In Silico Isomerization Produces Apt Negative Data for VHTS Validation

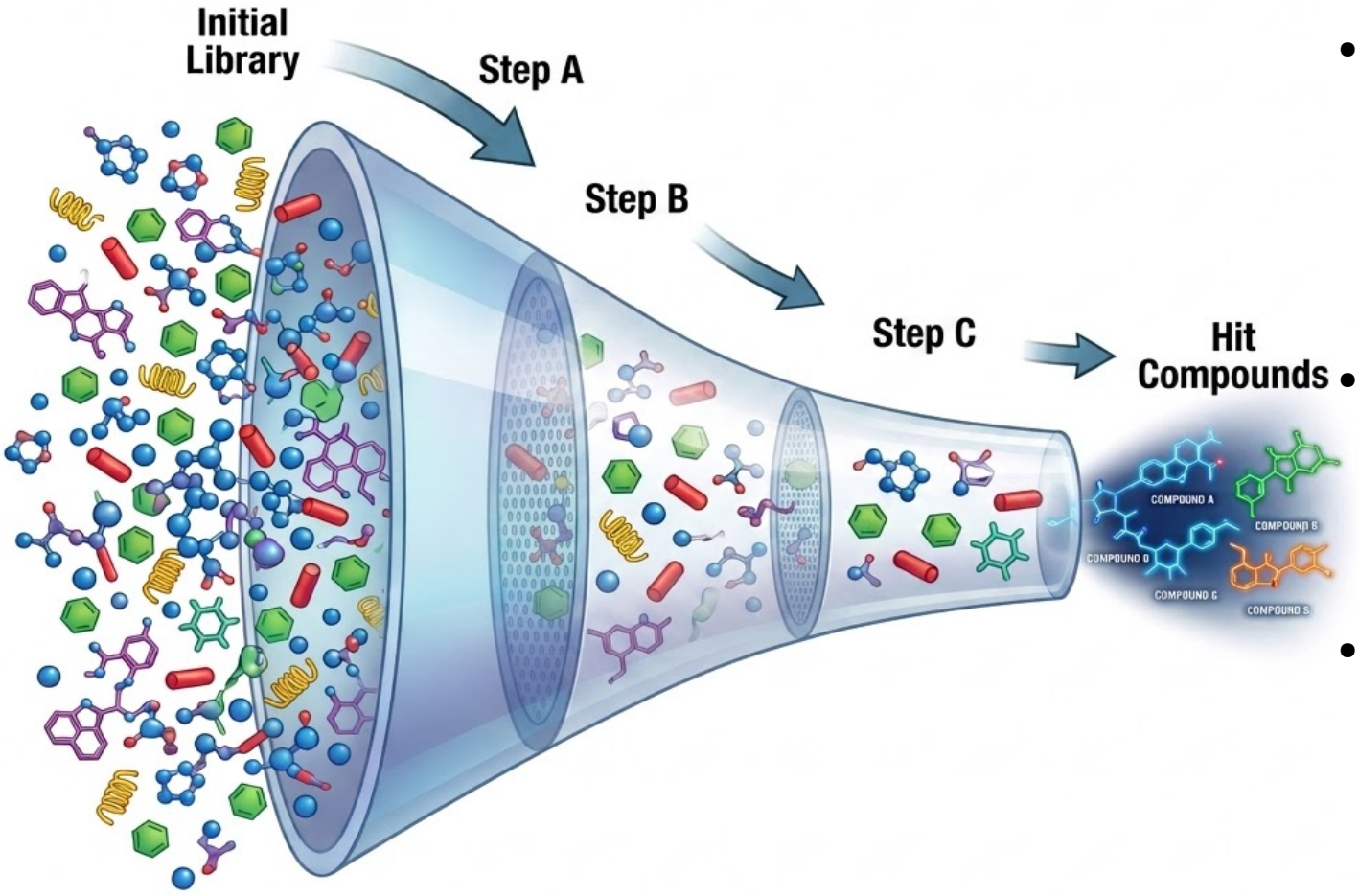
Stefan Ivanov, PhD

Medical University of Sofia
Bulgarian Academy of Sciences

Central & Eastern European Cheminformatics Conference 2026
07.09.2026, Prague, Czech Republic

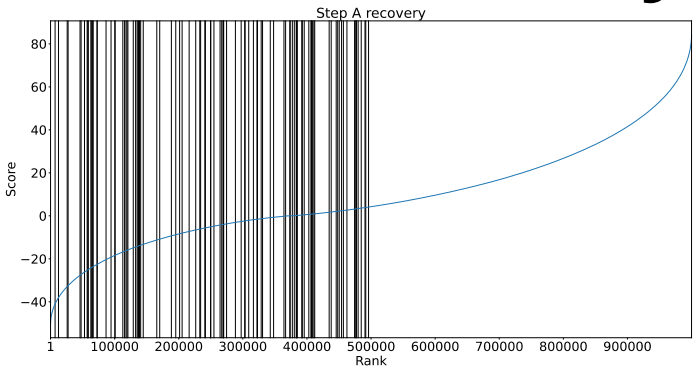
sivanov@ddg-pharmfac.net
stefan@biophys.bas.bg

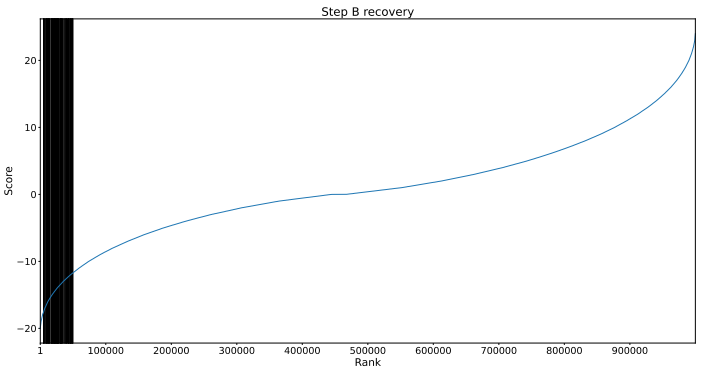
VHTS Funnel

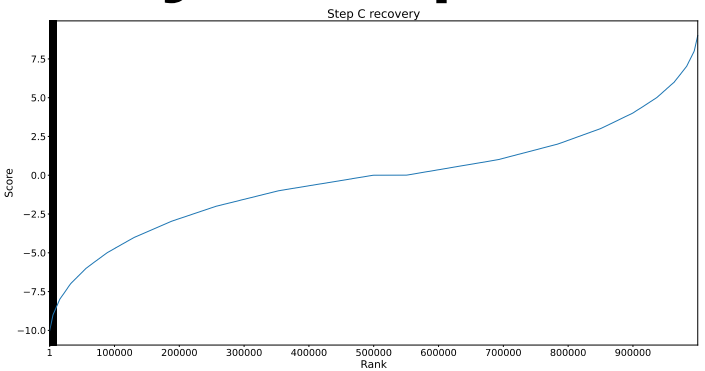


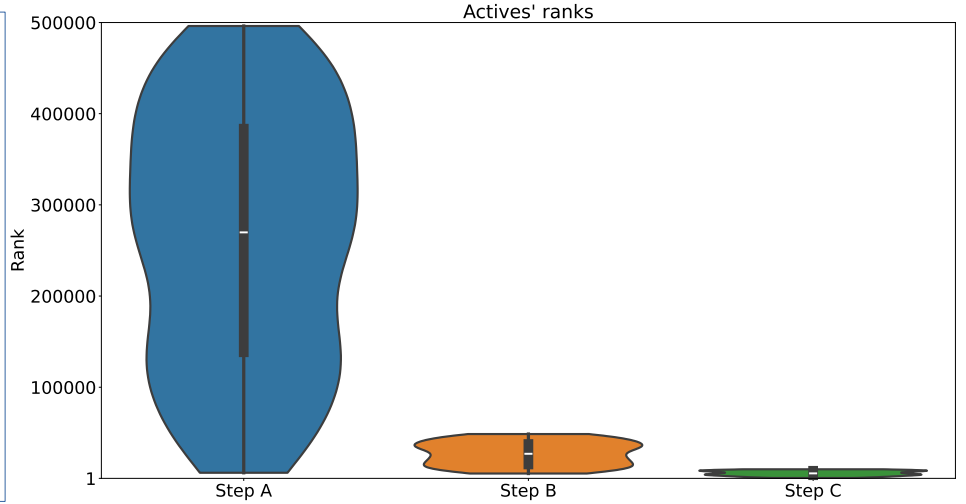
- VHTS pipelines typically have multiple steps, but performance is usually evaluated for the entire pipeline.
- Performance metrics are hard to specify because the random base rate for the initial set is unknown.
- Unproductive or even counterproductive steps or elements can easily avoid detection.

Recovery Analysis at Every Step









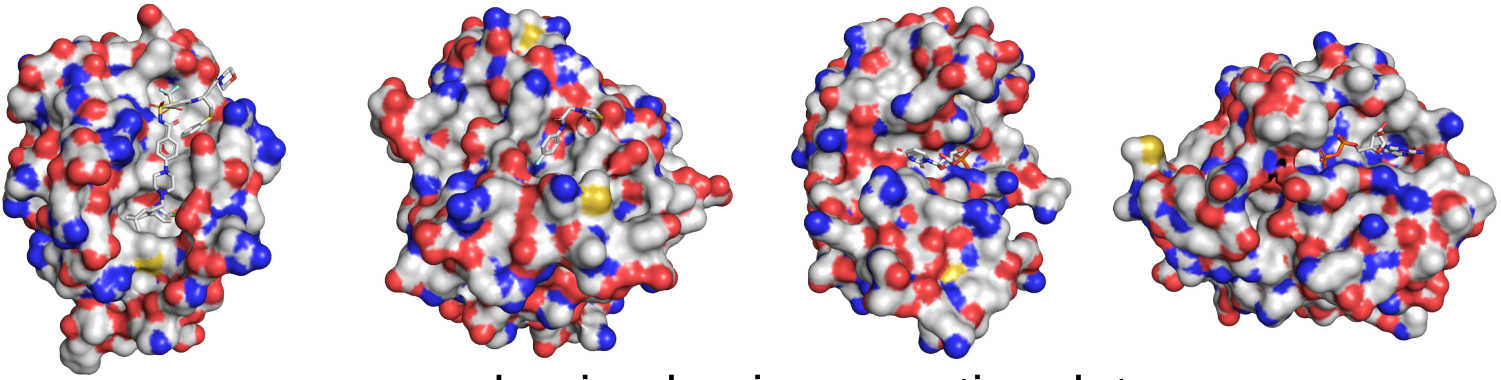
Ivanov S. M. (2026) An Efficient Computational Chemistry Approach to Generating Negative Data for Drug Discovery Pipeline Validation. *Front. Bioinform*. 6:1756279

Ivanov S. M. (2026). *In silico* isomerization produces apt negative data for VHTS validation. *JCIM.* 66(15):9246–9261

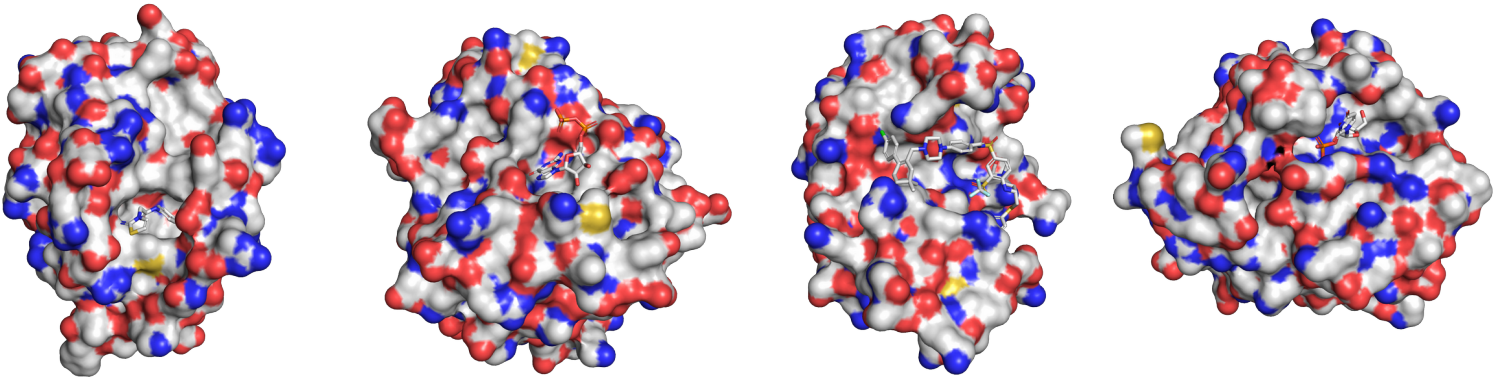
- Using sets of known binders and decoys, performance should be evaluated at every step.
- When using stochastic or quasi-stochastic techniques, estimates for ranking variation should be provided.

Negative Data From Ligand Randomization

cognate protein – ligand pairs, positive data



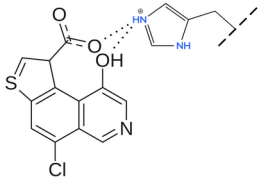
randomized pairs, negative data



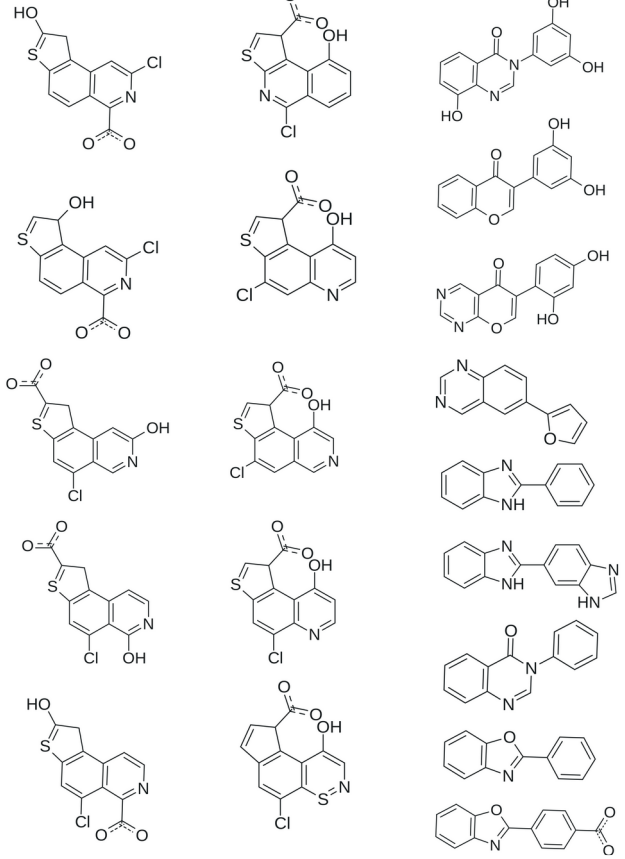
- Randomizing ligands across curated 3D structures provides ample negative data all the while ensuring a perfect match in properties between actives and decoys.
- No other means of decoy selection can provide such a match.

Negative Data From Ligand Isomerization

ligand from crystal structure



isomers, likely nonbinders

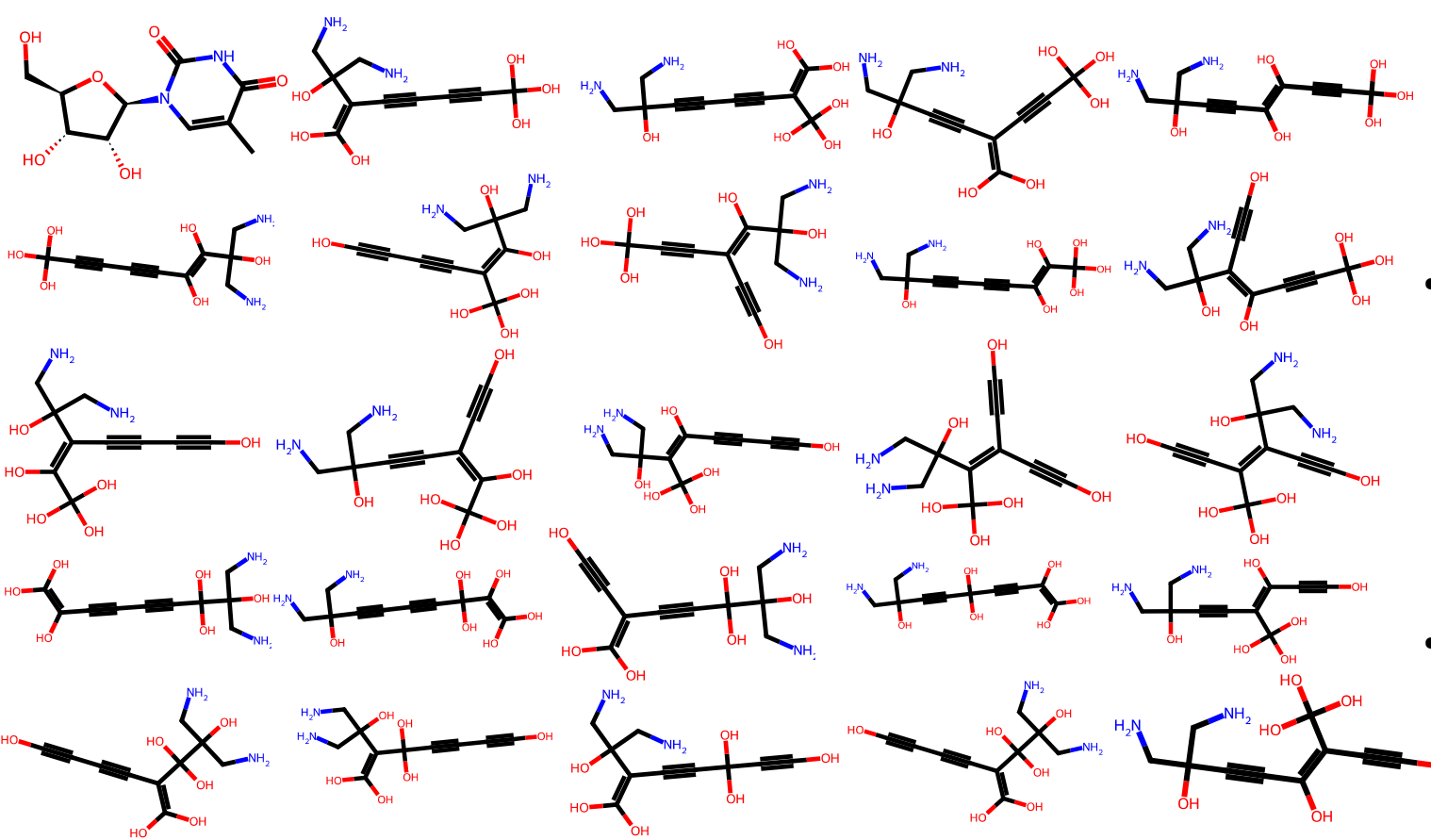


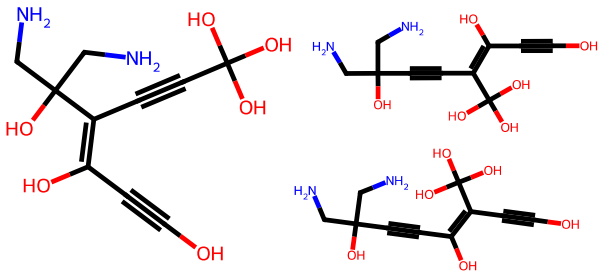
isomers, likely binders

library ligands

- Isomerization maintains atomic composition which is a major determinant of chemical properties. Therefore, isomerization is an astute approach to generating negative data from the positive data.

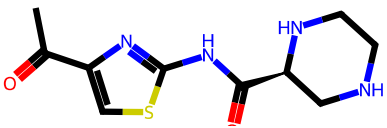
Negative Data From Ligand Isomerization

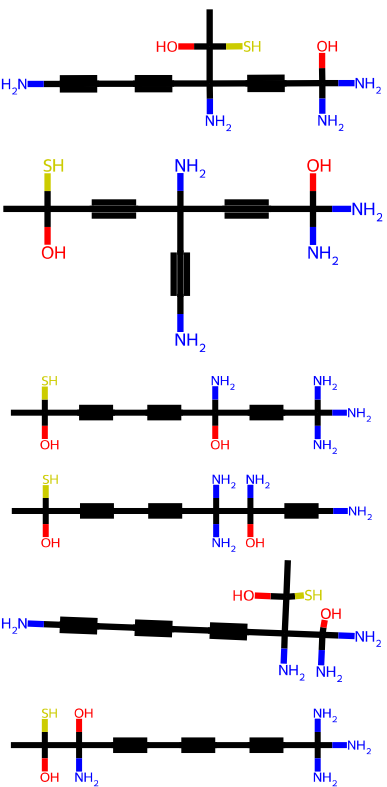




- Despite alternative conformations of the orthosteric binding site and the possibility of alternative/alosteric binding sites, overall random binding rates are low.
- The risk of false negatives is likely to be especially low in the low Tanimoto similarity regime.

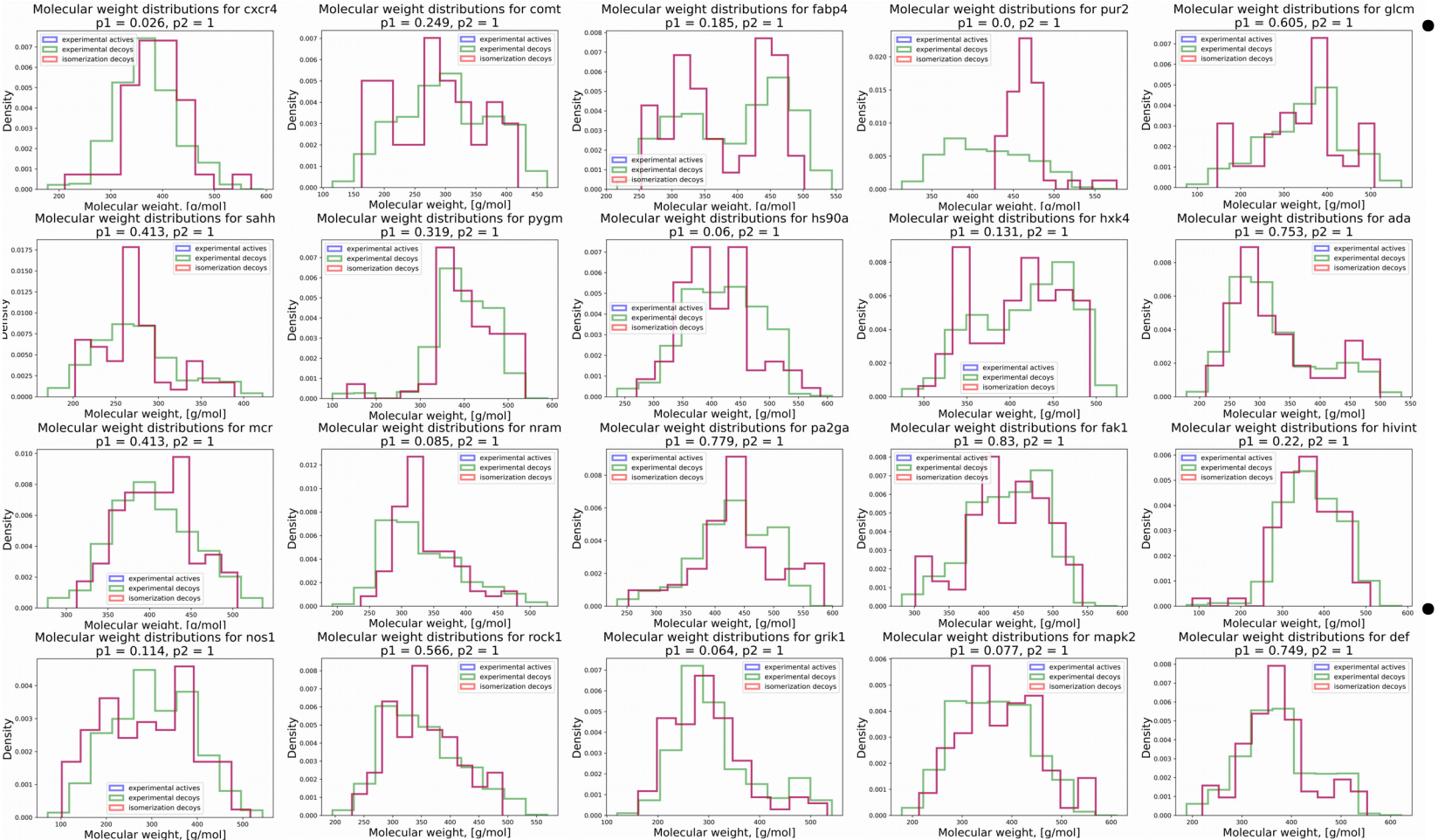
Negative Data From Ligand Isomerization





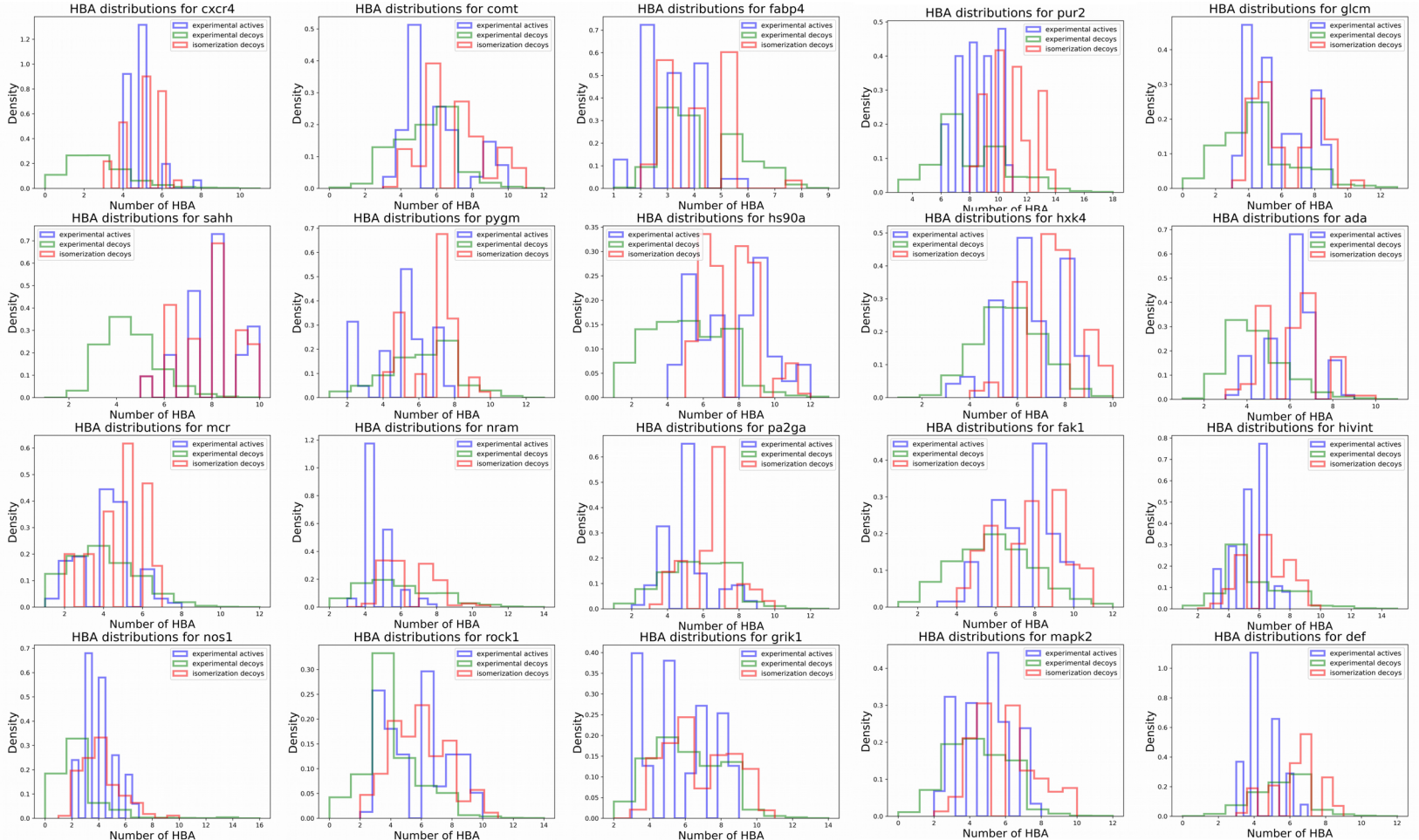
- Primary interest lies in relative performance; whatever false negatives there are, are treated as negatives. This applies to any negative sets, be it experimental or *in silico*.

Property Matching



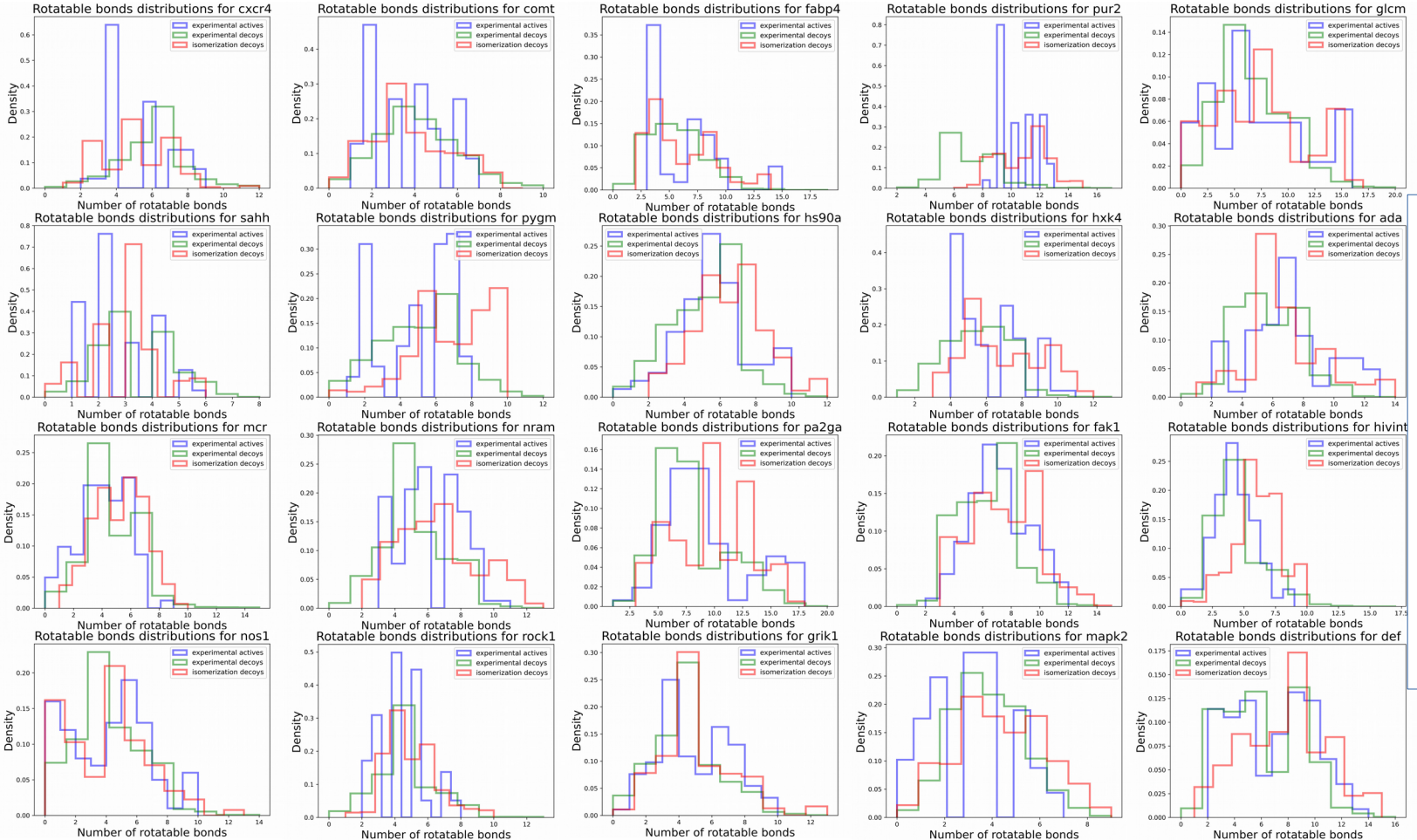
- Simple but important properties such as molecular weight, numbers of HBA, HBD, rotatable bonds, TPSA or total surface area can be more finely controlled and matched.
- Existing sets can have considerable mismatches between actives and decoys.

Property Matching



- Property matching and filtering can also be applied to existing collections.
- Isomerization preserves atomic composition and thereby offers a “shortcut” in property hyperspace to better matched decoys.

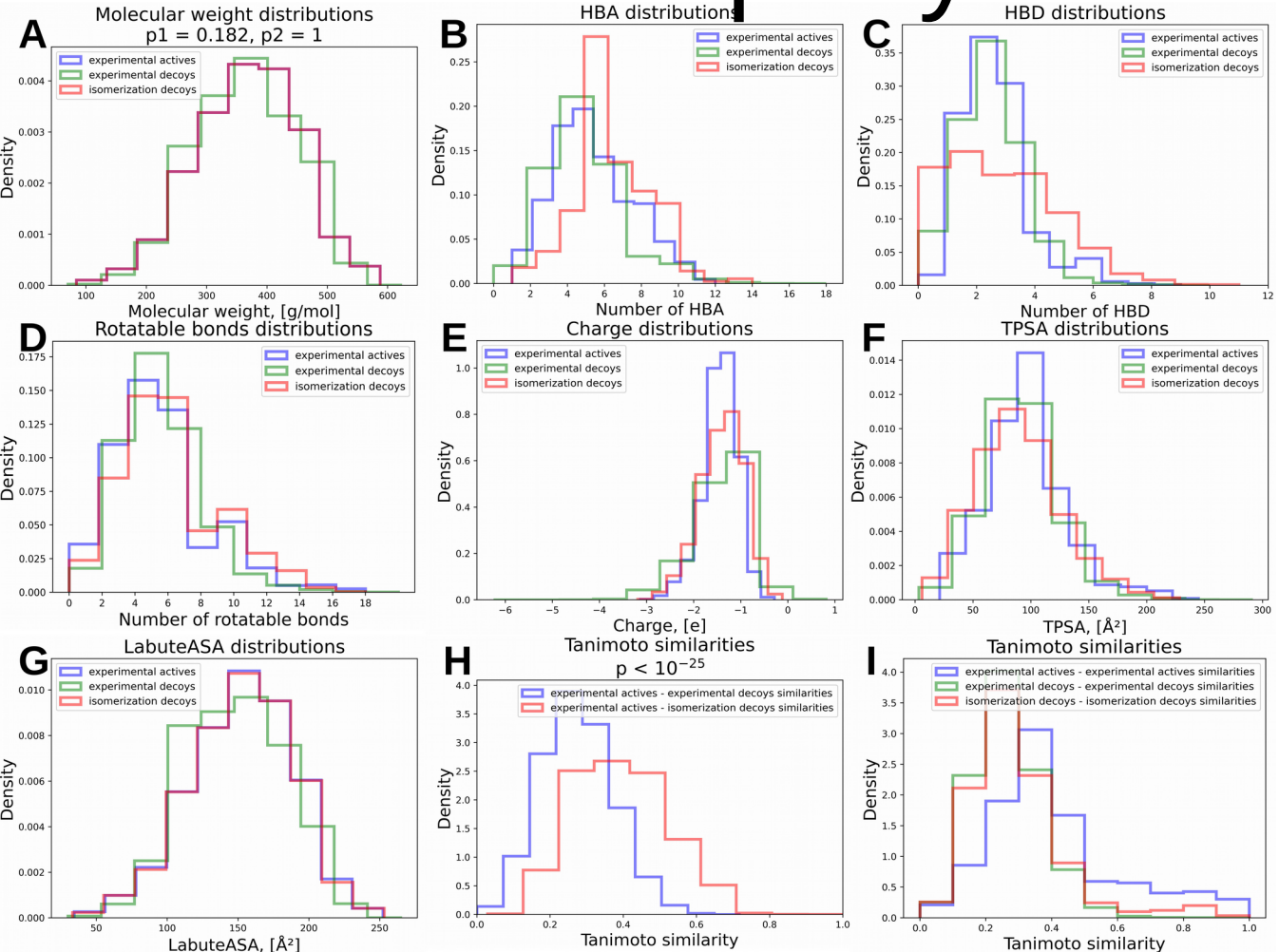
Property Matching



Ivanov S. M. (2026) An Efficient Computational Chemistry Approach to Generating Negative Data for Drug Discovery Pipeline Validation. *Front. Bioinform.* 6:1756279

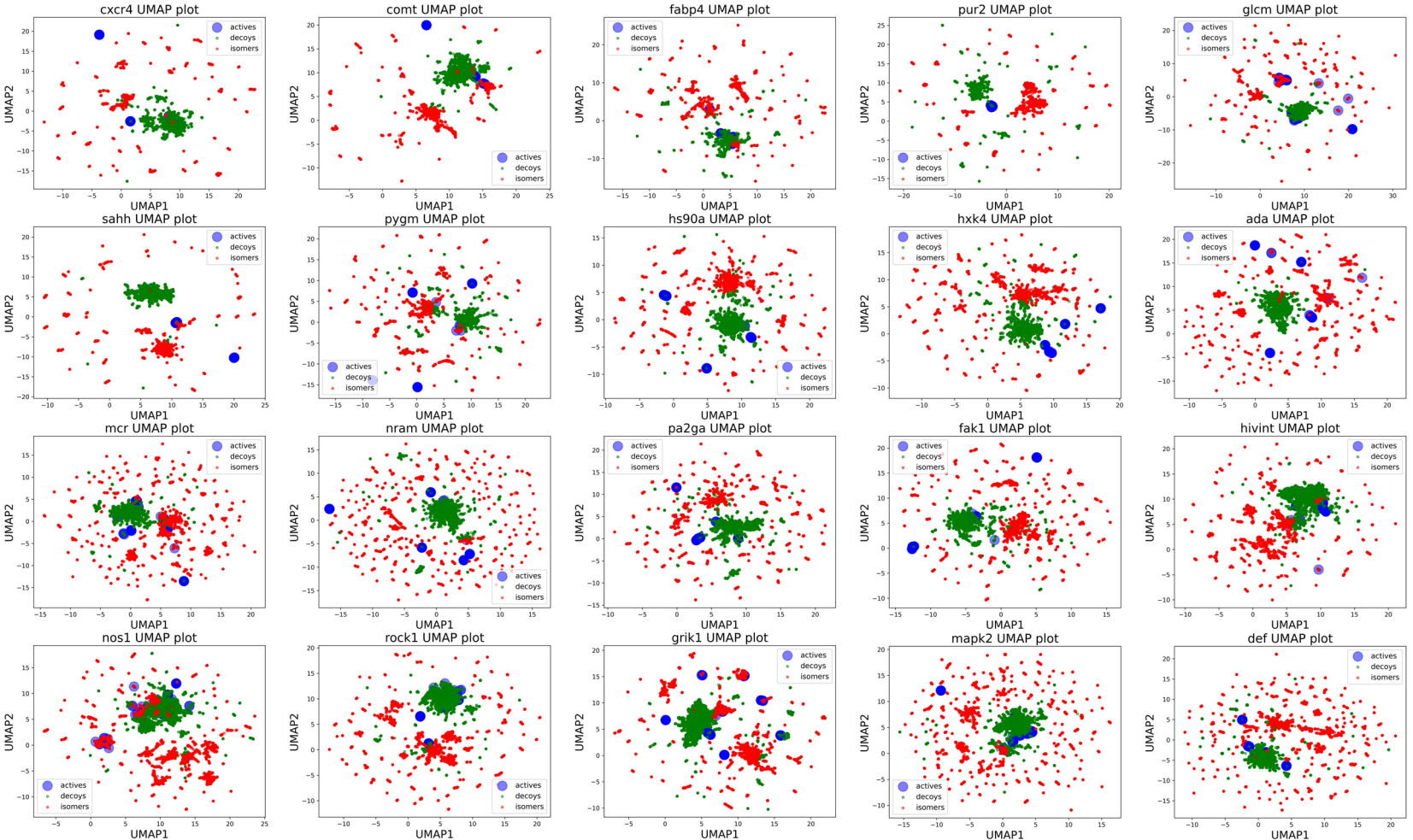
Ivanov S. M. (2026). *In silico* isomerization produces apt negative data for VHTS validation. *JCIM.* 66(15):9246–9261

Property Matching



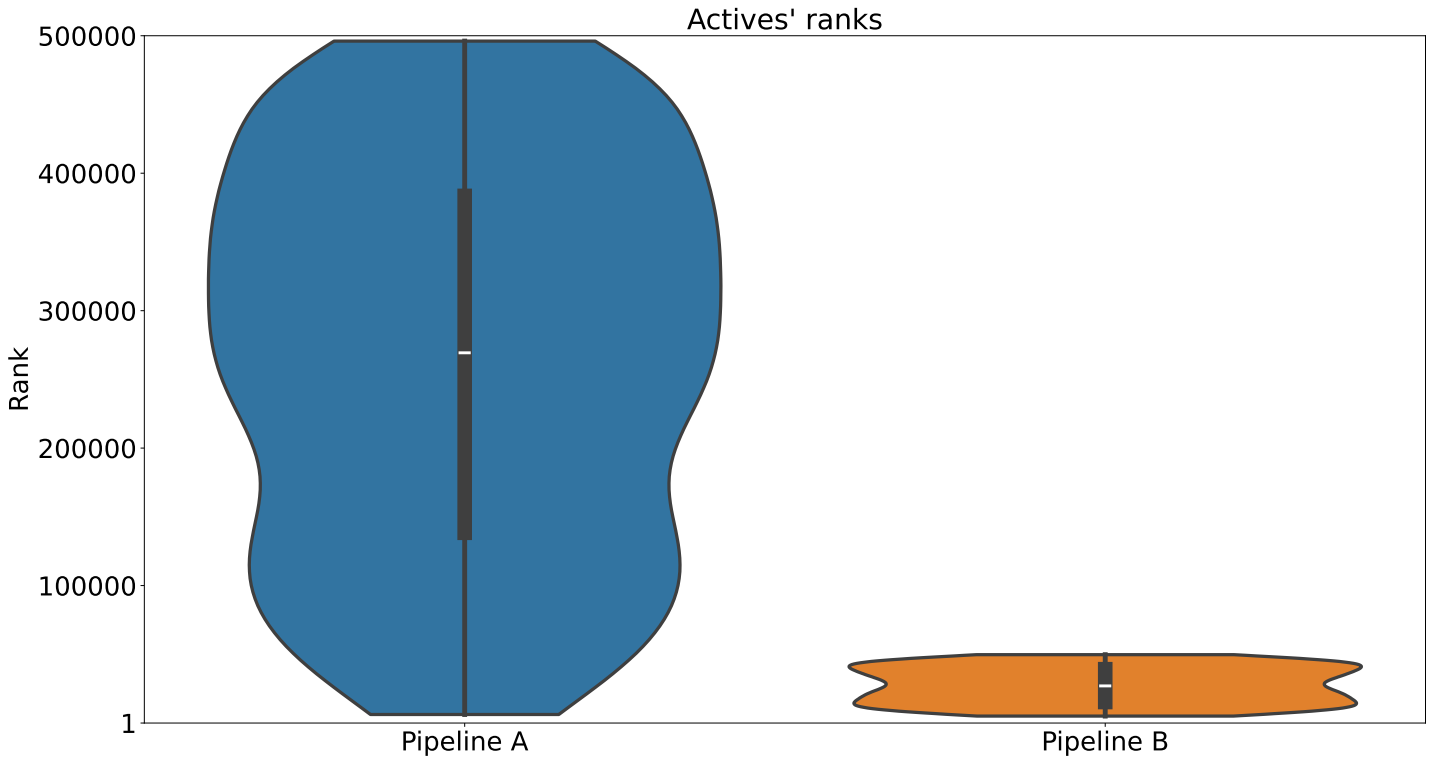
- Isomers perfectly match actives in molecular weight (MW) and are similarly or better matched in HBA, HBD, rotatable bonds, and molecular surface.
- Different properties vary in how strongly they depend on atomic connectivity. This work shows that for MW, TPSA, and LabuteASA, the ordering is MW < TPSA < LabuteASA.
- Crucially, isomers are more similar to the actives as evidenced by the Tanimoto distributions.
- Around half of isomers lie in a “Tanimoto Goldilocks zone’ between 0.4 and 0.8 whereas DUD-E decoys are often completely unrelated to the actives. This increases the risk of inflating performance metrics.

UMAP Plots



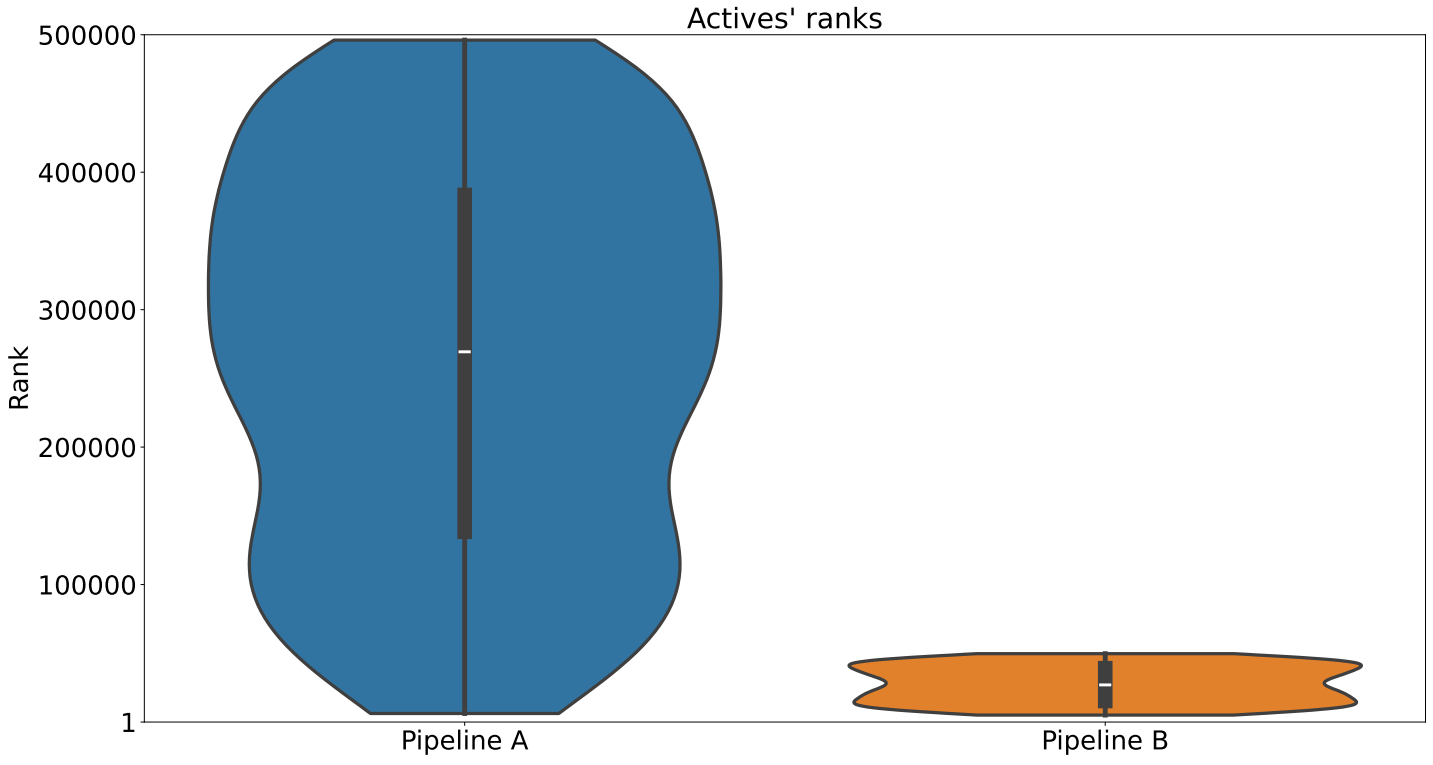
- In reality, there are single-digit numbers of active data points per target, slight variations of no more than a handful of scaffolds.
- Decoys are little better in this regard, often clustering closely in a narrow patch of property hyperspace.
- Isomer decoys tend to be much more diverse, i.e. much more spread out over that same hyperspace.

Implications for Method Development



- Do improvements to water models, force fields, or property estimates (e.g. entropy, solvation, RDFs, etc.) translate into improvements in VHTS? Why?
- Recovery is orthogonal to (most) parameters used in method development.

Implications for Metaphysics



- How interdependent are different properties with each other and with recovery?
- Is there more than one algorythm to obtain ΔG?
- Is there more than one algorythm to correctly rank ligands by affinity to a target?

Thank You. Q&A