Negative data sets

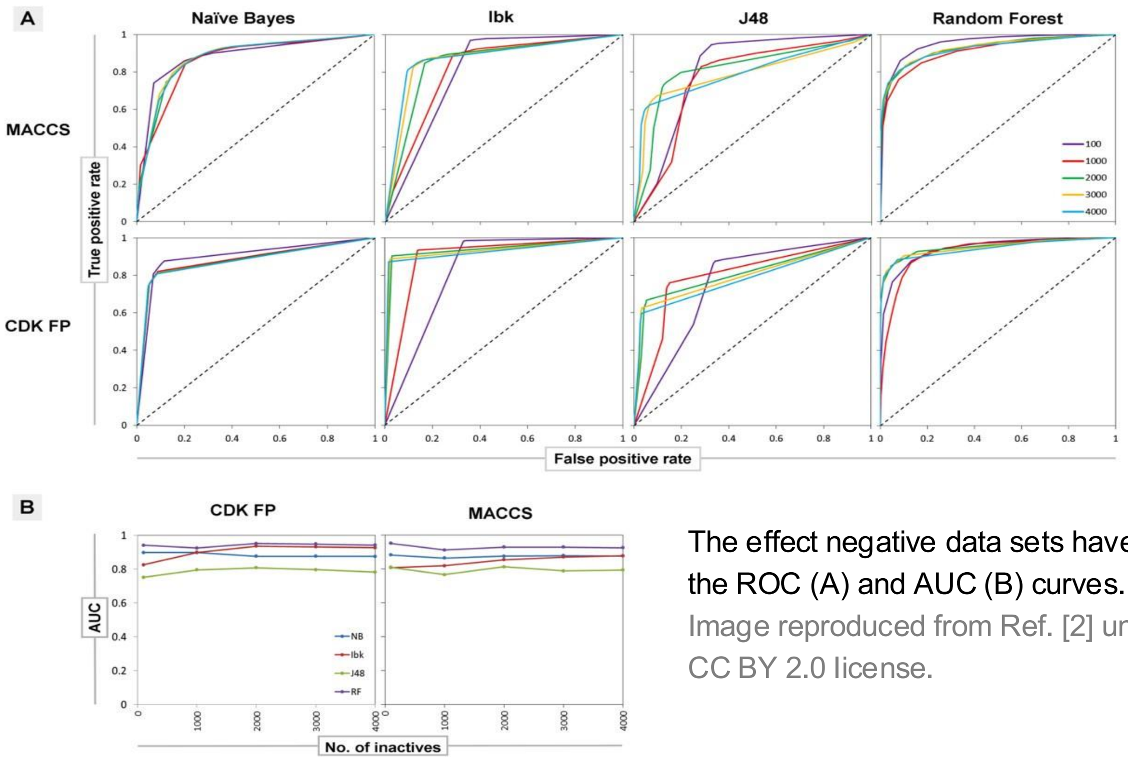
What are Negative Data Sets?

Negative datasets contain molecules that lack a target property or activity.[1]

When are they used?

They are crucial for developing accurate, unbiased, and generalised machine learning (ML) models.

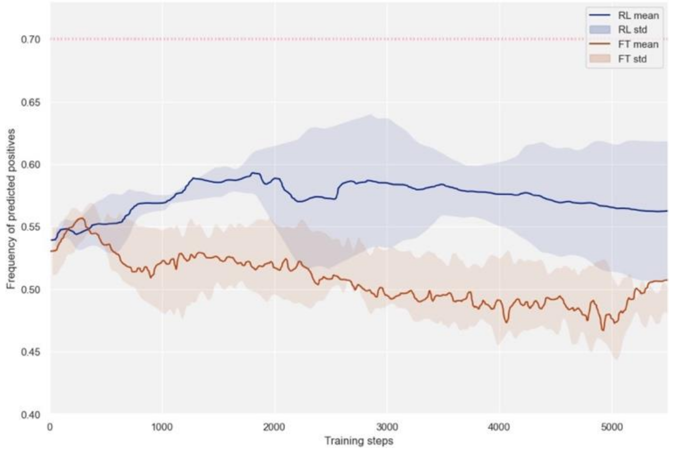
- Essential for active *vs* inactive
- Prevent overfitting
- Improve decision boundaries
- Critical in reaction predictions



The effect negative data sets have on the ROC (A) and AUC (B) curves. Image reproduced from Ref. [2] under CC BY 2.0 license.

Example:

predicting reaction outcome with language models.[3]



Frequency of correctly predicted positive reactions by the reinforcement learning and the fine-tuned models trained on hundreds of failed reactions and 20 positive reactions from the RegioSQM20 dataset. Image reproduced from Ref. [3] under CC BY 3.0 license.

Challenges

Most published papers only report

- Successful reactions
- High yielding conditions
- Active compounds
- Positive results

This happens because:

- Journals rarely accept "null results"
- Experimental time is limited; failures are not documented
- Labs don't want to publish unoptimised or low-yield data

Consequently, negative data is often synthetic, non-real data that can lead to poor real-world predictions. Such negative data is often not representative of real chemical space. They are often:

- "Easy" (decoys look obviously different from actives/positives);
- Unbalanced (orders of magnitude more negatives than positives);
- Homogeneous (same types of scaffolds repeated).

References

[1] *J. Org. Chem.* **2023**, *88*, 5239. [2] *J. Cheminform.* **2014**, *6*, 32. [3] *Sci. Adv.* **2025**, *11*, eadt5578.