Virtual Screening of Large Scale Libraries Guided with Docking and Deep Learning

Nicolas Drizard, Loic Chadoutaud, Yasmine Nahal, Stéphanie Labouille, Brice Hoffmann

- ❖ In Virtual Screening campaigns, physics-based simulations - like docking - enable to rank compounds against a target. Unfortunately, they are too expensive to be done on large scale datasets.
- ❖ Using deep learning and active learning, we can infer the most potent compounds in a given library without having to dock it entirely. We applied and evaluated different bayesian optimisation frameworks.

Problem Formulation

The goal is to find the top scoring molecules in docking with the minimum number of simulations:

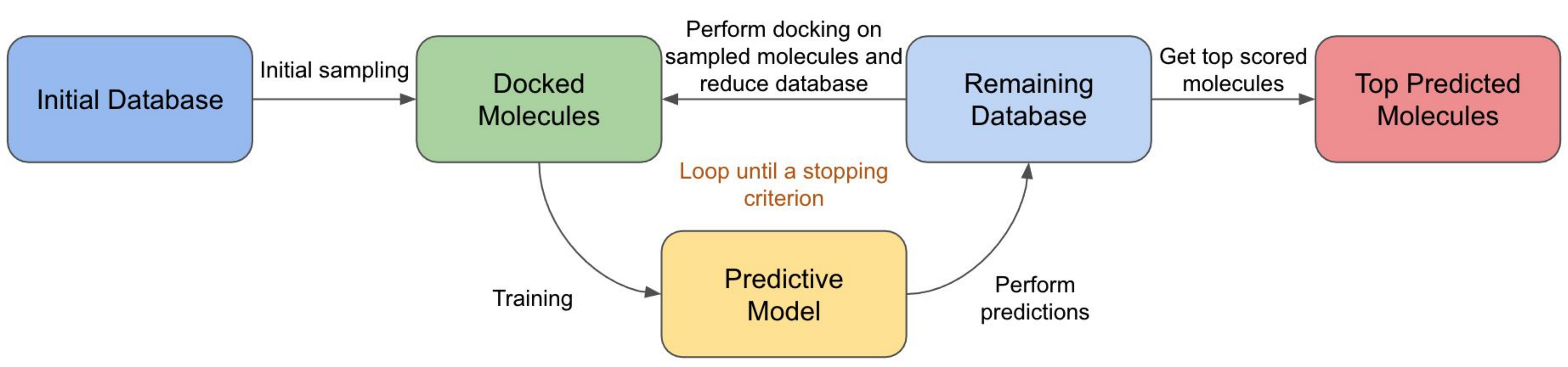
$$x^* = \underset{x \in \mathcal{A}}{\operatorname{argmax}} f(x)$$

where:

- A : large scale library (from 100M to 1B compounds)
- f : the docking scoring function, which is very expensive to compute, only a small fraction of the library can be evaluated, it can be seen as a *black box* function

Pipeline

Recent contributions[1,2] are tackling this problem with a bayesian optimization framework where a batch of molecules is sampled in an iterative fashion with an acquisition function, then docked and used to improve a surrogate model to predict the docking score.

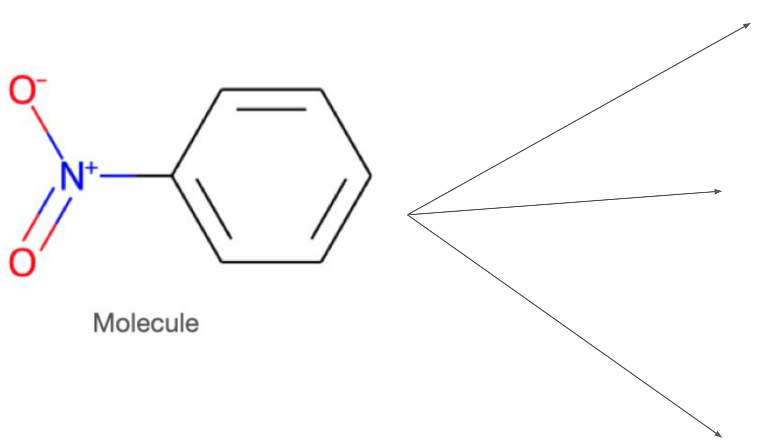


Key Elements

- Predictive Model
- Sampling Strategy
- Parameters: initial training size, sample size, number of iterations

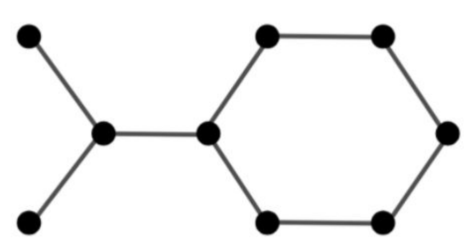
Predictive Models

Molecule Representation



1D Descriptors (molecular weight, number of bounds, …)

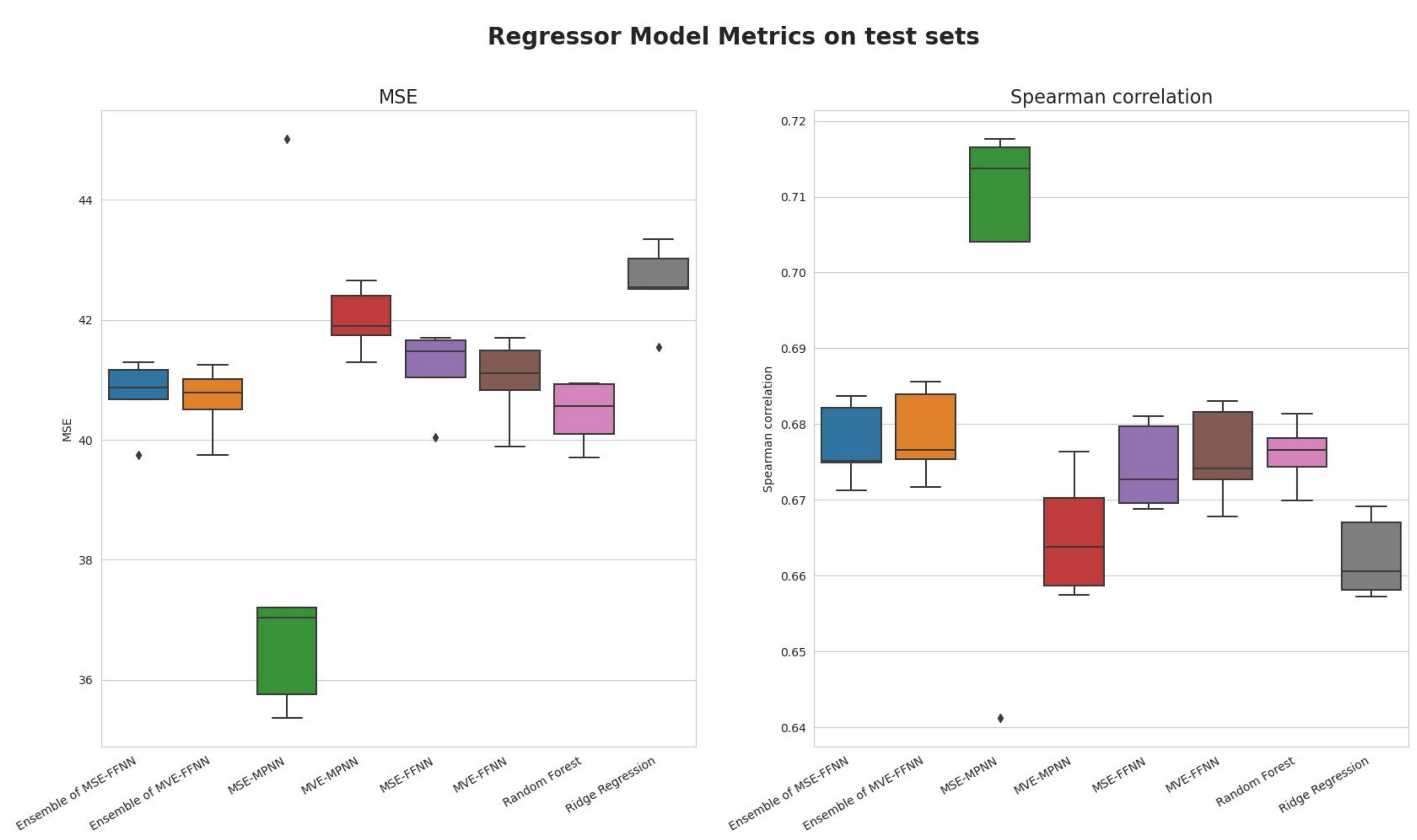
2D Descriptors: Morgan's Fingerprint



Graph Representation

Predictive Models

- Linear Models: Logistic Regression / Ridge Regression
- Non-linear Models: Random Forest, Gradient Boosting Trees
- Deep Learning Models
  - Feed Forward Neural Networks (FFNN)
  - Message Passing Neural Networks (MPNN)



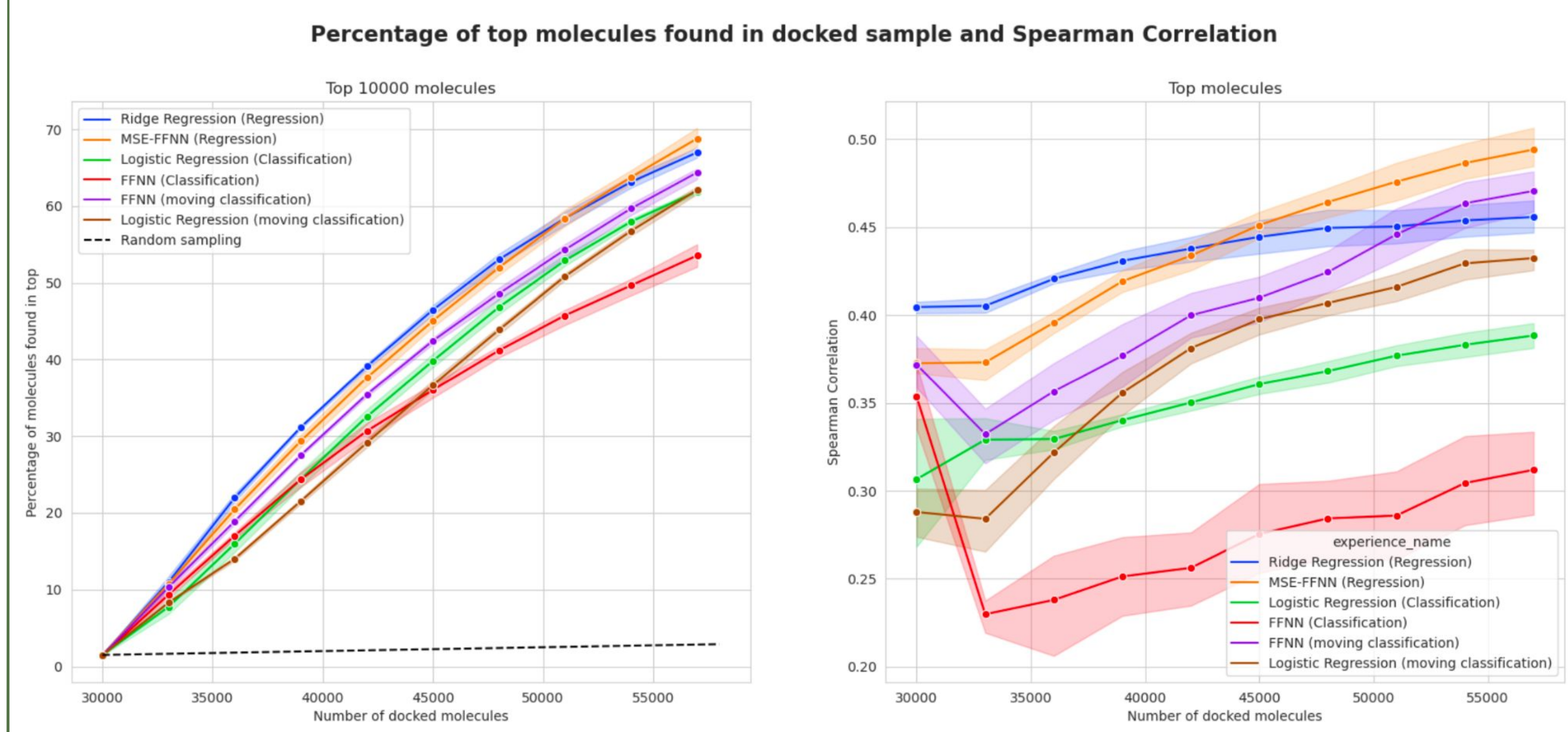
Sampling Strategies

Objective

At each iteration, the sample of molecules need to improve *both* the predictive model and the number of top scoring molecules.

Methods

- **Greedy**: select molecules with best predicted scores;
- **Random top**: select randomly molecules among top predicted molecules;
- **Uncertainty top**: select most uncertain molecules among top predicted molecules;
- **Diversity**[3] **top** select top molecule using a score depreciating molecules close to already sampled molecules



Diversity Strategies

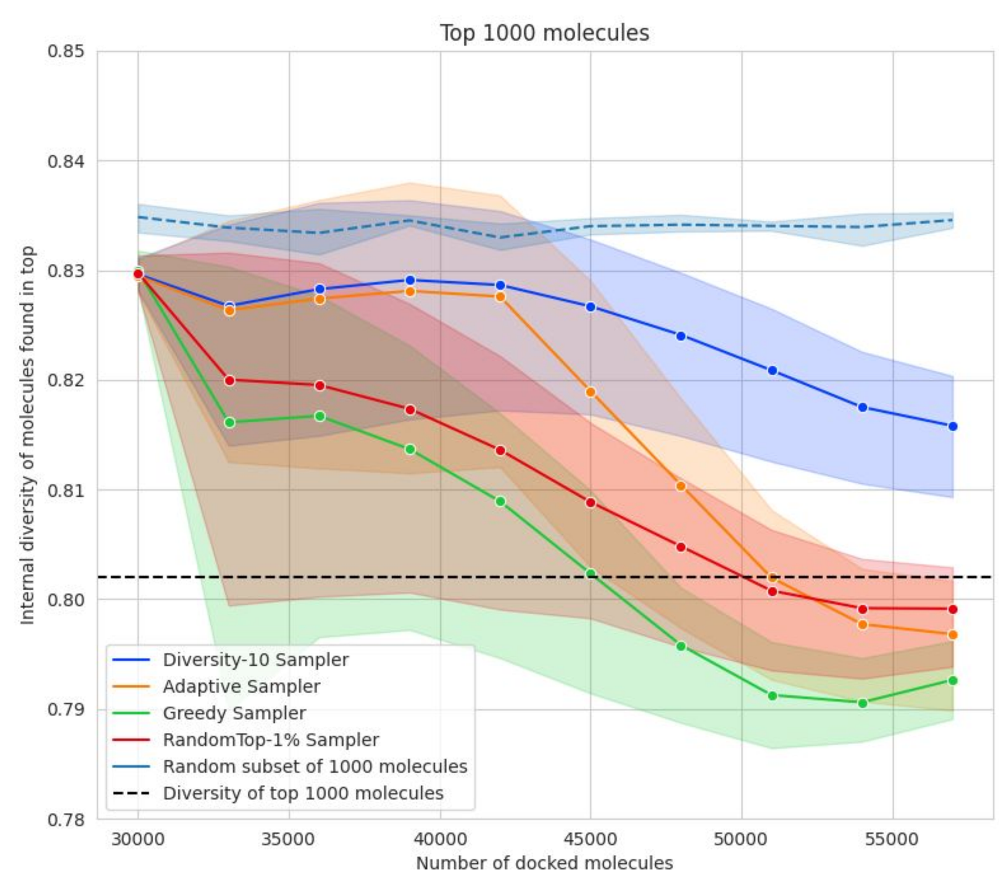
The goal is to obtain more diversity in the set of best molecules retrieved. Diversity strategy parameter *alpha* defines how much the score is decreased w.r.t the distance to already sampled molecules.

Settings

Experiment with a library of 2M molecules using a budget of 60k dockings (~ 3% of library size) using 10 sampling iterations. We measure the internal diversity as the mean of the pairwise tanimoto distance between all molecules of the set.

Results

The top molecules retrieved by the diversity sampler provides the most diverse set of best molecules..



Conclusion & Next Steps

Conclusions

- Regression frameworks retrieve a higher percent of top molecules than classification
- Greedy Sampling select more top molecules and diversity sampling select more diverse molecules

Next Steps

- Evaluate ensemble models which are more stable and generalize better
- Improve uncertainty quantification to evaluate uncertainty based methods

References

[1] Francois Berenger et al. Lean-Docking: Exploiting Ligands' Predicted Docking Scores to Accelerate Molecular Docking. Journal of Chemical Information and Modeling, 2021

[2] Francesco Gentile et al. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. ACS Central Science, 6(6), 2020.

[3] Thorsten Meinl et al. Maximum-score diversity selection for early drug discovery. Journal of Chemical Information and Modeling, 51(2):237–247, 2011.

Contact

Nicolas Drizard: nicolas.drizard@iktos.com