ESCOLA GAÚCHA DE BIOINFORMÁTICA

Molecular Docking in the Cloud:

Introduction to Molecular Docking

PhD. Pablo Ricardo Arantes
PhD. Conrado Pedebos

Porto Alegre, July 14th 2025

Molecular Docking Limitations

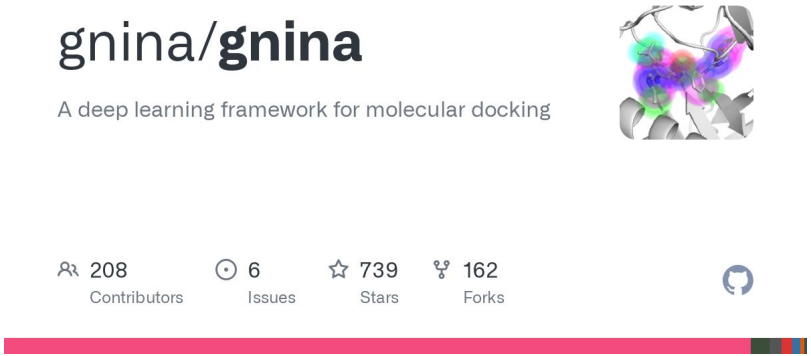
ESCOLA GAÚCHA DE BIOINFORMÁTICA

- Main limitations:
  - Rigid receptor (flexibility is possible, but limited and error-prone)
  - Computational cost for large datasets (important, but we'll talk more about this on Thursday)
  - Limited treatment of relevant effects (entropy, enthalpy, solvation)
  - **Scoring functions** (simplified approximations / many false positives and/or negatives / difficult to predict true experimental binding energies)

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

- Many new methods are being developed
- GNINA → Use of Convolutional Neural Networks (CNN) to evaluate poses
  - A fork of *smina*, which is a fork of *vina*
  - Increased accuracy for pose prediction and affinity scores
  - Trained using datasets of protein-ligand binding (PDBbind)
  - Autodock Vina is still used to generate initial poses → scoring with GNINA

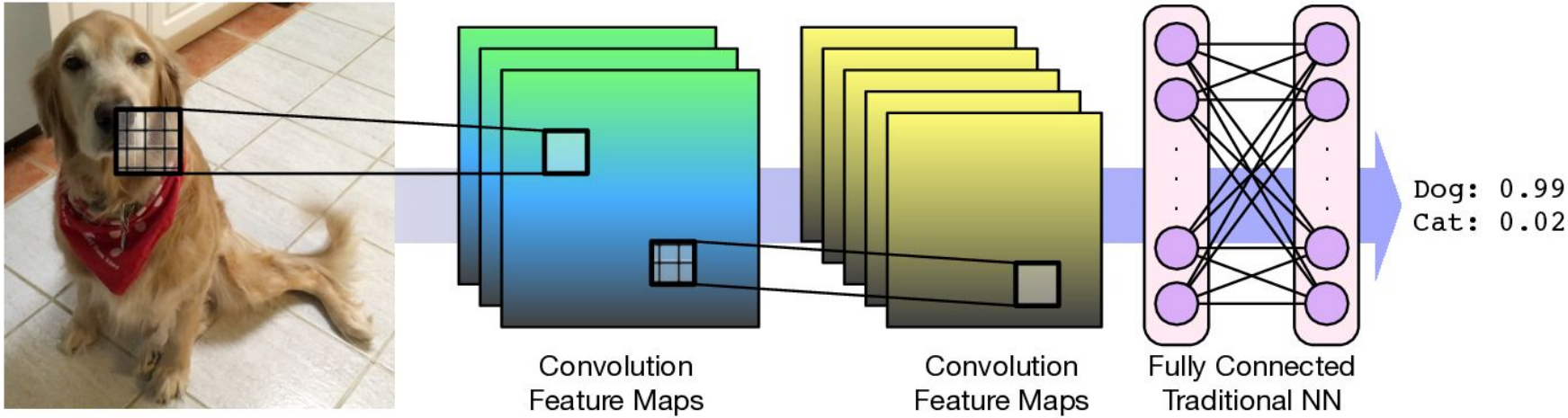


https://github.com/gnina/gnina

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Convolutional neural networks learn spatially related features of an input grid to generate a prediction.



Apply filters (kernels) that slide across input data to detect features (edges, shapes, patterns)

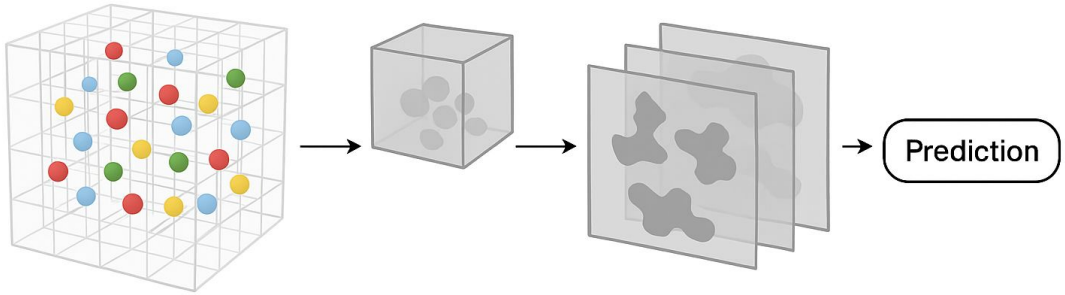
Interpret extracted features for classification or regression tasks

https://gnina.github.io/gnina/rsc_workshop2021/

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Convolution neural network (CNN) in a molecular context



3D Voxel Grid

Atom types

Convolution Filter

Solvent Environment
Residue-based Interactions
Flexibility

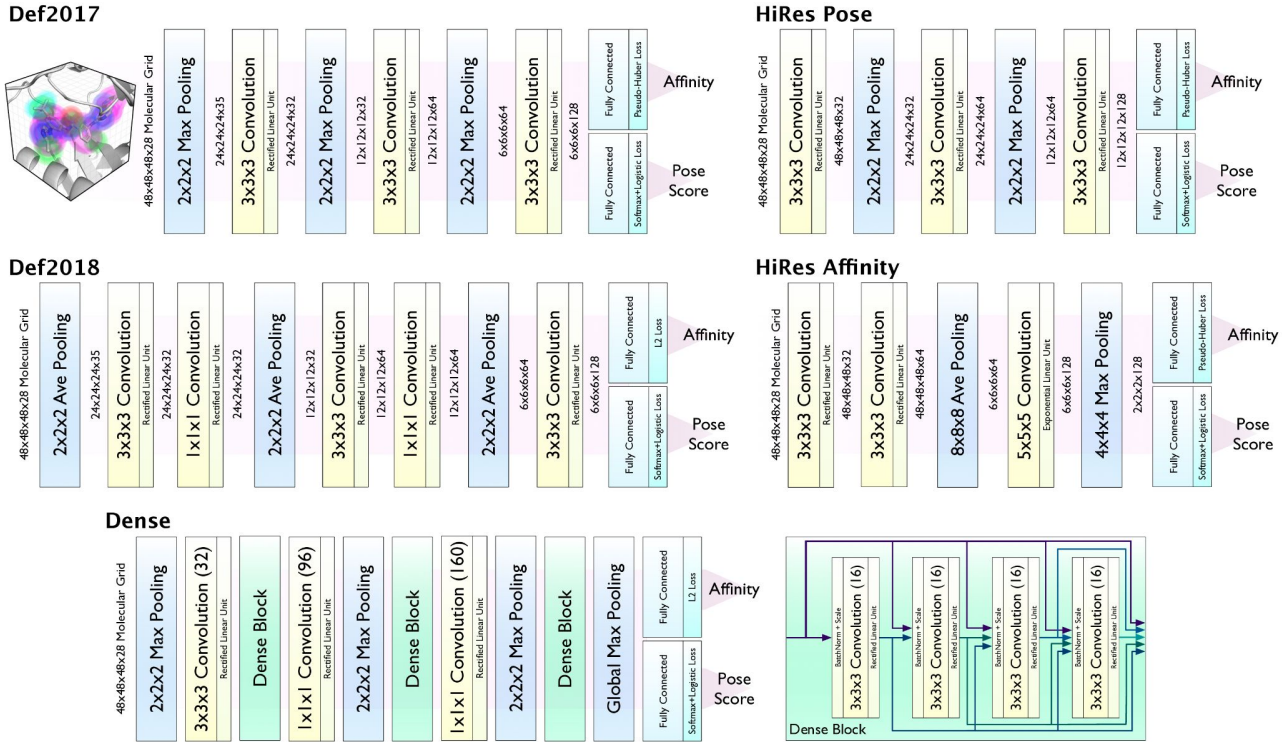
Feature Maps

Hydrophobic Patches
Electrostatic Interactions
Protein-Ligand Binding

AI-powered Molecular Docking

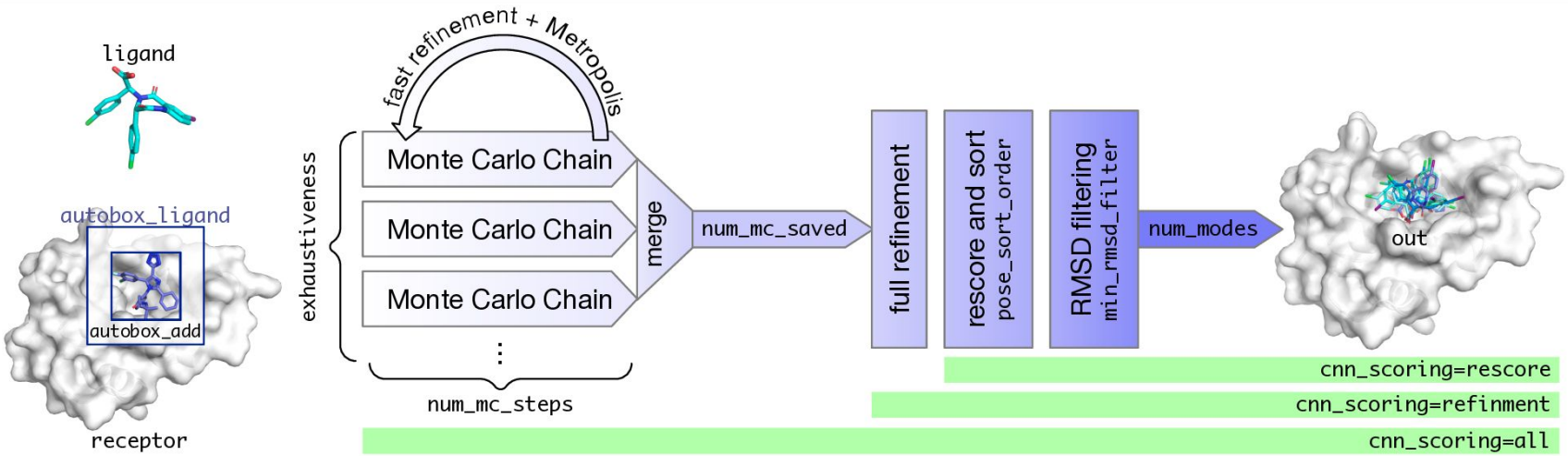
ESCOLA GAÚCHA DE BIOINFORMÁTICA

- Many different models can be employed
- A combination (ensemble) of five models was identified to give the best performance
- CNN model predicts both pose quality and binding affinity



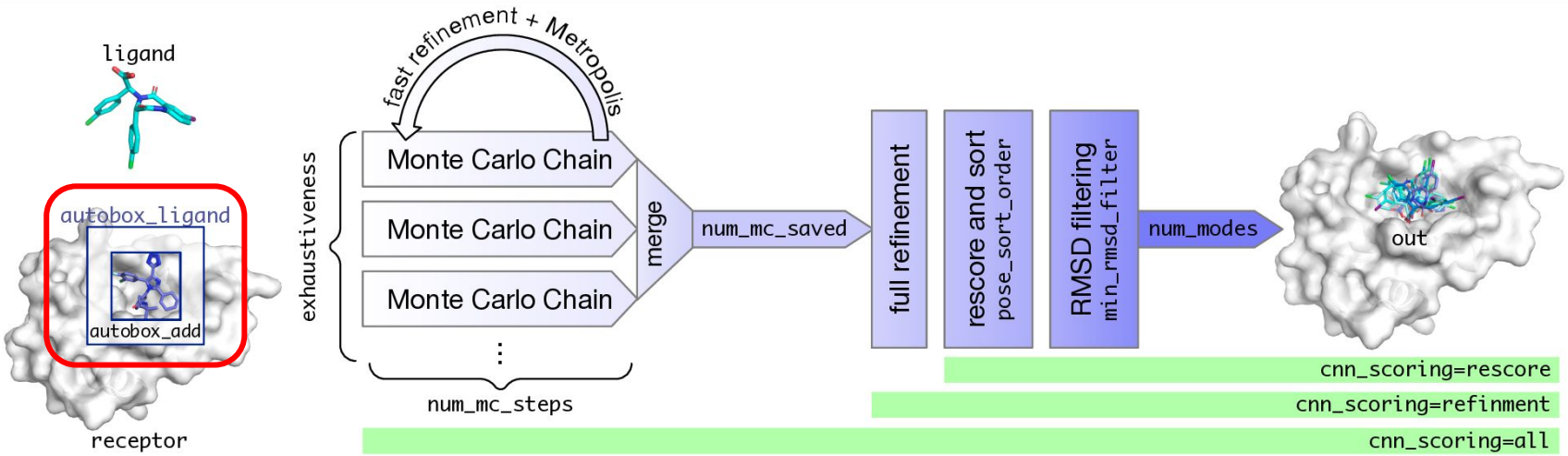
https://gnina.github.io/gnina/rsc_workshop2021/

AI-powered Molecular Docking



McNutt *et al. J. Cheminform*. (2021) 13, 43

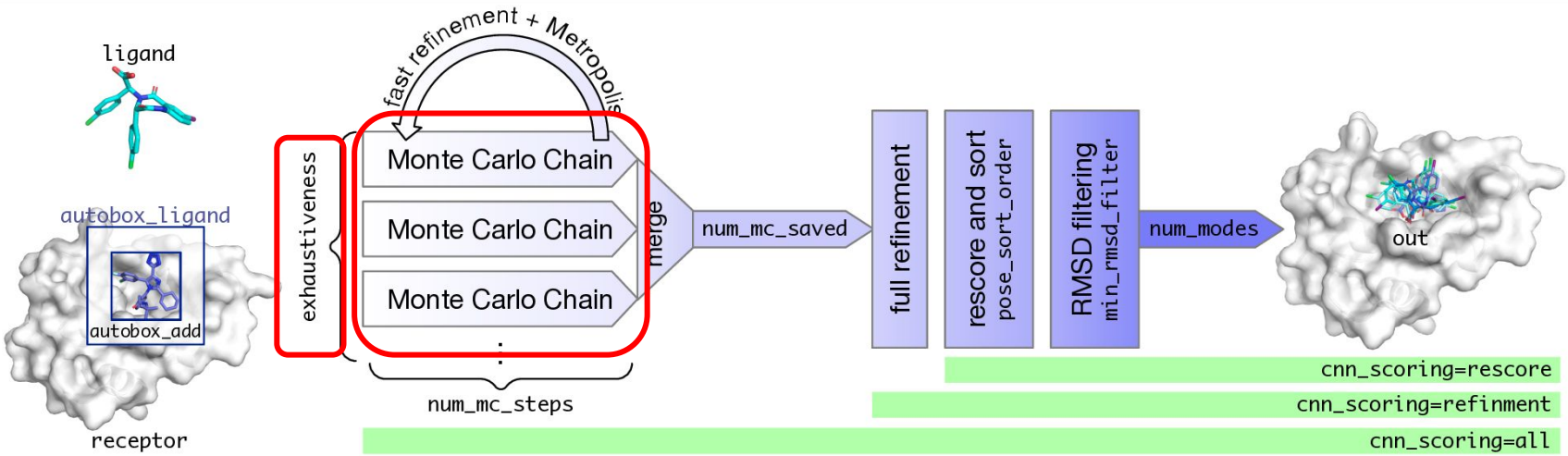
AI-powered Molecular Docking



autobox_ligand = creates the binding site box based on the ligand structure
autobox_add = adds the amount of buffer space

McNutt *et al. J. Cheminform*. (2021) 13, 43

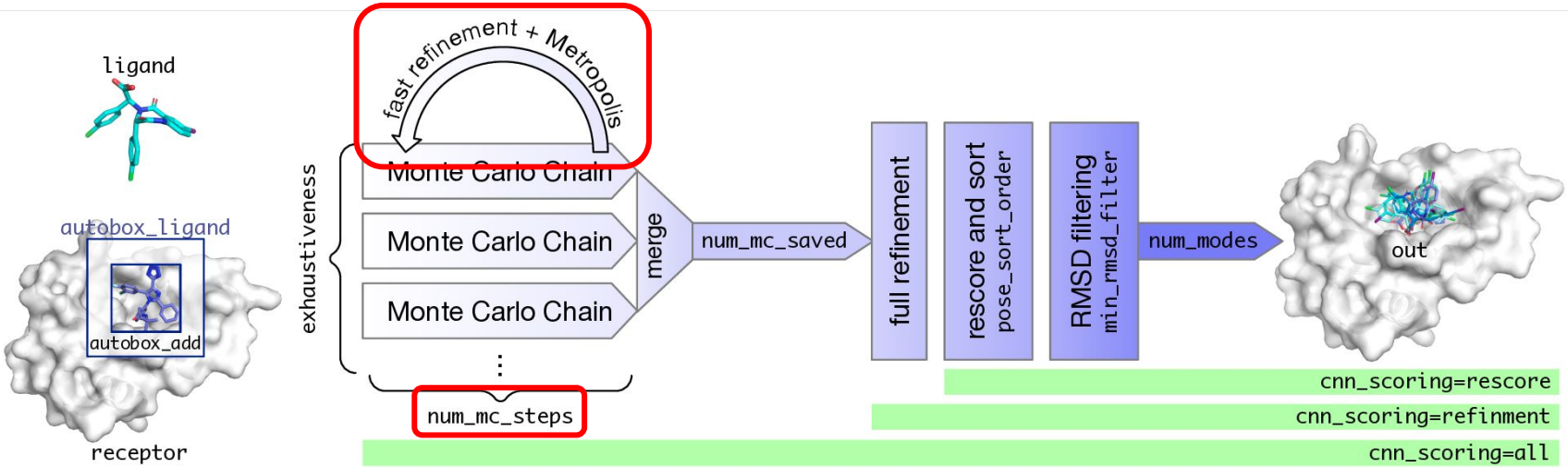
AI-powered Molecular Docking



Monte Carlo Chain = Monte Carlo sampling of the ligand conformations
exhaustiveness = number of chains running concomitantly

McNutt *et al. J. Cheminform*. (2021) 13, 43

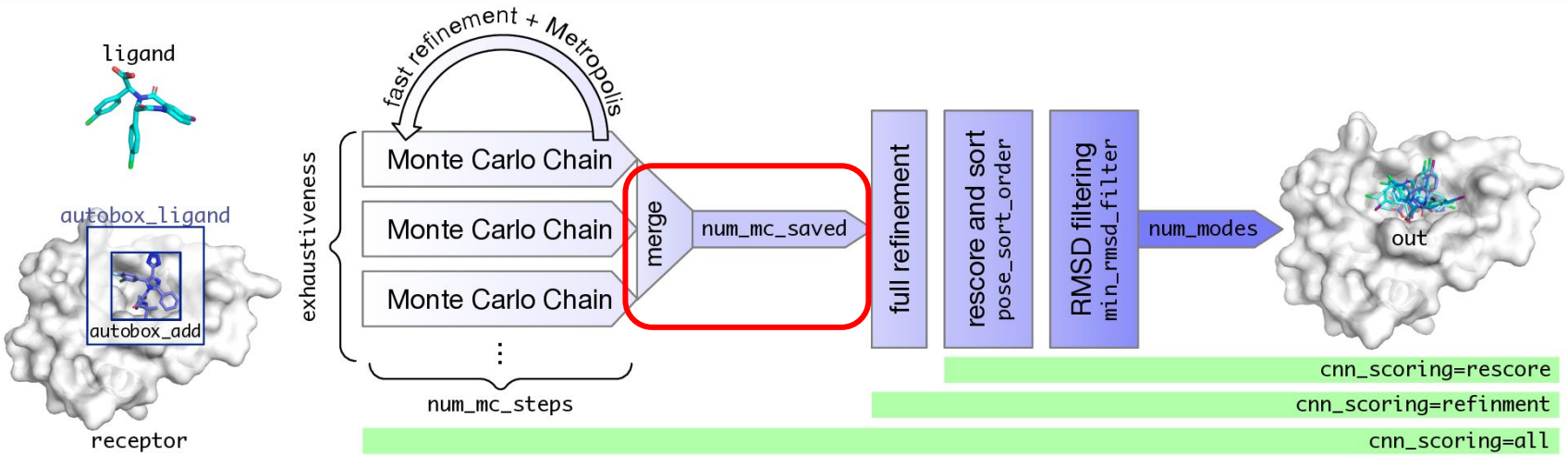
AI-powered Molecular Docking



fast refinement + Metropolis = energy minimization and criterion for acceptance of the new state
num_mc_steps = automatically determined (can be changed)

McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking



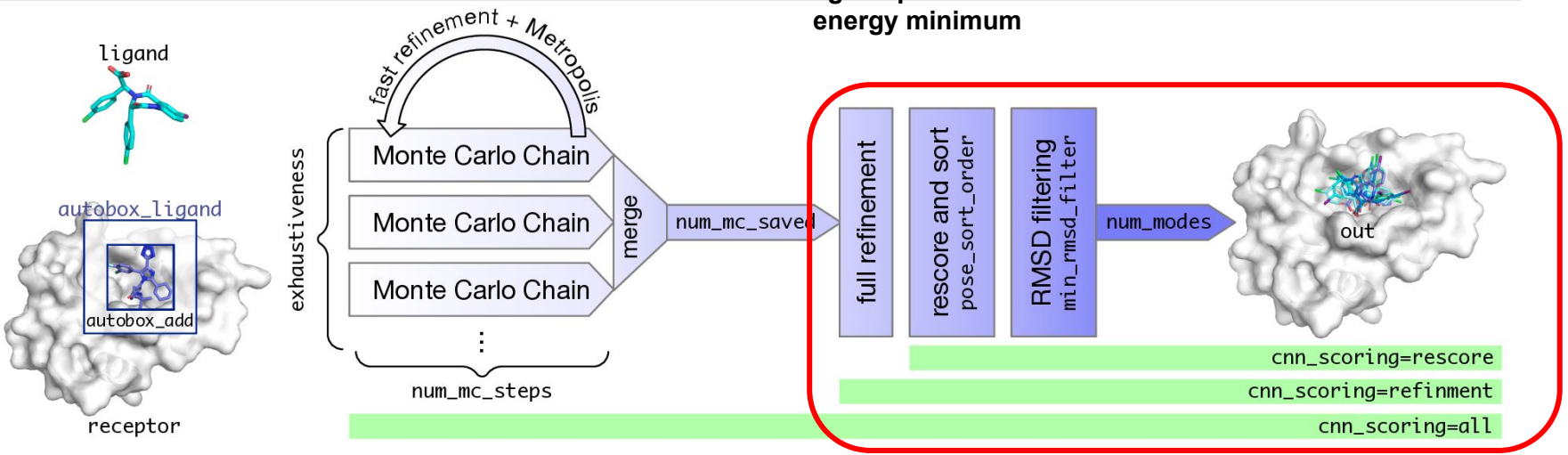
num_mc_saved = 50 (default), number of top conformations retained

McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

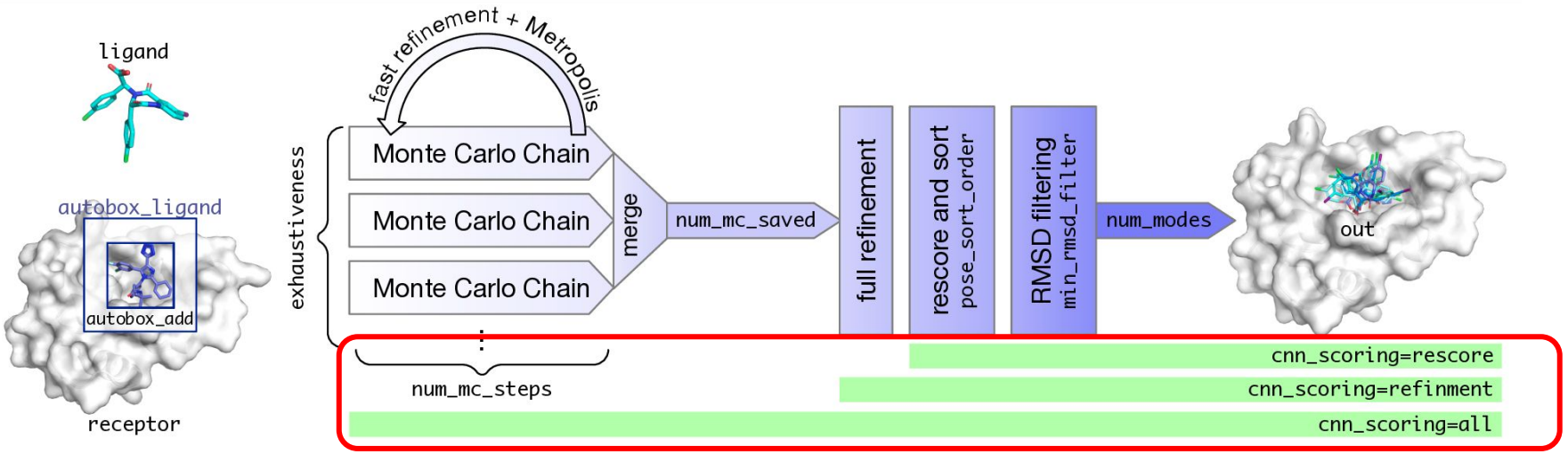
full refinement = shifts the ligand pose to a local energy minimum



pose_sort_order = how the top scoring poses are ranked
min_rmsd_filter = ligand conformations too similar are filtered out
num_modes = final number of poses for analysis (default = 10)

McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking



McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

```
gnina --receptor 1BCU_PROT.pdb --ligand 1BCU_LIG.sdf
    --out 1BCU_gnina_poses.sdf.gz --autobox_ligand
    1BCU_LIG.sdf --autobox_add 4 --cnn
    crossdock_default2018 dense_3 --cnn_scoring rescore
    --exhaustiveness 8 --num_mc_saved 50 --cnn_rotation 0
    --num_modes 9 --min_rmsd_filter 1
```

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

```
gnina --receptor 1BCU_PROT.pdb --ligand 1BCU_LIG.sdf
    --out 1BCU_gnina_poses.sdf.gz --autobox_ligand
    1BCU_LIG.sdf --autobox_add 4 --cnn
    crossdock_default2018 dense_3 --cnn_scoring rescore
    --exhaustiveness 8 --num_mc_saved 50 --cnn_rotation 0
    --num_modes 9 --min_rmsd_filter 1
```

Different rotations of the protein-ligand complex the CNN is able to see for each conformation (does not seem to alter the results, so it is set to 0)

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

- What's behind the workflow?
  - OpenBabel for input parsing (PDB, sdf, mol2, etc)
  - Binding site: cartesian coordinates or automatic box around the ligand (input ligand)
    - **First lecture**: specified manually
    - **autobox_ligand** is used to define the binding site, a rectangular prism is constructed using the minimum and maximum values for the x, y, and z coordinates of the ligand to which additional spacing (**autobox_add**) is added in every dimension (**ligand needs to rotate freely**)
  - Scoring functions:
    - If CNN for rescoring only: user-input or built-in (Vina, Vinardo)
    - CNN function: user-input or built-in (*crossdock_default2018, dense, general_default2018, redock_default2018, and default2017*) - each of which is trained using different training data and/or a different model architecture

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

- What's behind the workflow?
  - CNN calculations are performed using the cuDNN accelerated Caffe deep learning framework (GNINA 1.0) or PyTorch (GNINA 1.3)
  - CNN models predict:
    - Pose score (**CNNscore**) = probability that the pose has a low root mean square deviation (RMSD) to the binding pose
    - Binding affinity (pK) (**CNNAffinity**)
  - *cnn_scoring default* = CNN scoring for the final resorting of ligand conformations and the empirical scoring function everywhere else in the pipeline.
  - *cnn_scoring*:
    - "none" = smina pipeline
    - "rescoring" = CNN used in the final sorting of ligand conformations
    - "refinement" = same as the previous step but CNN is also used after Monte Carlo chain step for refining
    - "all" = CNN used at every step including energy minimization

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

```
                 _
                (_)
  __ _ _ __   _ _ __   __ _
 / _` | '_ \| | '_ \ / _` |
| (_| | | | | | | | | (_| |
 \__, |_| |_|_|_| |_|\__,_|
  __/ |
 |___/

gnina v1.0.1 HEAD:aa41230   Built Mar 23 2021.
gnina is based on smina and AutoDock Vina.
Please cite appropriately.

Commandline: ./gnina -r rec.pdb -l lig.pdb --autobox_ligand lig.pdb
Using random seed: -216854720

0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************

mode |  affinity  |    CNN     |   CNN
     | (kcal/mol) | pose score | affinity
-----+------------+------------+----------
    1       -8.51       0.8985      6.783
    2       -8.30       0.4491      6.450
    3       -6.80       0.3258      6.043
    4       -7.34       0.3023      6.230
    5       -5.90       0.1754      5.397
    6       -6.33       0.1679      5.559
    7       -6.98       0.1668      5.825
    8       -5.24       0.1607      5.505
    9       -7.00       0.1523      5.957
```

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

```
            _
           (_)
  __ _ _ __  _ _ __   __ _
 / _` | '_ \| | '_ \ / _` |
| (_| | | | | | | | | (_| |
 \__, |_| |_|_|_| |_|\__,_|
  __/ |
 |___/

gnina v1.0.1 HEAD:aa41230   Built Mar 23 2021.
gnina is based on smina and AutoDock Vina.
Please cite appropriately.

Commandline: ./gnina -r rec.pdb -l lig.pdb --autobox_ligand lig.pdb
Using random seed: -216854720

0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************

mode |  affinity  |    CNN     |   CNN
     | (kcal/mol) | pose score | affinity
-----+------------+------------+----------
    1       -8.51       0.8985      6.783
    2       -8.30       0.4491      6.450
    3       -6.80       0.3258      6.043
    4       -7.34       0.3023      6.230
    5       -5.90       0.1754      5.397
    6       -6.33       0.1679      5.559
    7       -6.98       0.1668      5.825
    8       -5.24       0.1607      5.505
    9       -7.00       0.1523      5.957
```

Affinity is the original scoring by *vina* (or *vinardo*) scoring function

CNN pose score is a value between 0 and 1 that is used to rank the poses of the ligand, where a score of 1 denotes a perfect ligand pose

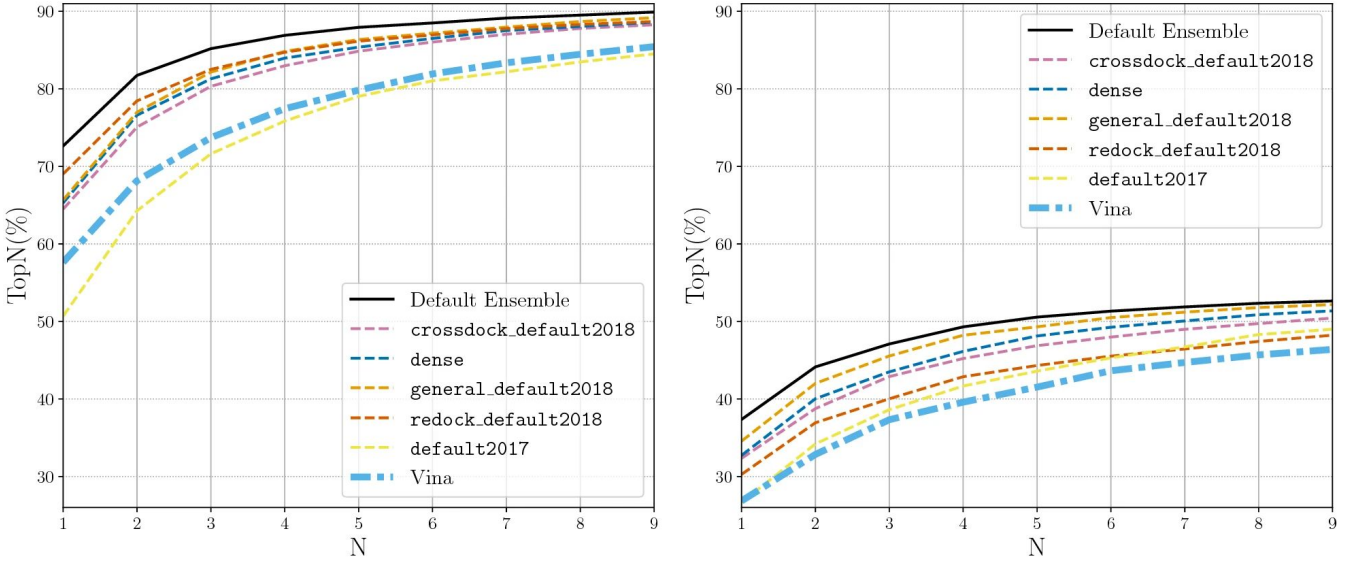
CNNaffinity is the affinity of the docked complex as determined by the CNN

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Default ensemble = *dense, general_default2018_3, dense_3, crossdock_default2018, and redock_default2018*

(only “rescoring” option used)



(a) Redocking (b) Cross-docking

Top1 = >70%

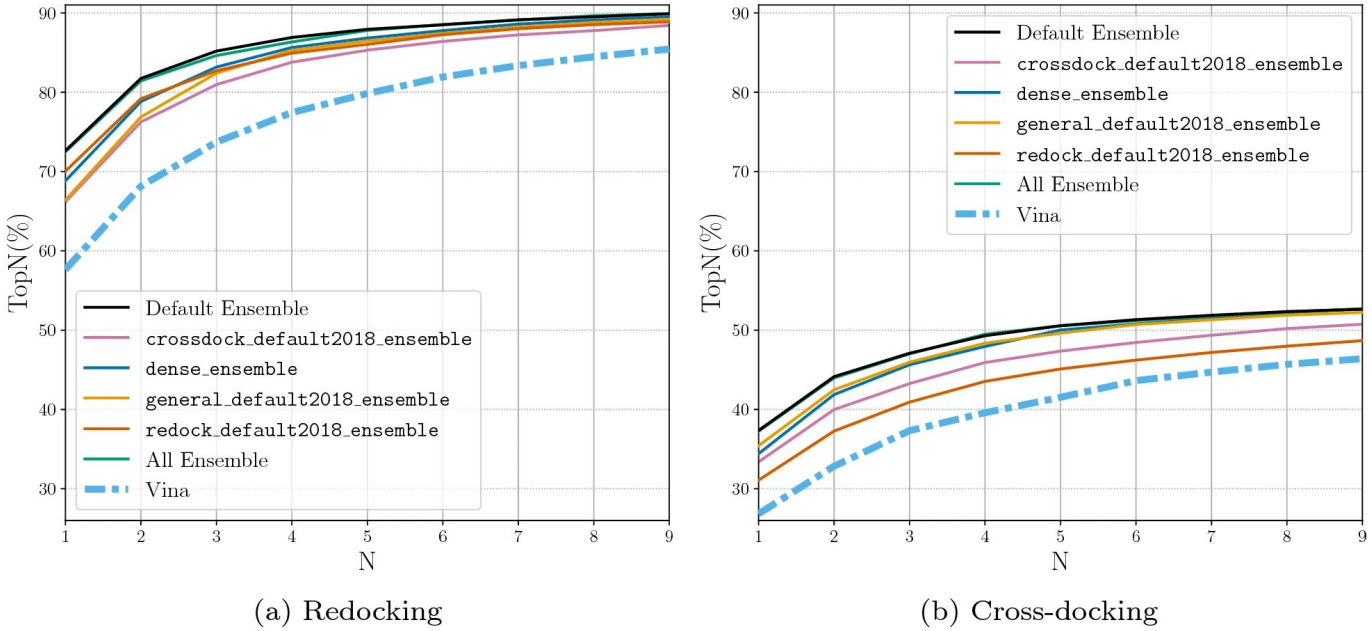
In 70% of the protein-ligand docking targets tested, the top 1 pose had an RMSD < 2.0 Å from the crystal pose.

McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Using All Ensembles achieve similar results as the Default (not worth it!)



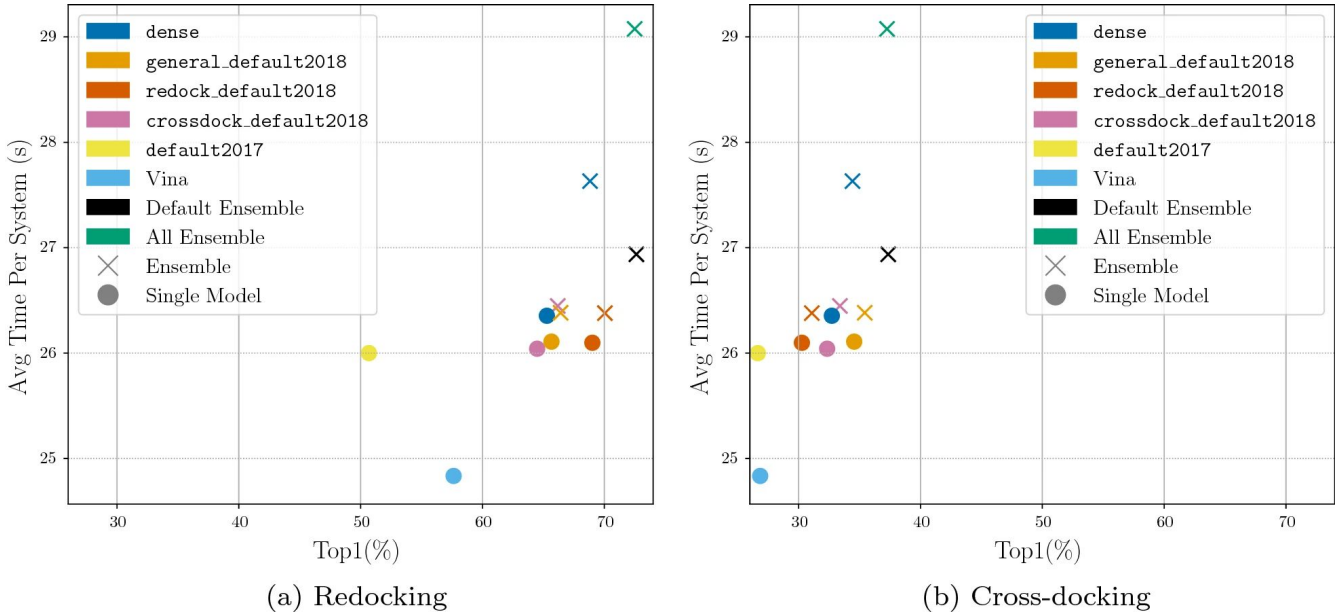
(a) Redocking (b) Cross-docking

McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Default is not the fastest, but achieves the best results

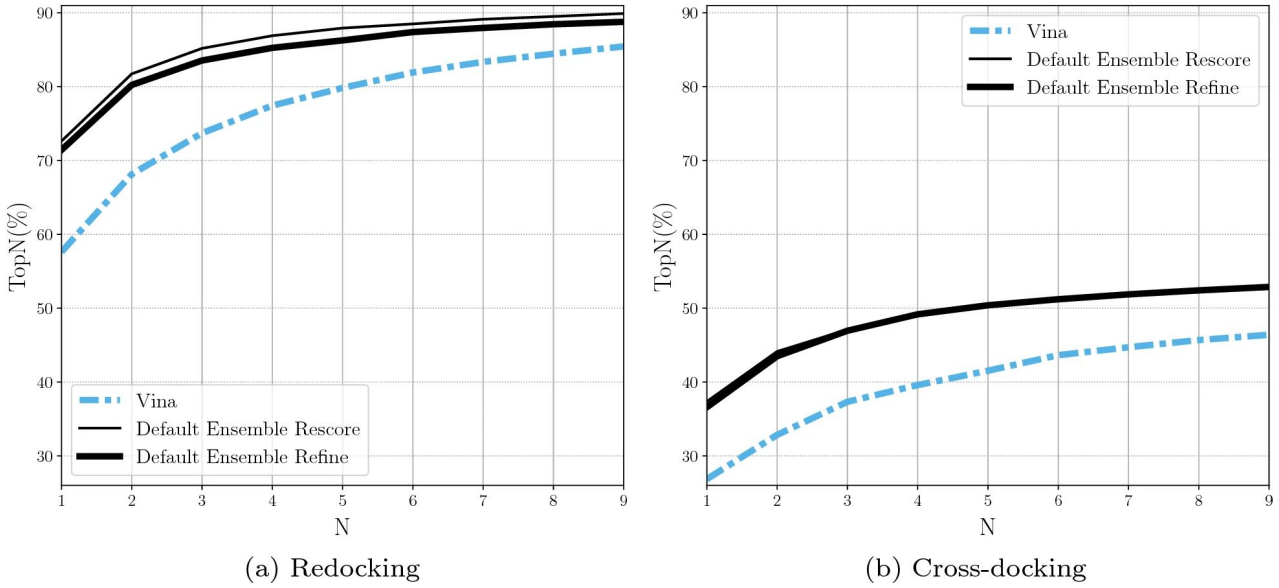


McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Increasing the CNN level to Refine does not achieve better results

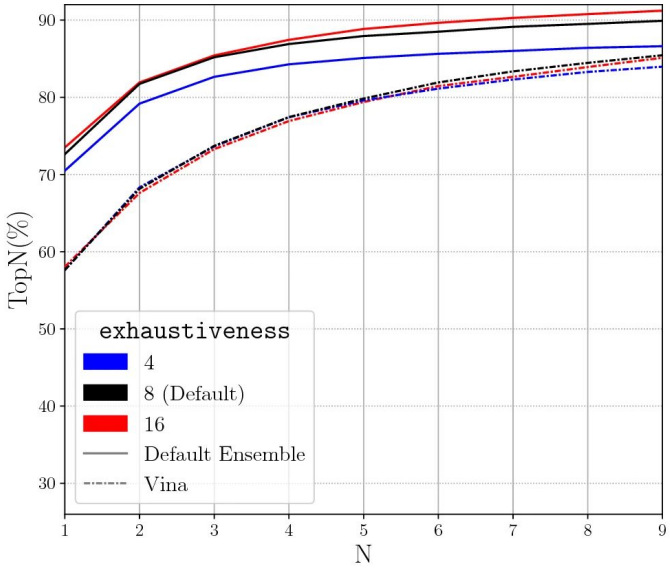


McNutt *et al. J. Cheminform*. (2021) 13, 43

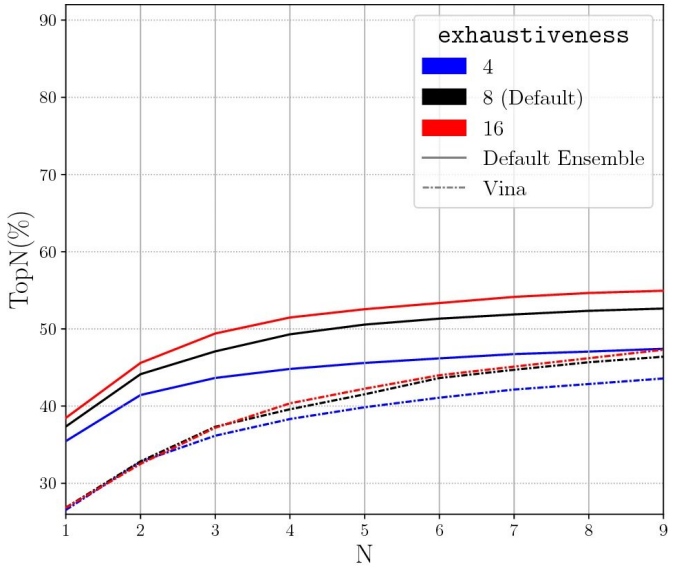
AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

In typical docking calculations, increasing the exhaustiveness parameters yields better results, but it might not be worth it…



(a) Redocking



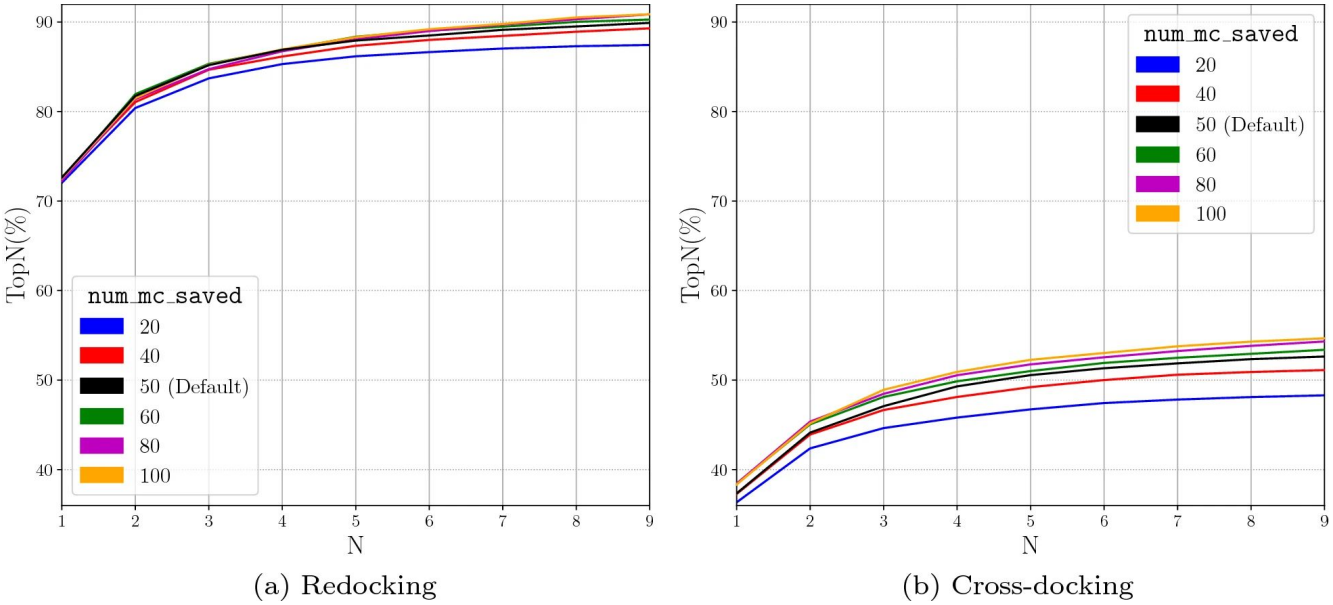
(b) Cross-docking

McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

default (50) num_mc_saved is probably the best cost-benefit scenario

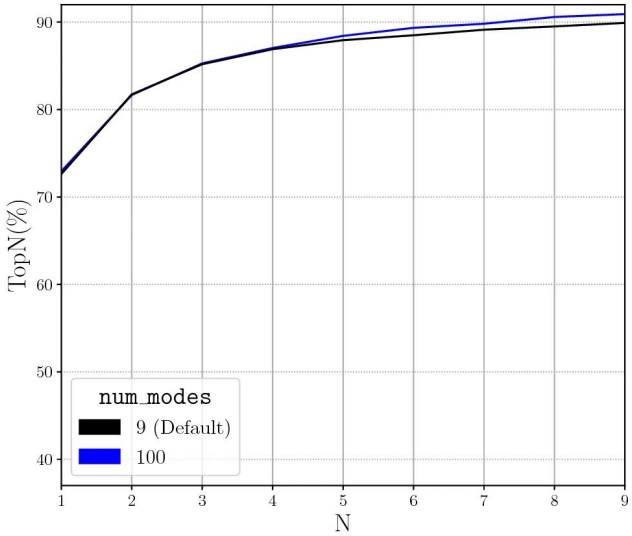


McNutt *et al. J. Cheminform*. (2021) 13, 43

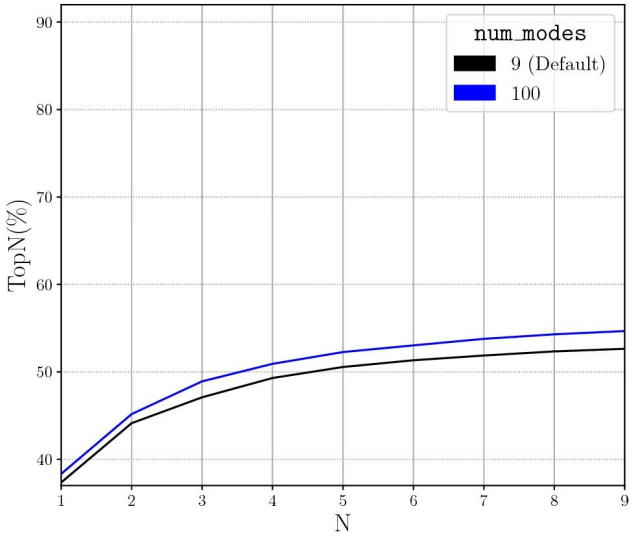
AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

The first 9 poses are usually enough to get achieve 90% TopN (at least one pose below 2 Å of RMSD)



(a) Redocking



(b) Cross-docking

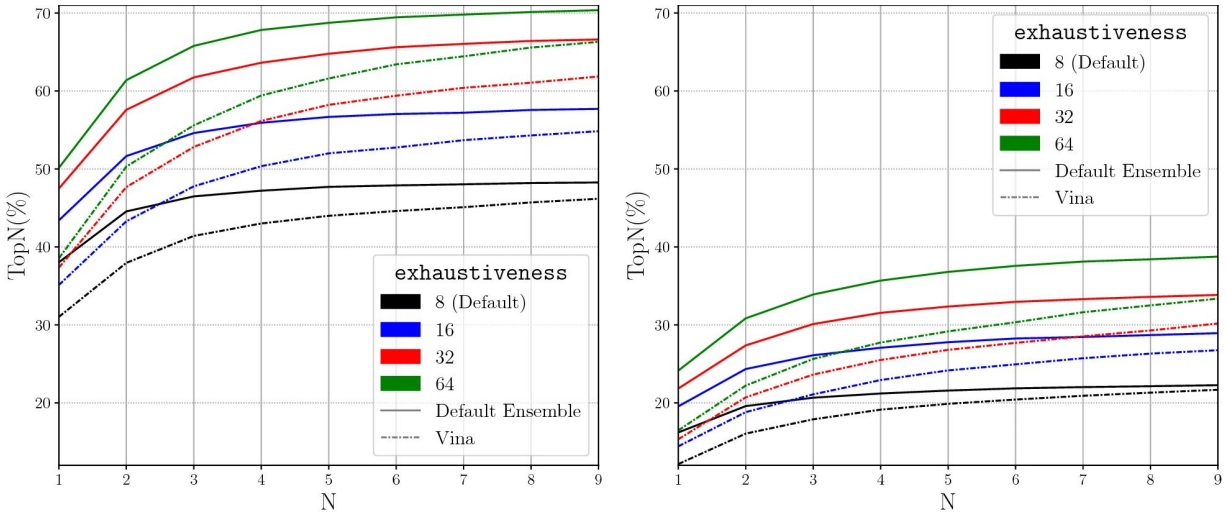
McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Things change when we do a blind docking run!

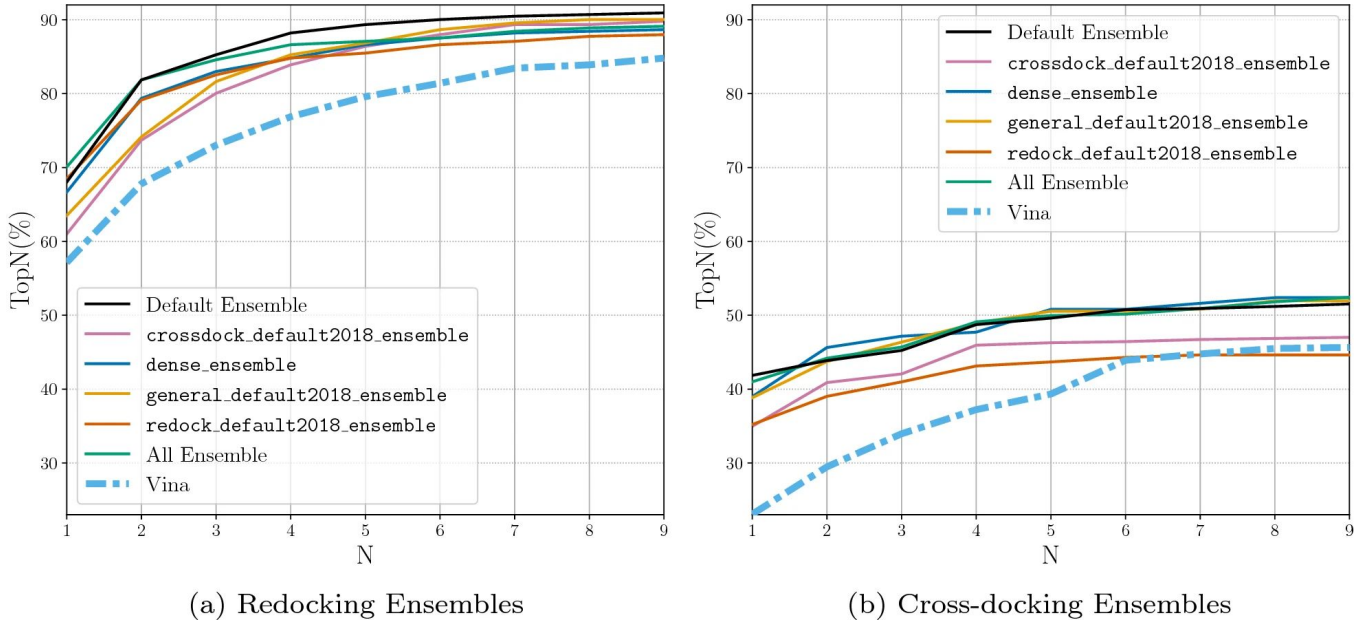
Use max. exhaustiveness available



(a) Redocking (b) Cross-docking

McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking



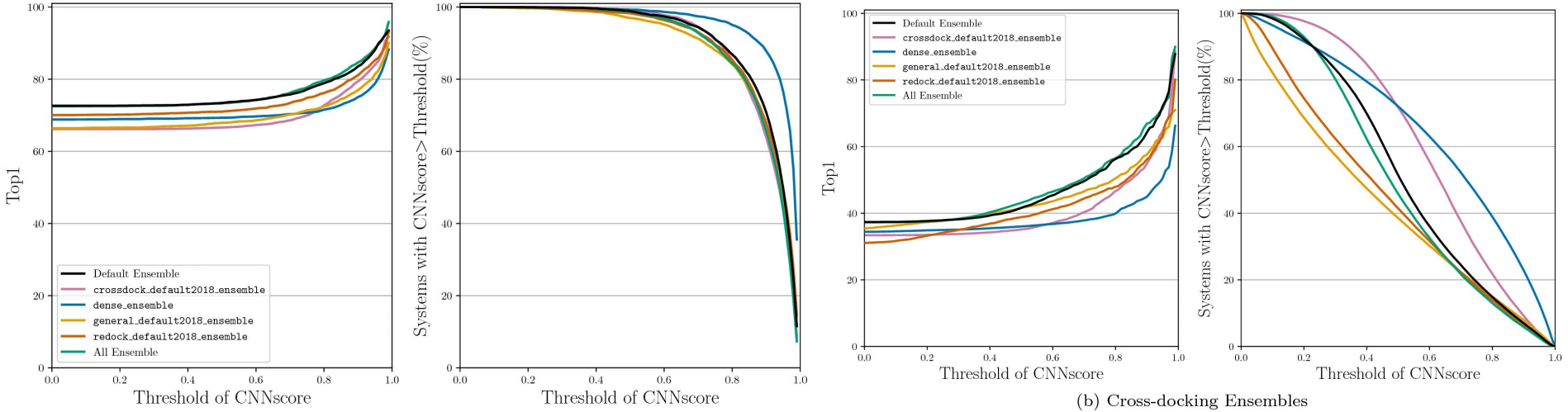
(a) Redocking Ensembles

(b) Cross-docking Ensembles

Testing against data not used in the training set

McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking



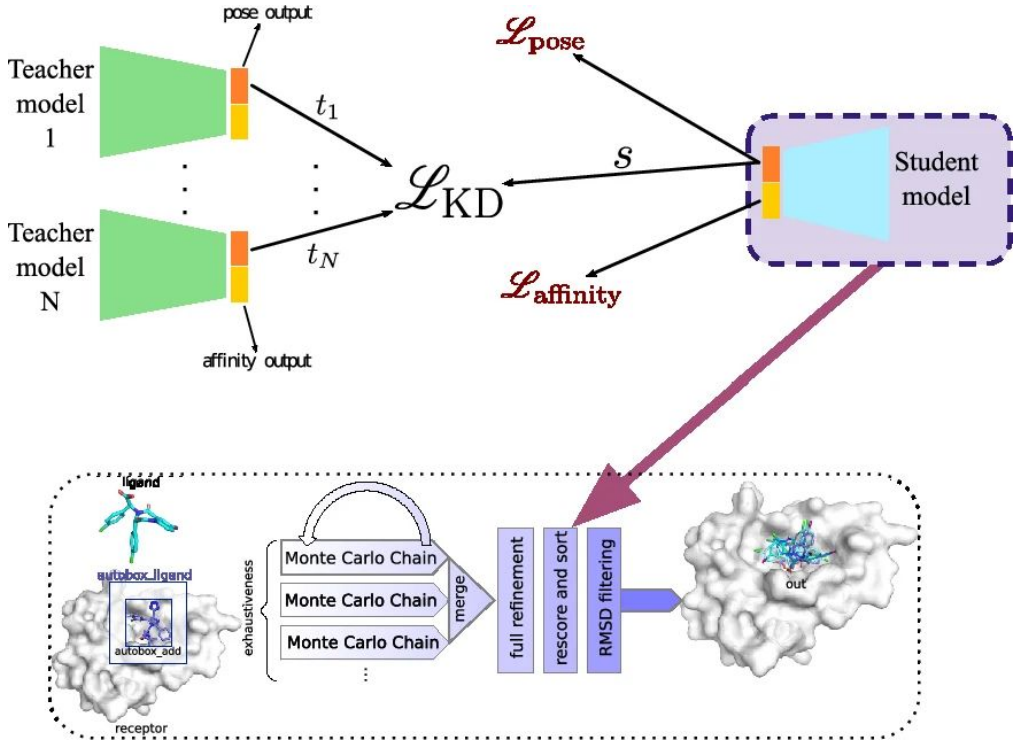
McNutt *et al. J. Cheminform*. (2021) 13, 43

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Gnina 1.3 updates

Faster calculations



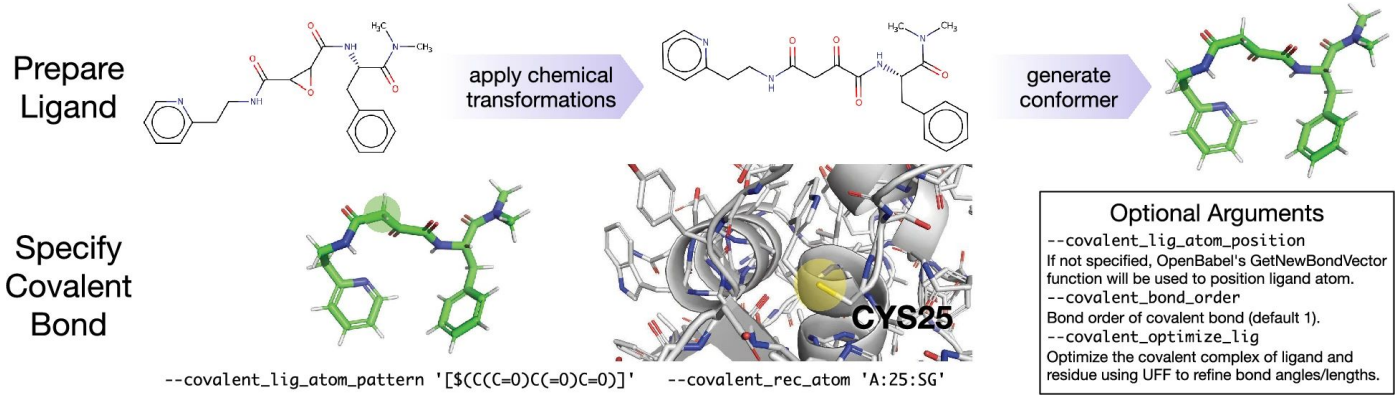
McNutt *et al. J. Cheminform*. (2025) 17, 28

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Gnina 1.3 updates

Covalent docking (atom of the ligand is covalently bound to an atom of the receptor)



McNutt *et al. J. Cheminform*. (2025) 17, 28

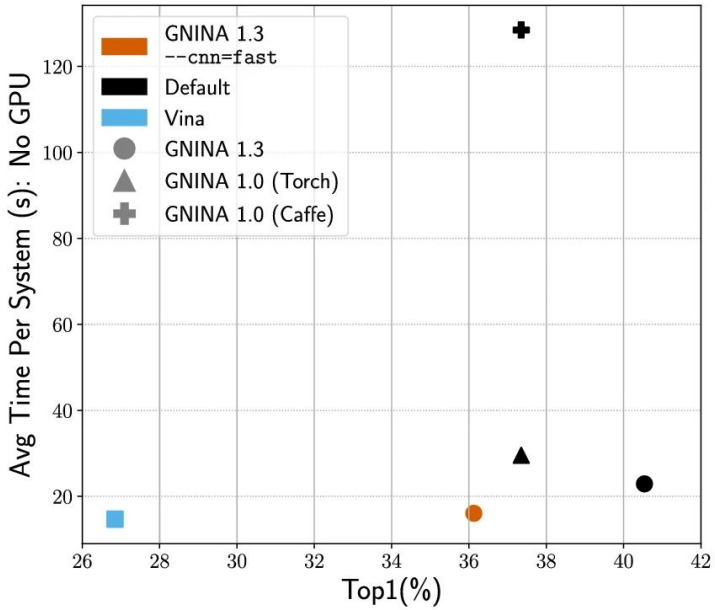
AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

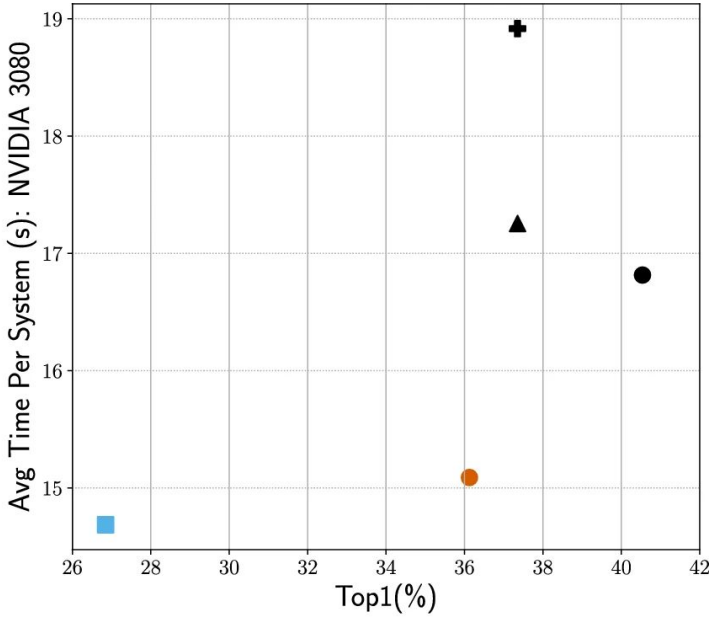
Gnina 1.3 updates

Old default ensemble is almost as fast as Vina now! (--cnn=fast)

Good for high throughput VS



(a) No GPU



(b) GPU

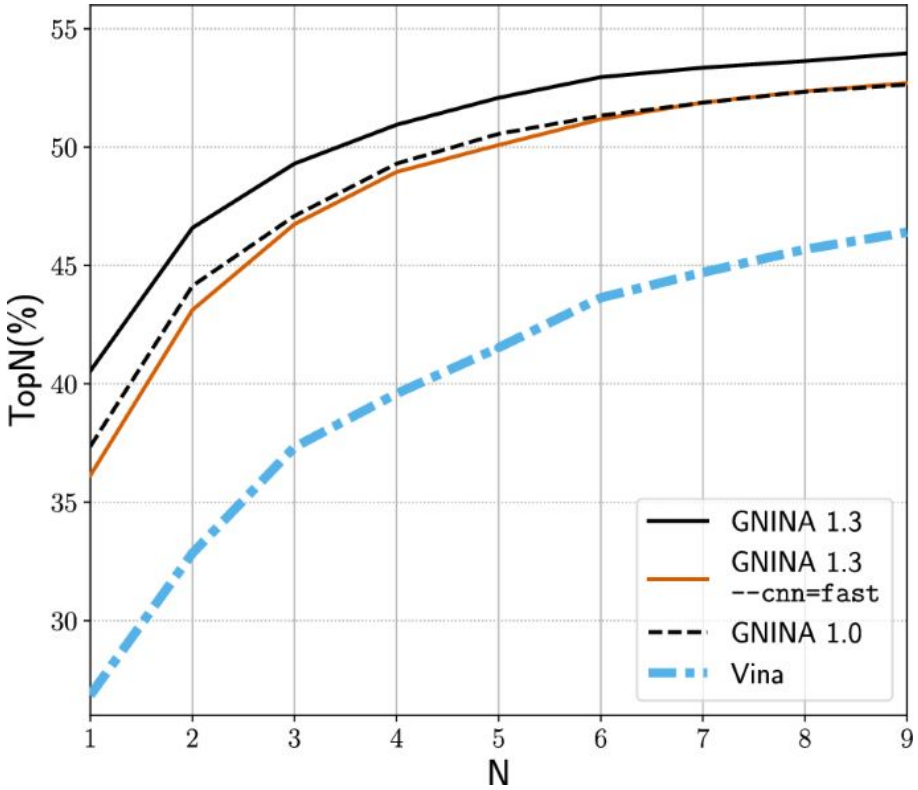
McNutt *et al. J. Cheminform*. (2025) 17, 28

AI-powered Molecular Docking

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Gnina 1.3 updates

New default ensemble performs better than the rest

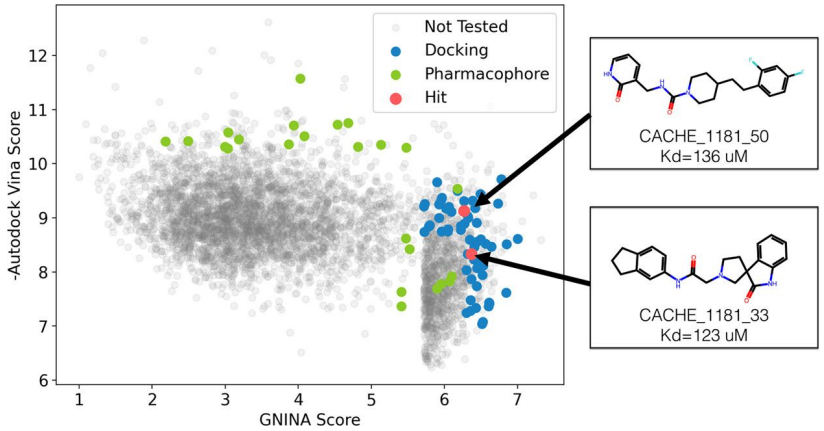


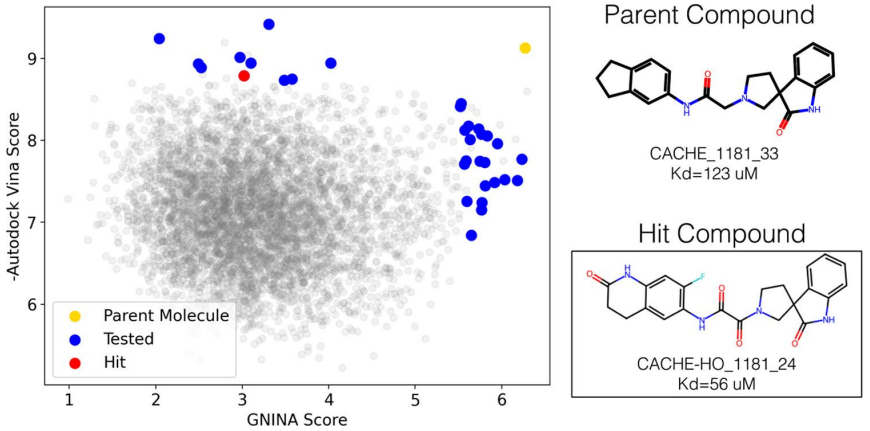
McNutt *et al. J. Cheminform*. (2025) 17, 28

CACHE Challenge #1

ESCOLA GAÚCHA DE BIOINFORMÁTICA

Critical Assessment of Computational Hit-finding Experiments: A public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding





Gnina tied in the first place in the Challenge #1!

"...we find it noteworthy that a purely docking-based approach (with Gnina) was not obviously distinguishable from more complex approaches that used molecular simulation or were guided by expert medicinal chemist."

Dunn *et al. J. Chem. Inf. Model*. (2024) 64 (24), 9388–9396

AI-powered Molecular Docking - Practical Lesson

ESCOLA GAÚCHA DE BIOINFORMÁTICA

- We’re using Colab notebooks to perform molecular docking using Gnina!
- You can access the notebooks via the link below or the QR code:

https://github.com/pablo-arantes/EGB2025-MC14