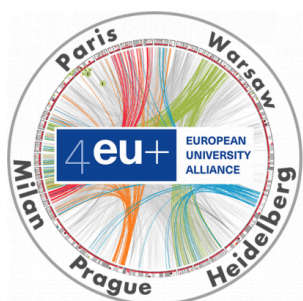


# Generative ML pipeline for screening process

Jakub Adamiak, Maciej Bielecki, Marcin Potkański,  
Paweł Nagórko, Piotr Trzaskowski

Supervisors: Joanna Sułkowska, PhD, DSc, Prof. Tit,  
Wanda Niemyska, PhD



**European Bioinformatics  
Masters Network**

<b>1. Introduction.....</b>	<b>3</b>
1.1. Generative models as alternative to virtual screening.....	3
1.2. Punicalagin as control Nsp13 inhibitor.....	3
1.3. Project workflow.....	4
<b>2. Materials and methods.....</b>	<b>5</b>
2.1. Generation of ligands.....	5
2.2. Filtering of generated ligands.....	5
2.3. Analog lookup.....	5
2.4. Assessment of analogs.....	6
<b>3. Results.....</b>	<b>6</b>
3.1. Generated ligands.....	6
3.2. Database analogs.....	7
<b>4. Discussion.....</b>	<b>9</b>
<b>5. References.....</b>	<b>10</b>

# 1. Introduction

This year's topic was the same as last year, namely, the development of an original computational procedure for the identification of potential inhibitors of Nsp13, the RNA helicase of SARS-Cov-2.

In the project, we were paired with the teams from Sorbonne University (1) and Milano University (1).

## 1.1. Generative models as alternative to virtual screening

Ordinarily, the first step of modern drug discovery is virtual screening (VS) of databases in an attempt to find ones likely to be effective at binding to a given receptor. This is typically accomplished by the molecular docking of each ligand to the structure of the receptor, which, while being an accurate method of filtering ligands based on their affinity to the receptor, has the demerit of being quite slow.

A potential alternative to molecular docking are generative models (GM), which use machine learning algorithms to generate all-new ligands, giving consideration to the receptor's structure. One such GM is Pocket2Mol, an autoregressive generative graph neural network model.

Although GMs are promising in drug discovery, they have limitations. Sometimes they generate unrealistic and unsynthesizable molecules - their output contains molecular structures that are chemically infeasible. This brings challenges to people who want to synthesize and validate potential drugs.

Our idea to address these limitations involves leveraging existing databases to identify and validate generated molecules. Rather than immediately proceeding to the costly and often unsuccessful process of synthesizing molecules generated by a GM, we decided to cross-reference these virtual creations with established databases of confirmed compounds.

The idea is that the heuristic nature of this approach will allow us to search much bigger datasets (even in billions of compounds) with relatively low costs both time and computation.

## 1.2. Punicalagin as control Nsp13 inhibitor

As input, Pocket2Mol requires the coordinates of a pocket within the receptor's structure, to which it will attempt to generate fitting ligands. We determined the coordinates for such a pocket within Nsp13's structure based on a publication by Lu et al. (2022), in which punicalagin (PUG) is found to be an inhibitor of Nsp13, and an effective *in vitro* antiviral against SARS-CoV-2.

In addition, we've also chosen to use PUG itself as a control inhibitor of Nsp13, to serve as a reference to which generated ligands would be compared.

### 1.3. Project workflow

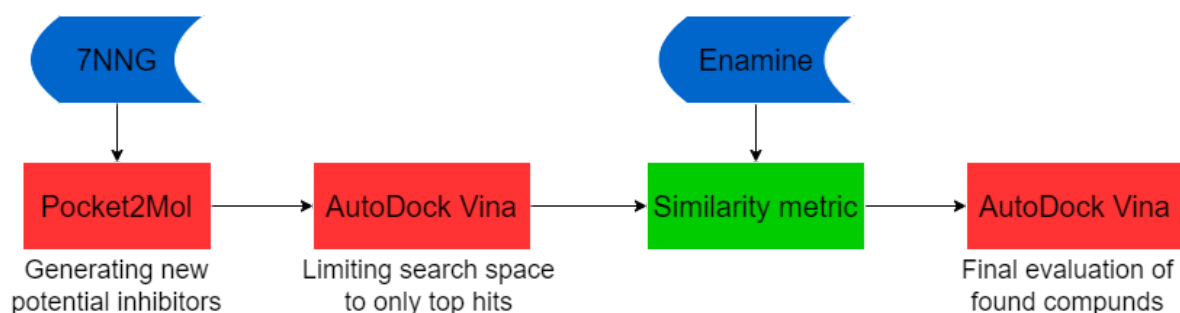


Figure 1. Workflow schematic.

The project workflow begins by employing Pocket2Mol to generate molecules specifically targeting a pocket in Nsp13's structure (see §1.2).

Subsequently, the generated molecules undergo assessment using AutoDock Vina (Eberhardt et al., 2021; Trott and Olson, 2010) to evaluate their binding affinities to Nsp13. This reduces the number of molecules to process in the next step.

To identify analogs and assess their similarity, the Tanimoto coefficient is applied to Morgan fingerprints, using Enamine's *REAL* Diversity Set (50M compounds). This step aims to discover existing compounds that share structural similarities with the generated molecules.

Finally, the identified analogs are likewise subjected to a second round of assessment using AutoDock Vina, ensuring a comprehensive evaluation of their binding properties and potential as promising candidates for further drug development.

This workflow integrates computational methods, molecular generation, and database analysis to streamline the identification of potential candidates in big data sets.

## 2. Materials and methods

### 2.1. Generation of ligands

To generate potential ligands, we used pretrained Pocket2Mol model (Peng et al. 2022). We targeted Nsp13's binding site of Punicalagin, with the size of the generation space set to 25 Å.

### 2.2. Filtering of generated ligands

Given the number of ligands generated by Pocket2Mol (see §3.1.), we found it necessary to limit the search space. To that end, each ligand generated by Pocket2Mol was docked to Nsp13. The structure of Nsp13 we used was PDB ID 7NNG (Newman et al., 2021), as the one that Lu et al. used in their publication. As mentioned in §1.3, docking was performed using AutoDock Vina.

Aside from Pocket2Mol's generated ligands, we also docked PUG to obtain a reference score.

Given a ligand-receptor pair, AutoDock Vina generates a number of binding poses which the ligand may take, along with a table of binding energy of each pose, and the RMSD (in Å) between each pose and the lowest-energy pose. Using these values, the average binding energy and average RMSD was calculated for each docked ligand (including PUG).

While these values can be used as a docking score as-is, we found doing so to be unintuitive, as the desired ligands would minimize both values. To remedy this, two more sets of values were calculated, those being each ligand's protein-ligand affinity (PLA) and locality of binding (LoB), respectively as the opposite of average binding energy and the reciprocal of average RMSD.

Finally, filtering of ligands proper was performed by comparing individual ligands' PLA and LoB to the PLA and LoB of PUG, and if both of these values for a given ligand exceeded those of PUG's, the ligand would be selected for further processing, i.e. lookup of analogs.

### 2.3. Analog lookup

For assessing similarity between two molecules we used the Tanimoto coefficient calculated between Morgan fingerprints of two molecules. The challenge we faced was the size of the database which was 50 million compounds. So our algorithm had to be optimized.

For the Tanimoto coefficient it is considered that two molecules are similar with a score of at least 0.6.

We used a multithreading approach to speed up the calculation and to work on small portions of the database at once. Another advantage of this approach is that our algorithm is easily scalable for more complicated architectures.

### 2.4. Assessment of analogs

For each molecule we found that had high enough Tanimoto coefficient we used Vina to assess its binding affinity. This step was similar to one described in §2.2.

## 3. Results

### 3.1. Generated ligands

In total, we generated 1592 potential ligands of varying sizes (Fig. 2). Of these, 33 had higher PLA and LoB than PUG (Table 1). The average PLA for all generated ligands was  $6.73 \pm 0.91$  kcal/mol, and the average LoB was  $0.19 \pm 0.05$  Å<sup>-1</sup>. The PLA and LoB of PUG was 7.96 kcal/mol and 0.142 Å<sup>-1</sup>, respectively.

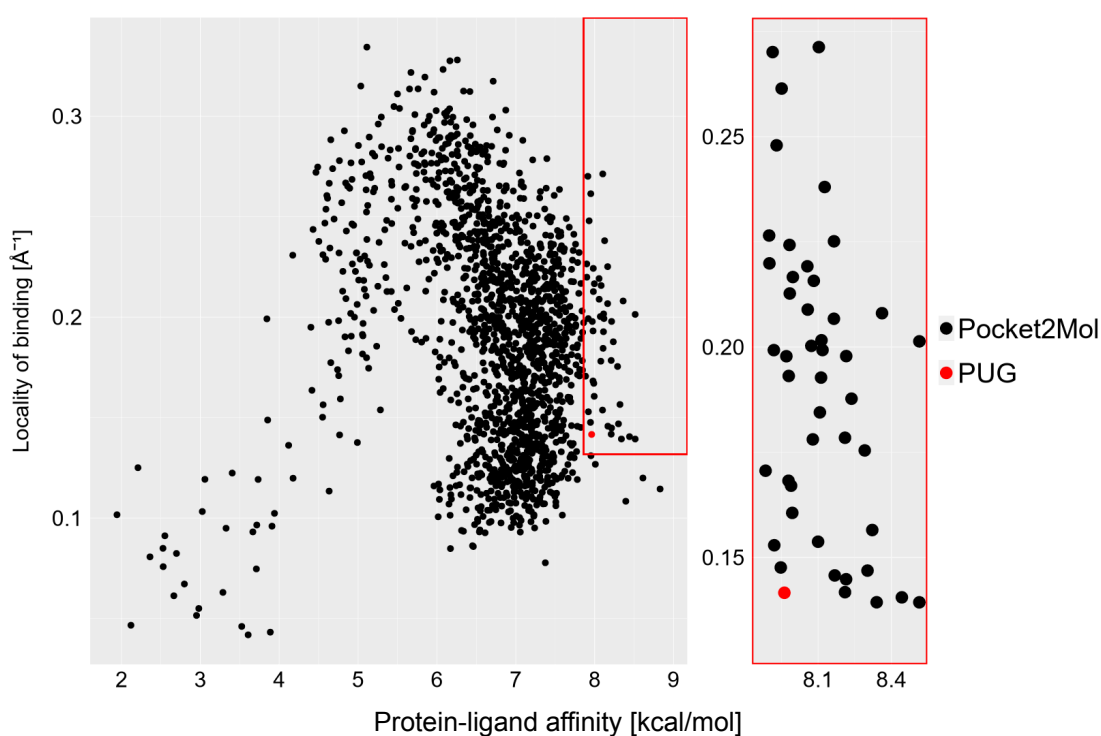


Figure 2. The PLA and LoB of PUG and all generated ligands (left), or of PUG and ligands with higher scores than PUG (right).

Molecule	PLA [-kcal/mol]	LoB [Å-1]
<chem>Cc1cc(CC2=CC=NC2)cc(O)c1CC1=CC(=CCC1)C1=C(C(=O)NC(=O)[C@H](O)[C@H](CNC(=O)O)COP(=O)(O)O)CCC1</chem>	8.514	0.201
<chem>Cc1cc(CC2=CCNC2=O)cc(O)c1CC1=CC(=CCC1)C1=C(C(=O)NC(=O)[C@@H](O)[C@H](CNC(=O)O)COP(=O)(O)O)CCC1</chem>	8.362	0.208
<chem>Cc1cc(CC2=CCCC2)cc(O)c1CC1=CC(=CCC1)C1=C(C(=O)NC(=O)[C@H](O)[C@@H](CNC(=O)O)COP(=O)(O)O)CCC1</chem>	8.322	0.157
<chem>CC1=C(Cc2cccc(C3=C(C(=O)NC(=O)[C@H](O)[C@@H](CNC(=O)O)COP(=O)(O)O)CCC3)c2)CCCC1</chem>	8.302	0.147
<chem>O=C(O)Nc1cc(CC2=CCN=C2O)c(F)c(O[C@@H]2C[C@H](O)[C@H](O)[C@@H](O[P@](=O)(O)OP(=O)(O)O)[C@@H]2O)c1</chem>	8.291	0.175
<chem>O=C(O)NC[C@@H](COP(=O)(O)O)[C@H](O)C(=O)NC(=O)C1=C(C2=CCCC(=C2)Cc2ccc(CC3=CCCC3)cc2O)CCC1</chem>	8.237	0.188
<chem>O=C(O)NC[C@@H](COP(=O)(O)O)[C@H](O)C(=O)NC(=O)C1=C(C2=CCCC(=C2)Cc2ccc(CC3=CCCC3)cc2O)CCC1</chem>	8.215	0.198
<chem>CC(C)Cc1ccc(CC2=CC(=CCC2)C2=C(C(=O)NC(=O)[C@H](O)[C@H](CNC(=O)O)COP(=O)(O)O)CCC2)c(O)c1</chem>	8.215	0.145
<chem>O=C(O)NC[C@H](COP(=O)(O)O)[C@H](O)C(=O)NC(=O)C1=C(c2cccc(CC=C3CCCC3)c2)CCC1</chem>	8.210	0.142
<b>Punicalagin</b>	<b>7.962</b>	<b>0.142</b>

Table 1. The 10 best ligands generated by Pocket2Mol, ranked by PLA.

## 3.2. Database analogs

With the script we created we searched for analogs of generated molecules in the database. The results we obtained were not ideal. For the best generated compounds most similar molecules in database had the Tanimoto coefficient of at most 0.34 which means they are very dissimilar. Tanimoto coefficient of above 0.6 was obtained for molecules with docking scores less than Punicalagins.

Avg docking score	best Tanimoto coefficient
8.4	0.3
7.7	0.45
7.2	0.64

Tanimoto coefficient for different docking scores, important to notice is that the only coefficient higher than 0.6 has docking score less than Punicalagin which we wanted to surpass

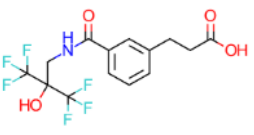
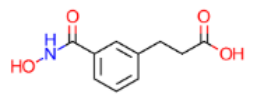
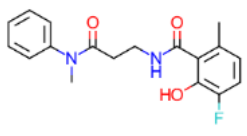
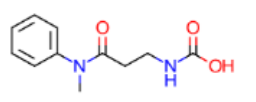
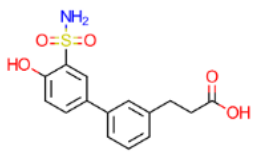
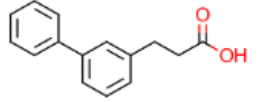
Generated molecule	Docking score	Analog from Enamine	Docking score
	-7.9		-7.2
	-7.9		-6.5
	-7.4		-6.9

Table showing best analogs for each generated molecule



## 4. Discussion

To conclude, we would like to say that as we believe, our approach is viable. It is not computationally demanding. The programs that we use can be run on a home computer and the results can be obtained within no more than a few days. All software is open source. Most programs are multithreaded, some of them as for example Pocket2Mol can be run on GPU. Moreover, we believe that we can expand the database that we use and our approach will remain fast. That is, we think about using the whole database provided by ENAMINE, the database of over 1.2B molecules. Thanks to scalability of multithreaded architectures using this approach will remain fast with not that powerful computer clusters.

Secondly, analogs of molecules generated by Pocket2Mol, can be synthesized by ENAMINE with a great chance of success (80%+ stated by ENAMINE) so this solves biggest problem with generative models which is cost of synthesis (mostly failures during that process) and risk of complete failure.

Apart from that generative ML models are rapidly developing field. Any improvements to the existing models or creation of new ones means that this approach will yield better results.

The problem that we found is that the best generated molecules are the biggest ones, but they have worse analogs in the database. It is caused by the fact that search space of molecules grows exponentially with them increasing in size and the database does not reflect that. In fact there are less big molecules than smaller ones.

## 5. References

- Eberhardt J, Santos-Martins D, Tillack AF, Forli S (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling*, 61(8), 3891-3898.  
DOI: 10.1021/acs.jcim.1c00203
- Lu L, Peng Y, Yao H, Wang Y, Li J, Yang Y, Lin Z (2022). Punicalagin as an allosteric NSP13 helicase inhibitor potently suppresses SARS-CoV-2 replication in vitro. *Antiviral Research*, 206:105389.  
DOI: 10.1016/j.antiviral.2022.105389
- Newman JA, Douangamath A, Yadzani S et al (2021). Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nature Communications*, 12, 4848.  
DOI: 10.1038/s41467-021-25166-6
- Trott O, Olson AJ (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2), 455-461.  
DOI: 10.1002/jcc.21334
- Peng X, Luo S, Guan J et al. (2022) Pocket2mol: efficient molecular sampling based on 3D protein pockets. In: *International Conference on Machine Learning*. PMLR, Baltimore, Maryland, USA. pp. 17644–17655
- Qian H, Zhou J, Tu S, Xu L. DrugGen: a database of de novo-generated molecular binders for specified target proteins. *Database (Oxford)*. 2023 Dec 27;2023:baad090.  
doi: 10.1093/database/baad090. PMID: 38150626; PMCID: PMC10752461.