Predicting hits with ML and limited data: 3 new tricks

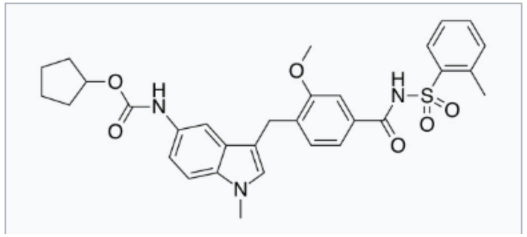
Jan H. Jensen

Department of Chemistry,
University of Copenhagen

@janhjensen

Target: some membrane protein (no X-ray structure)

Zafirlukast

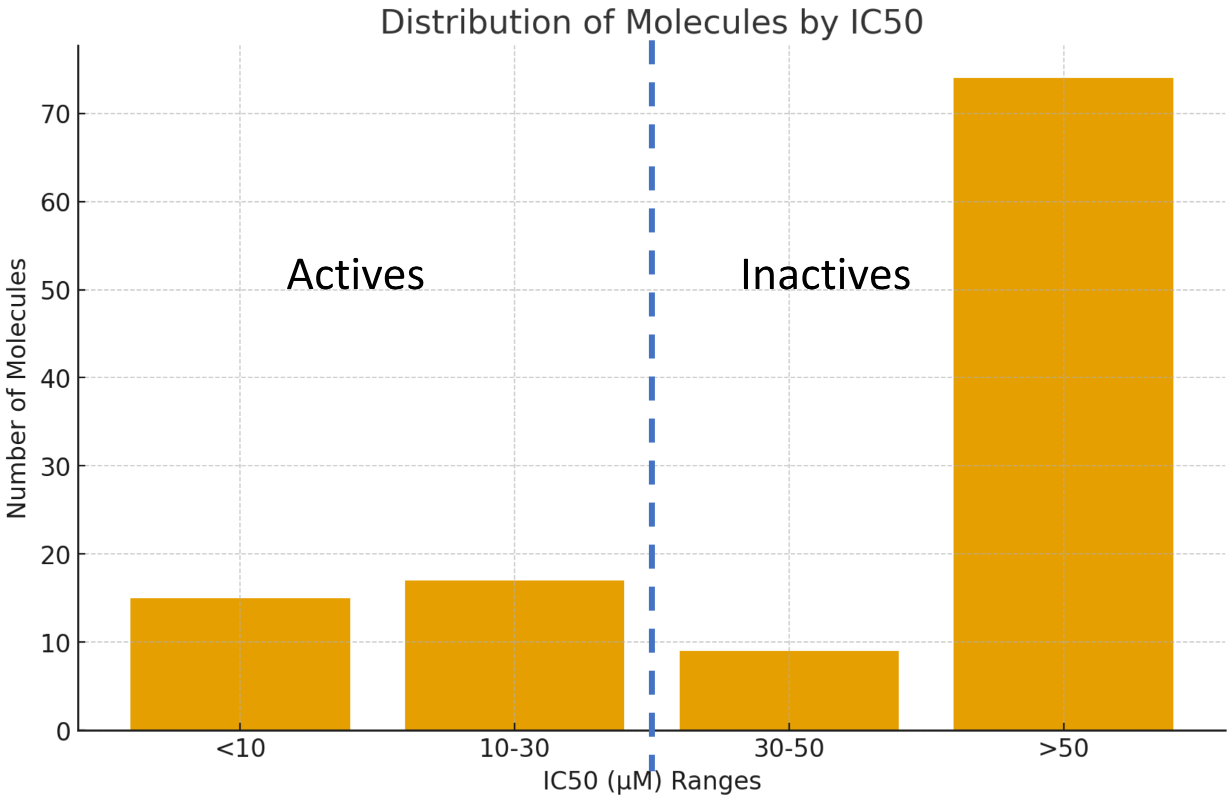


IC_{50} 2 µM

~5 years
Some Pharma goodwill
BS and MSc students

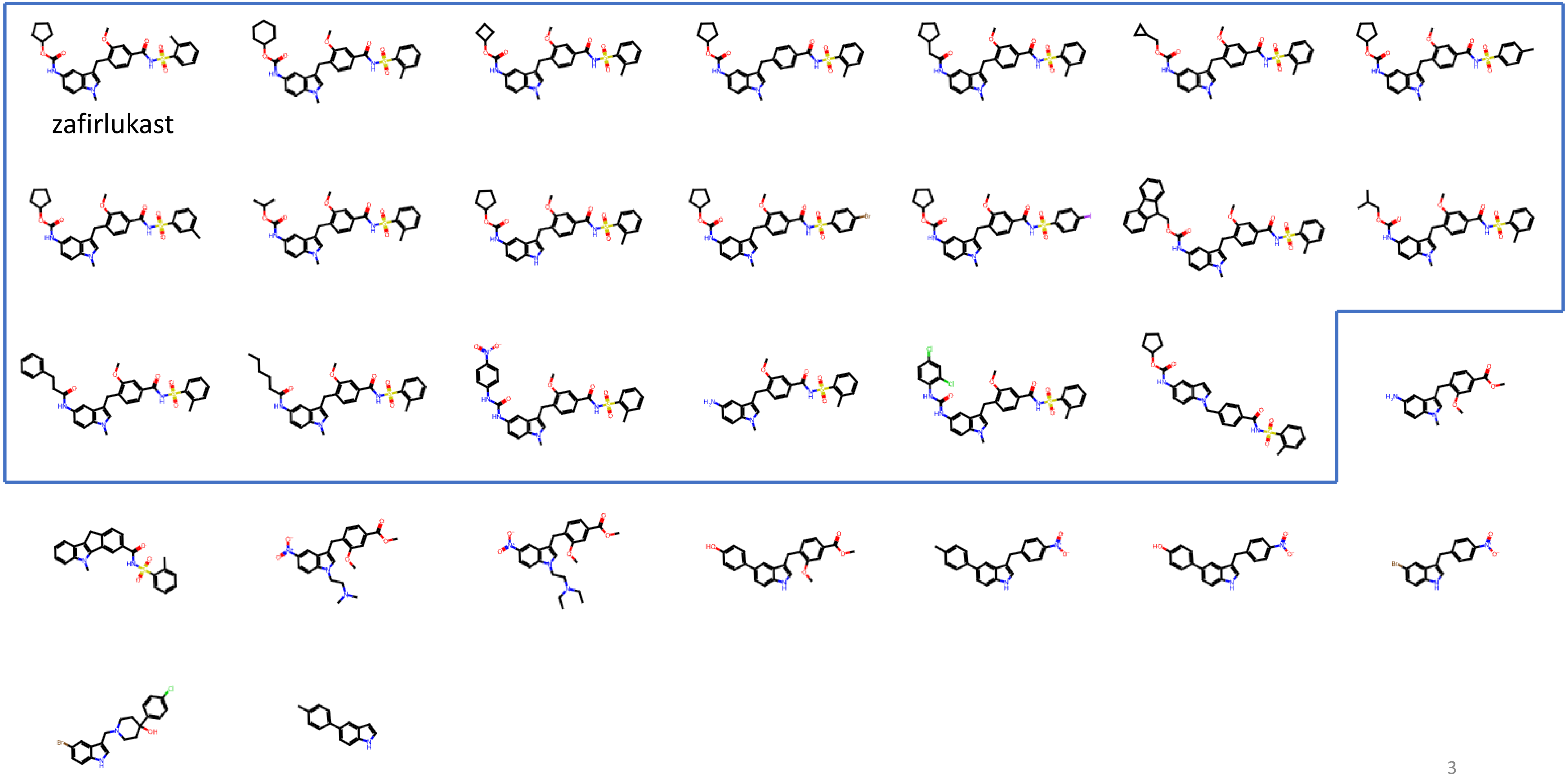
1st-year PhD student*

115 indoles



***Niels Guldager**
Christian M. Pedersen
Anders Aa. Rehfeld

66% actives are zafirkulast derivatives



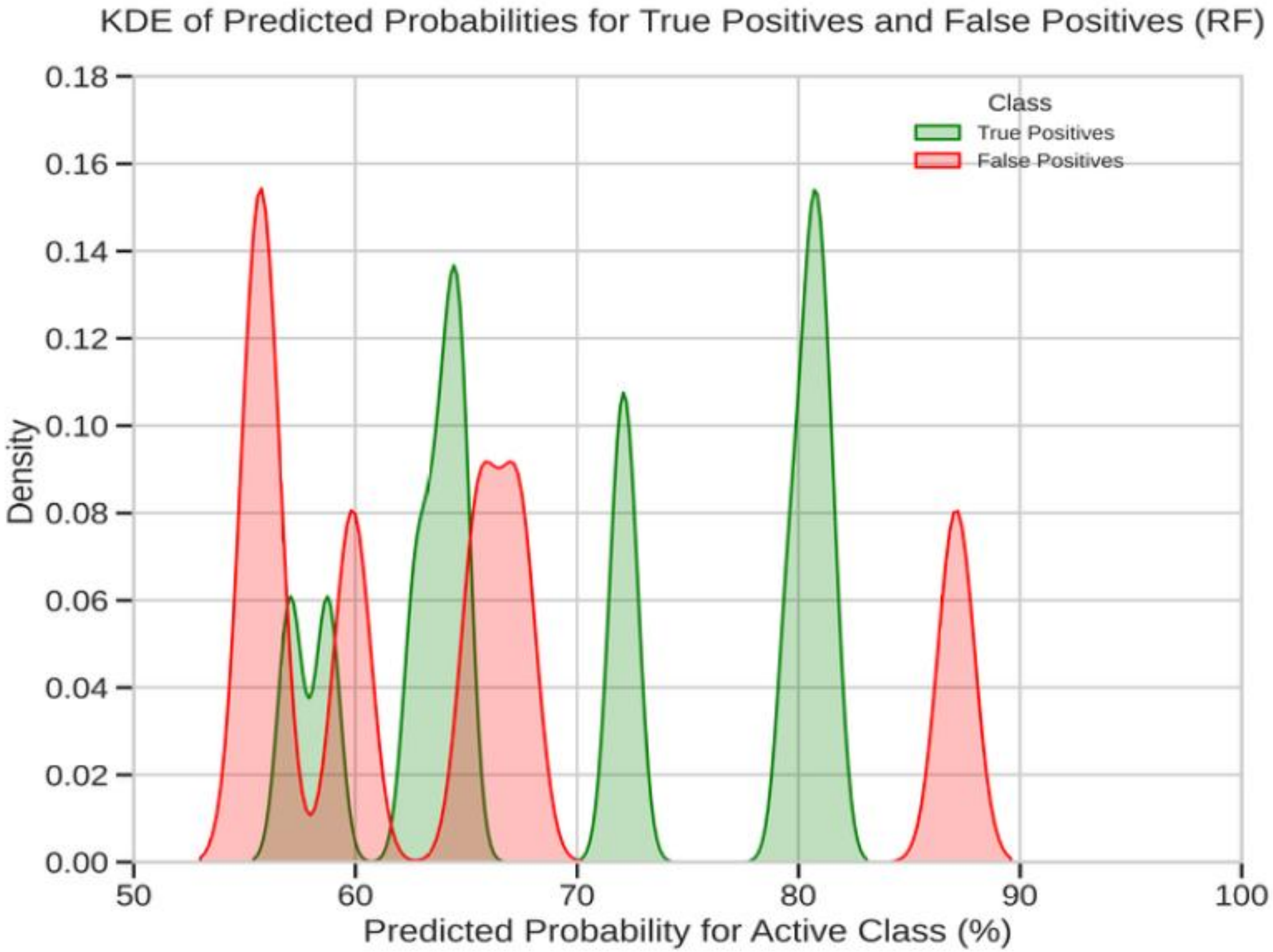
Very challenging data set: Businness as usual doesn’t work

ECFP4-based RF* classifier

Henry Teahan

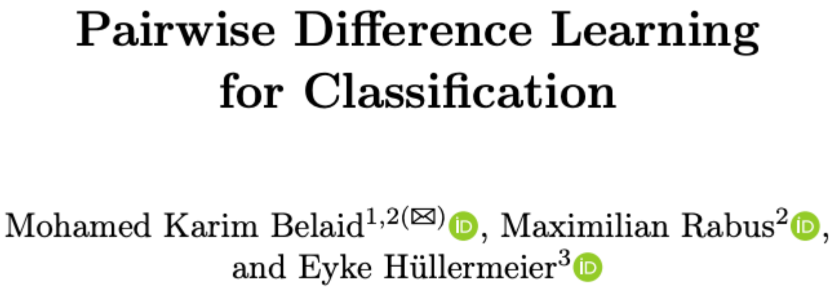
Maria H. Rasmussen

Leave-one-out



*bagged trees

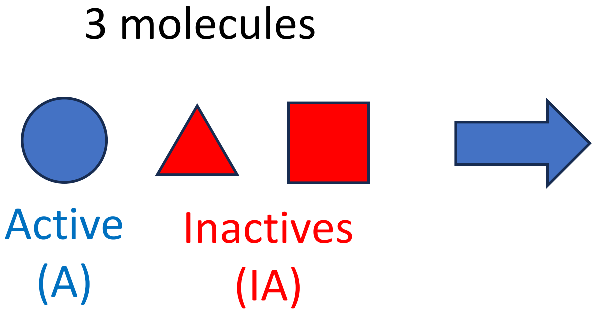
Trick # 1: PDL



Pairwise Difference Learning for Classification

Mohamed Karim Belaid[1,2(✉)], Maximilian Rabus[2], and Eyke Hüllermeier[3]

Training



9 pairs

●	●	1 (same category)
●	▲	0 (different category)
●	■	0
▲	●	0
▲	▲	1
▲	■	1 (same category)
■	●	0
■	▲	1
■	■	1

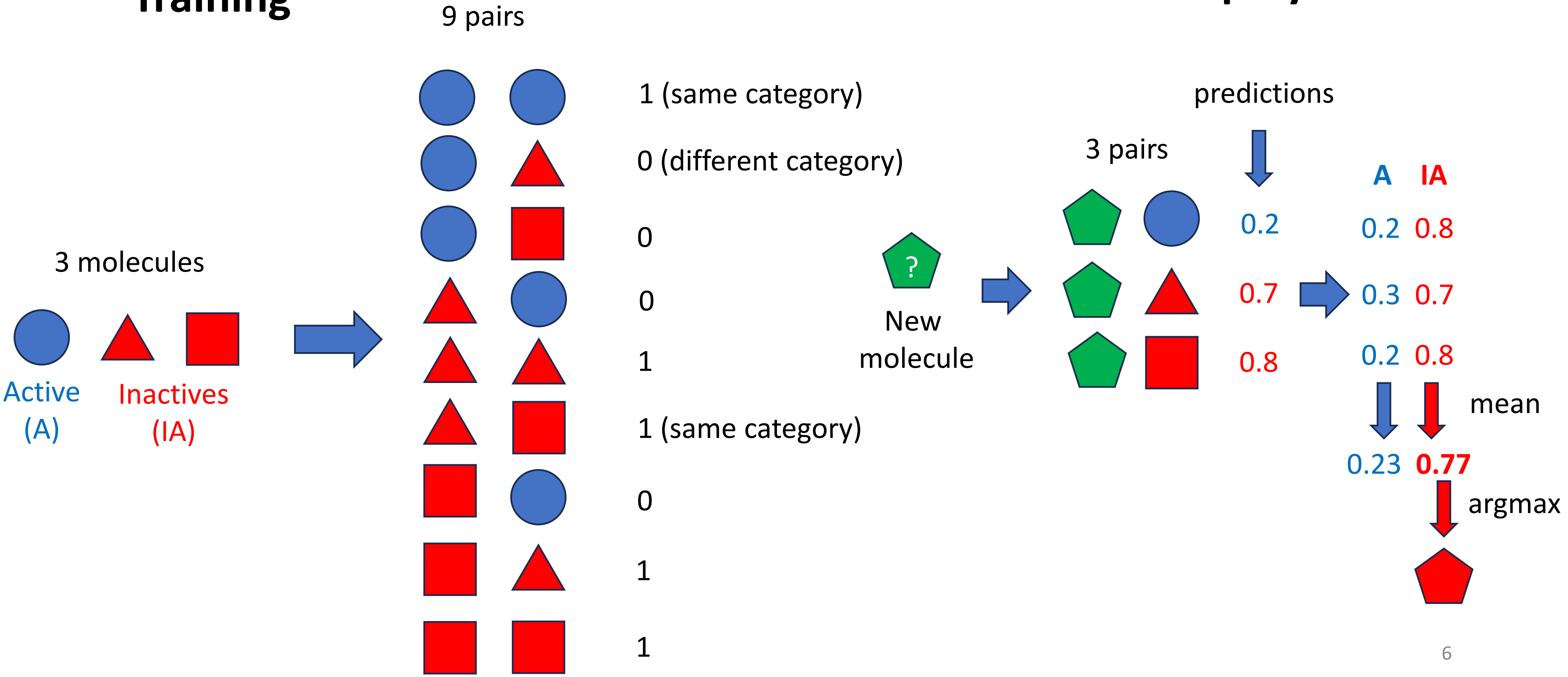
Trick # 1: PDL

Pairwise Difference Learning for Classification

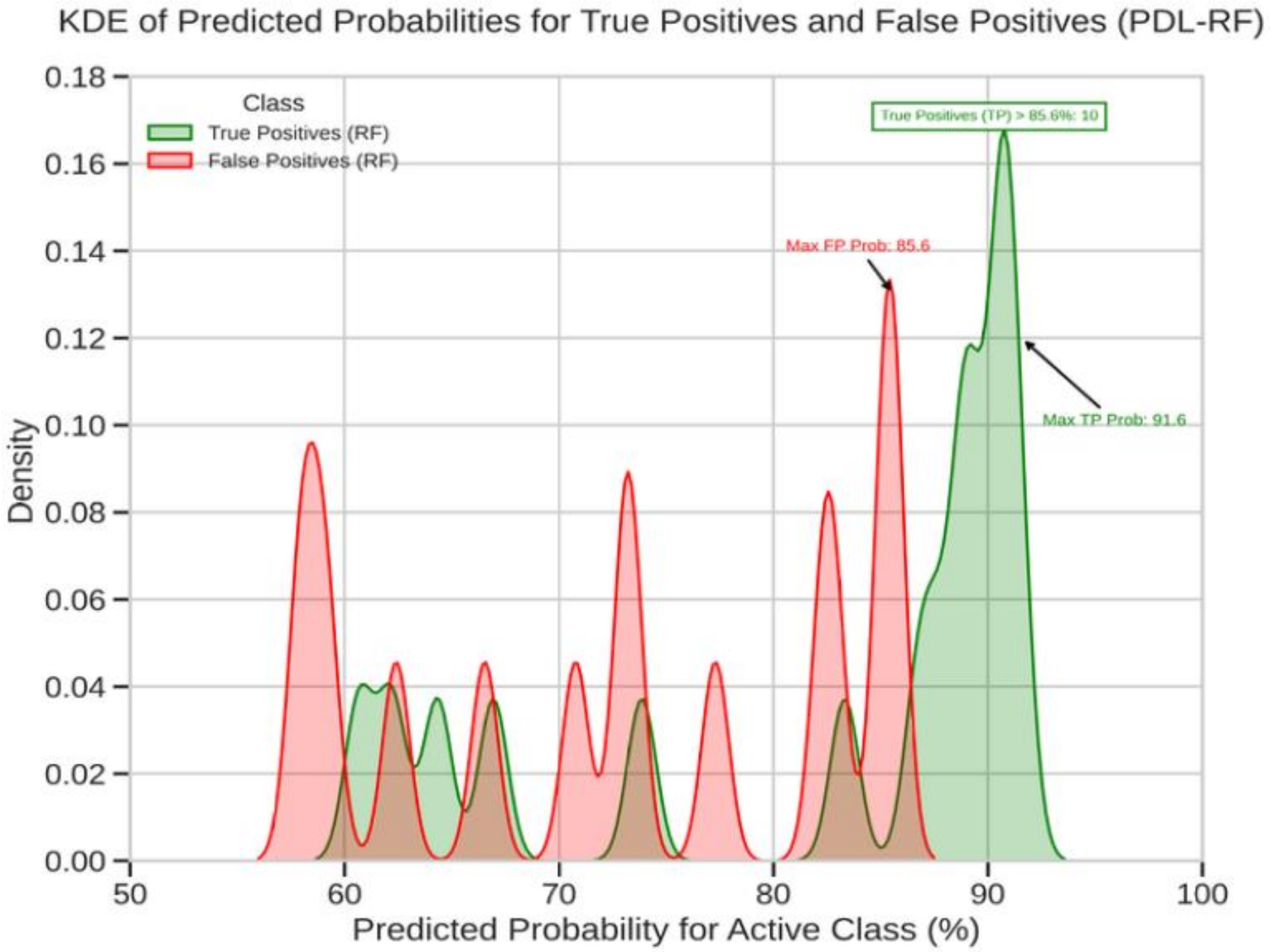
Mohamed Karim Belaid[1,2(✉)], Maximilian Rabus[2], and Eyke Hüllermeier[3]

Training

Deployment



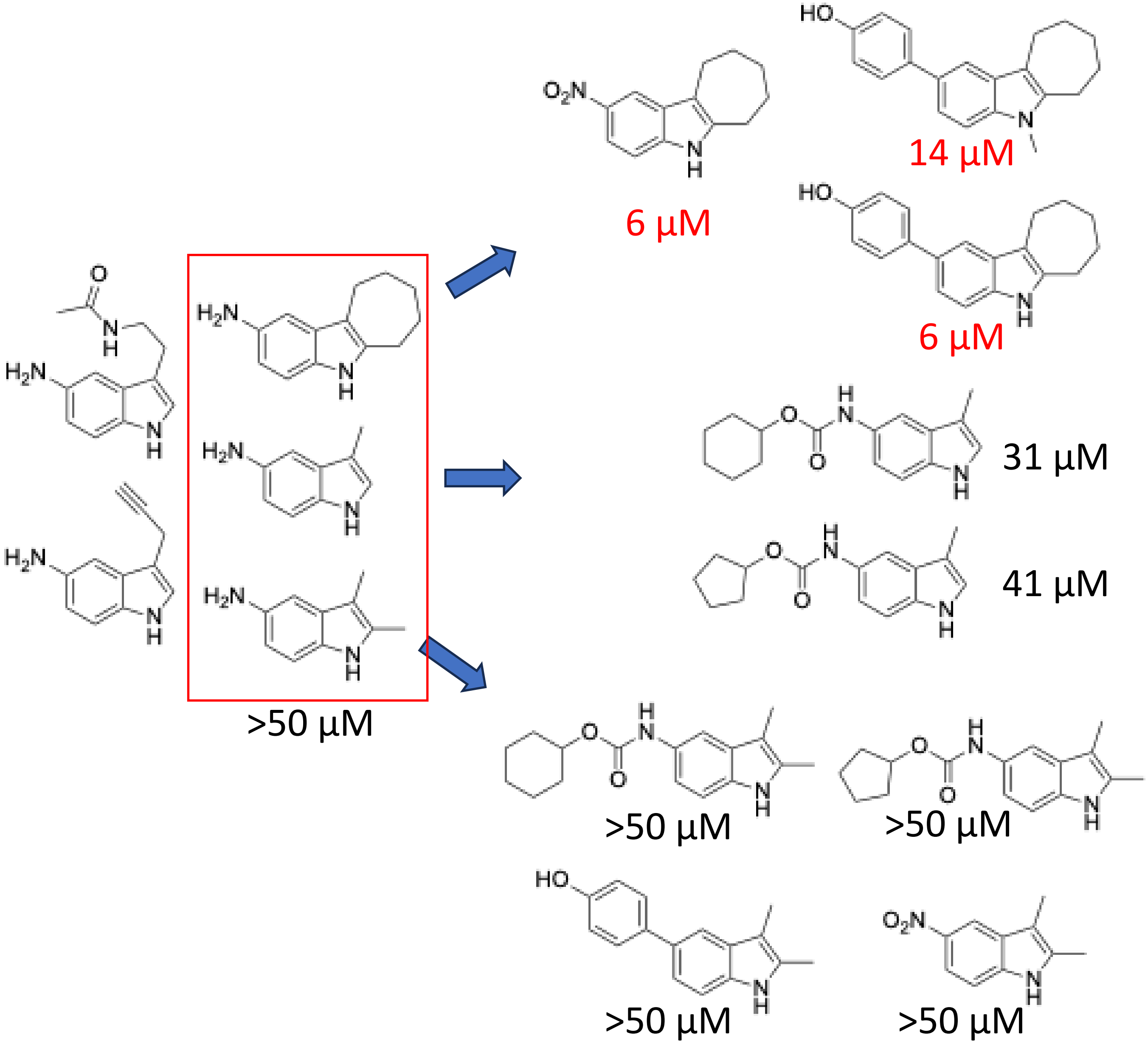
PDL works much better


KDE of Predicted Probabilities for True Positives and False Positives (PDL-RF)

PDL fingerprint **FP1** + **FP2** + **(FP1-FP2)**

Screening Enamines Building Blocks set (1288 indoles)

Predicted



6 µM

14 µM

6 µM

31 µM

41 µM

>50 µM

>50 µM

>50 µM

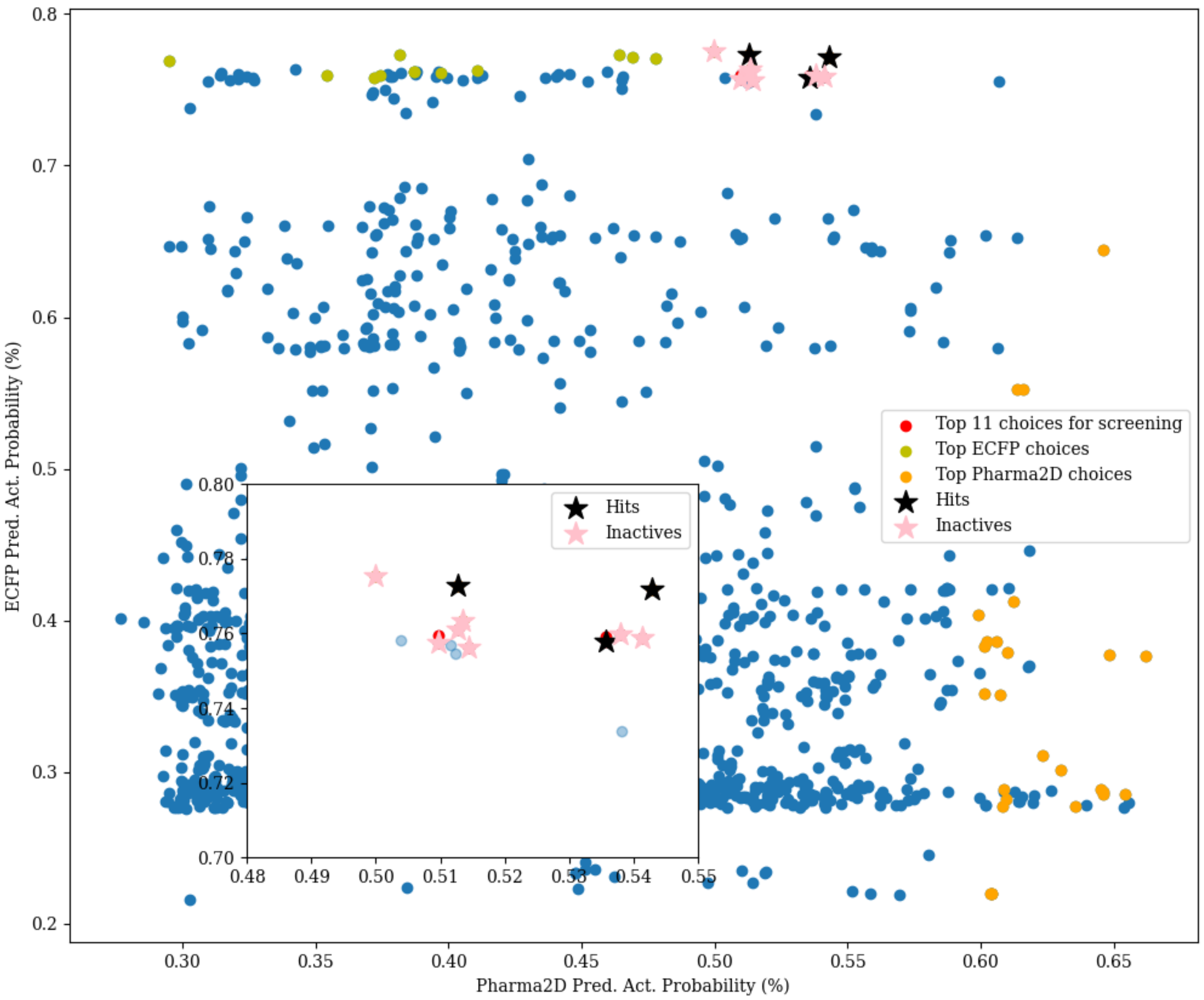
>50 µM

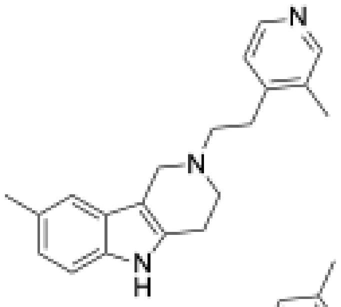
>50 µM

Tested

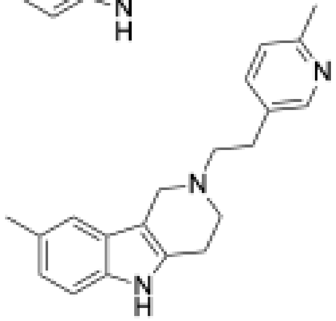
11 new predictions from Enamine Hit Locator Set (7649 indoles)

(10 were tested -> 0 hits)

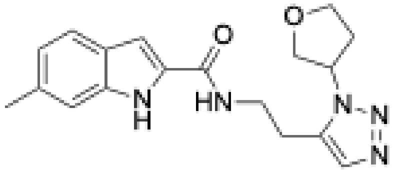




46 µM



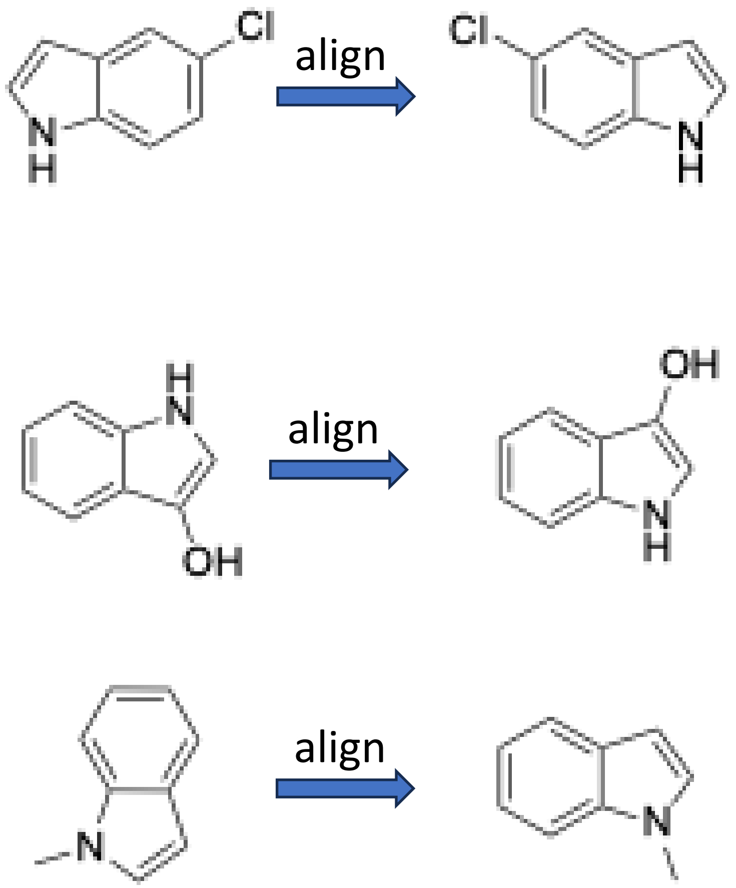
47 µM



47 µM

Trick # 2 The "2D atomic pharmacophore model"

Training set



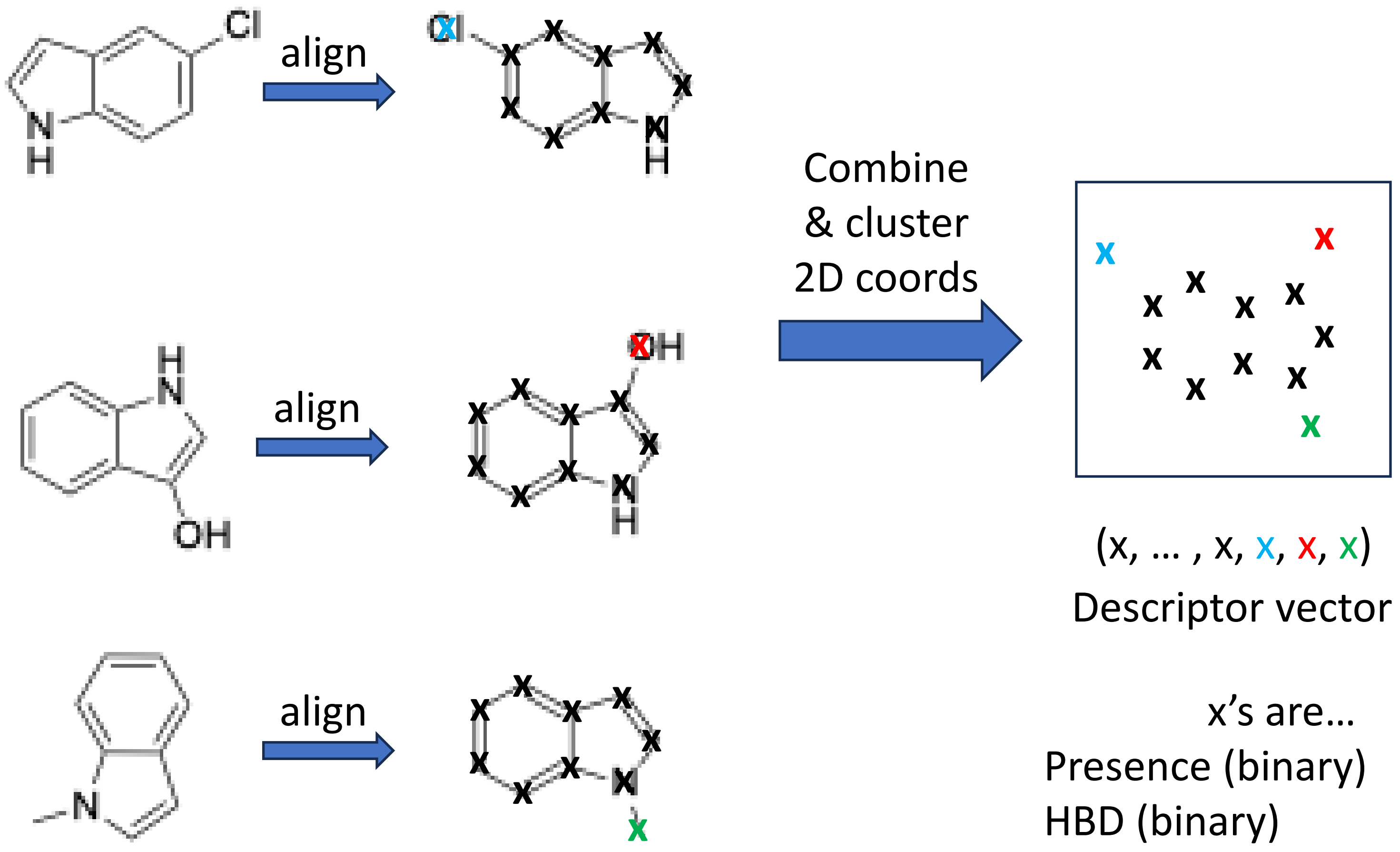
align

align

align

Trick # 2 The “2D atomic pharmacophore model”

Training set



Combine
& cluster
2D coords

(x, … , x, x, x, x)

Descriptor vector

x’s are…

Presence (binary)

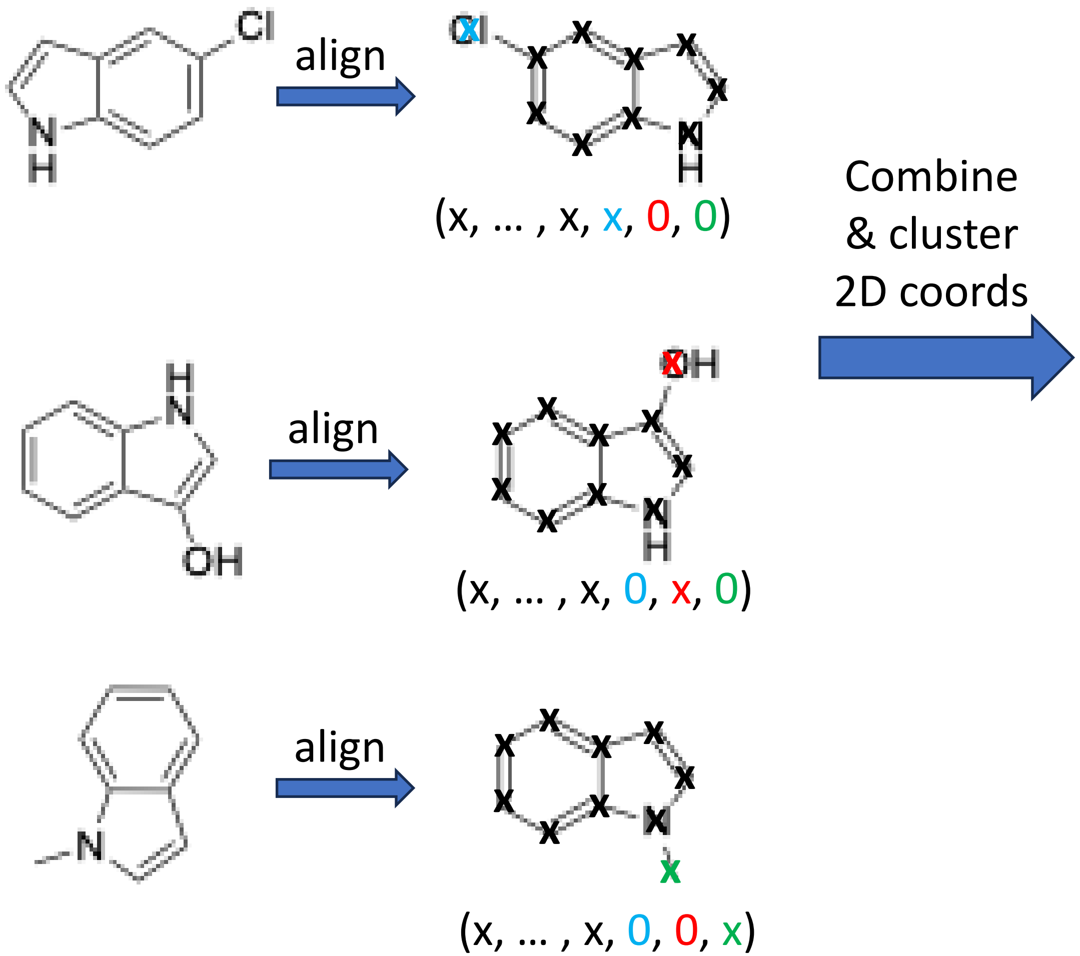
HBD (binary)

HBA (binary)

Atomic logP values (float)

Trick # 2 The “2D atomic pharmacophore model”

Training set

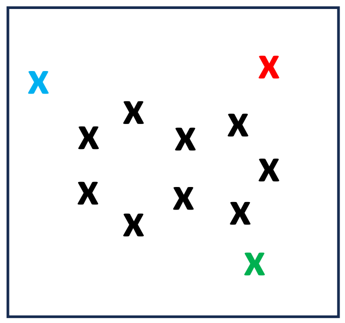


(x, … , x, x, 0, 0)

(x, … , x, 0, x, 0)

(x, … , x, 0, 0, x)

Combine & cluster 2D coords



(x, … , x, x, x, x)

Descriptor vector

x’s are…

Presence (binary)

HBD (binary)

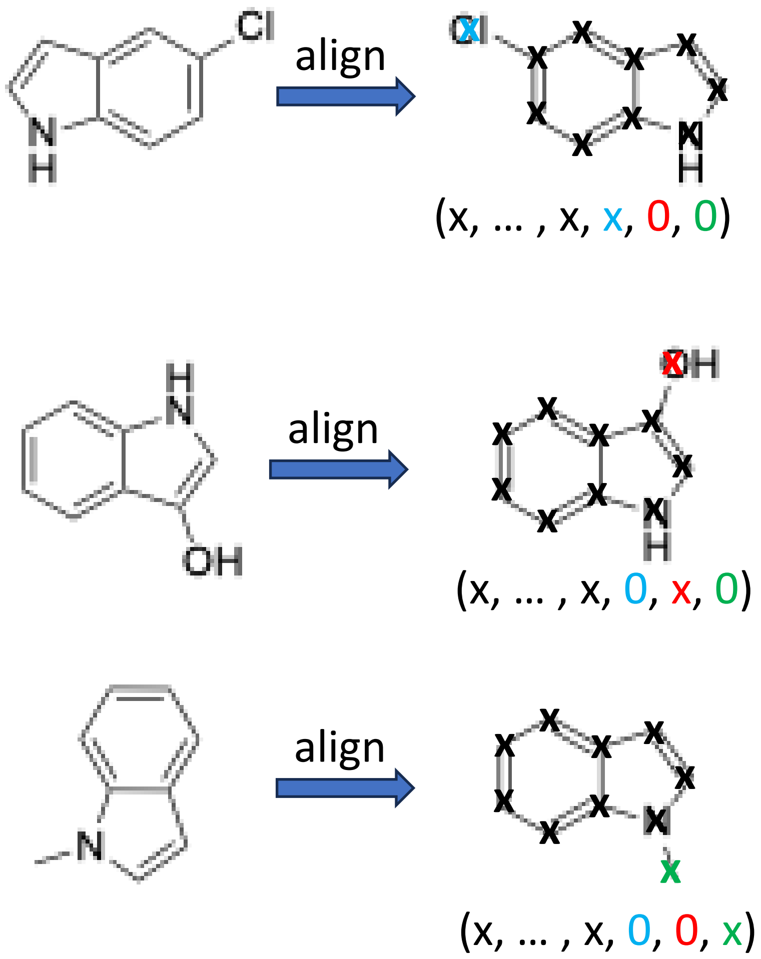
HBA (binary)

Atomic logP values (float)

Trick # 2

The “2D atomic pharmacophore model”

Training set

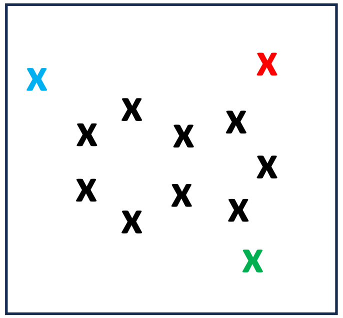


(x, … , x, x, 0, 0)

(x, … , x, 0, x, 0)

(x, … , x, 0, 0, x)

Combine & cluster 2D coords



(x, … , x, x, x, x)

Descriptor vector

x’s are…

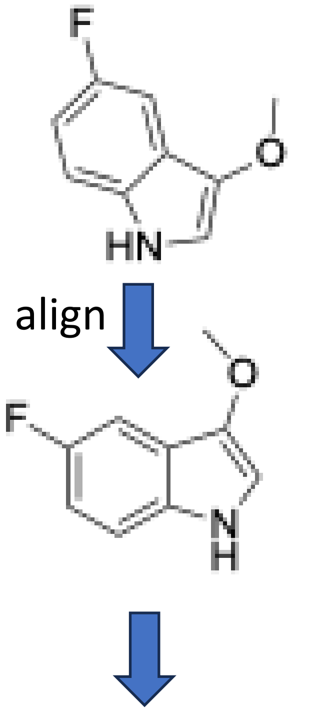
Presence (binary)

HBD (binary)

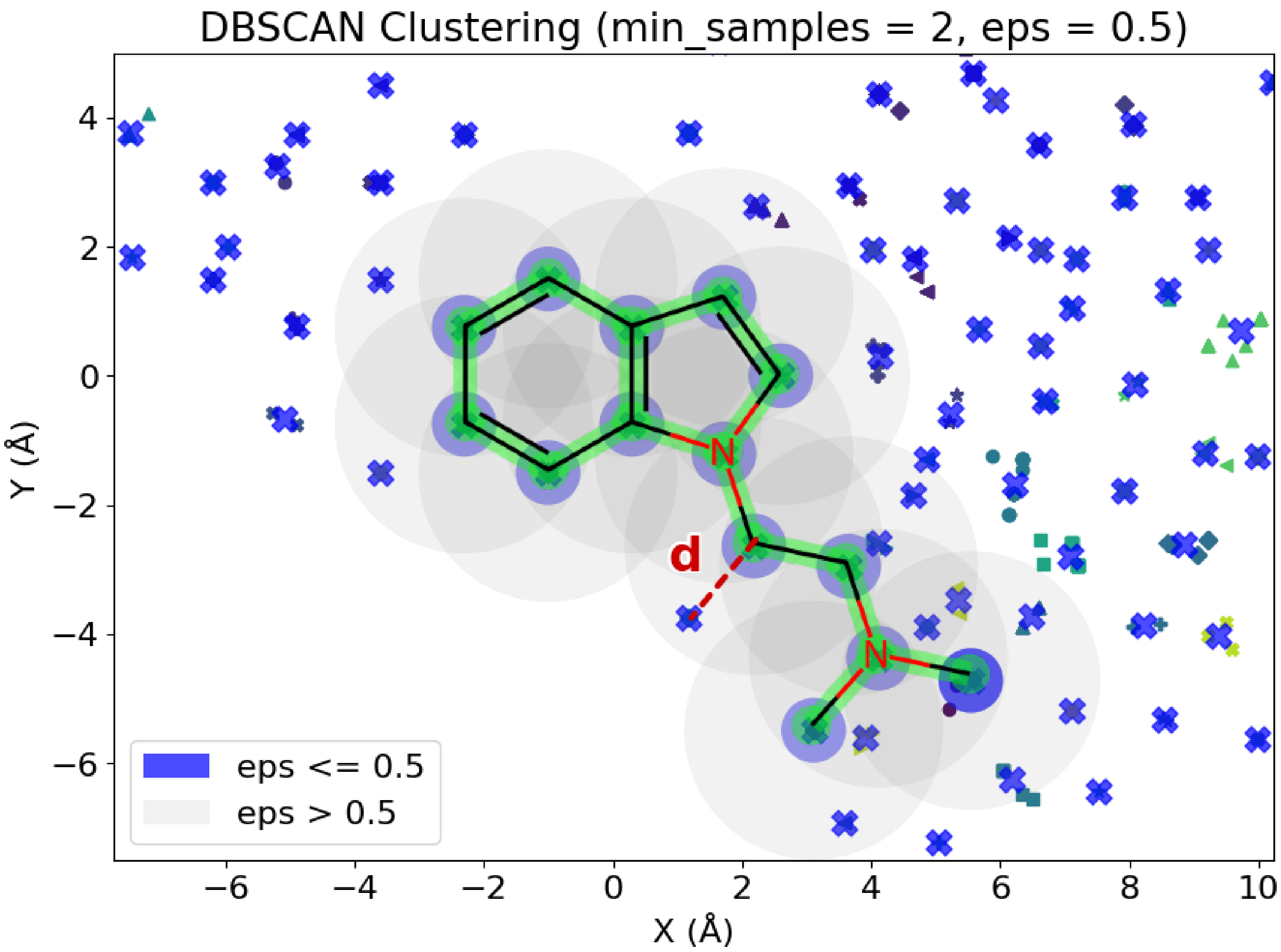
HBA (binary)

Atomic logP values (float)

New molecule



(x, … , x, x, x, 0)



eps > 0.5 → property scaled with distance

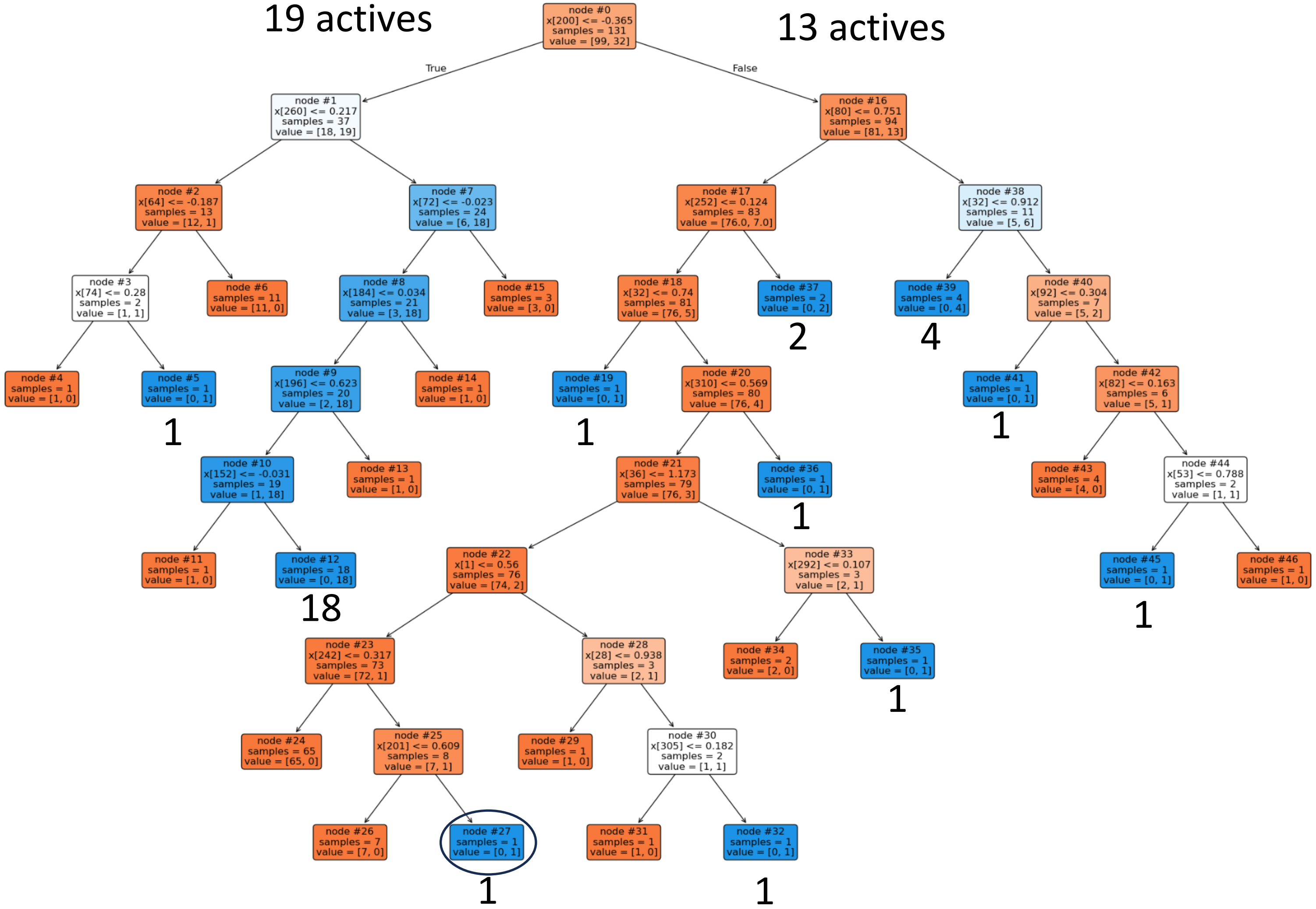
Trick # 3

Better descriptor → simpler model

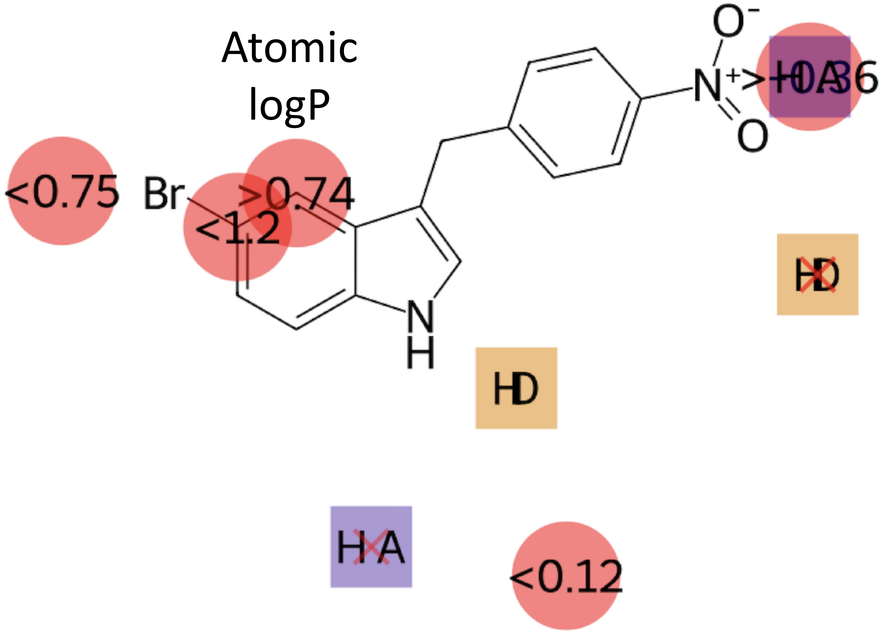
Decision tree offers interpretability

Left side:
19 actives

Right side:
13 actives

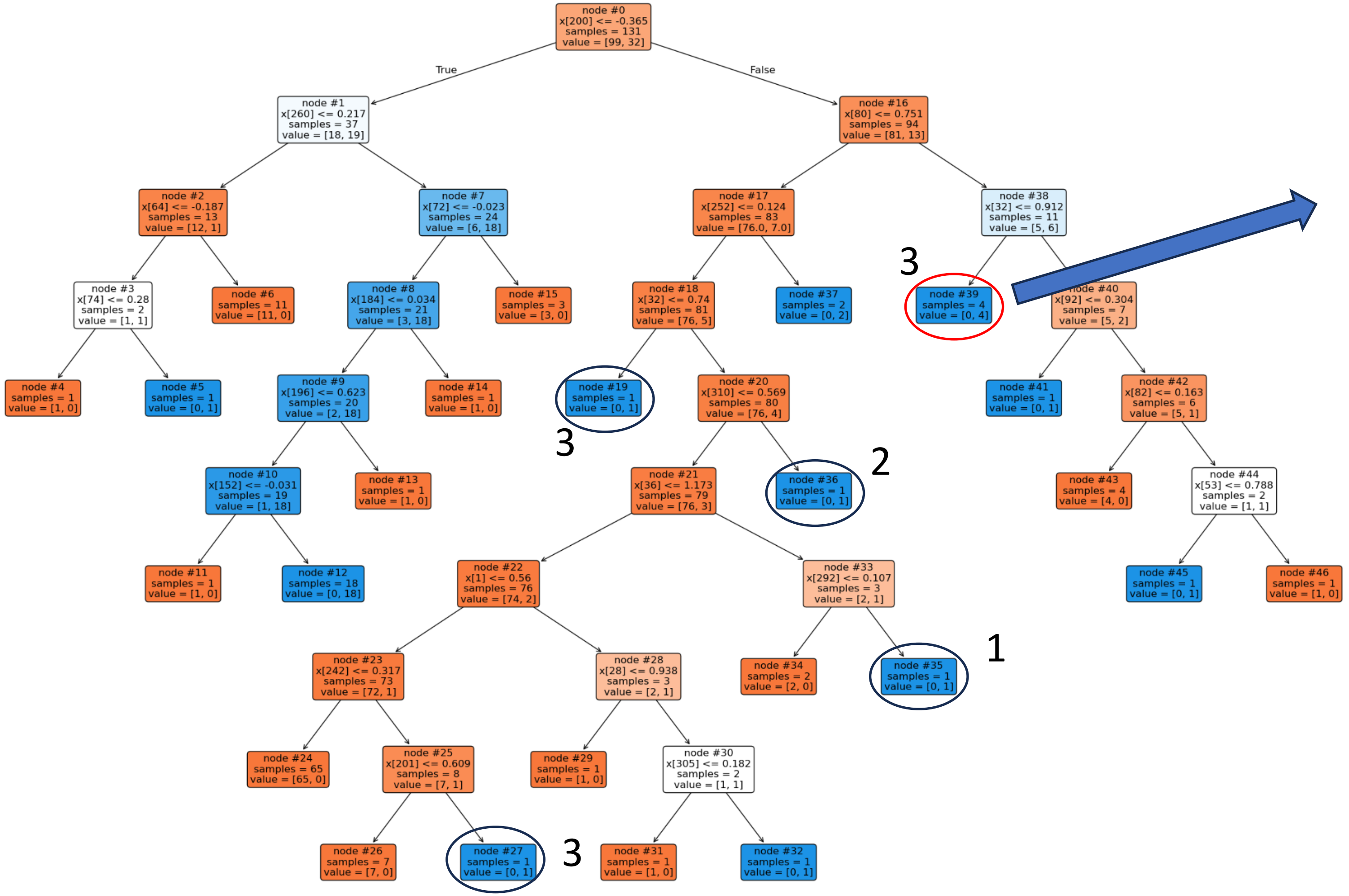


Atomic
logP

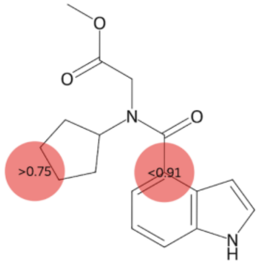


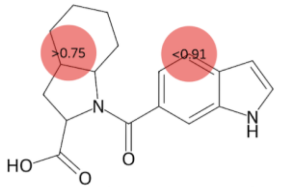
12 new predictions from Enamine Hit Locator Set

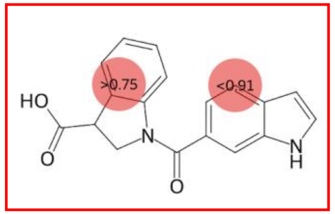
(11 were tested → 1 hit)


node #0 x[200] <= -0.365 samples = 131 value = [99, 32]
True
False
node #1 x[260] <= 0.217 samples = 37 value = [18, 19]
node #16 x[80] <= 0.751 samples = 94 value = [81, 13]
node #2 x[64] <= -0.187 samples = 13 value = [12, 1]
node #7 x[72] <= -0.023 samples = 24 value = [6, 18]
node #17 x[252] <= 0.124 samples = 83 value = [76.0, 7.0]
node #38 x[32] <= 0.912 samples = 11 value = [5, 6]
node #3 x[74] <= 0.28 samples = 2 value = [1, 1]
node #6 samples = 11 value = [11, 0]
node #8 x[184] <= 0.034 samples = 21 value = [3, 18]
node #15 samples = 3 value = [3, 0]
node #18 x[32] <= 0.74 samples = 81 value = [76, 5]
node #37 samples = 2 value = [0, 2]
3
node #39 samples = 4 value = [0, 4]
node #40 x[92] <= 0.304 samples = 7 value = [5, 2]
node #4 samples = 1 value = [1, 0]
node #5 samples = 1 value = [0, 1]
node #9 x[196] <= 0.623 samples = 20 value = [2, 18]
node #14 samples = 1 value = [1, 0]
node #19 samples = 1 value = [0, 1]
3
node #20 x[310] <= 0.569 samples = 80 value = [76, 4]
node #41 samples = 1 value = [0, 1]
node #42 x[82] <= 0.163 samples = 6 value = [5, 1]
node #10 x[152] <= -0.031 samples = 19 value = [1, 18]
node #13 samples = 1 value = [1, 0]
node #21 x[36] <= 1.173 samples = 79 value = [76, 3]
node #36 samples = 1 value = [0, 1]
2
node #43 samples = 4 value = [4, 0]
node #44 x[53] <= 0.788 samples = 2 value = [1, 1]
node #11 samples = 1 value = [1, 0]
node #12 samples = 18 value = [0, 18]
node #22 x[1] <= 0.56 samples = 76 value = [74, 2]
node #33 x[292] <= 0.107 samples = 3 value = [2, 1]
node #45 samples = 1 value = [0, 1]
node #46 samples = 1 value = [1, 0]
node #23 x[242] <= 0.317 samples = 73 value = [72, 1]
node #28 x[28] <= 0.938 samples = 3 value = [2, 1]
node #34 samples = 2 value = [2, 0]
node #35 samples = 1 value = [0, 1]
1
node #24 samples = 65 value = [65, 0]
node #25 x[201] <= 0.609 samples = 8 value = [7, 1]
node #29 samples = 1 value = [1, 0]
node #30 x[305] <= 0.182 samples = 2 value = [1, 1]
node #26 samples = 7 value = [7, 0]
node #27 samples = 1 value = [0, 1]
3
node #31 samples = 1 value = [1, 0]
node #32 samples = 1 value = [0, 1]

We picked these molecules ...

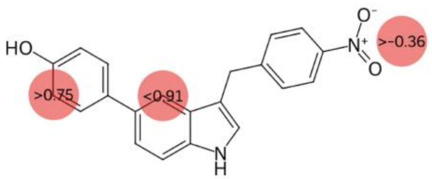




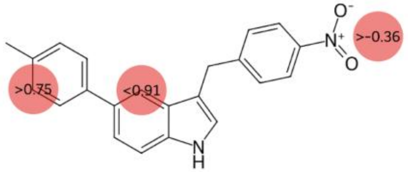


6 µM

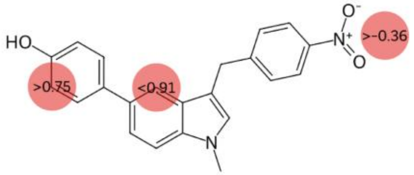
... because of these molecules



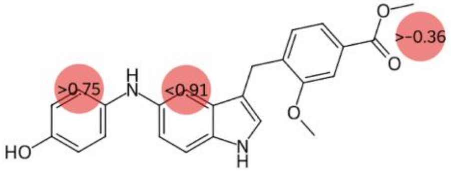
2 µM



24 µM



8 µM



19 µM

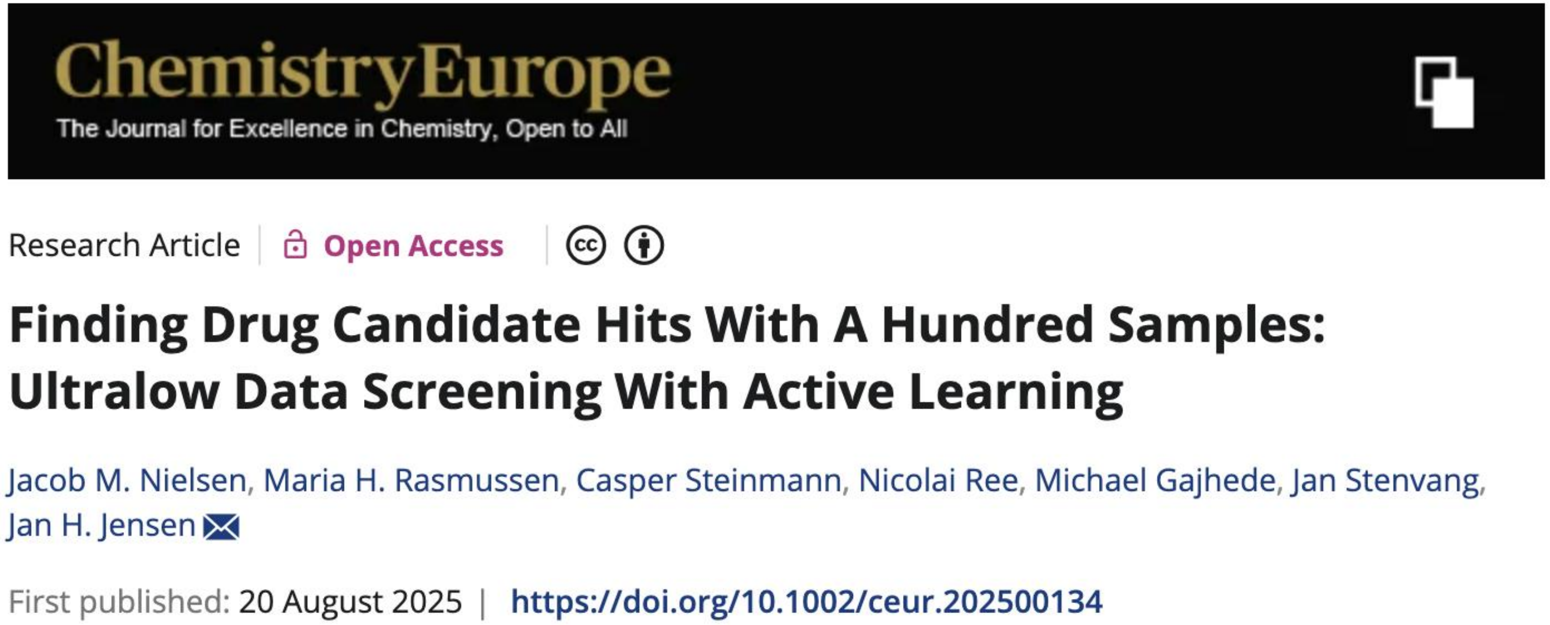
3 tricks

Pair difference learning for classification*

2D atomic pharmacophore model

Decision tree for insight/analysis

*also works for regression



ChemistryEurope

The Journal for Excellence in Chemistry, Open to All

Research Article | Open Access

Finding Drug Candidate Hits With A Hundred Samples: Ultralow Data Screening With Active Learning

Jacob M. Nielsen, Maria H. Rasmussen, Casper Steinmann, Nicolai Ree, Michael Gajhede, Jan Stenvang, Jan H. Jensen

First published: 20 August 2025 | https://doi.org/10.1002/ceur.202500134