

# 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*

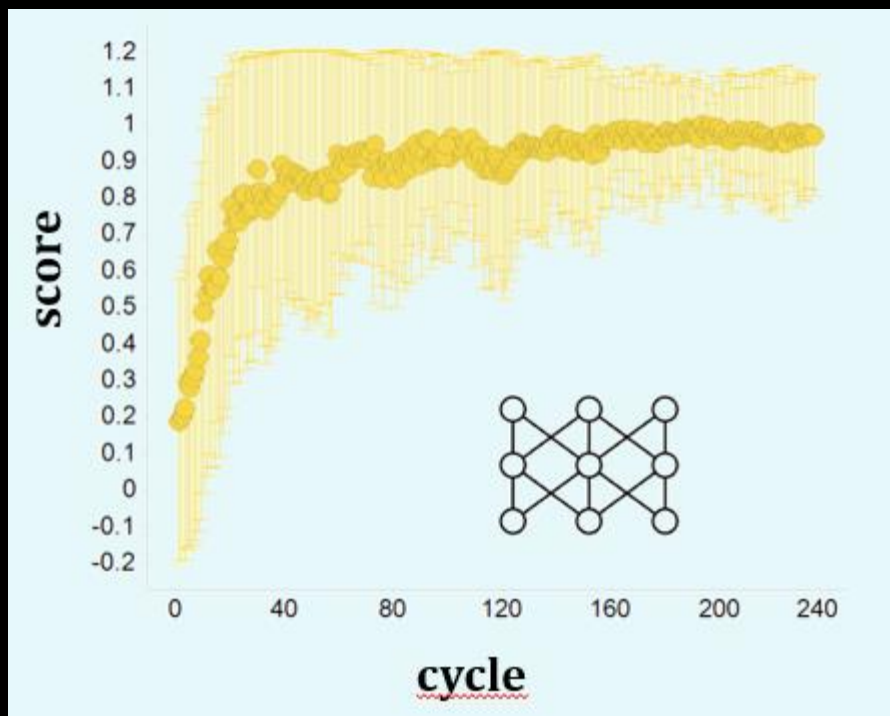
*[email: jonas.bostrom@astrazeneca.com](mailto:jonas.bostrom@astrazeneca.com)*

*Twitter: @DrBostrom*

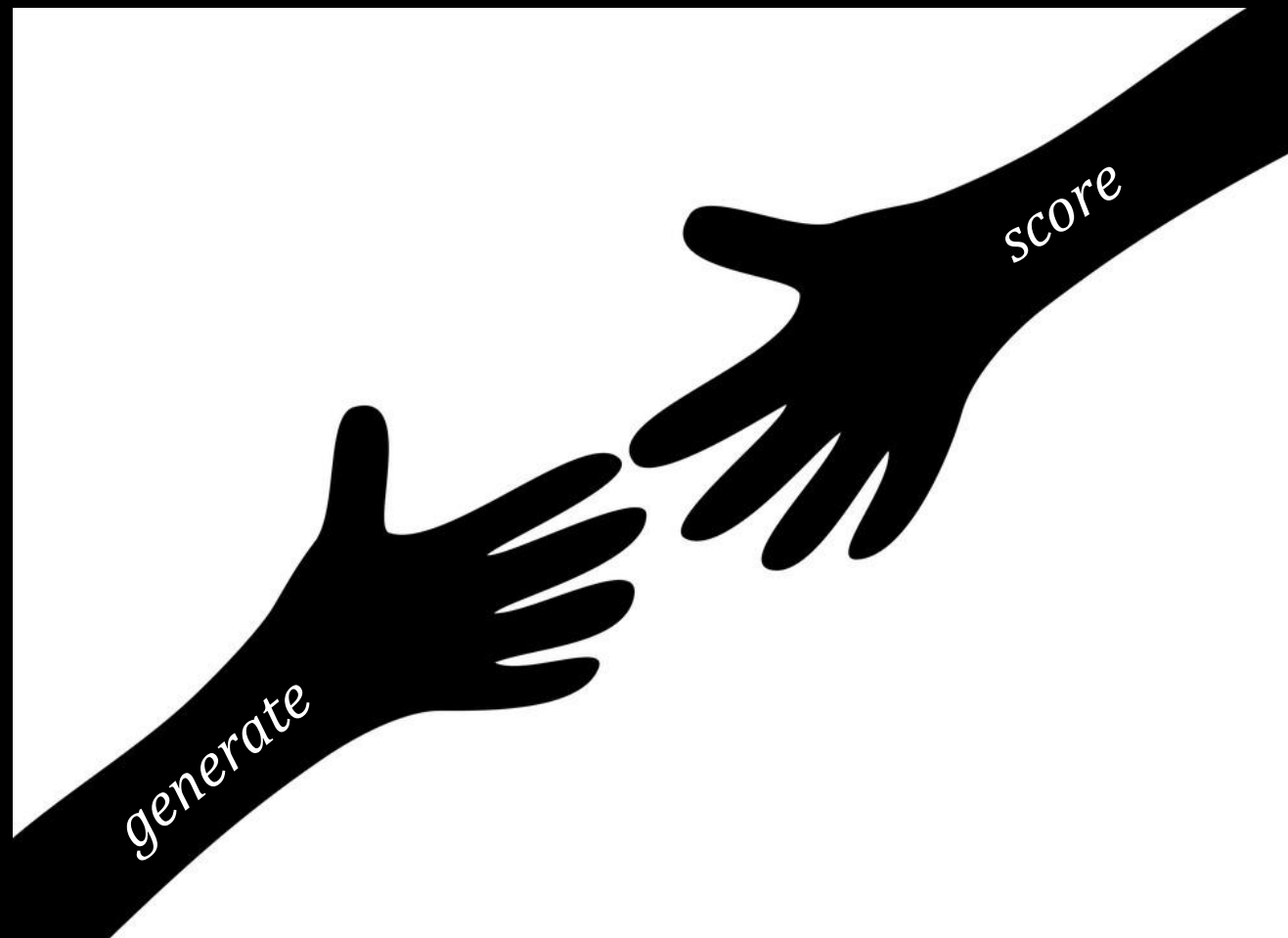
*RDK UGM 2021, October 14<sup>th</sup>*

# 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.



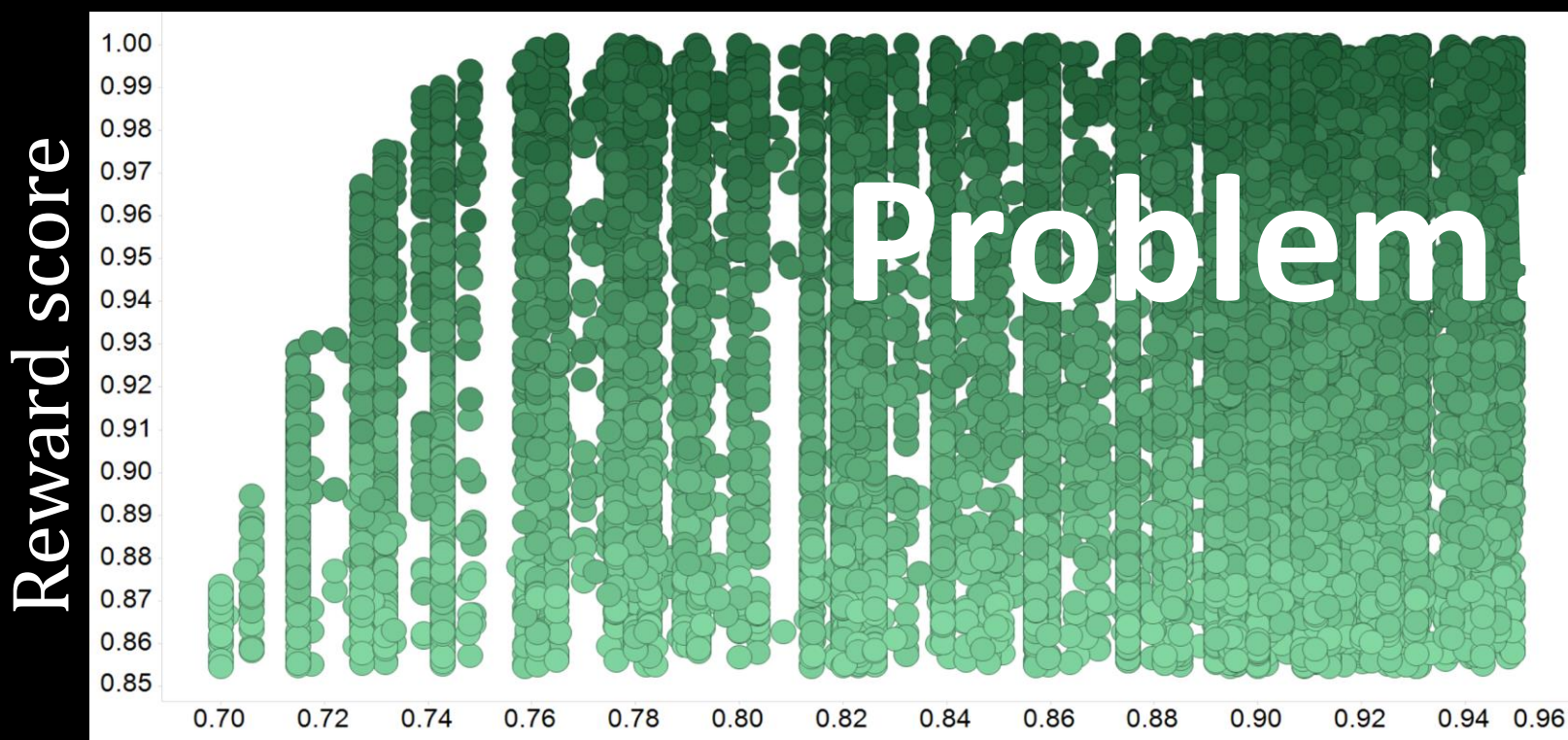
reinforcement learning



# 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?*

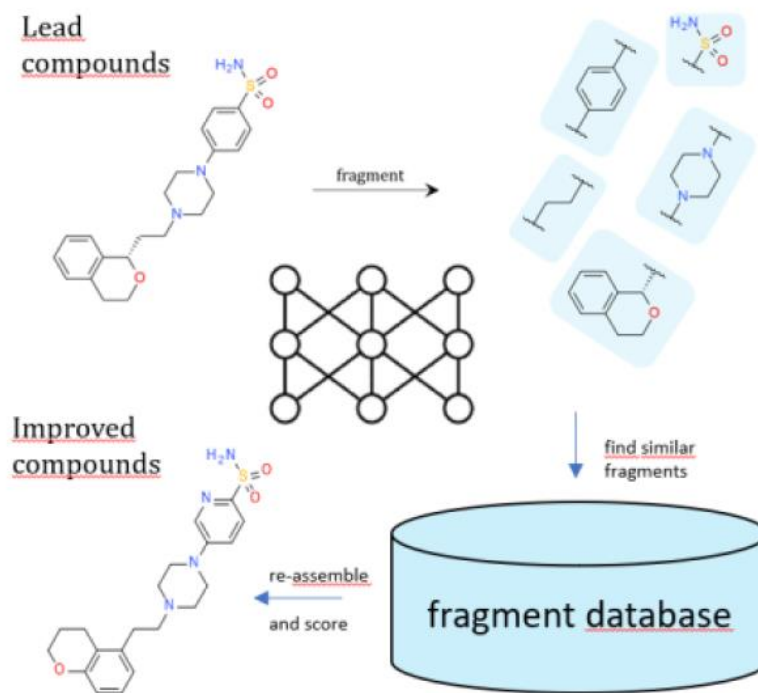


We are (generally) *not* at the stage where we can single out a few compounds with precision from the many.

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

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

Generative approach using similar fragments, instead of sequences of letters (SMILES).



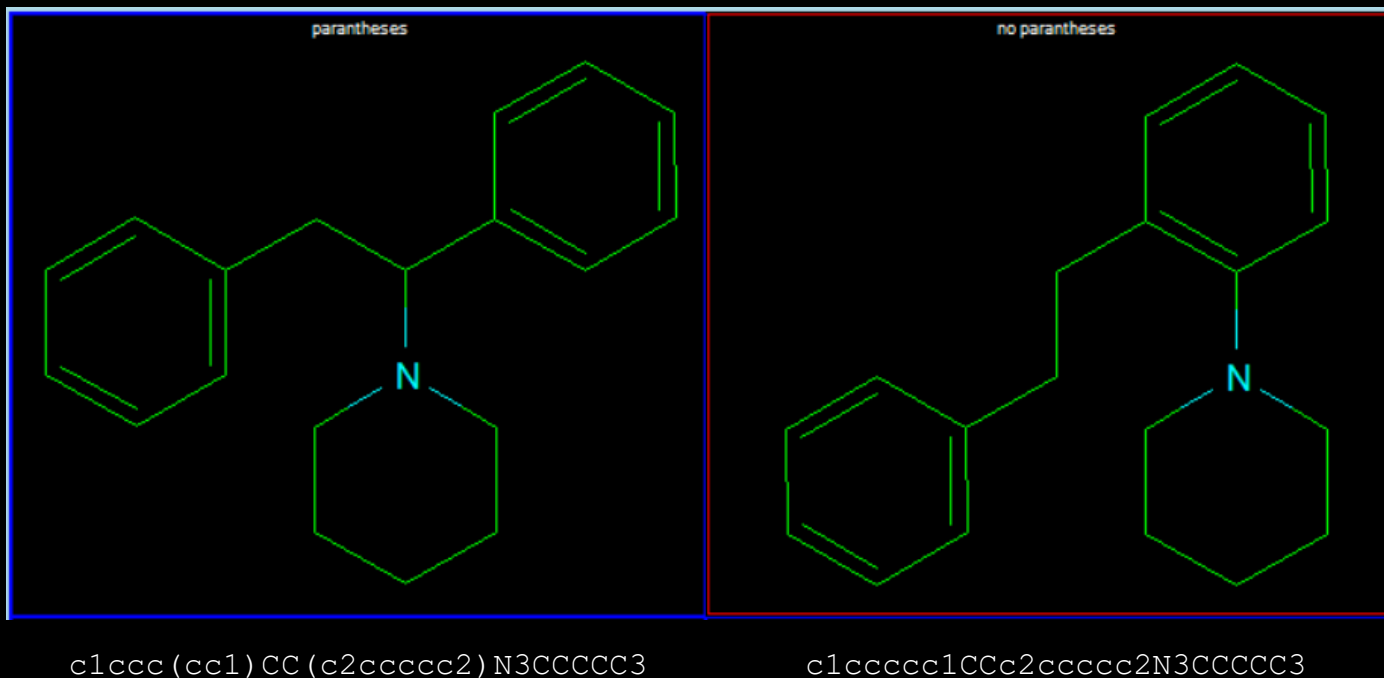
Medicinal chemists tend to retain certain functional groups in designs for many good reasons

# Problem 2. Molecular Representations

**SMILES strings** are often used

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

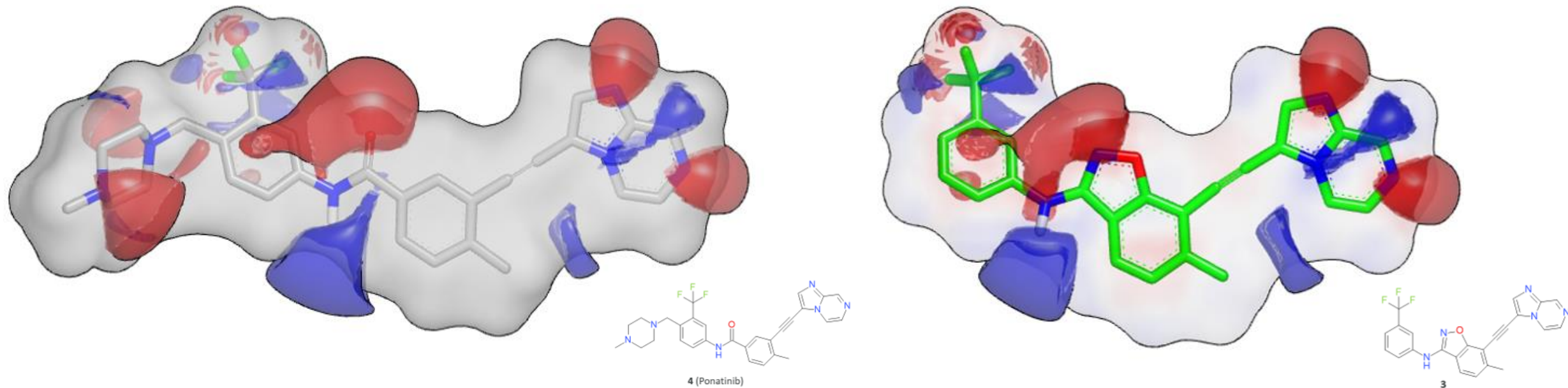
However, a conservative modification to a SMILES string may cause a large effect in their 3D structure.  
(e.g. removal of brackets denoting substitution, and a Y-shaped compound becomes linear)



I know, I know....there are graph-based and other useful methods like DeepSMILES and SELFIES....but

# Molecules are three-dimensional

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



...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.



# We developed a fragment-based method

The use of “DeepFMPO” and similarity has been successful in AZ drug hunting projects.



JOURNAL OF  
CHEMICAL INFORMATION  
AND MODELING

Cite This: J. Chem. Inf. Model. 2019, 59, 3166–3176

Article

pubs.acs.org/jcim

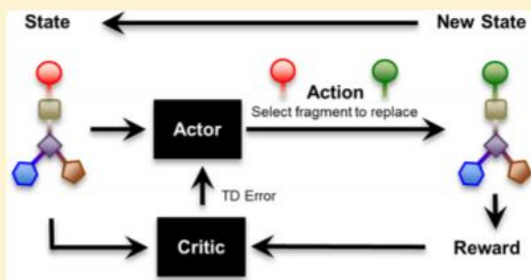
## Deep Reinforcement Learning for Multiparameter Optimization in *de novo* Drug Design

Niclas Ståhl,<sup>\*,†</sup> Göran Falkman,<sup>†</sup> Alexander Karlsson,<sup>†</sup> Gunnar Mathiason,<sup>†</sup> and Jonas Boström<sup>‡</sup>

<sup>†</sup>School of Informatics, University of Skövde, 541 28 Skövde, Sweden

<sup>‡</sup>Medicinal Chemistry, Early Cardiovascular, Renal and Metabolism, R&D BioPharmaceuticals, AstraZeneca, 431 83 Mölndal, Sweden

**ABSTRACT:** In medicinal chemistry programs it is key to design and make compounds that are efficacious and safe. This is a long, complex, and difficult multiparameter optimization process, often including several properties with orthogonal trends. New methods for the automated design of compounds against profiles of multiple properties are thus of great value. Here we present a fragment-based reinforcement learning approach based on an actor–critic model, for the generation of novel molecules with optimal properties. The actor and the critic are both modeled with bidirectional long short-term memory (LSTM) networks. The AI method learns how to generate new compounds with desired properties by starting from an initial set of lead molecules and then improving these by replacing some of their fragments. A balanced binary tree based on the similarity of fragments is used in the generative process to bias the output toward structurally similar molecules. The method is demonstrated by a case study showing that 93% of the generated molecules are chemically valid and more than a third satisfy the targeted objectives, while there were none in the initial set.



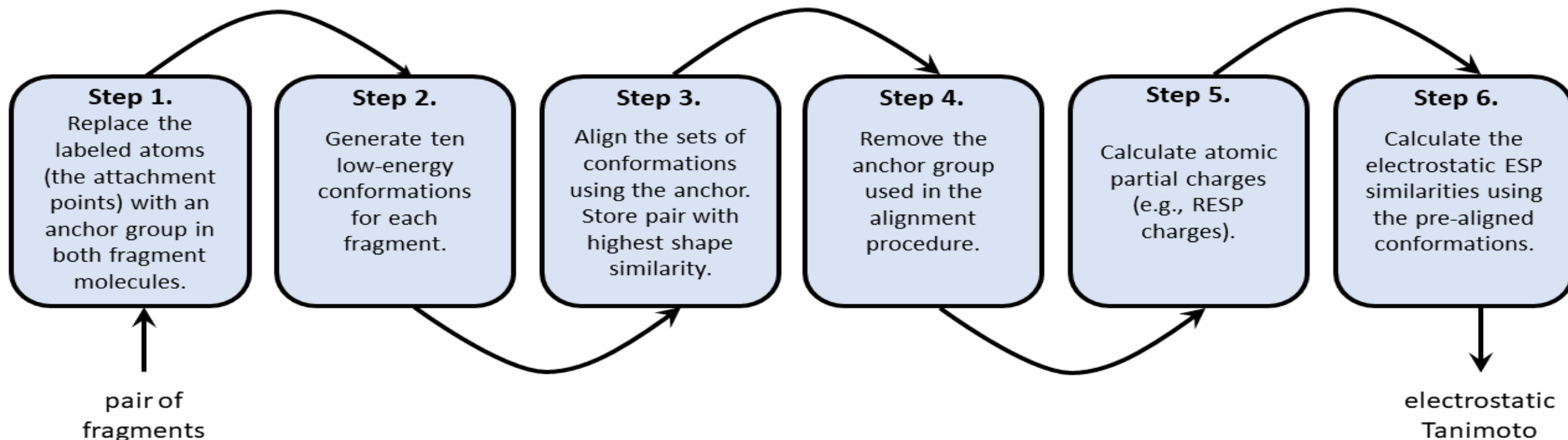
# But it's not 3D

# 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.

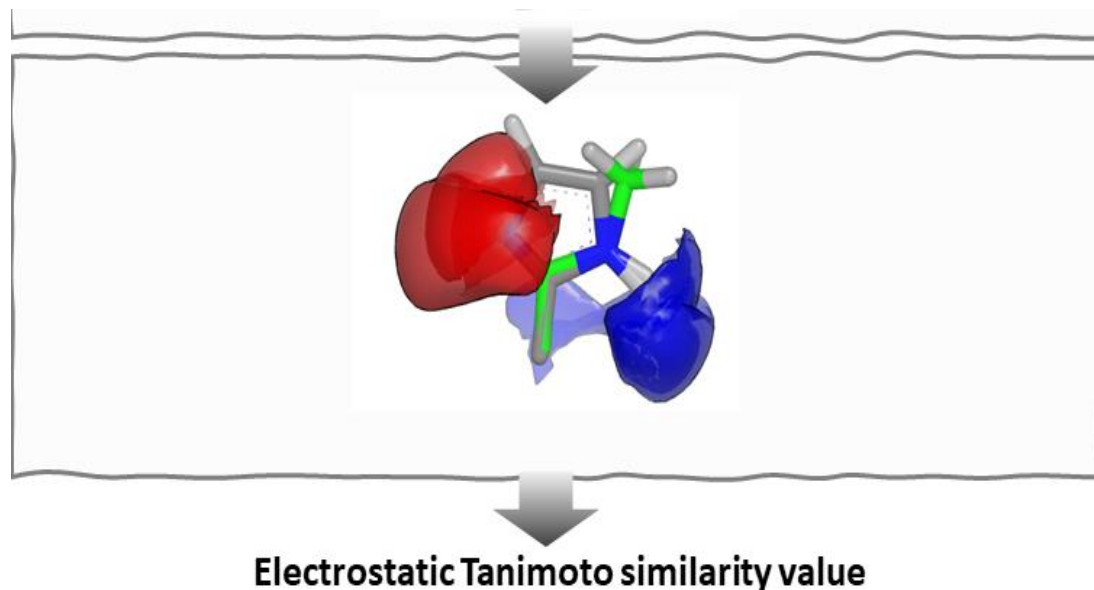
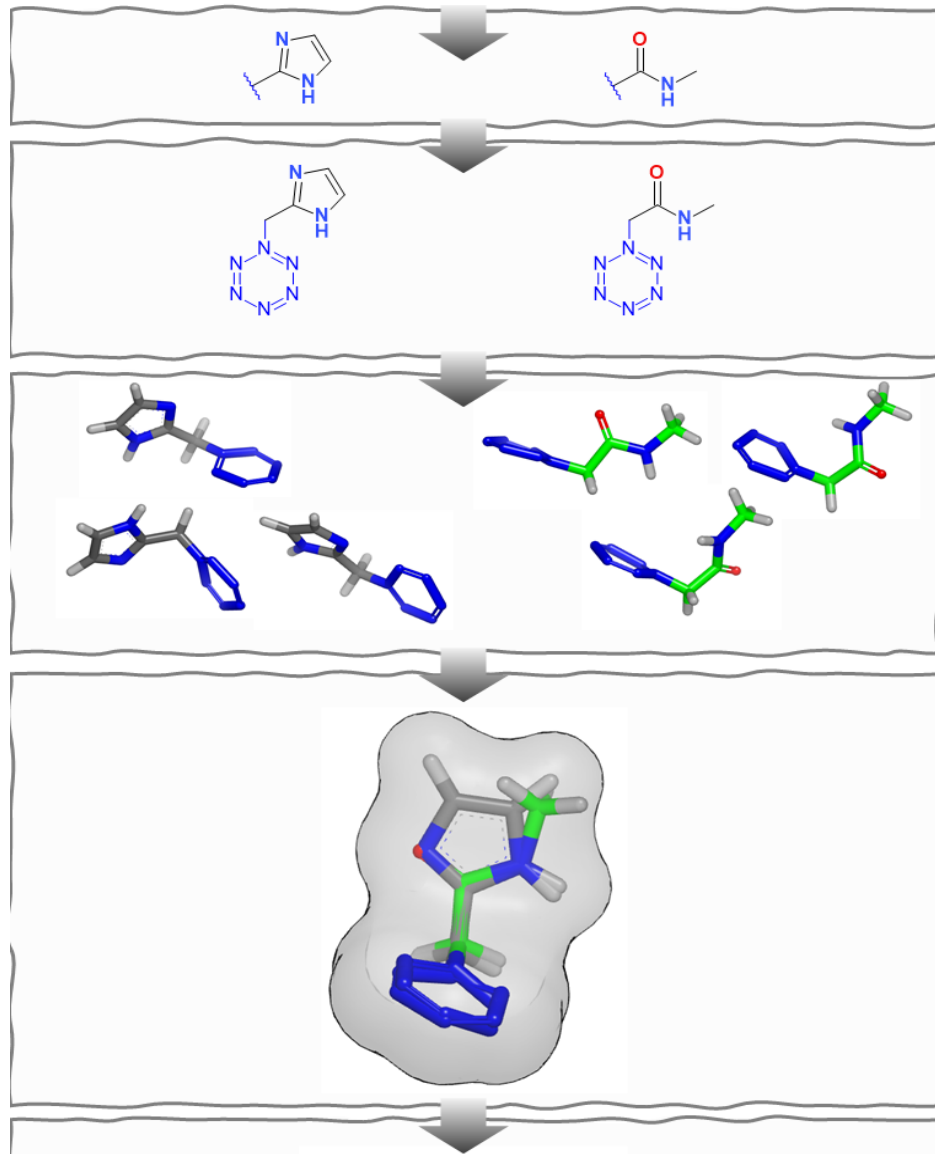
It can be described in six steps...





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

Fragment pair: \*c1[nH]ccn1 vs \*C(=O)NC



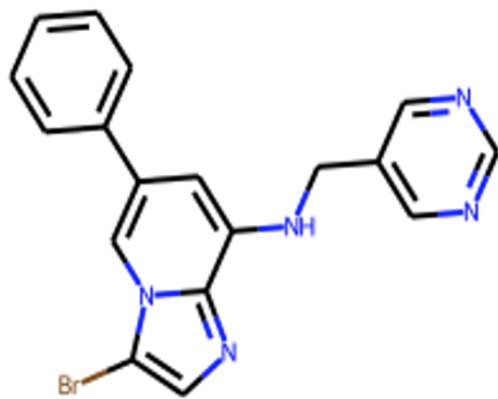
Accordingly, the pairwise similarity between all fragments is calculated, and used to construct the balanced binary tree

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

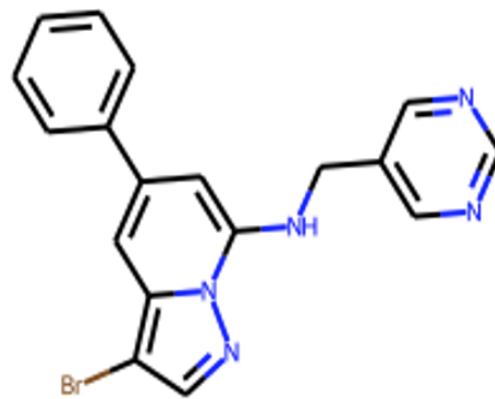
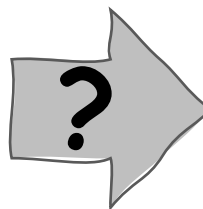
# Scaffold hopping exercise

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?

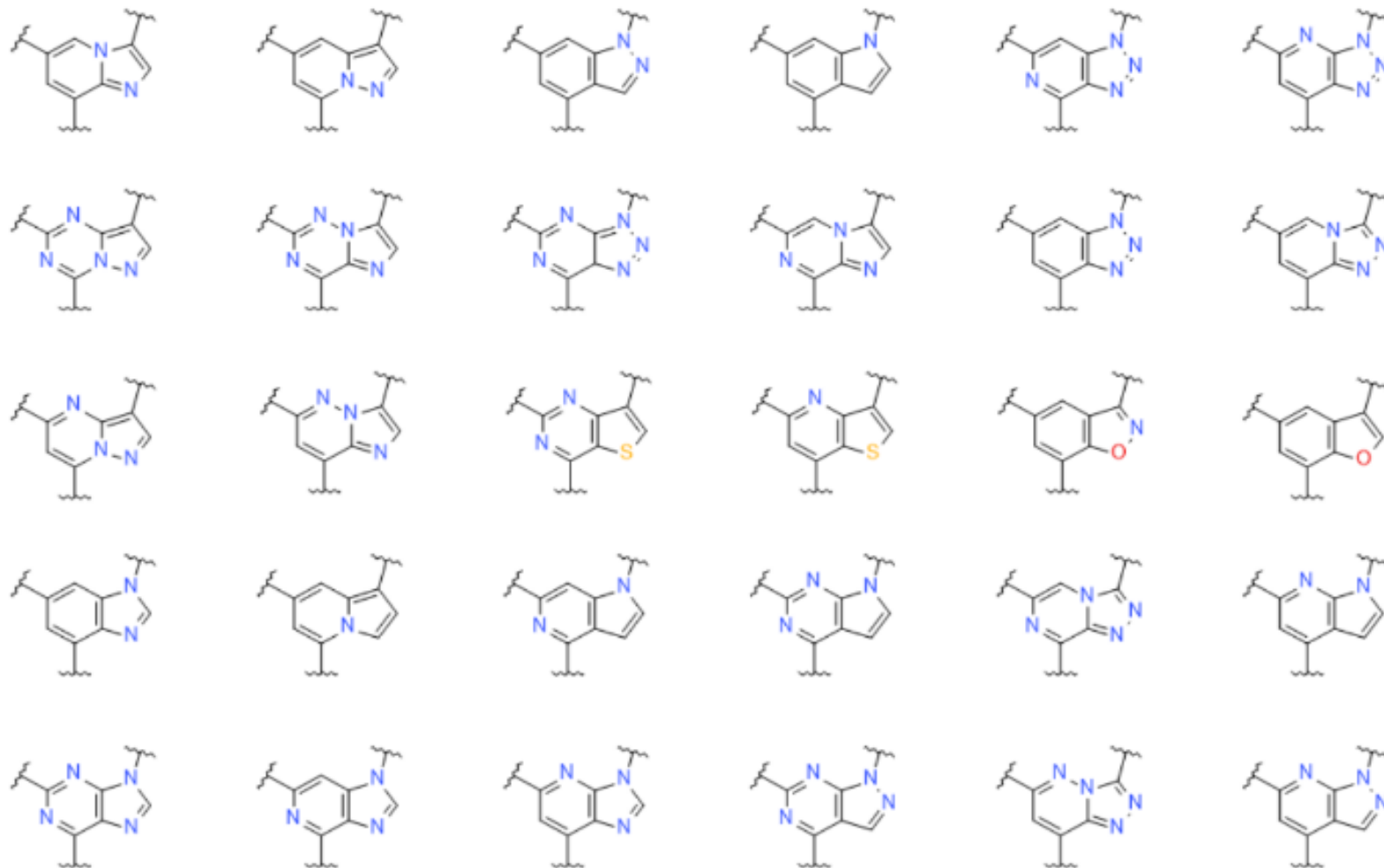
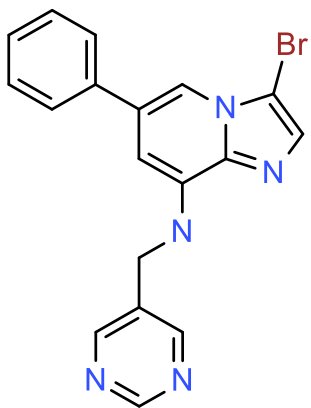


cpd 1  
CDK2: 100nM

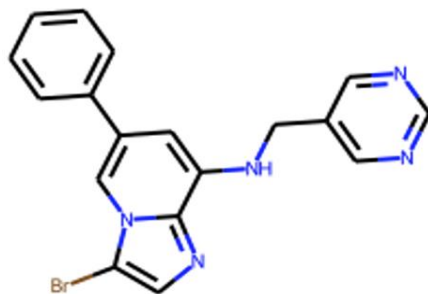


cpd 2  
CDK2: 100nM

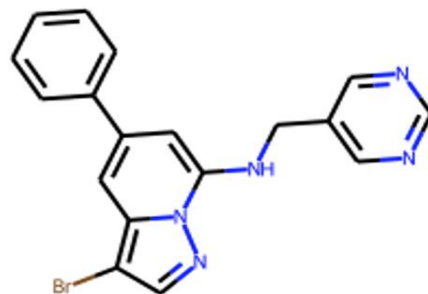
# Retrieve all 9-membered heterocycles in ChEMBL v28



# 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

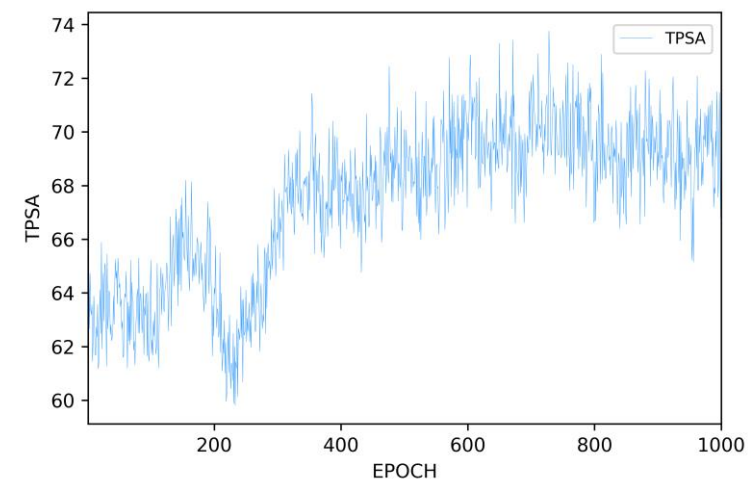
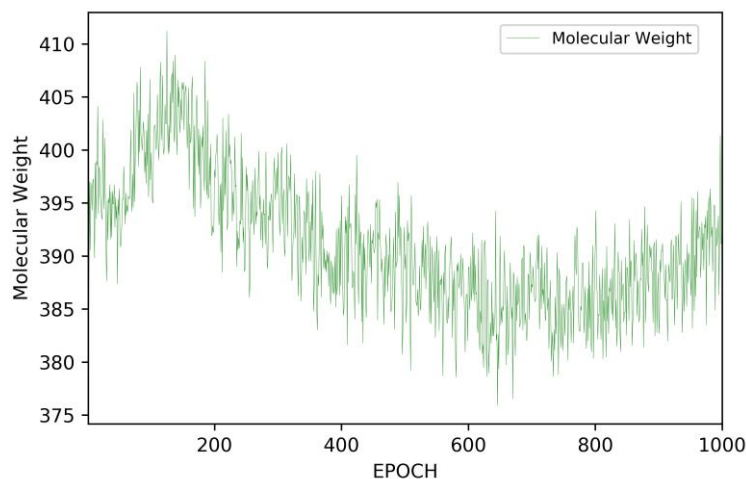
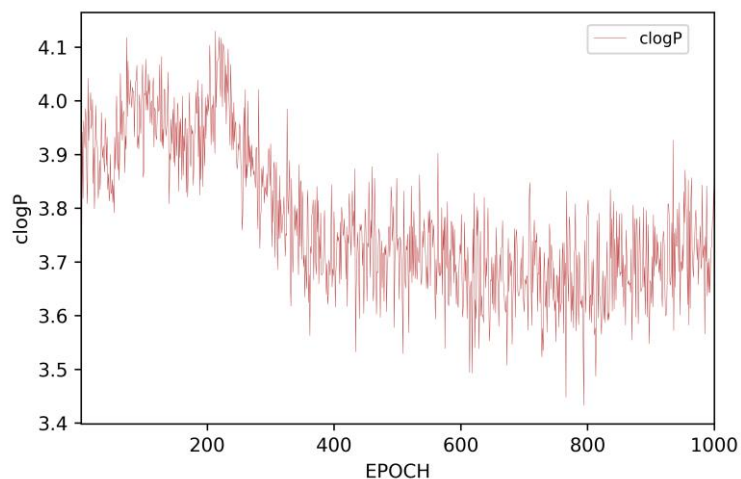
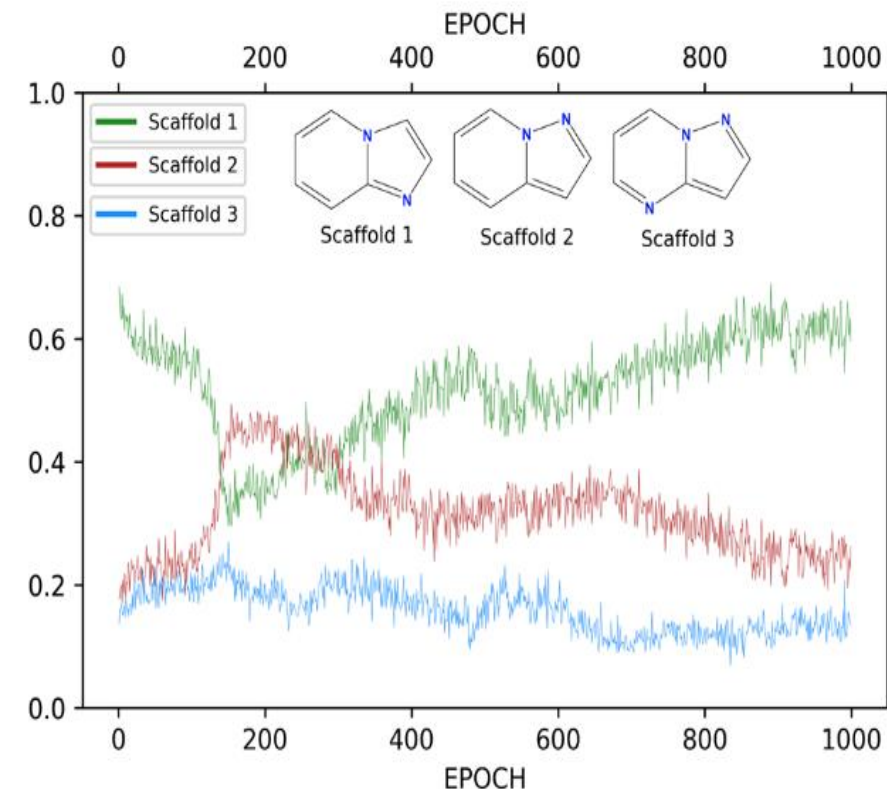
The lead series compounds were obtained by a substructure search using the (imidazo(1,2-a)pyridine) central scaffold of compound **1** on the [surechembl.org](https://www.ebi.ac.uk/chembl/) website and yielded 138 close analogs.

Sweetspot:  $320 < \text{MW} < 420$ ,  $2.3 < \text{clogP} < 4.3$  and  $45 < \text{PSA} < 75$

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/hester/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/deepFMP0v3D>  
(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

Paper soon available on <https://chemrxiv.org>

**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!