The value of shape and electrostatic similarities in deep generative methods

*Dr **Jonas Boström***

Medicinal Chemistry, Research and Early Development Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

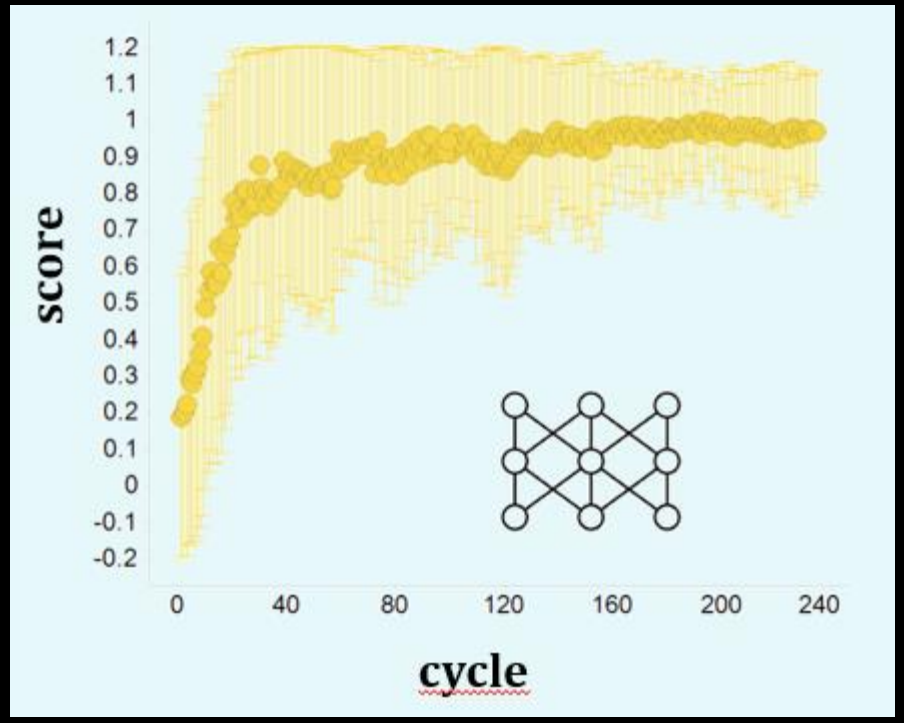
email: jonas.bostrom@astrazeneca.com

Twitter: @DrBostrom

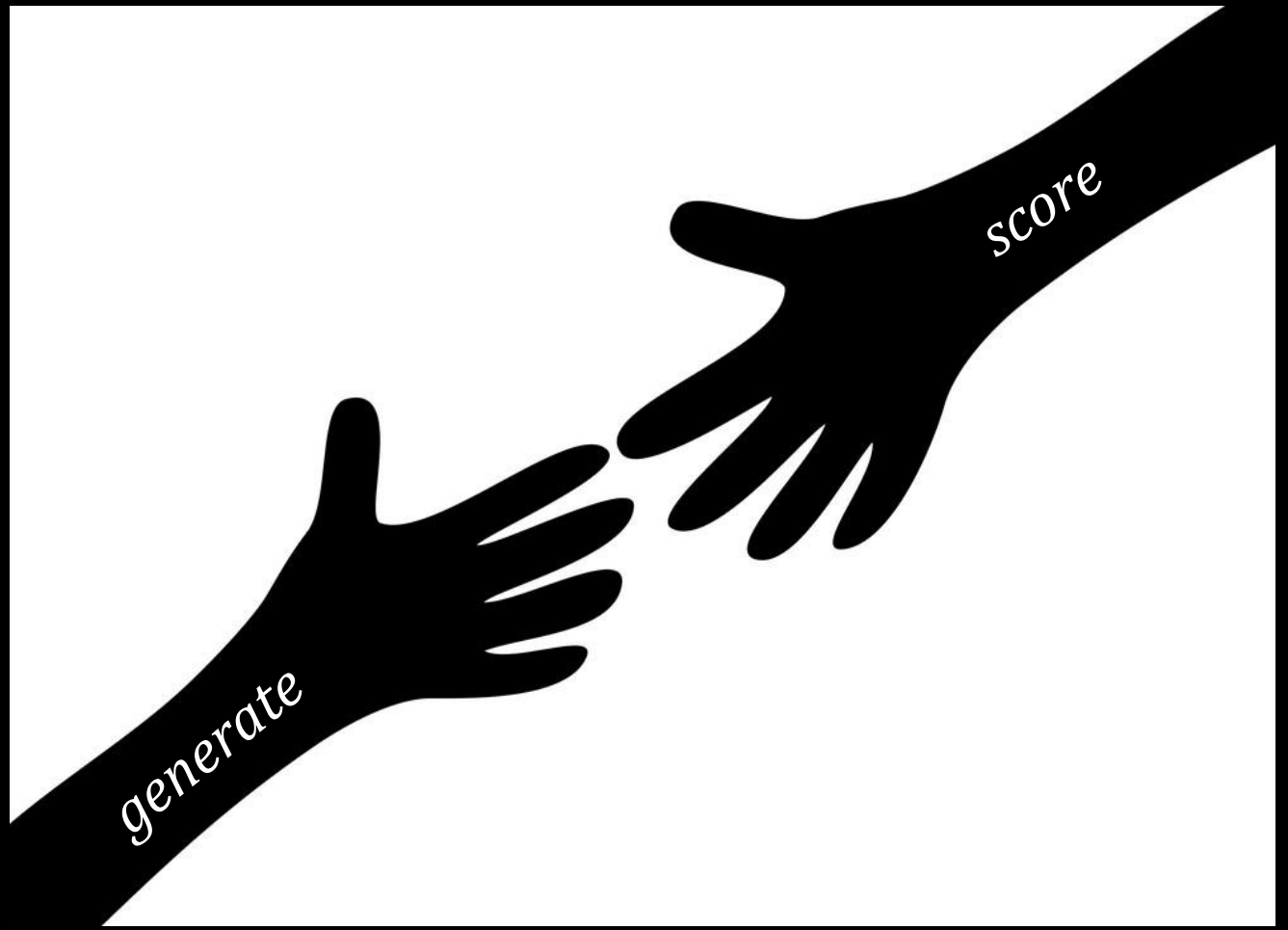
RDK UGM 2021, October 14th

Problem 1. Scoring

There are now many, many new cool ways of generating molecules.
But…that's the "easy" part of the problem.
Scoring remains a major challenge,
and is less studied.



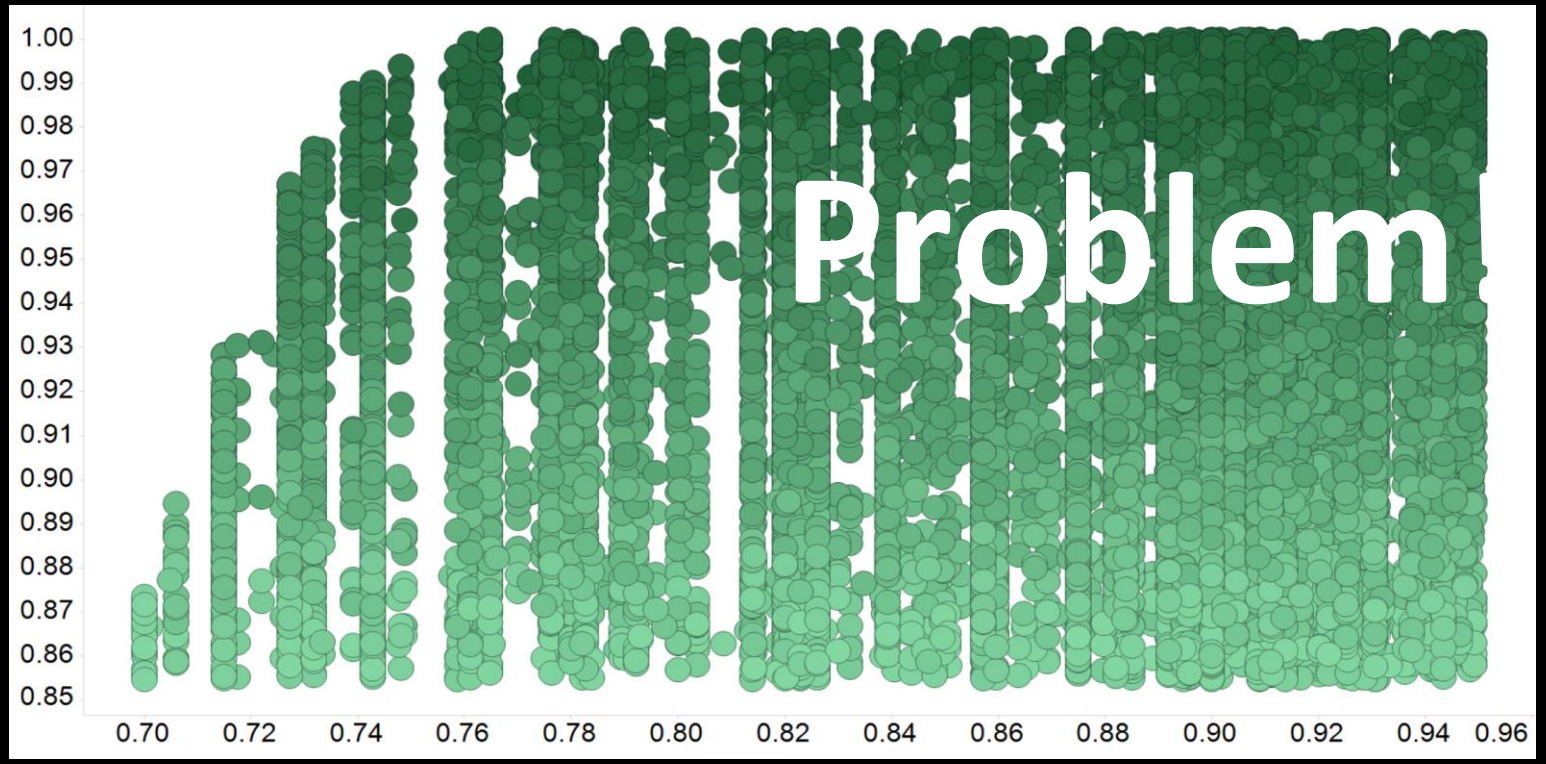
reinforcement learning



This is what typically happens in ”real-life”

Tens of thousands of novel high-scoring compounds are ”AI”-generated.

How to select which to make in the lab?



We are (generally) *not* at the stage where we can single out a few compounds with precision from the many.

Biasing the generation towards compounds “similar” to known actives?

Similar molecules tend to have similar properties – use similarity as score.

Many *de novo* methods focus on “diversity” and “novelty”.

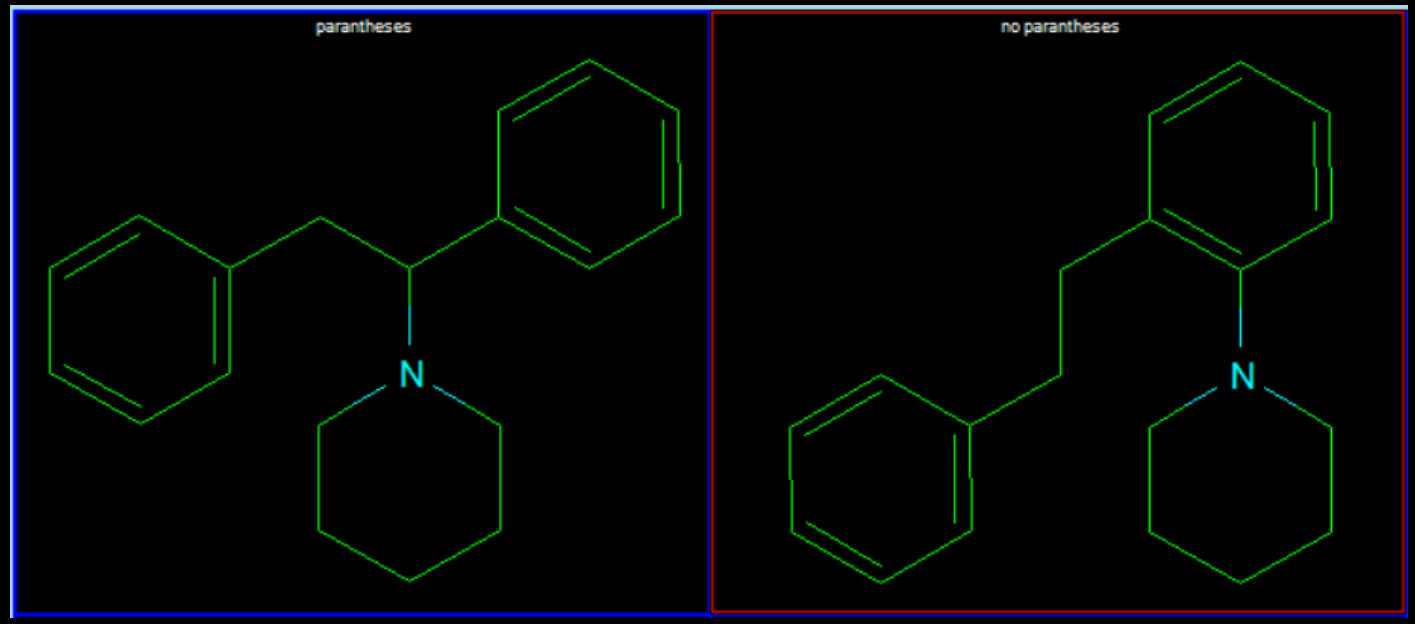
Focus on similarity implies interpolation. Providing higher confidence in predictions.
Large structural changes (extrapolation) are more uncertain.

Problem 2. Molecular Representations

SMILES strings are often used

….convenient for the AI algorithm, since a string is trivial to manipulate and transform.

However, a conservative modification to a SMILES string may cause a large effect in their 3D structure.
(e.g. removal of brackets denoting substitution, and a Y-shaped compound becomes linear)



`c1ccc(cc1)CC(c2ccccc2)N3CCCCC3`

`c1ccccc1CCc2ccccc2N3CCCCC3`

There are graph-based and other useful methods like DeepSMILES and SELFIES….but

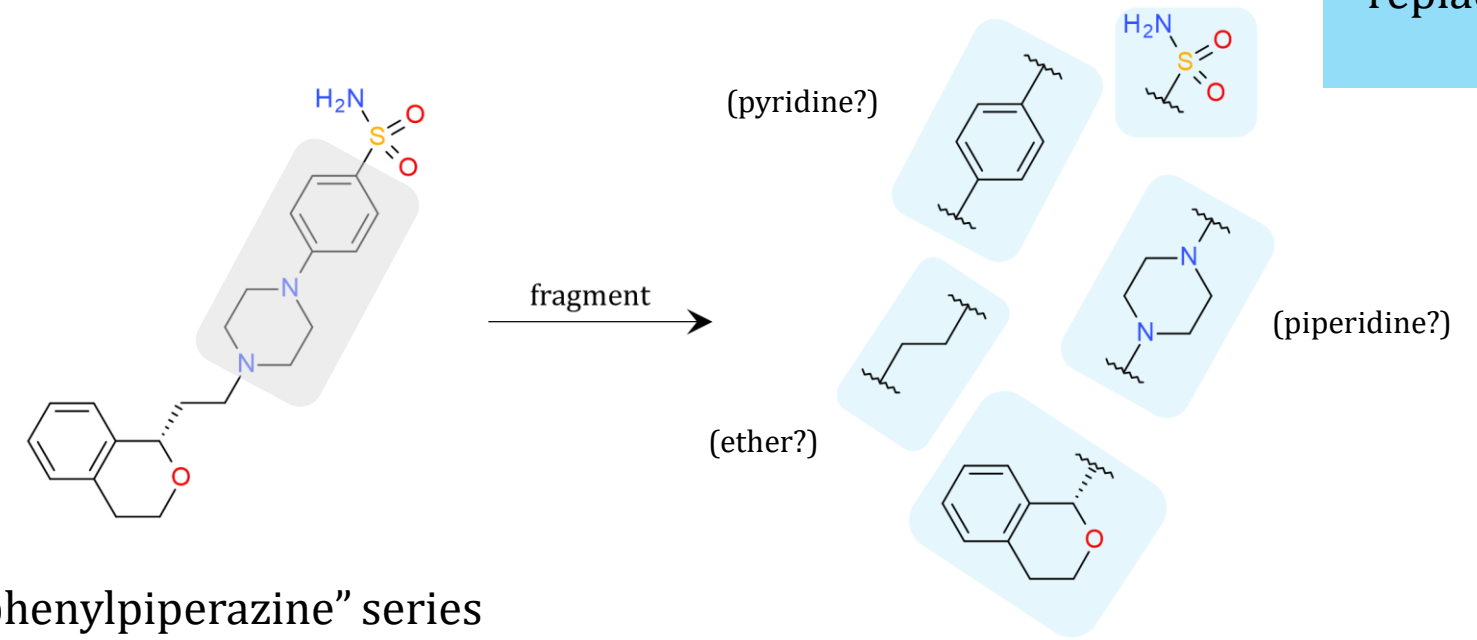
How about using Fragments?

Drug hunters tend to retain certain functional groups our designs for good reasons like

- maintaining essential ligand-protein interactions and
- for being able to use established synthetic routes with available chemical intermediates
- …

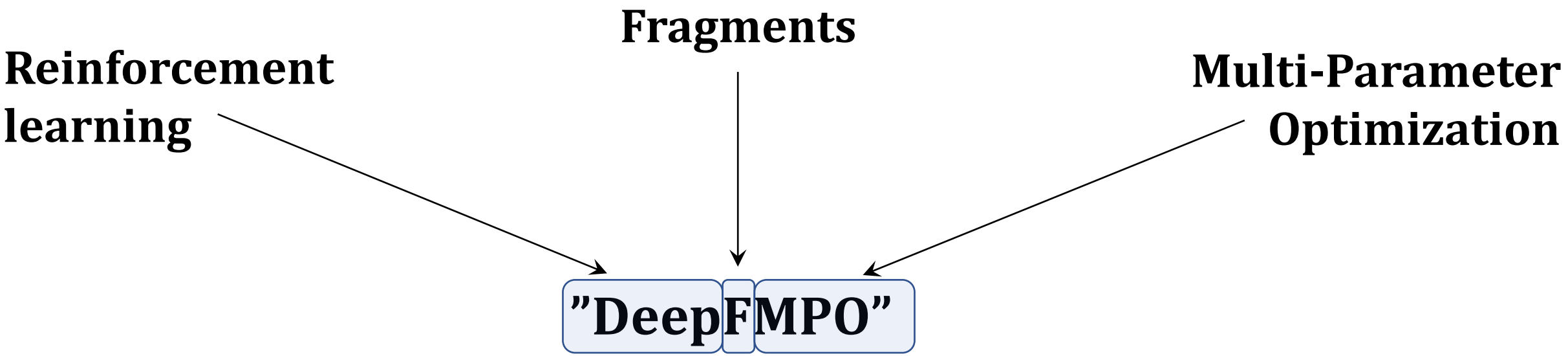
In line with the concept of structural series

An automated method replacing with similar fragments?



The “phenylpiperazine” series

We developed a fragment-based method



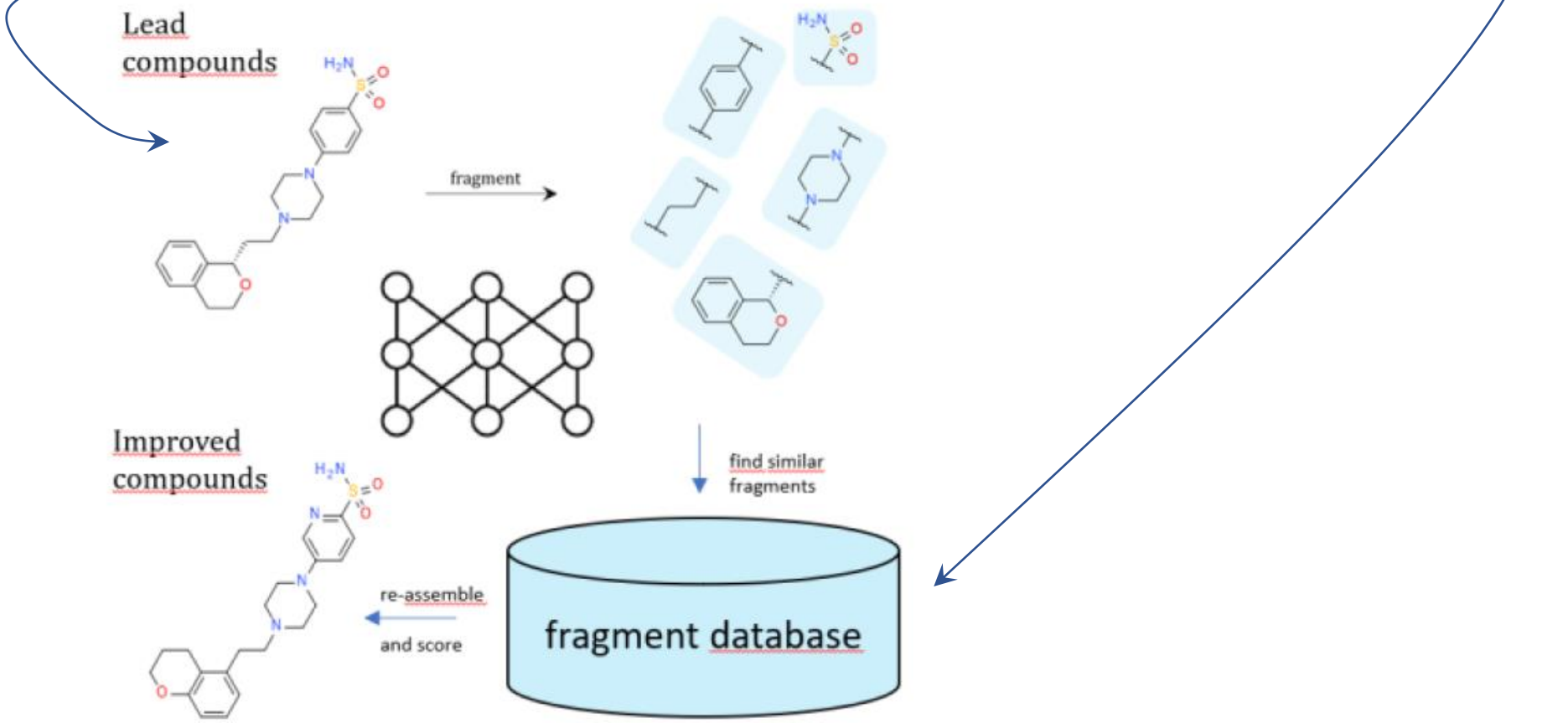
… aims to be a user-friendly way for the automatic generation
of structurally similar compounds fulfilling a set of given (sweet-spot) criteria.

But it’s not 3D

https://pubs.acs.org/doi/abs/10.1021/acs.jcim.9b00325

How DeepFMPO works

```
python Main.py fragment_molecules.smi lead_file.smi
```



The balanced binary tree...

...is generated from the fragment_file

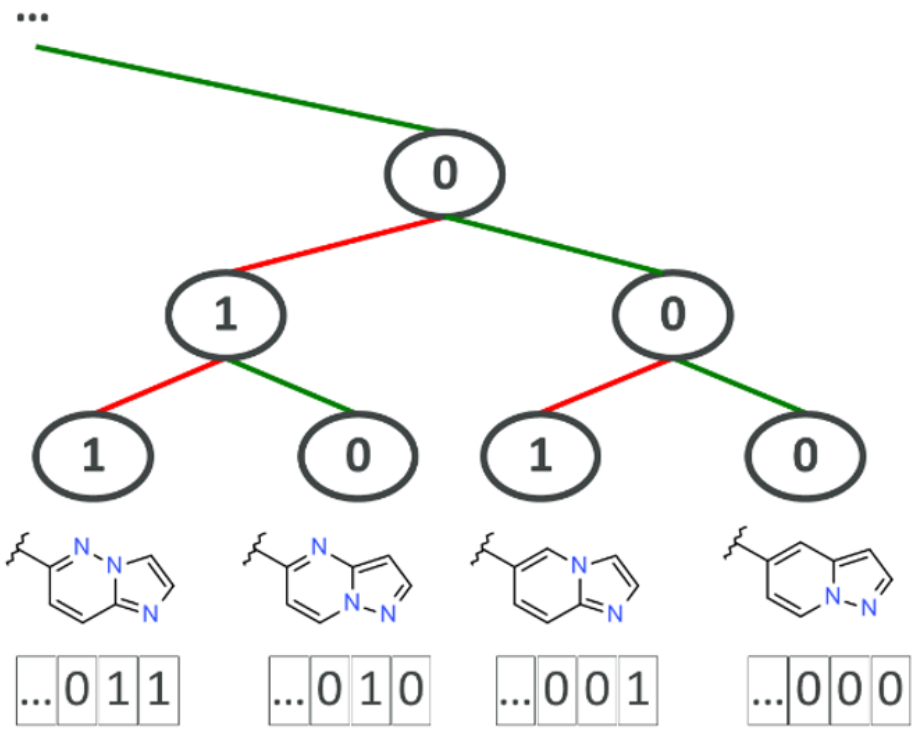
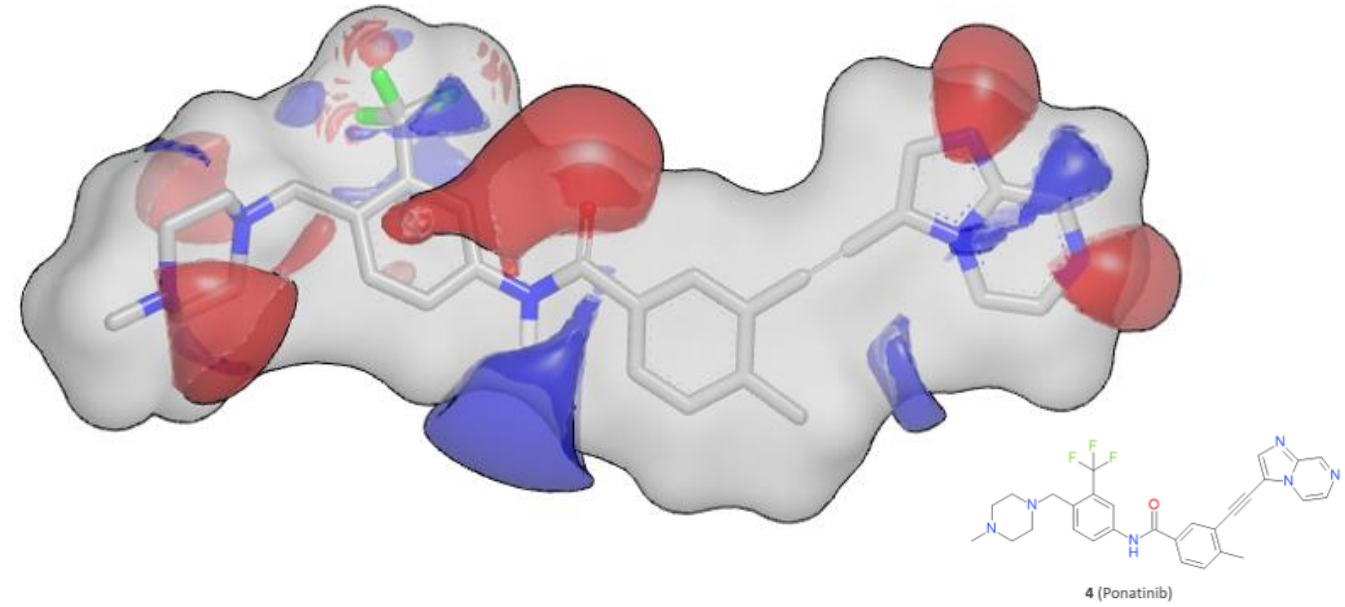


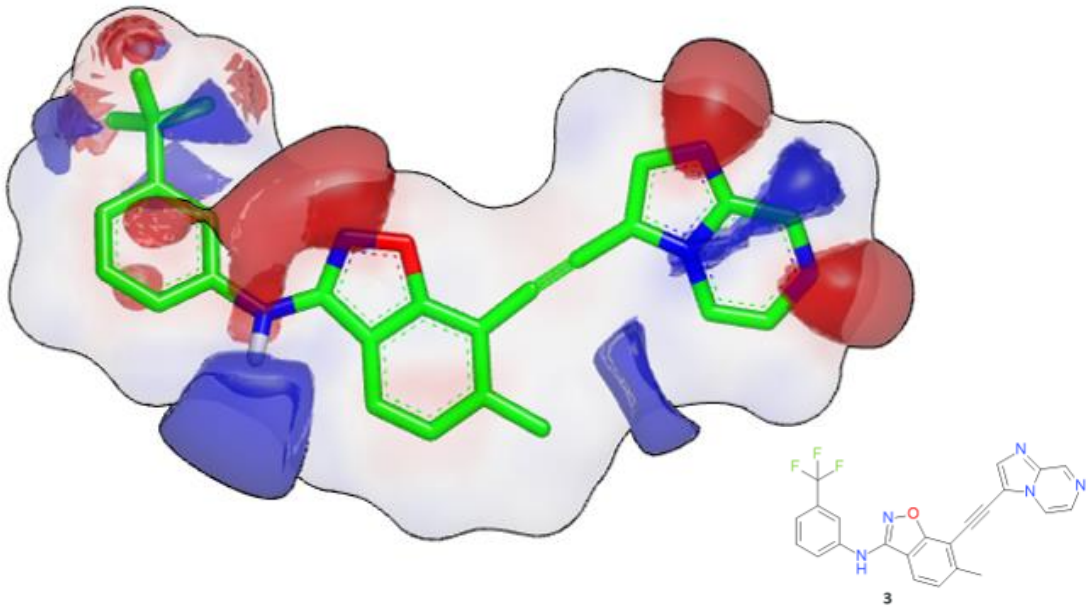
Figure 1. A snippet of the balance binary tree used in DeepFMPO. Fragments that are similar are placed close to each other. The encoding of a fragment is determined by the path from the root to the leaf. Every branching to the left adds a "1" to the end of the encoding and a branching to the right adds a "0".

Molecules are three-dimensional

Shape and electrostatic properties of molecules are primary determinants of molecular recognition – 3D *should* be better!



4 (Ponatinib)



3

…it has proven problematic to fully realize advantages of 3D based techniques.

mainly due to the difficulty of obtaining accurate descriptions of molecules bioactive conformations, as well as the subsequent molecular alignments.

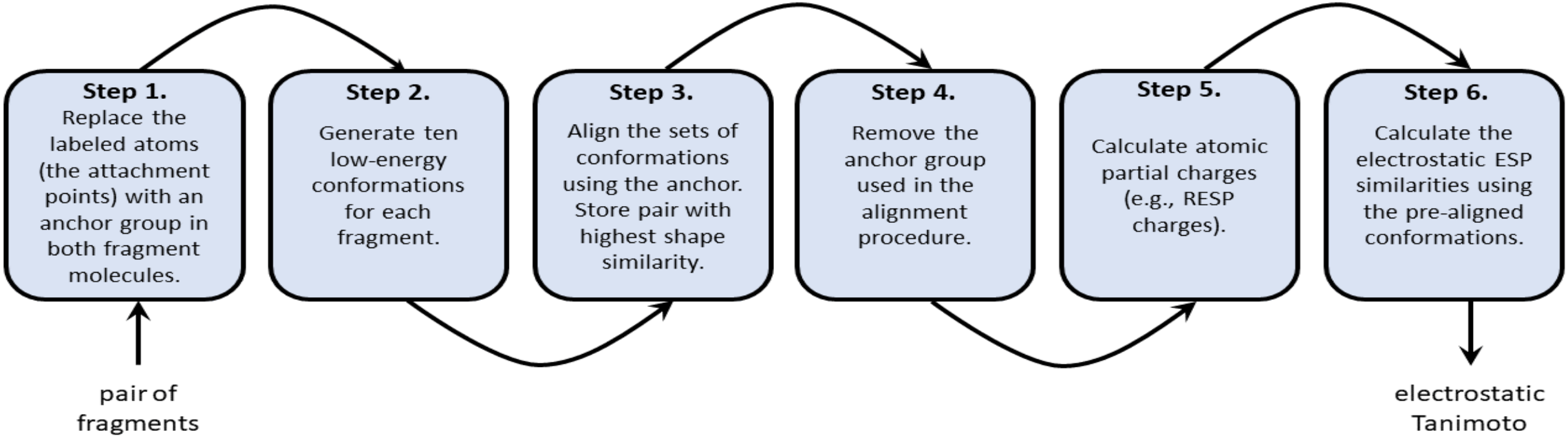
2D depictions from Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 2019, 37, 1038.

Replace the similarity calculations (2D->3D)

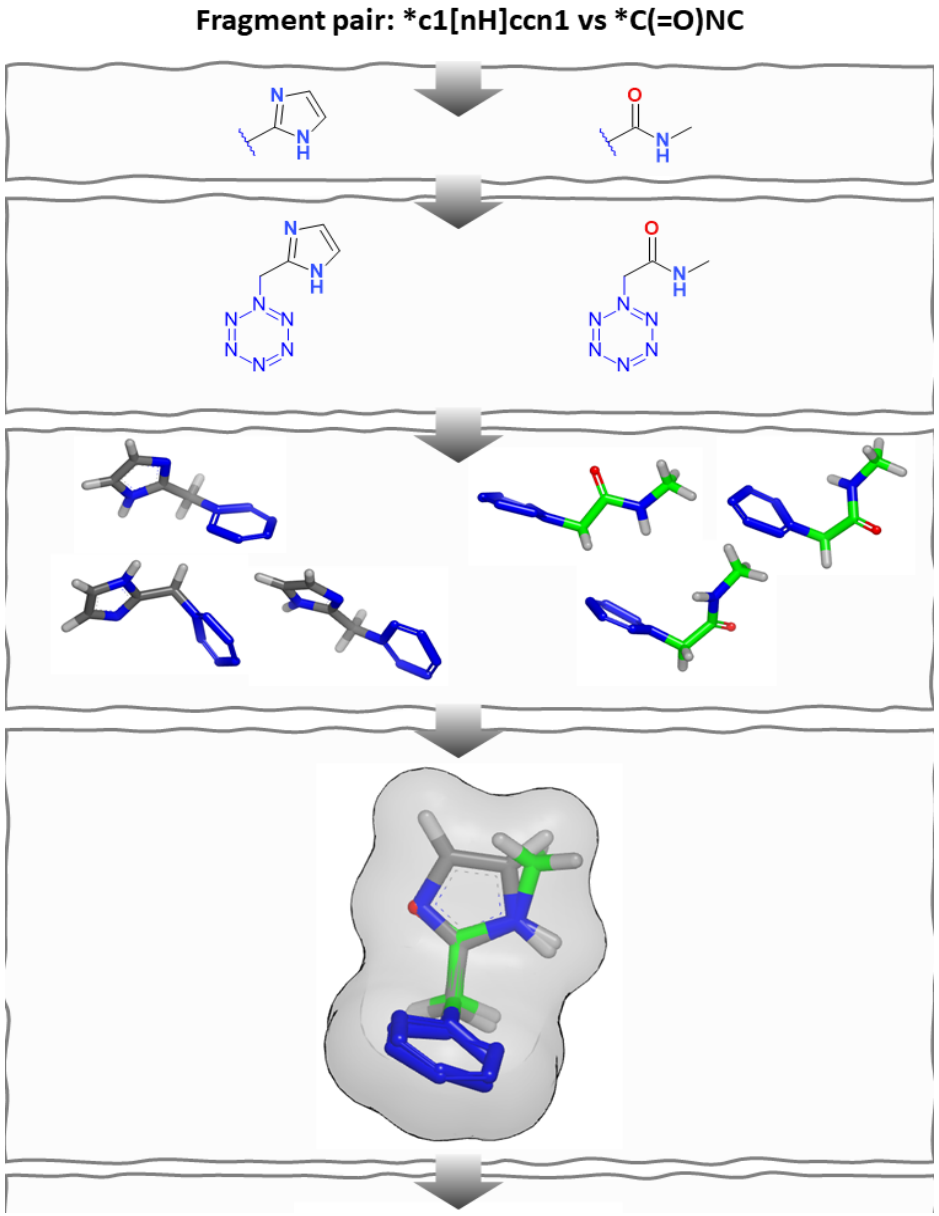
As before, the generative process outputs optimized molecules similar to the input structures.

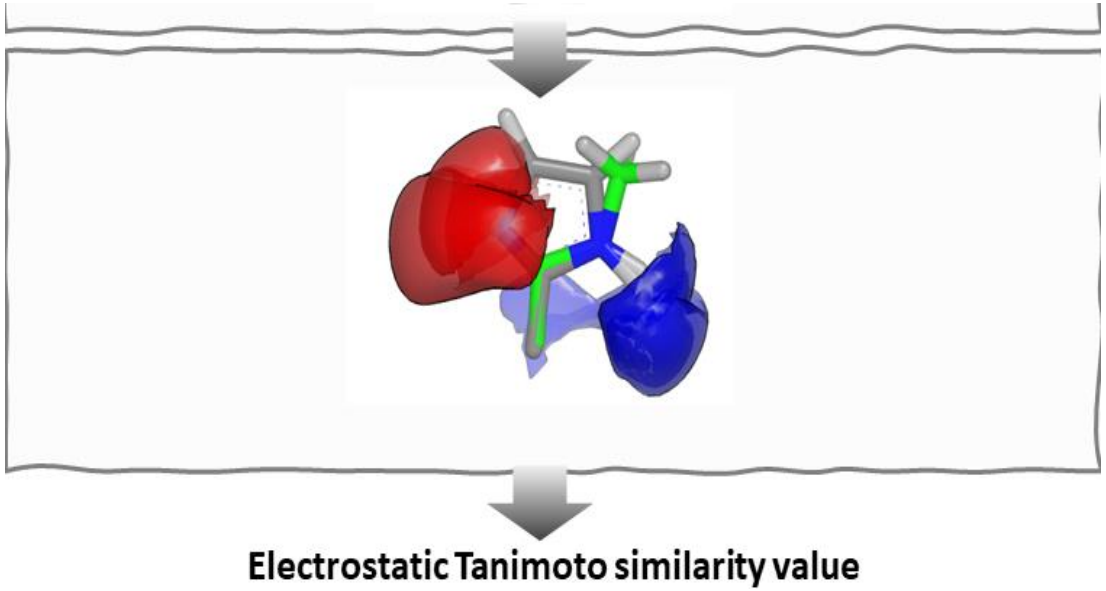
But now replacing parts of these molecules with fragments of similar 3D-shape and electrostatics.

It can be described in six steps…



Replace the similarity calculations (2D->3D)





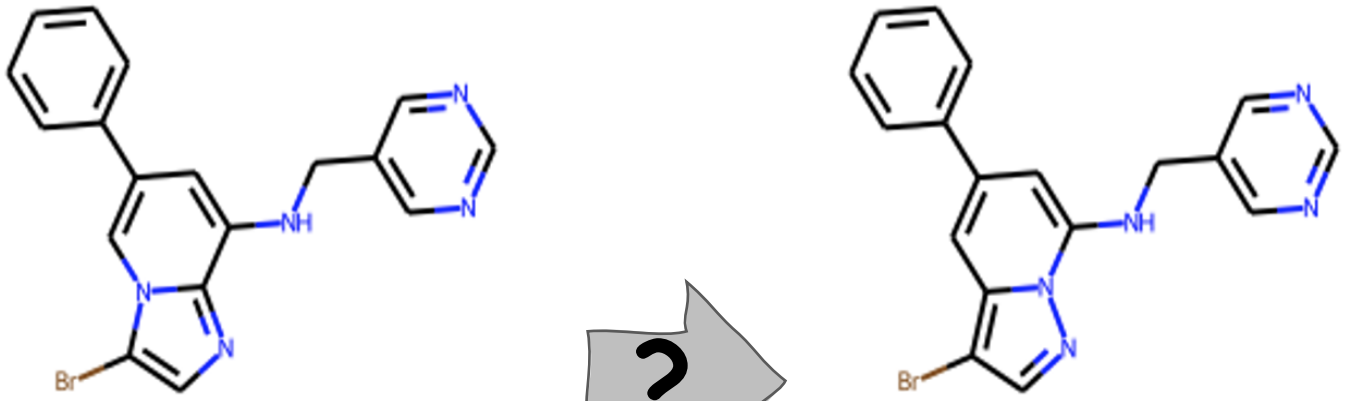
Accordingly, the pairwise similarity between all fragments is calculated, and used to construct the balanced binary tree

In this fashion we can simulate 3D properties while overcoming the notoriously difficult step of accurately describing bioactive conformations.

Scaffold hopping exercise

It is often the case that a drug hunting team has identified a promising lead compound that needs optimization.

Can we use DeepFMPO to come up with ideas of analogs compounds, with optimal properties, with novel central rings to be introduced as scaffold replacements?

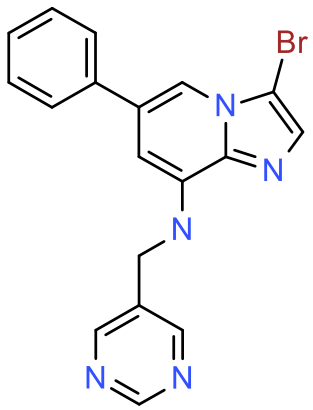


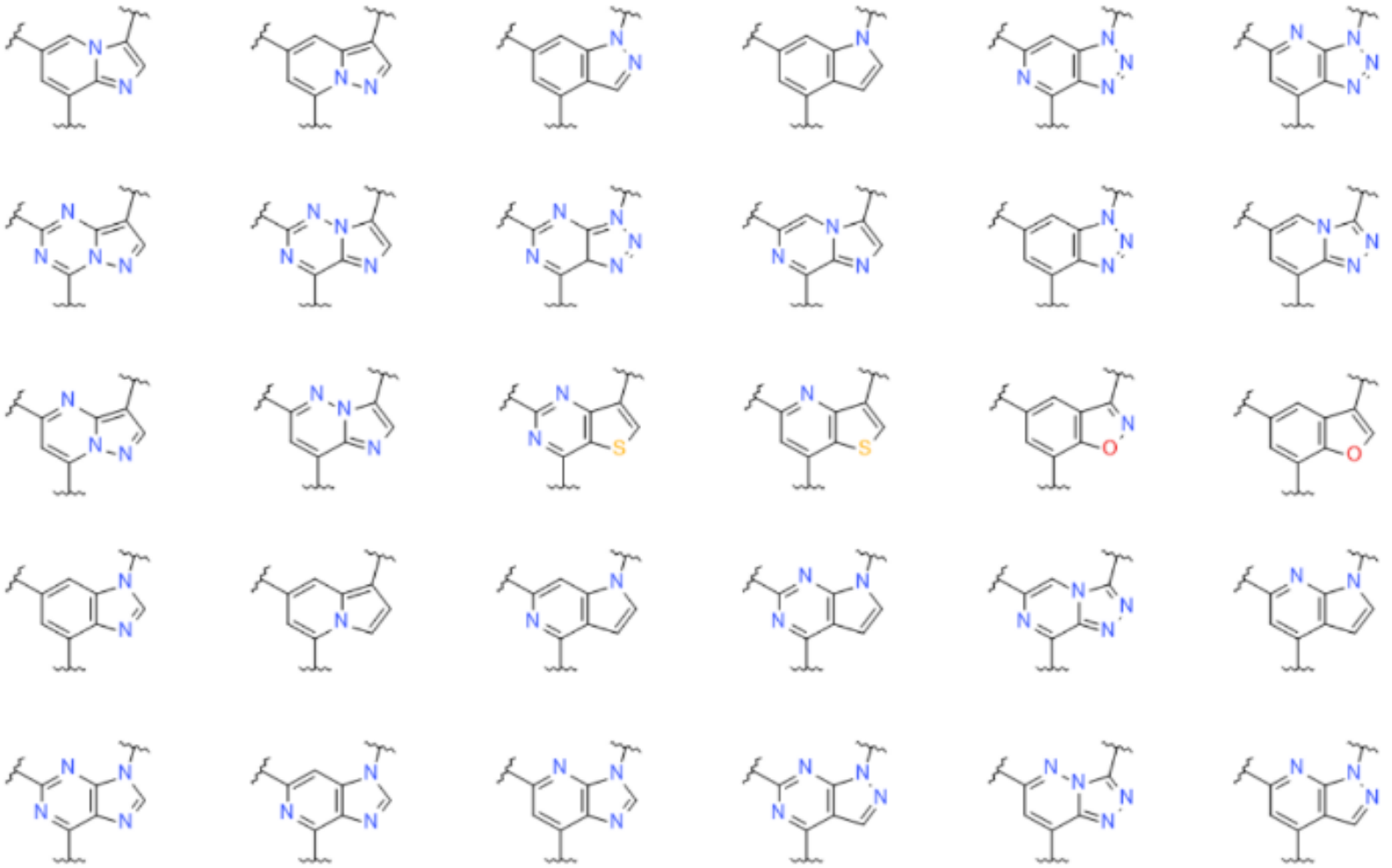
cpd **1**
CDK2: 100nM

cpd **2**
CDK2: 100nM

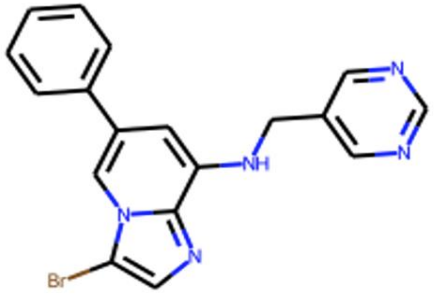
Fischmann, T.O. *et al* Structure-guided discovery of cyclin-dependent kinase inhibitors. *Biopolymers*, 2008, 89, 372.

Retrieve all 9-membered heterocycles in ChEMBL v28

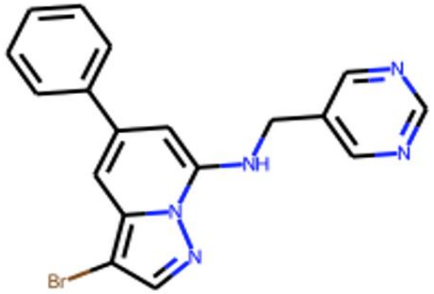




Use ESP-Sim to rank



cpd **1**
CDK2: 100nM



cpd **2**
CDK2: 100nM

Table 1. The rankings for the 1 vs 2 fragment pair, among pair-wise comparison of 30 different heterocyclic rings. The rankings, and Tanimoto value, using a range of different 2D similarity methods available through RDKit [19] and the new ESP-Sim measure are reported.

Method	Rank (max = 30)	Tanimoto
ESP-Sim (B3LYP/6-31G**)	1	0.88
Morgan fingerprint (radius 2)	5	0.44
Morgan fingerprint (radius 3)	5	0.31
MACCS keys fingerprints	17	0.72
MCS Tanimoto	21	0.50
Topological fingerprints	22	0.23

Also….
MACCSkeys FPs gave identical values for many scaffolds suggesting that it is not sufficient for capturing subtle differences.

A couple of clearly structurally dissimilar fragments in (e.g. 1,4,6-trimethylpyrazolo[5,4-b]pyridine vs 2,4,7-trimethylimidazo[2,1-f][1,2,4]triazine) that are ranked low when using ESP-Sim (as they should), but top-ranked when using Morgan 2D-fingerprints.

Implement in DeepFMPO

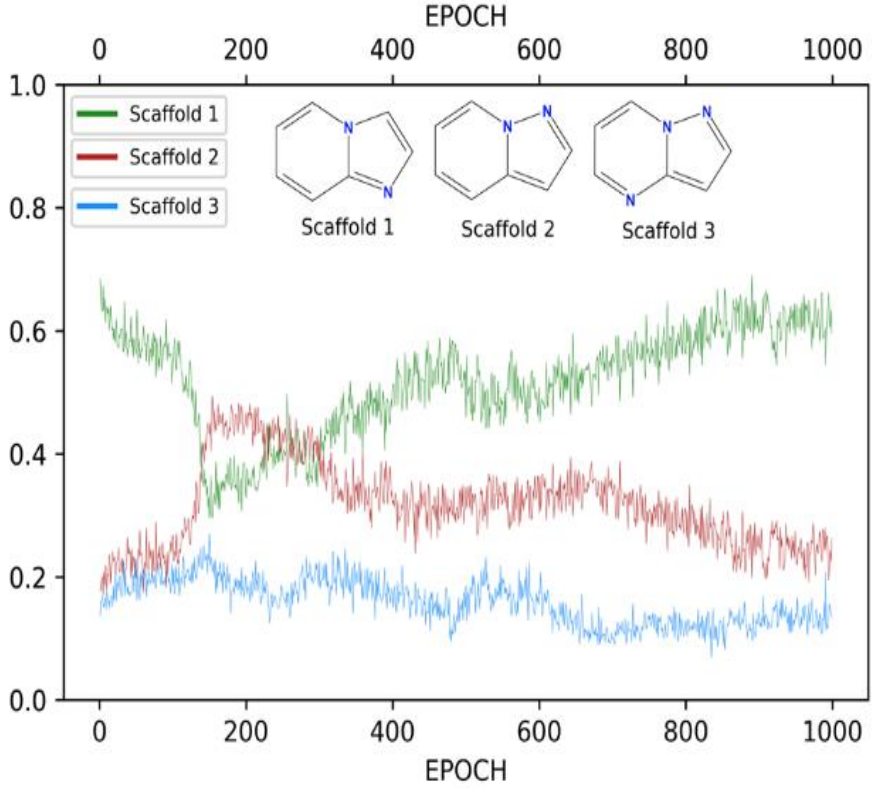
The lead series compounds were obtained by a substructure search using the (imidazo(1,2-a)pyridine) central scaffold of compound **1** on the surechembl.org website and yielded 138 close analogs.

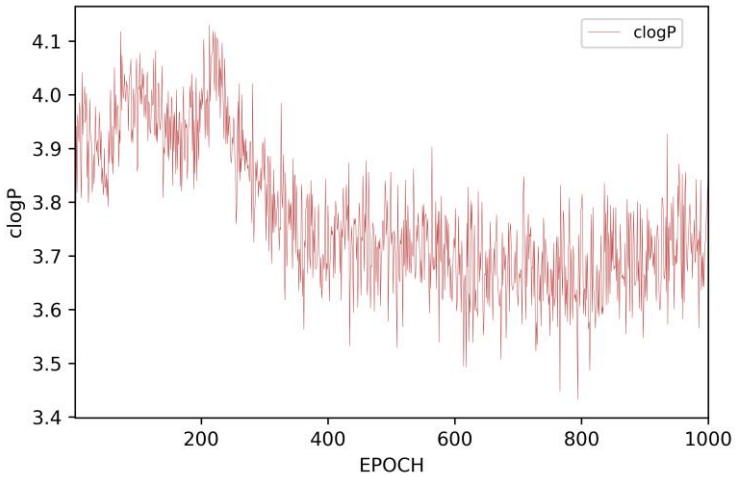
Sweetspot: 320<MW<420, 2.3<clogP<4.3 and 45<PSA<75

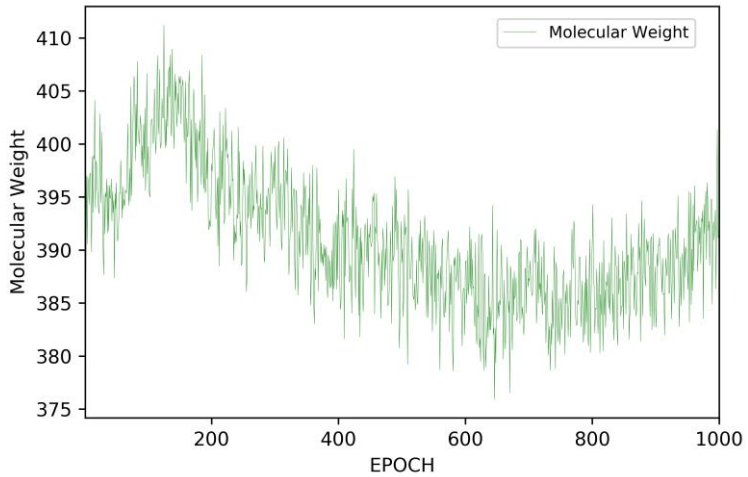
In total 6359 new unique molecules are generated.

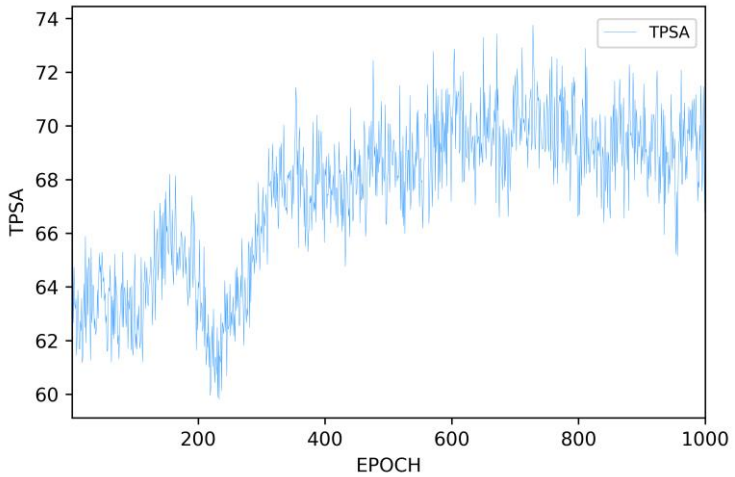
Two-thirds are sweet-spot compounds.

Scaffolds of compound 1 and 2 are the most represented.









Some details

- The code is based on Python 3, NumPy, SciPy, Sklearn, Keras, Theano and **RDKit.**
- The comparison of electrostatic potentials and molecular shape is performed using the ESP-Sim python package (https://github.com/hesther/espsim). Rank orders similar to EON
- Partial charges are calculated using the open-source quantum chemistry program Psi4 (https://psicode.org), allowing the calculation of state-of-the-art partial charges. For example, RESP with B3LYP/6-31G**.
- Dask (https://dask.org), a library for parallel computing, is used to speed up the process.

Code and data sets are here: https://github.com/giovanni-bolcato/deepFMPOv3D
(README to come)

Try it out?

The value of shape and electrostatic similarities in deep generative methods

DeepFMPO v3D

What? a fragment based generative method using shape and electrostatic similarities

Why? an attempt to address current issues with generative methods

Preprint available here:
http://doi.org/10.33774/chemrxiv-2021-sqvv9

Feedback appreciated!

Acknowledgements

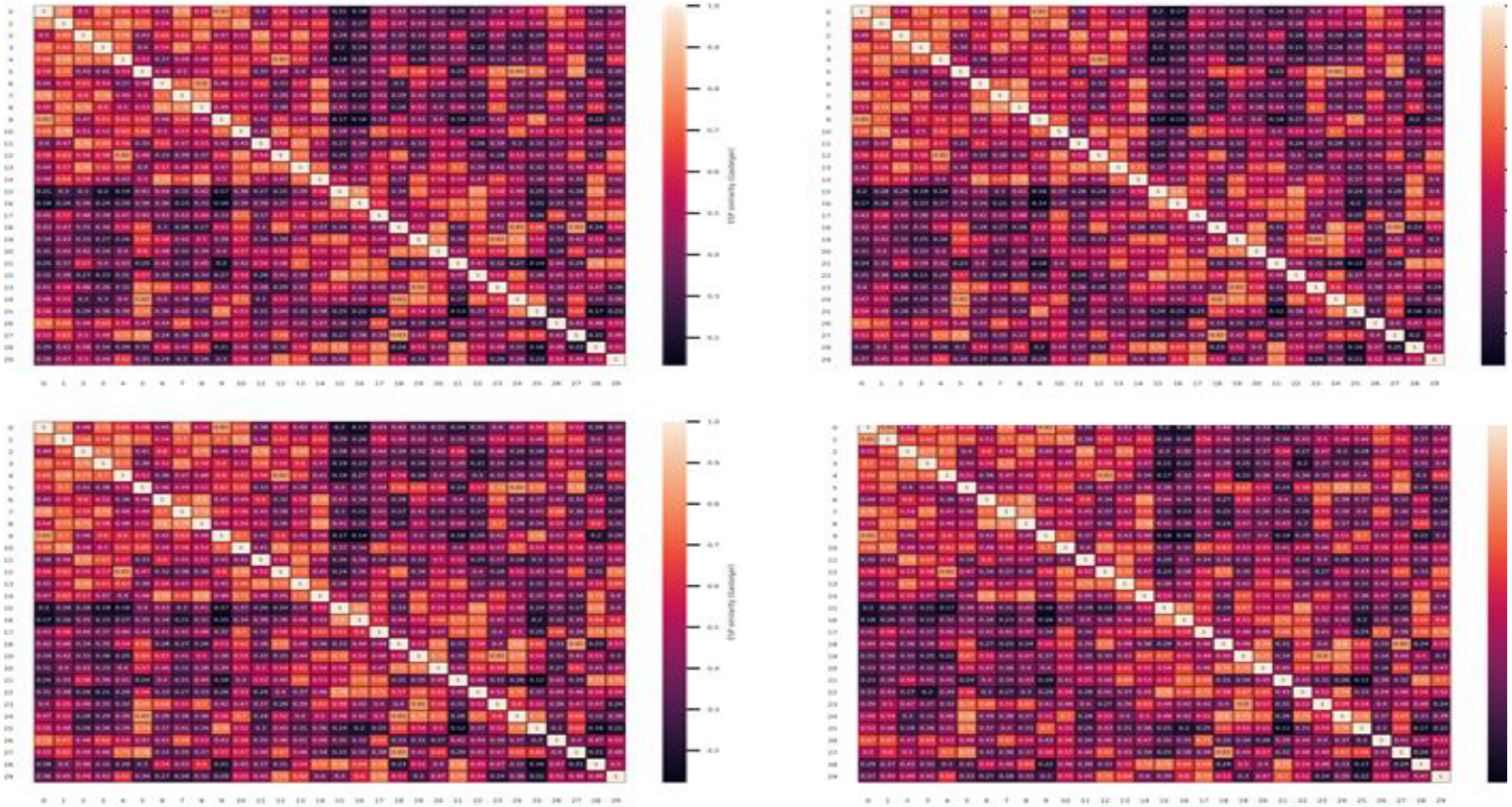
Giovanni Bolcato, University of Padova, Italy

Esther Heid, Massachusetts Institute of Technology (MIT), US

Niclas Ståhl, Jönköping University, Sweden

Greg and RDKit community!

Different anchors



All-against-all comparison experiments were conducted with four structurally different anchor fragments (top-left: hexazine, top-right: carboxylic acid, bottom-left: piperidine, bottom-right: iodine). The different anchors give essentially the same results.