Previously…

Intro to python

- Numpy for math and computations
- Pandas for dataframes
- Sklearn for ML

Concepts in ML

- Train & test data
- Different problems require different modelling approaches
- Preparing your data carefully is crucial

A peek at our data: DEL

- One protein, many molecules screened
- Training data is represented as molecular fingerprints
- Class (im)balance: most molecules don’t bind

Notebooks this week

Representing molecules

- Review molecular fingerprints
- Discuss how representation choices impact modelling
- Data shift at test time

Training on large data

- You have ~375K training points
- We really want to scale up to screen ~37 billion 😱
- What types of models are most suited to our setting?
- Evaluation & picking the best model

Crosstalk in practice

Enabling Open Machine Learning of DNA Encoded Library Selections to Accelerate the Discovery of Small Molecule Protein Binders

18 October 2024, Version 1

Working Paper

James Wellnitz, Shabbir Ahmad, Nabin Begale, Jermiah Joseph, Hong Zeng, Albina Bolotokova, Aiping Dong, Shaghayegh Reza, Pegah Ghiabi, Gibson Elisa, Xuemin Cheng, Guiping Tu, Xianyang Li, Jian Liu, Dengfeng Dou, Jin Li, Rachel J. Harding, Aled M. Edwards, Benjamin Haibe-Kains, Levon Halabelian, Alexander Tropsha, Rafael Couñago

Show author details

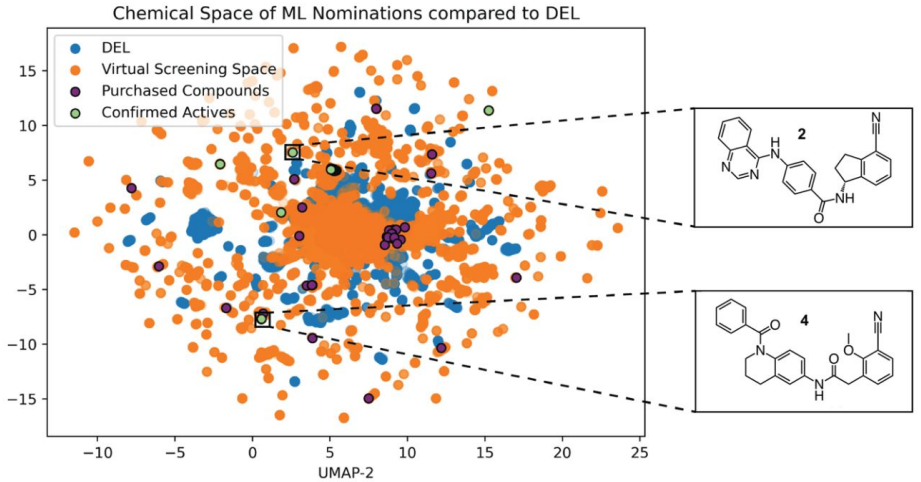

Figure 4: *Chemical Space occupied by the DEL training data from AIRCHECK (blue) and the virtual screening library (orange). The 50 ML model nominations are highlighted purple. All 7 confirmed actives overlap the space occupied by DEL. Fingerprint representations were reduced to two dimensions using UMAP*[38].