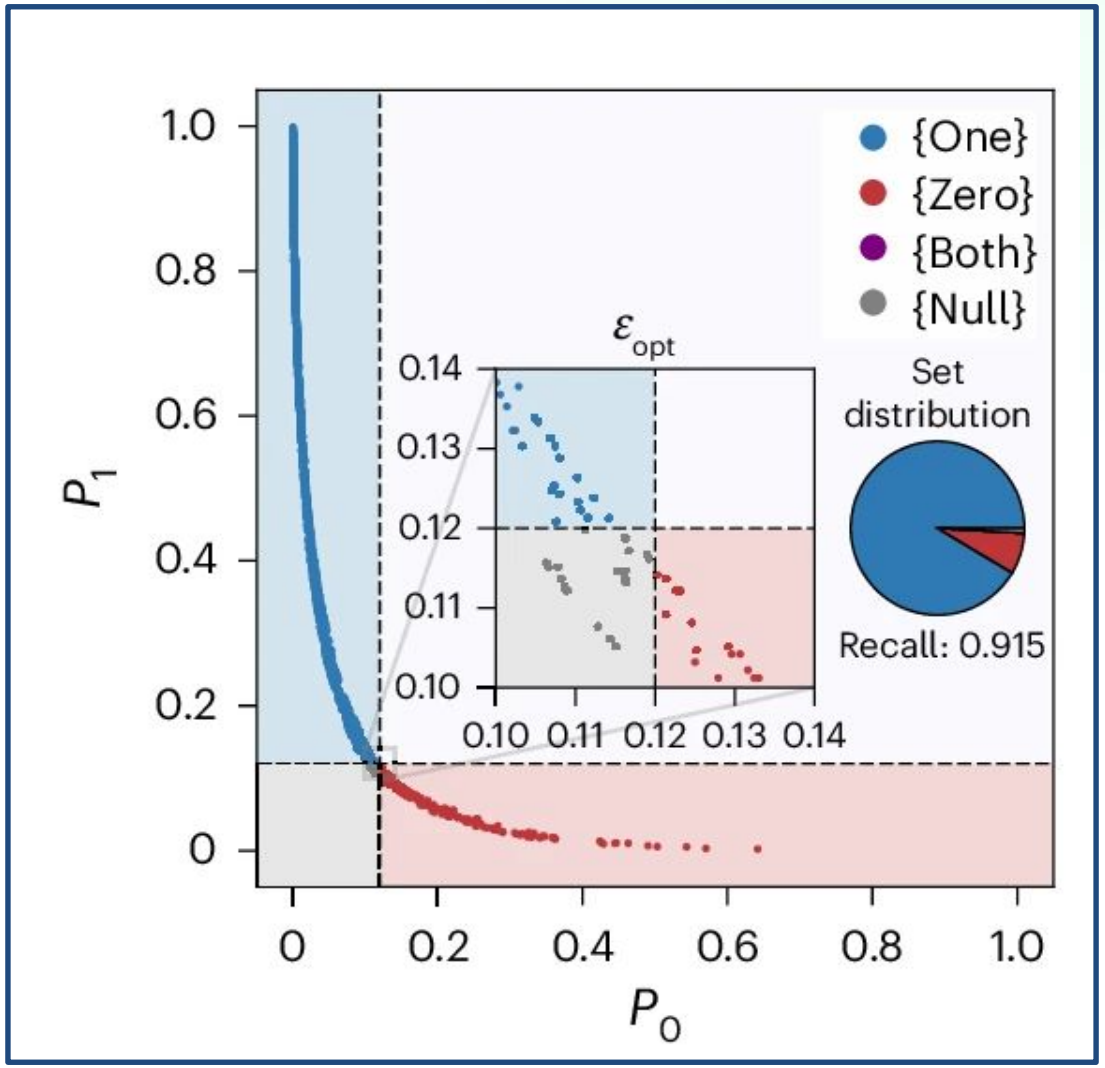
- The ML approach found fewer unique Bemis-Murcko scaffolds in the top 1% (13% vs. 23% from full docking).
- **However:** Pairwise Tanimoto analysis showed the compounds were **not significantly less diverse**.

Final Validation: Finding Known Actives

Test: Can a model trained **only on docking scores** find **experimentally confirmed** ligands from ChEMBL?

Result: The model correctly classified **91.5%** ($A_{2A}R$) and **86% (D_2R)** of known ligands as virtual actives, validating the workflow.





Outline

3

Benchmarking of conformal predictors

Goal: Identify the optimal workflow using 11 M random compounds from Enamine REAL chemical space.

4

Optimized workflow for ultralarge chemical libraries

Goal: Check the workflow in two targets using 235 M compounds docked data.

5

Prospective virtual screen of a multi-billion-scale library

Goal: Apply prospectively the optimal workflow using 3.5 B compounds from the Enamine REAL chemical space.

The 3.5B Compound Screen

The Library

A prospective screen of 3.5 Billion compounds from the **Enamine REAL space** database was initiated **D_2R** target.

Ro4:

- MW < 400 Da
- clogP < 4

The Prediction

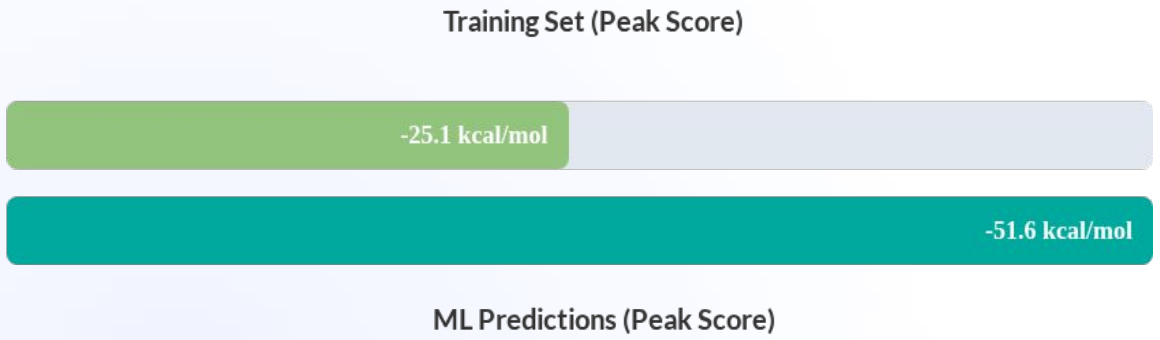
A strict significance level (**ε=0.005**) was set, and the ML model predicted **~25 million** virtual actives.

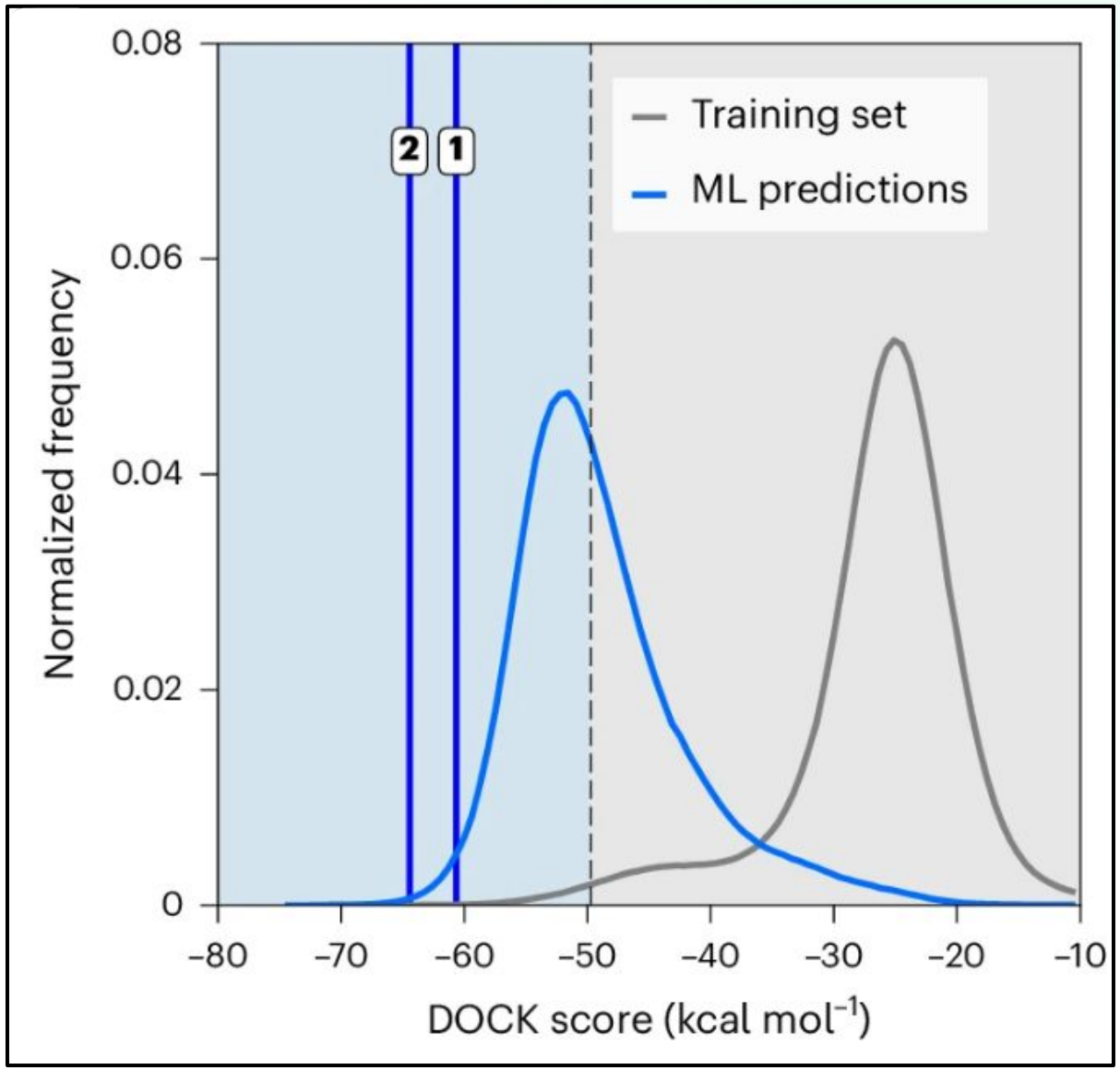
The Docking

The compounds were ranked by "quality of information", and the **top 5 million** were prioritized for full docking calculations.

Prospective virtual screen of a multi-billion-scale library

For **D_2R**, the docking score distribution shifted dramatically. The **5M prioritized** compounds showed a **49-fold enrichment** of virtual actives (scores better than -49.7 kcal/mol)





A Massive Reduction in Cost

Efficiency Gains

The ML-guided workflow achieved a 568-fold reduction in compute cost versus explicitly docking the full 3.5 billion compound library.

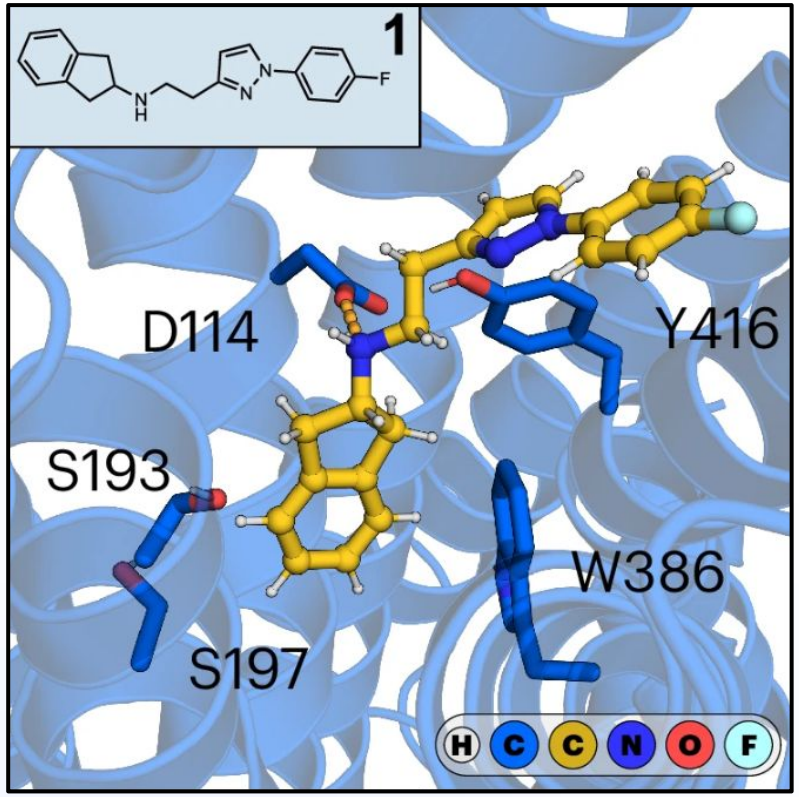
- **700-fold Library Reduction:** 3.5B compounds down to 5M docked.
- **ML Cost:** ~2,500 core-hours for docking, training and prediction.

568x

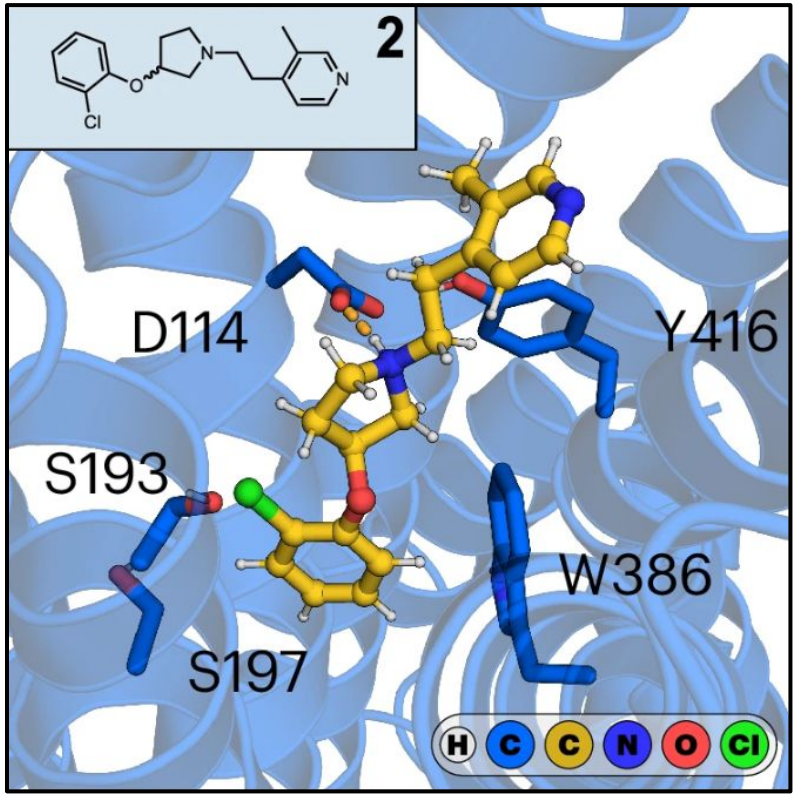
Compute Cost Reduction

Experimental Validation: D_2R Agonists Found

31 top-ranked compounds were **synthesized** and **tested**



Compound 1: A novel D_2R ligand was identified and validated, showing a K_i value of **3.0 µM**.



Compound 2: A second validated hit with a distinct chemical structure showed a K_i value of **3.8 µM**.

Radioligand displacement and affinity values (Ki)

Both compounds were **functionally tested** quantifying D_2R-mediated changes in intracellular cAMP

Result: Compounds 1 and 2 were **full agonists** of the D_2R with potency values (EC_{50}) values:

- Cmpd 1: 10 µM, Emax = 99%
- Cmpd 2: 14 µM, Emax = 100%

Maximal effect (Emax) relative to dopamine

A Harder Challenge: Polypharmacology

The workflow is proven for a single target. But can it find ligands with **complex properties**, such as hitting **multiple** targets at once?

Advanced Application: A Dual-Target Goal for Parkinson's Disease

The Parkinson's Disease

Parkinson's Disease treatment can involve modulating multiple G protein-coupled receptors (GPCRs) in the central nervous system.

The Dual-Target Goal

Find a **single molecule** that acts as:

- An **agonist** for the D_2R
- An **antagonist** for the $A_{2A}R$

The Challenge

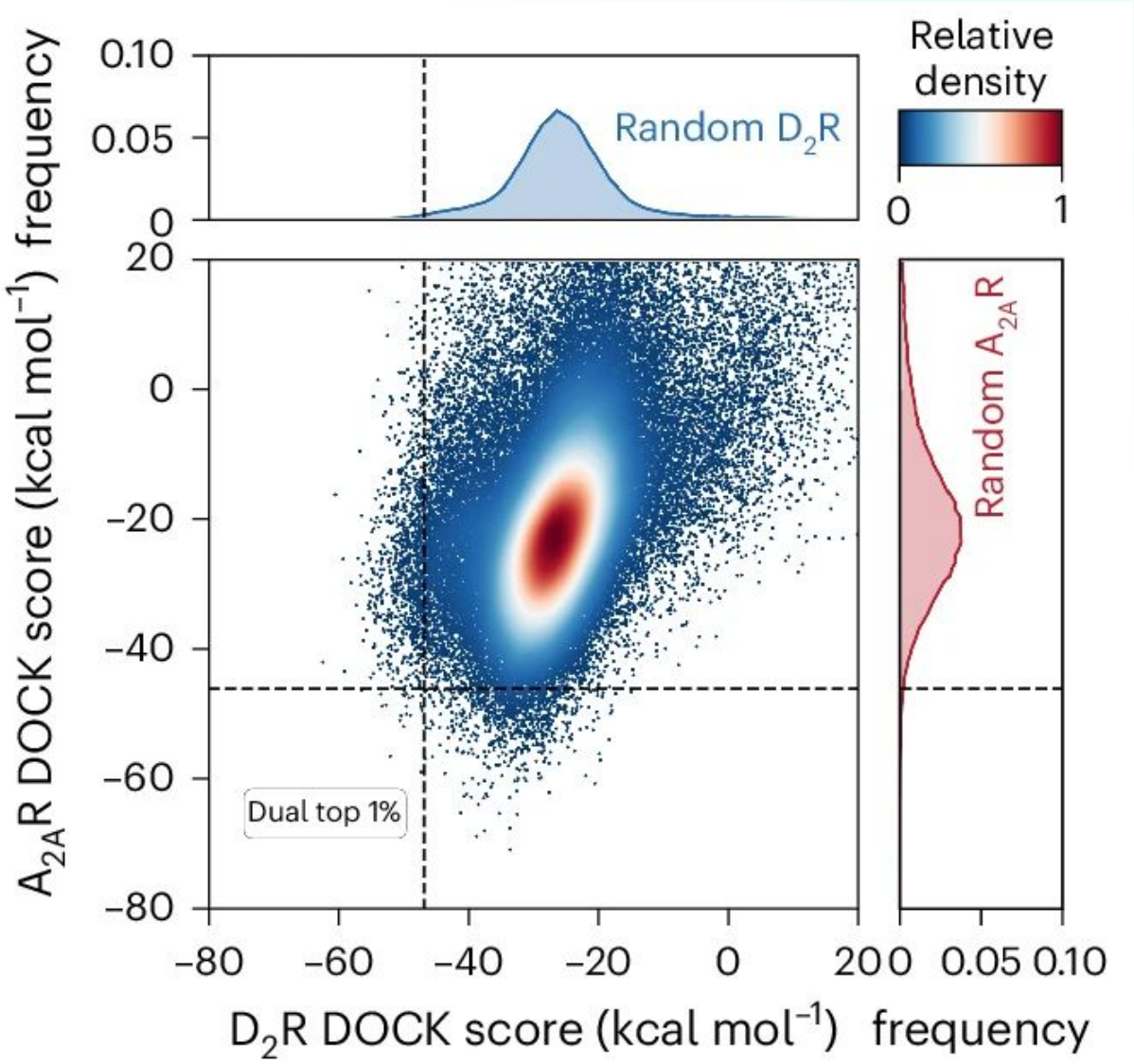
The binding sites are very different. The overlap of random top-scoring hits for **both** targets is **<0.02%**. It's like finding a needle in two haystacks.

The Polypharmacology Challenge

Dual-Target Ligands for Parkinson's Disease

One advantage of multi-billion-scale libraries is the **potential** to find ligands with complex "polypharmacology."

The overlap of random top-scoring hits for **both** targets is **<0.02%**.



A New ML-Guided Strategy

Train Dual Models

Two separate Conformal Predictors were trained (one for $A_{2A}R$, one for D_2R) on a 1 million compound docking set.

Predict 3.5B Library

The trained models were used to predict the class (active/inactive) for all 3.5 billion compounds in the Enamine REAL space library.

Prioritize Dual-Confidence

Compounds were ranked by the **sum** of their "quality of information" for **both** targets. The top 5 million were selected for docking.

$$\Delta P = \Delta P_{A2AR} + \Delta P_{D2R}$$

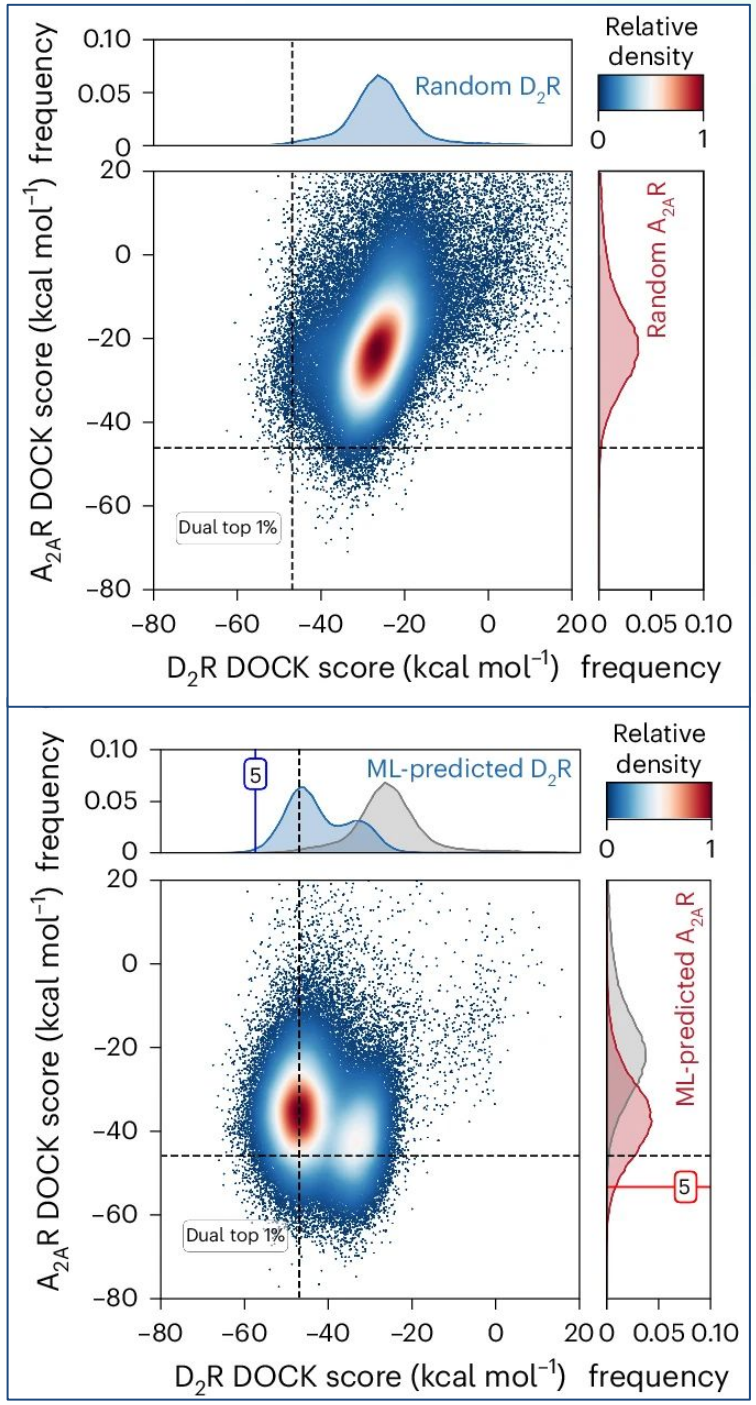
ML Prioritization Works

The 5M prioritized compounds were massively enriched for dual-target hits compared to a random screen

- **3.8%** of the 5M compounds were in the top 1% for **both** targets (a "dual virtual active")
- This represents a 191-fold enrichment
- Substantial score shifts for both $A_{2A}R$ (17-fold) and D_2R (34-fold)

191x

Enrichment of Dual Actives



Experimental Validation: Compound 5

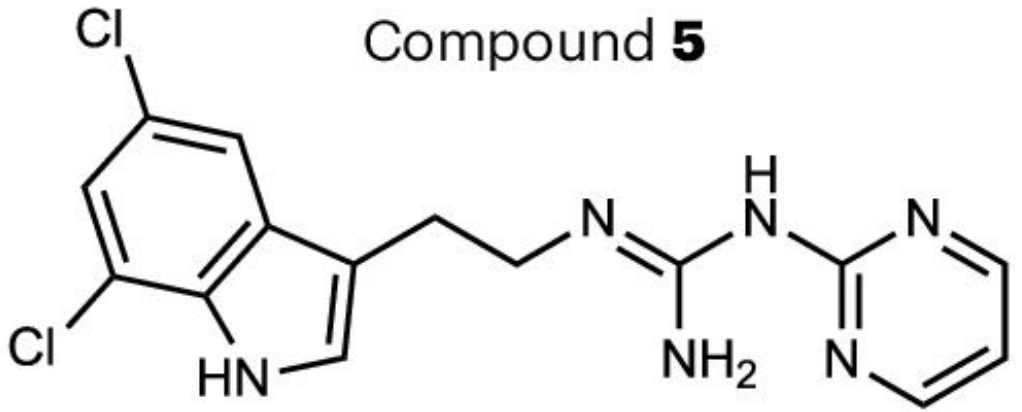
A Dual-Target Hit Found

From the 5M docked compounds, 45 were selected, synthesized, and tested.

Compound 5 was successfully identified and confirmed as a dual-target ligand:

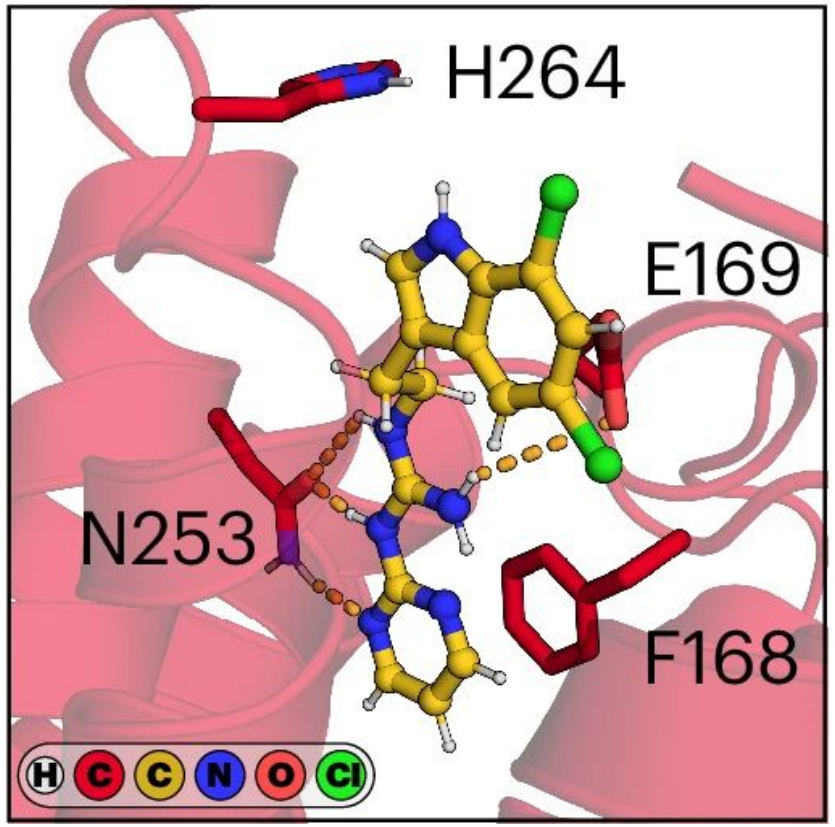
$A_{2A}R$ Affinity (K_i): **20 μM**

D_2R Affinity (K_i): **14 μM**



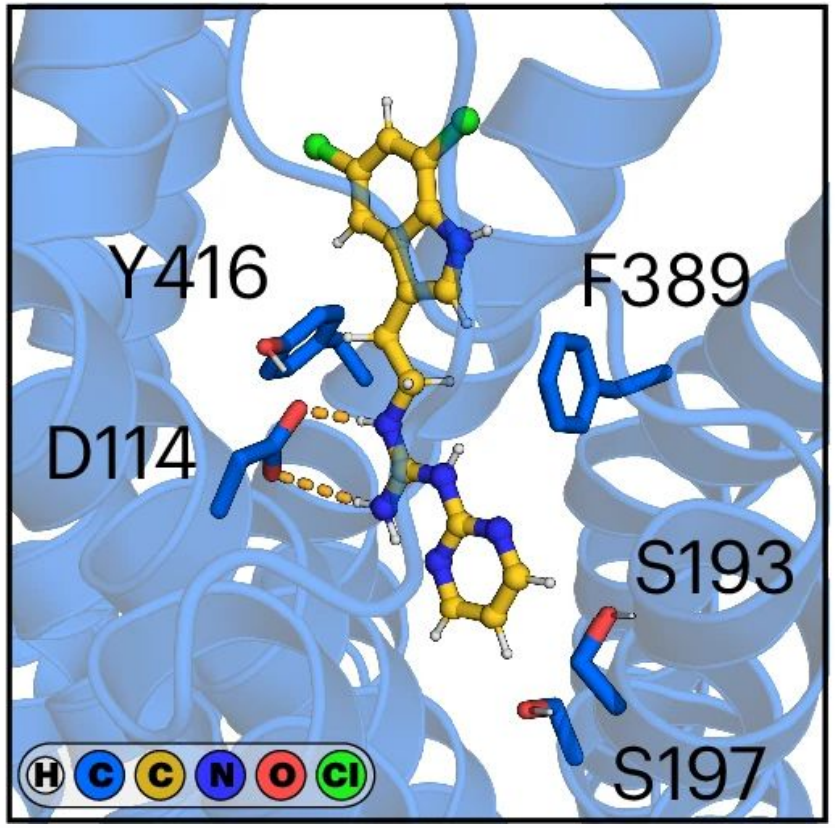
	$A_{2A}R$	D_2R
Score	−53.38	−57.32
P_1	0.740	0.924
P_0	0.007	0.001
K_i	20	14

Predicted Binding Modes



$A_{2A}R$ Binding Site

The model correctly predicted that **Compound 5** forms a key hydrogen bond with the orthosteric site residue **Asn253**.



D_2R Binding Site

Simultaneously, the model predicted the critical hydrogen bond with the D_2R key residue **Asp114**.

Outline

1. Introduction
2. Conformal predictor
3. Benchmarking of conformal predictors
4. Optimized workflow for ultralarge chemical libraries
5. Prospective virtual screen of a multi-billion-scale library
6. Machine learning-guided design of polypharmacology
7. **Conclusions**

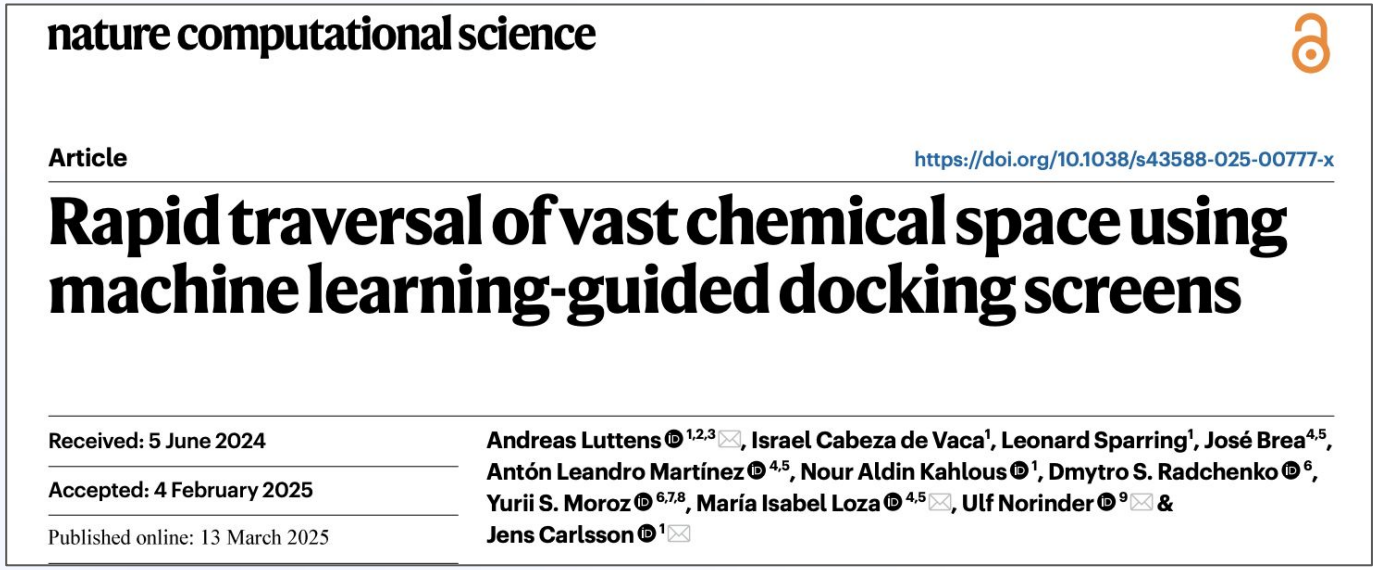
Key Conclusions

1. Works: The ML-guided workflow (CatBoost + CP) successfully screens multi-billion-scale libraries at a fraction of the cost.

2. Efficient: Reduces **significantly** the computational cost, enabling virtual screens that were previously impossible.

3. Powerful: The method can find novel single-target ligands and, more importantly, complex, multi-target polypharmacology.

4. Open: The code and benchmarking datasets are shared to catalyze further development.



nature computational science

Article

https://doi.org/10.1038/s43588-025-00777-x

Rapid traversal of vast chemical space using machine learning-guided docking screens

Received: 5 June 2024

Accepted: 4 February 2025

Published online: 13 March 2025

Andreas Luttens[1,2,3], Israel Cabeza de Vaca[1], Leonard Sparring[1], José Brea[4,5], Antón Leandro Martínez[4,5], Nour Aldin Kahlous[1], Dmytro S. Radchenko[6], Yurii S. Moroz[6,7,8], María Isabel Loza[4,5], Ulf Norinder[9] & Jens Carlsson[1]

GitHub

https://github.com/Carlssonlab/conformalpredictor.git

zenodo

https://doi.org/10.5281/zenodo.7953917

Acknowledgements

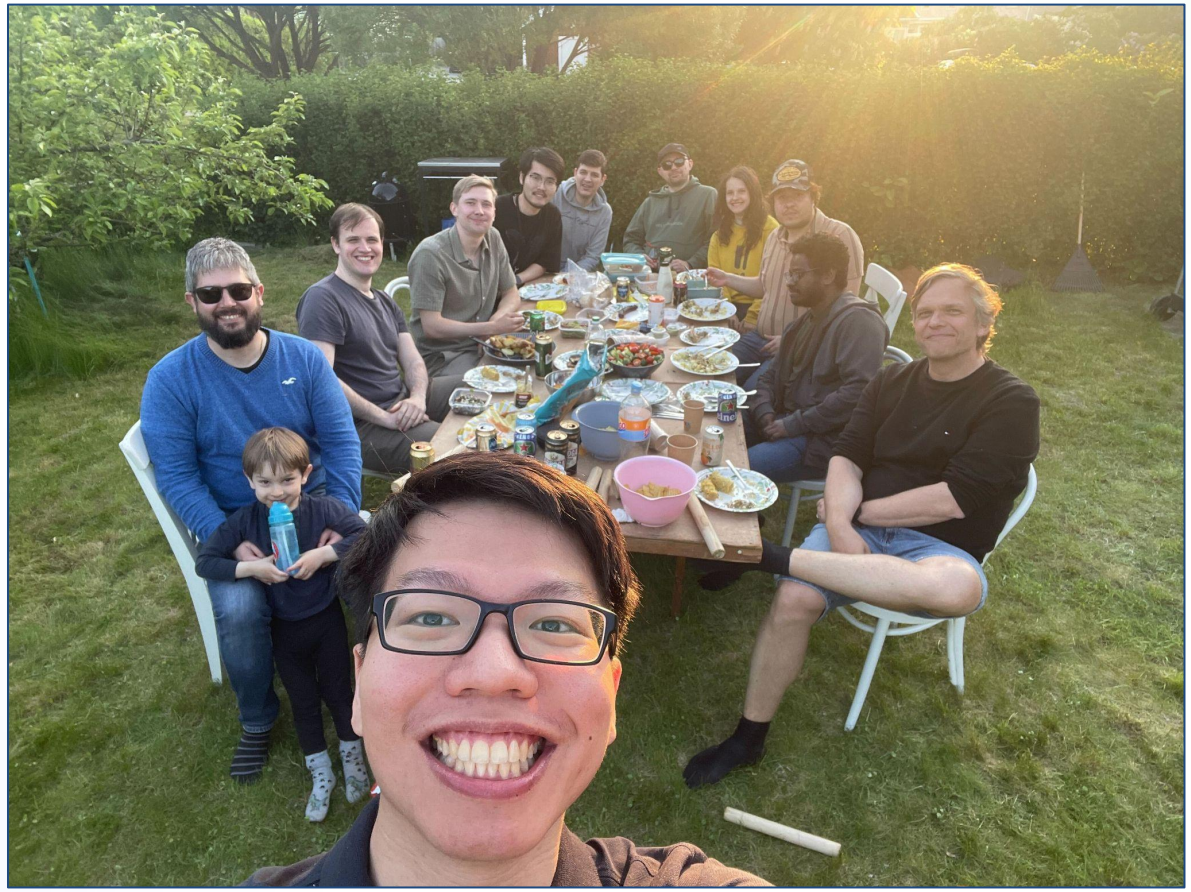
Uppsala University

Andreas Luttens
Leonard Sparring
Nour Aldin Kahlous

Jens Carlsson
Ulf Norinder

Enamine Ltd

Dmytro S. Radchenko
Yurii S. Moroz



University of Santiago de Compostela

José Brea
Antón Leandro Martínez
María Isabel Loza

NAISS

Vetenskapsrådet

CANCERFONDEN

Knut and Alice Wallenberg Foundation

OLLE ENGKVISTS STIFTELSE

erc

Sven och Lilly Lawskis fond för naturvetenskaplig forskning

Thank You

Questions?

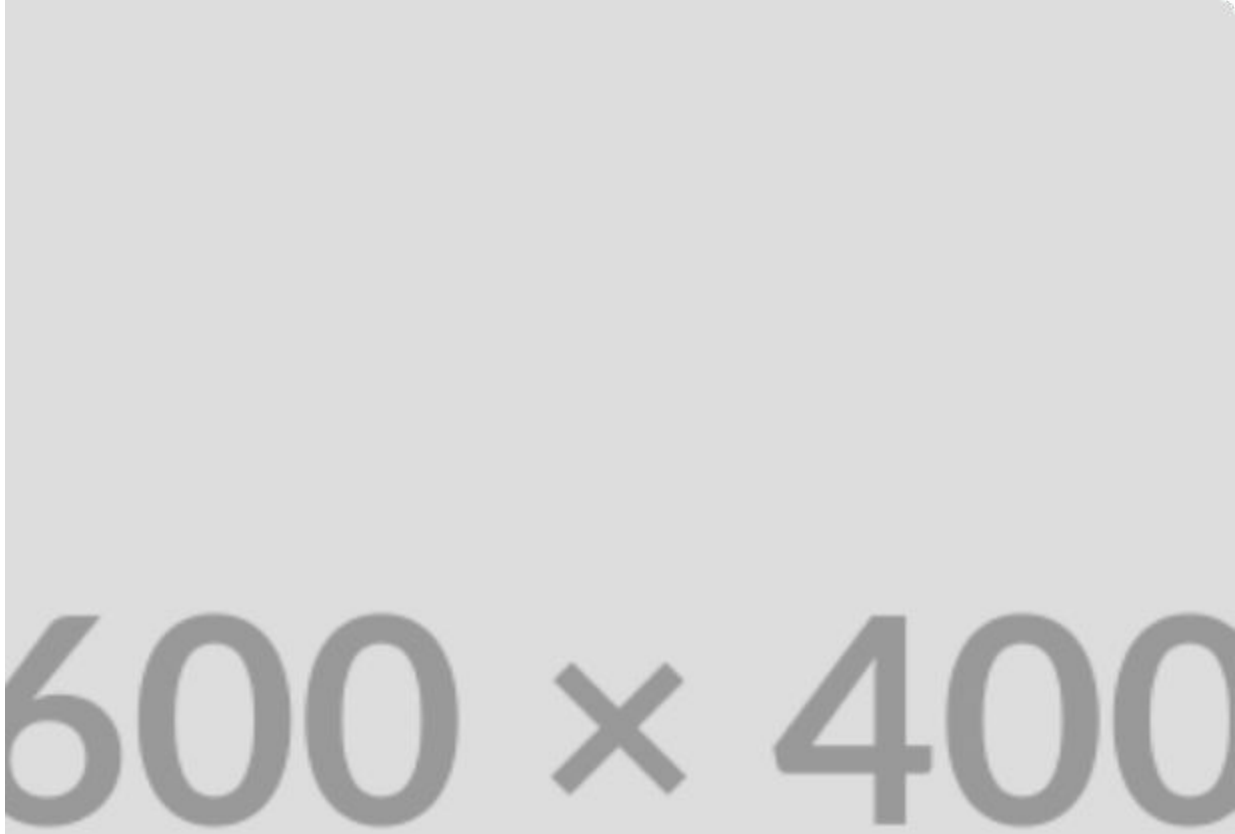
Case Study 1: D_2 Receptor Screen

Goal: Find New D_2R Ligands

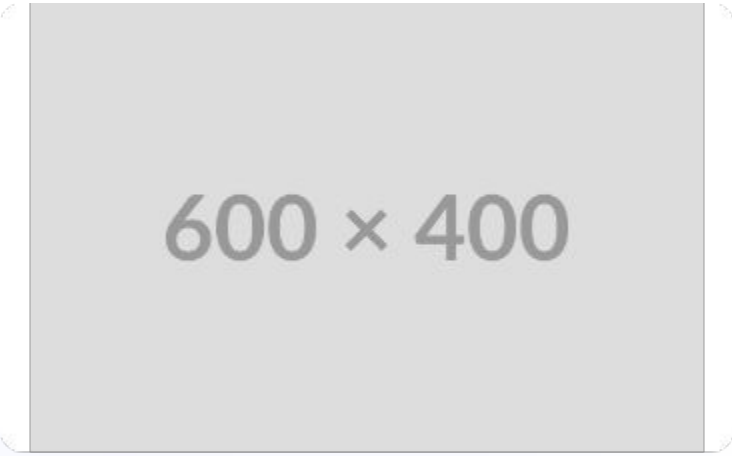
The team performed a prospective screen to find new ligands for the D_2 dopamine receptor, an important target for neuropsychiatric disorders.

- **Library:** 3.5 *billion* compounds.
- **Training:** ML model trained on 1M docked compounds.
- **Prediction:** ML model classified all 3.5B compounds.
- **Selection: 5 million** compounds were prioritized by the model and selected for actual docking.

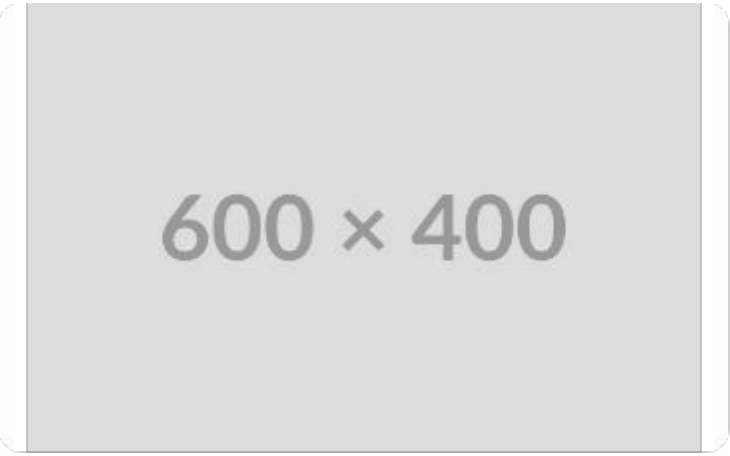
This is a 700-fold reduction of the library's computational cost!



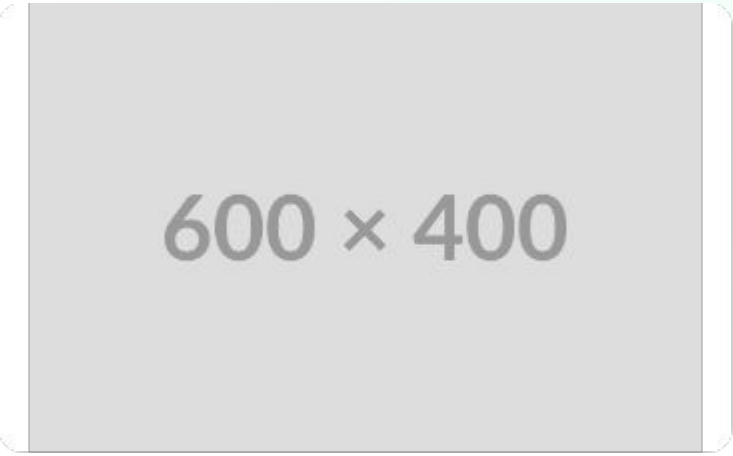
Case Study 1: New D_2R Agonists Discovered



Compound 1: A novel indole-based ligand identified by the screen. Showed significant binding (Ki = 3.0 μM).



Experimental Validation: 31 top-ranked compounds were synthesized and tested. Both compounds 1 & 2 were confirmed as **full agonists** of the D_2R.



Compound 2: A novel thiophene-based ligand identified by the screen. Showed significant binding (Ki = 3.8 μM).