

# DockTDesign

## Deep Generative Models for *de novo* Drug Design

Matheus M. P. da Silva

PhD Defense in Computational Modeling  
National Laboratory for Scientific Computing (LNCC/MCTI)

Advisors: Laurent E. Dardenne & Isabella A. Guedes

October 10, 2025

# Summary

---

## 1. Introduction and Motivation

1.1 Objetivo geral

## 2. Fundamentação Teórica

2.1 Machine Learning and Deep Learning

2.2 Multi and *Many*-Objective Optimization

## 3. Literature Review

## 4. DockTDeep

4.1 Methodology

4.2 Results

## 5. DockTDesign

5.1 Methodology

5.2 Results

## 6. Conclusions and Perspectives

# Motivation

---

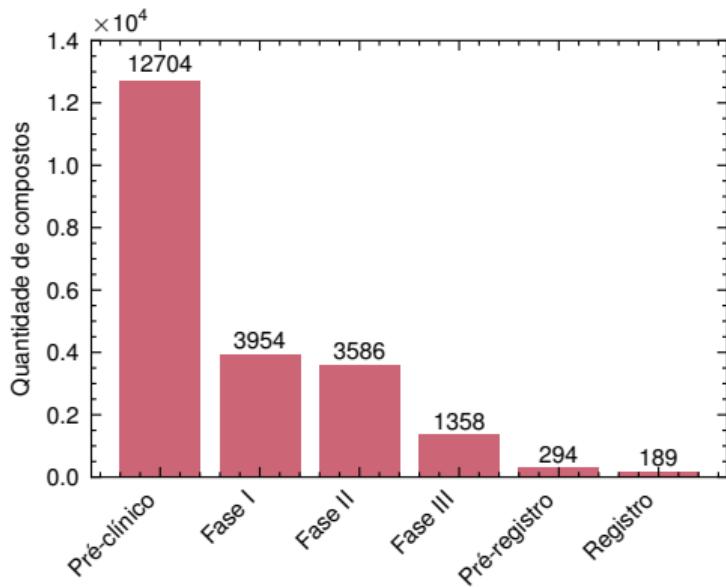


Figura: Number of compounds per development phase in 2025. Source: [Pharmaprojects](#).

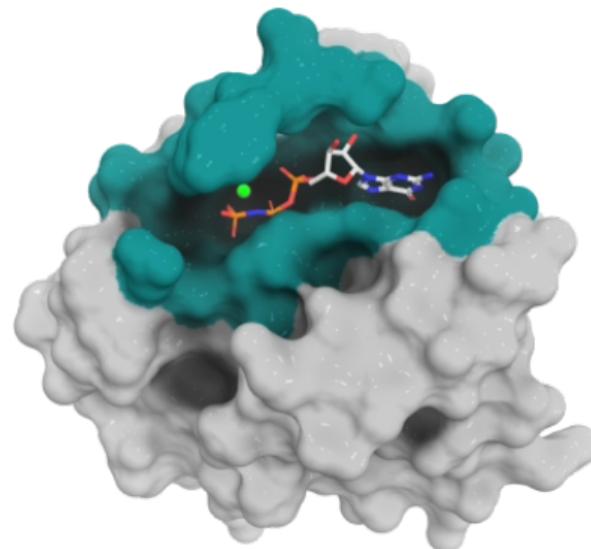
- Drug development: long, costly, and with low success rates ( $\leq 5\%$  in the preclinical phase).
- Increasing the success rate in early phases can have a large economic and public-health impact.
- Computational methods are essential for the early identification of promising molecules (*hits*).

## Objectives

- Pose prediction
- Binding affinity prediction

## Components

- Search algorithm
- Scoring function:
  - pose
  - binding affinity



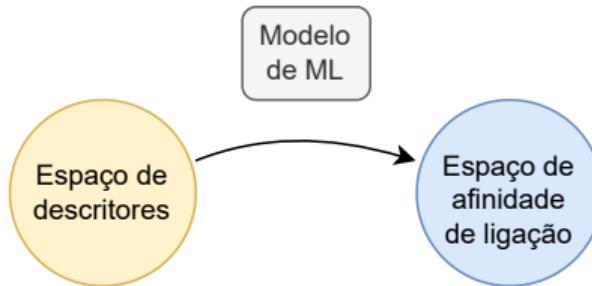
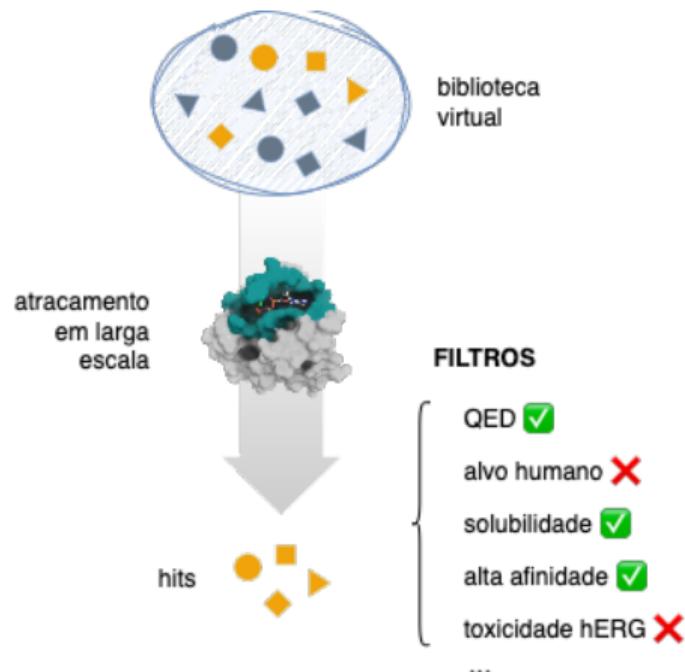


Figura: Scoring functions based on machine learning techniques (MLSFs).

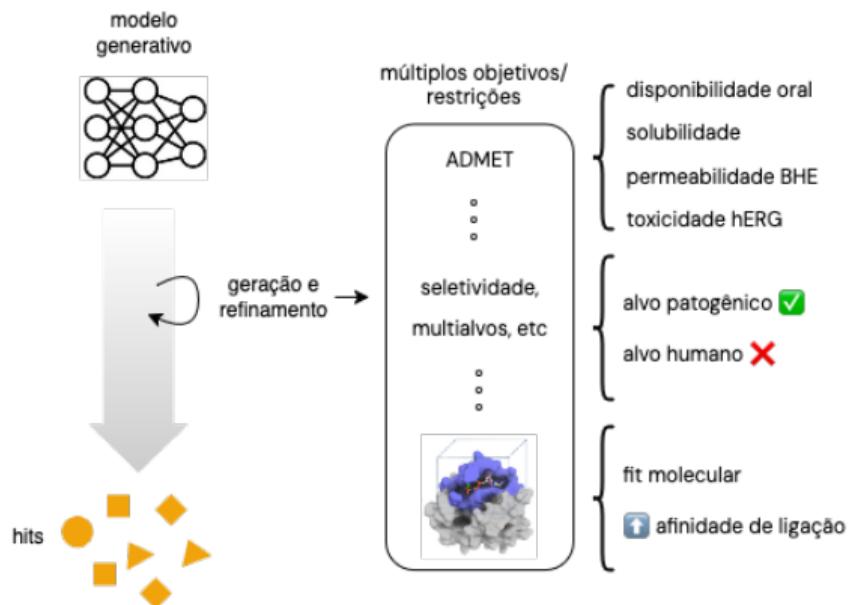
- Seek to reproduce experimental binding affinity values.
- Predictions based on a single pose of the receptor-ligand complex.
- Use datasets with structural and affinity information in their construction.

<sup>1</sup>machine learning (*machine learning*).

- Identify *hits* with high binding affinity from large virtual compound libraries.
- Fast and accurate scoring functions are essential.



- Generate molecules with desirable properties without the need for pre-defined structures and fixed databases.
- Essentially multi-objective problem.
- **Fast and accurate scoring functions are essential.**



# Thesis objectives

---

## General objective

Development of methodologies based on artificial intelligence (AI) for hit discovery, focusing on:

1. prediction of receptor-ligand affinity via 3D convolutional neural networks;
2. *de novo* generation of molecules using generative ML models and *many-objective* evolutionary algorithms.

# Theoretical Foundation

# Machine Learning

Machine learning consists of:

1. obtaining a representative dataset of the problem of interest;
2. constructing a statistical model based on the dataset (training).

**Supervised learning:**

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n,$$

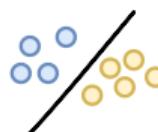
$$f^* = \arg \min_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim P} [\ell(f(x), y)].$$

**Unsupervised learning:**

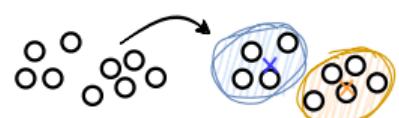
$$\mathcal{D}' = \{x_i\}_{i=1}^n,$$

$$h^* = \arg \min_{h \in \mathcal{H}} \mathbb{E}_{x \sim P} [\ell(h(x))].$$

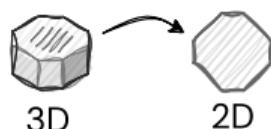
1. classificação (superv.)



2. clusterização (n. superv.)



3. redução dim. (n. superv.)



4. generativo (n. superv.)



Figura: Learning tasks.

# Deep Learning

---

Deep learning refers to ML models with multiple layers of transformations:

$$f_{\theta}(x) = f_{\theta_L}^{(L)} \circ f_{\theta_{L-1}}^{(L-1)} \circ \cdots \circ f_{\theta_1}^{(1)}(x).$$

Computation performed by a neuron:

$$a = g(w^T x + b),$$

where  $g(\cdot)$  is a non-linear activation function,  $w \in \mathbb{R}^m$  are the weights,  $b \in \mathbb{R}$  is the bias, and  $x \in \mathbb{R}^m$  is the input.

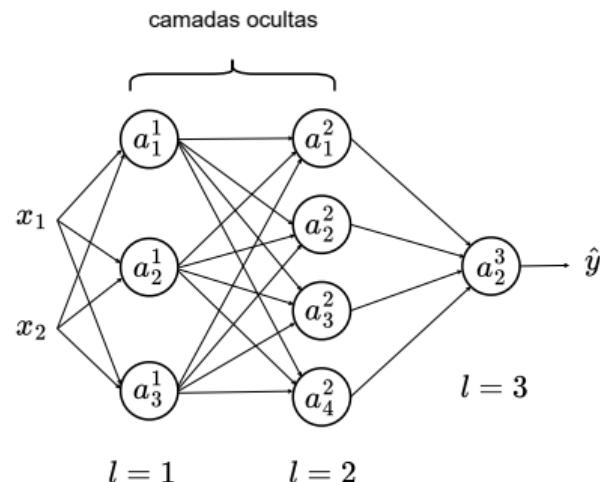


Figura: Architecture of an artificial neural network.

# Convolutional Neural Networks (CNNs)

- Local receptive fields.
- Weight sharing.
- Learning hierarchical representations.

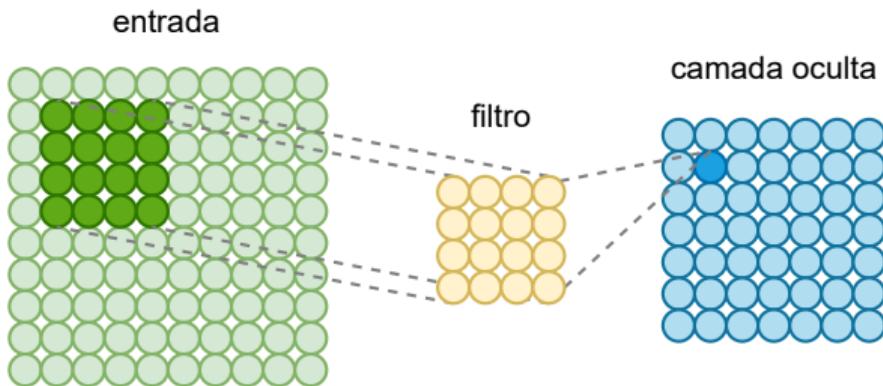
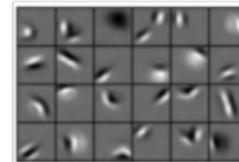
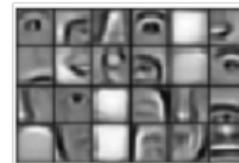


Figura: Convolution operation ( $X * F = Z$ ).

características baixo nível



características intermediárias



características alto nível



Figura: Hierarchical representations.

# Variational Autoencoders (VAEs)

- VAEs learn a **continuous** and **regular** latent space by modeling latent variables as probabilistic distributions.
- Latent space acquires “semantics”, favoring interpretation, manipulation, and robustness.

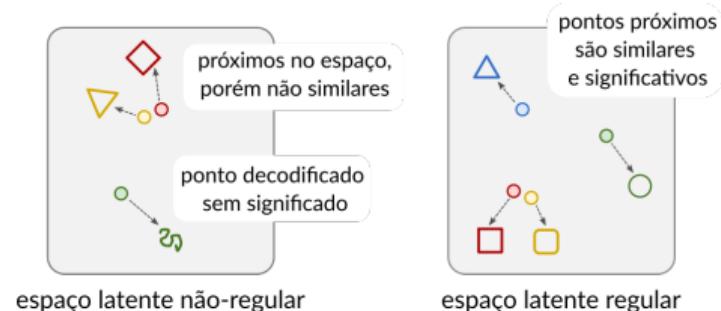
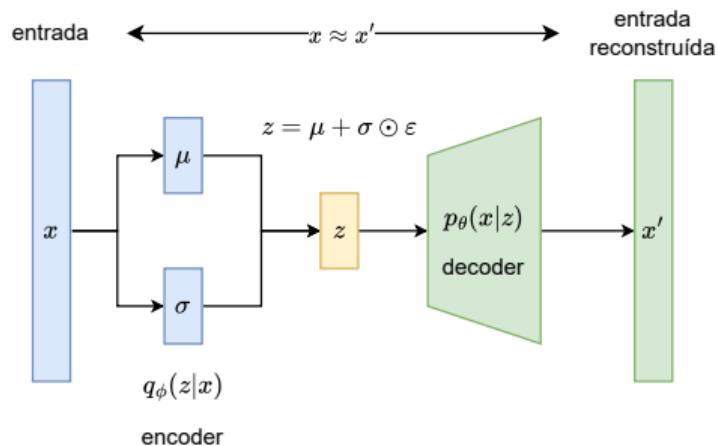


Figura: Properties of the latent space.

Figura: Variational Autoencoder (VAE).

# Multi-objective optimization

$$\begin{array}{ll}\text{minimize} & F(x) = (f_1(x), f_2(x), \dots, f_k(x))^T \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, 2, \dots, m \\ & h_j(x) = 0, \quad j = 1, 2, \dots, p\end{array}$$

$$f_i : \mathbb{R}^n \rightarrow \mathbb{R}$$

## Intuitive resolution

Convert to single-objective problem via weighted sum:

$$\min \sum_{i=1}^k w_i f_i(x), \quad \sum_{i=1}^k w_i = 1, \quad w_i \geq 0.$$

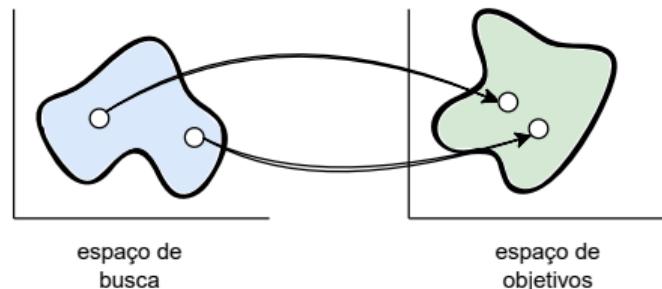


Figura: Different spaces in multi-objective optimization.

Objective space:

$F(x) = z = (z_1, z_2, \dots, z_k)^T$ , where  $z_i = f_i$ .

# Dominance and Pareto Front

---

We say that  $x_1$  dominates  $x_2$  if:

- $f_i(x_1) \leq f_i(x_2)$  for all  $i$  (i.e.,  $x_1$  is better or equal in all objectives)
- $f_j(x_1) < f_j(x_2)$  for at least one  $j$  (i.e.,  $x_1$  is strictly better in at least one objective)

In multi-objective optimization, we seek solutions as close as possible to the Pareto front (**convergence**) and as diverse as possible along it (**diversity**).

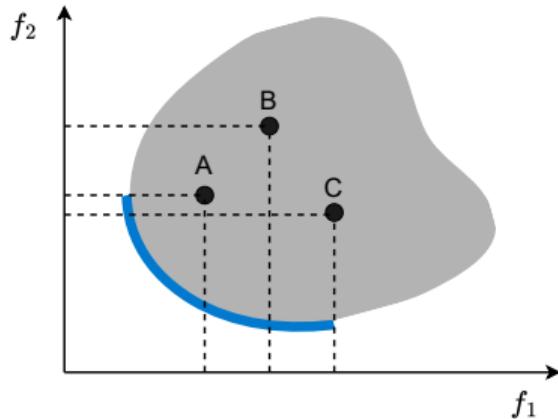


Figura: Pareto Front.

# Many-objective optimization

---

Problems with optimization with  $k > 3$  objectives are called *many-objective* and present additional challenges:

- Curse of dimensionality: the volume of the search space grows exponentially with  $k$ .
- Dominance resistance: most solutions tend to be non-dominated.
- Visualization of solutions: difficult to represent and interpret.

## State of the art

Specific techniques for multi-objective optimization maintain the objective functions separate. Examples include evolutionary algorithms, such as **NSGA-II** and **NSGA-III**.

# **Literature Review**

MLSFs<sup>a</sup> show superior performance to classical functions in various benchmarks.

<sup>a</sup>machine learning-based scoring functions (MLSFs).

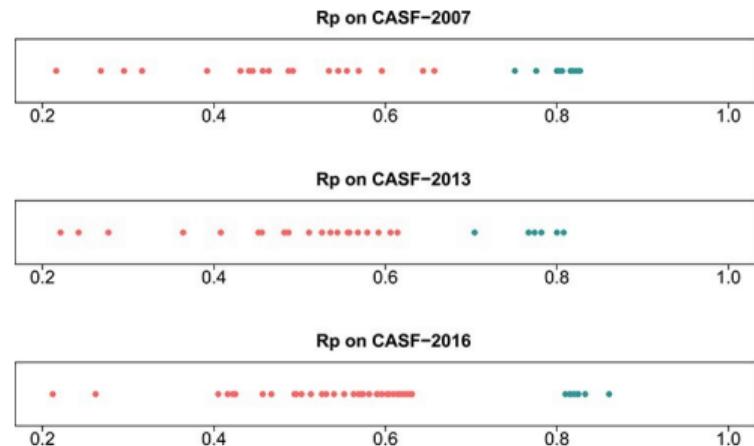


Figura: Comparative performance between classical functions (red) and MLSFs (green) [1].

# Literature Review: MLSFs

DockTDeep

- Doubts about the learning of intermolecular interactions and the real **generalization capacity** for new targets and ligands.
- “Out-of-domain” datasets have emerged as a more challenging evaluation.
- Strategies proposed to mitigate these biases: *docking*, *decoys*, and *crossdocking*.

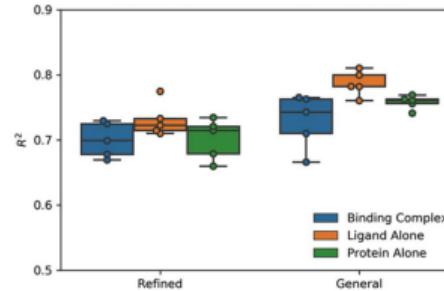


Figura: Ligand and protein bias [2].

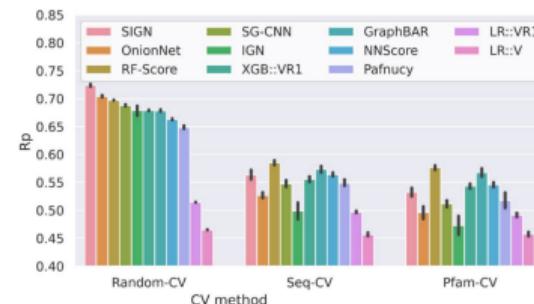


Figura: Performance on out-of-domain sets [3].

- CNNs are not **rotation invariant**.
- Two data augmentation approaches [4]:
  - 90° rotations (widely used);
  - random rotations in the complex.

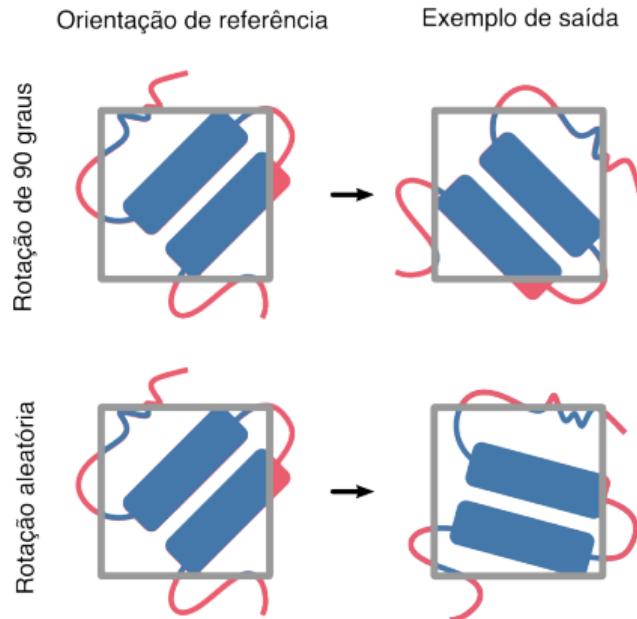


Figura: Different rotations applied to the protein-ligand complex.

# Literature Review: *de novo* design

DockTDesign

Three main approaches are employed for the *de novo* design of molecules using generative models:

- Distribution learning
- Conditional generation
- Objective-guided learning

## Current limitations

Methodologies that handle  $k \geq 4$  objectives are still little explored in the literature, with few studies considering more than three objectives, especially integrating appropriate techniques and generative models [5].

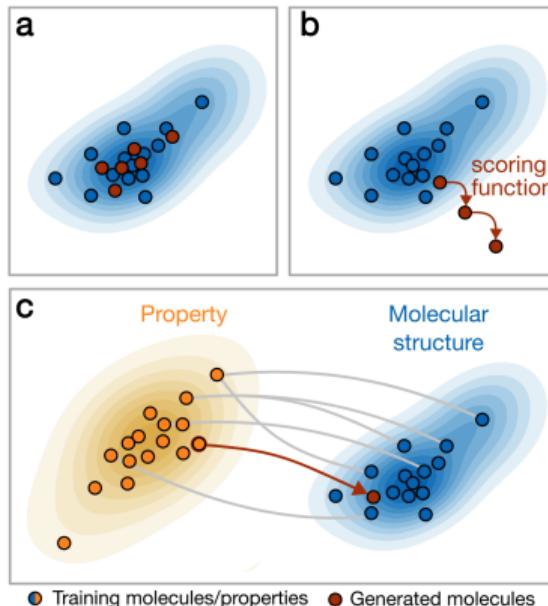


Figura: Generative approaches: (a) distribution learning, (b) objective-guided learning, and (c) conditional generation [6].

DockTDeep

# Dataset: training and validation

DockTDeep

## PDBbind v.2020

General set: **19.443** cpxs

Refined set: **5.316** cpxs

Coreset v.2013: **170** cpxs

Coreset v.2016: **261** cpxs

## Hyperparameter Search

Using the refined set (random split)

- Training: **3.599** cpxs
- Validation: **904** cpxs

## Activity ranges

The validation set data (**904** cpxs) were divided into activity ranges:

**Strong:**  $\Delta G_{\text{bind}} \leq -9.981 \text{ kcal/mol}$  (45 nM).

**Moderate:**  $-9.981 \text{ kcal/mol} < \Delta G_{\text{bind}} \leq -7.395 \text{ kcal/mol}$  (3.6  $\mu\text{M}$ ).

**Weak/inactive:**  $\Delta G_{\text{bind}} > -7.395 \text{ kcal/mol}$ .

Performance on test sets was evaluated using the *general set*, in different splits:

- Random (15.699 train/3.404 test).
- Temporal (13.317 train/1.786 test).
- Coreset v.2013 (18.933 train/170 test).
- Coreset v.2016 (18.842 train/261 test).
- Pfam-CV protocol (*general set*):
  - Evaluation of out-of-domain generalization.
  - Grouping by structural similarity of the binding site (Pfam database).
  - Repeated cross-validation 30x.
  - Proportion:  $\frac{2}{3}$  training and  $\frac{1}{3}$  test.

- Voxel grid  $24 \text{ \AA}^3$  and discretization  $1 \text{ \AA}$ .
- Channels for elements: C, H, O, N, S, and X (others).
- 21 channels: 6 protein, 6 ligand, 6 cpx, and 3 total volume.
- Representations generated with DockTGrid [7].
- GitHub: [github.com/gmmsb-lncc/docktgrid](https://github.com/gmmsb-lncc/docktgrid).

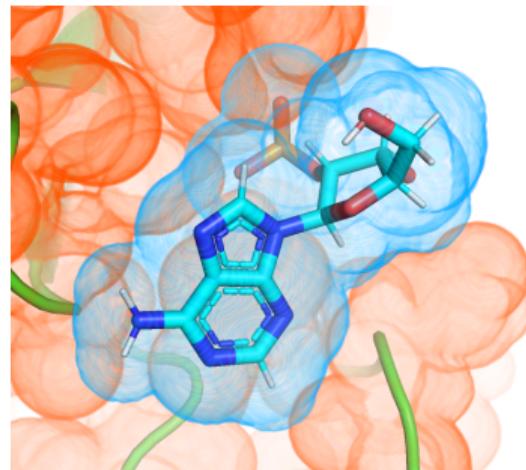


Figura: Illustrative example of the voxel representation used.

# Neural network architecture

DockTDeep

- 3D CNN with three convolutional layers.
- One dense layer of 1000 neurons.
- Linear output layer (regression).
- $\sim 2M$  trainable parameters.
- Learning rate:  $8,74 \times 10^{-4}$
- Training epochs: 1500.
- GitHub: [github.com/gmmsb-Incc/docktdeep](https://github.com/gmmsb-Incc/docktdeep) and preprint [4].

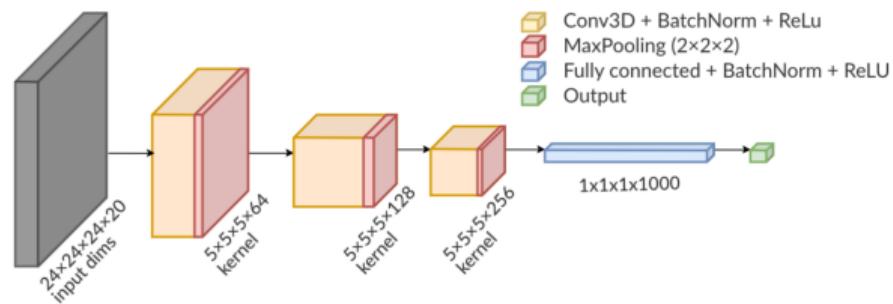


Figura: CNN network architecture.

# Data augmentation strategies

DockTDeep

Two data augmentation strategies were compared:

- 90° rotations;
- random rotations in the complex (applied at each new epoch).

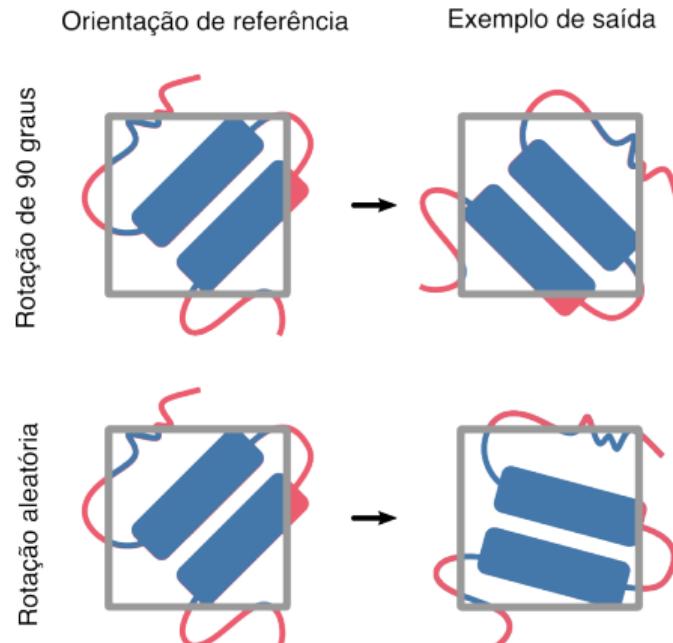
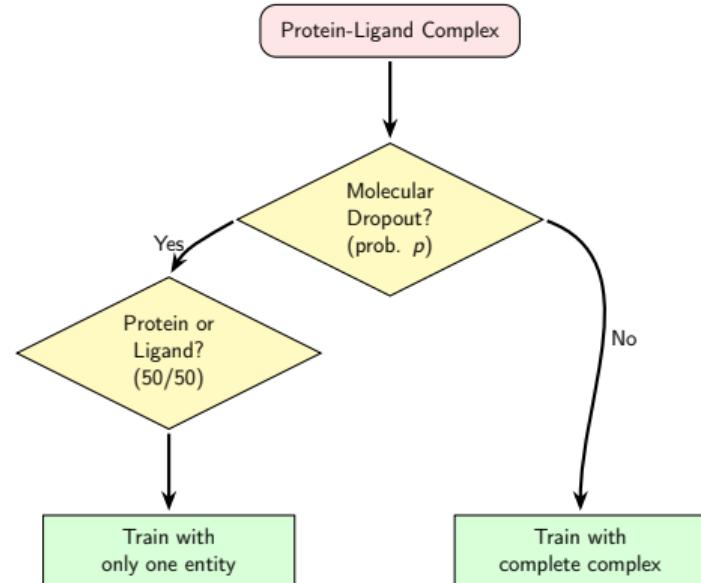


Figura: Different rotations applied to the protein-ligand complex.

# Molecular dropout

DockTDeep

- Regularization technique that avoids overfitting by randomly removing protein *or* ligand during training and setting affinity to zero.
- Each complex has chance  $p = 0,06$  of undergoing dropout per epoch.
- Seeks to force the model to learn relevant interactions, not depending on a single component.



# Results

# Results: rotational variance

DockTDeep

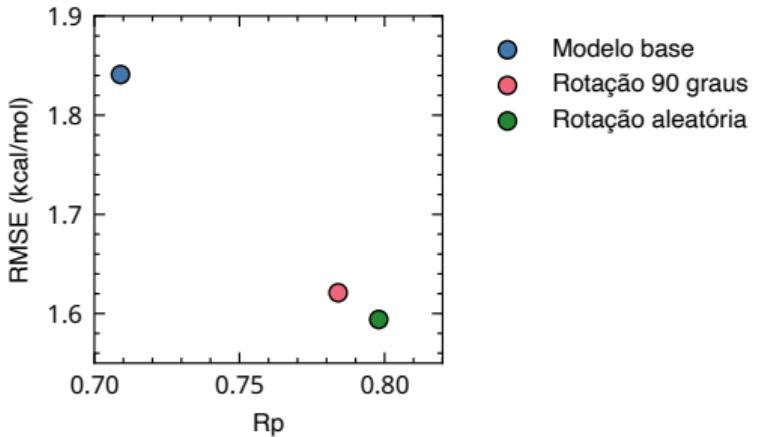


Figura: Comparison of predictive performance using RMSE (lower is better) and Rp (higher is better) metrics for each strategy.

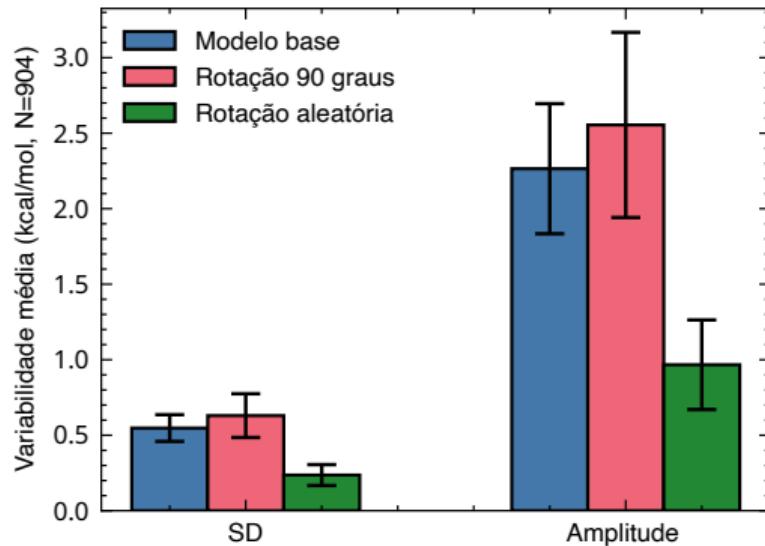
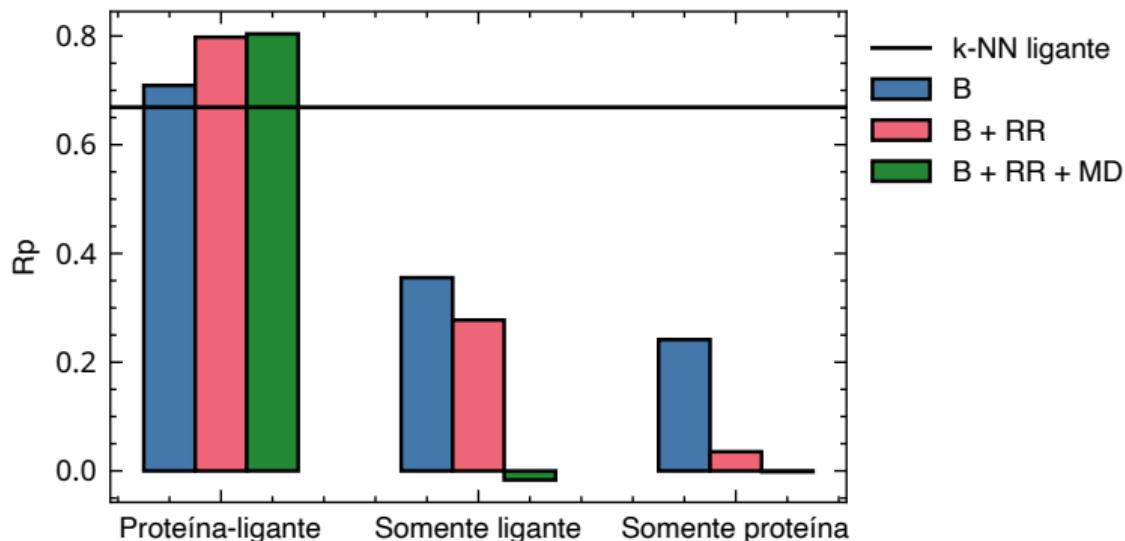


Figura: Mean standard deviation (SD) and range of predictions (kcal/mol) for the validation set (N=904).

# Results: protein/ligand bias

DockTDeep



**Figura:** Comparison of  $R_p$  values between three models: base model (B), using random rotation (B + RR), and using random rotation and molecular dropout (B + RR + MD), in addition to the k-NN model based only on the ligand. Evaluation was performed in three distinct scenarios: complete protein-ligand complex, ligand only, and protein only; keeping affinity labels unchanged.

# Results: protein/ligand bias

DockTDeep

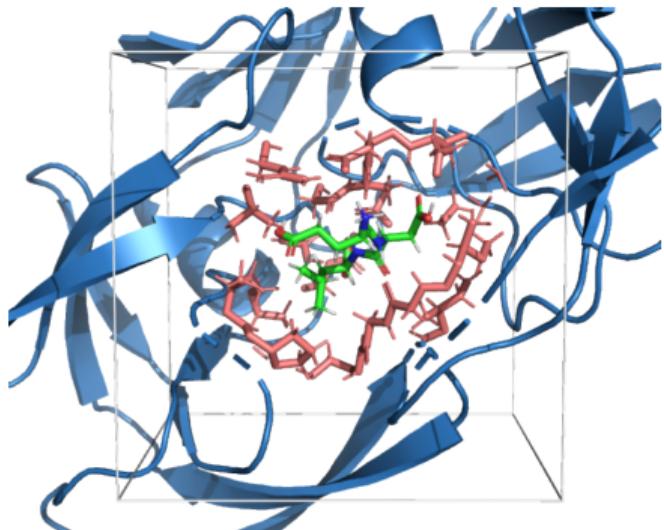


Figura: Visualization of removed protein atoms (highlighted in red), which are up to 5 Å from the ligand.

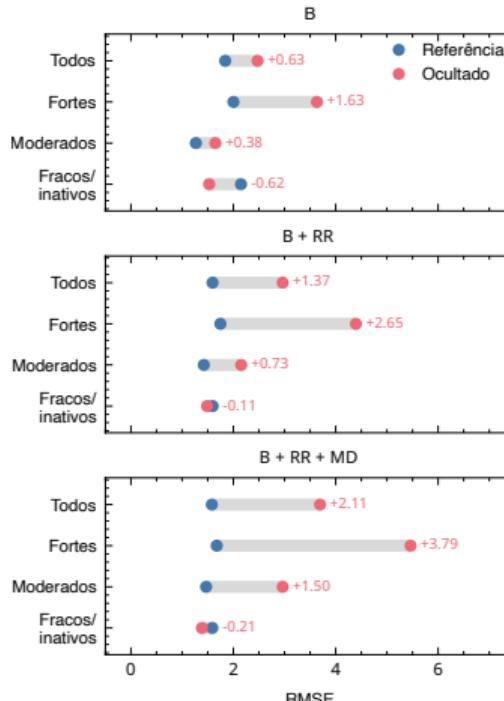
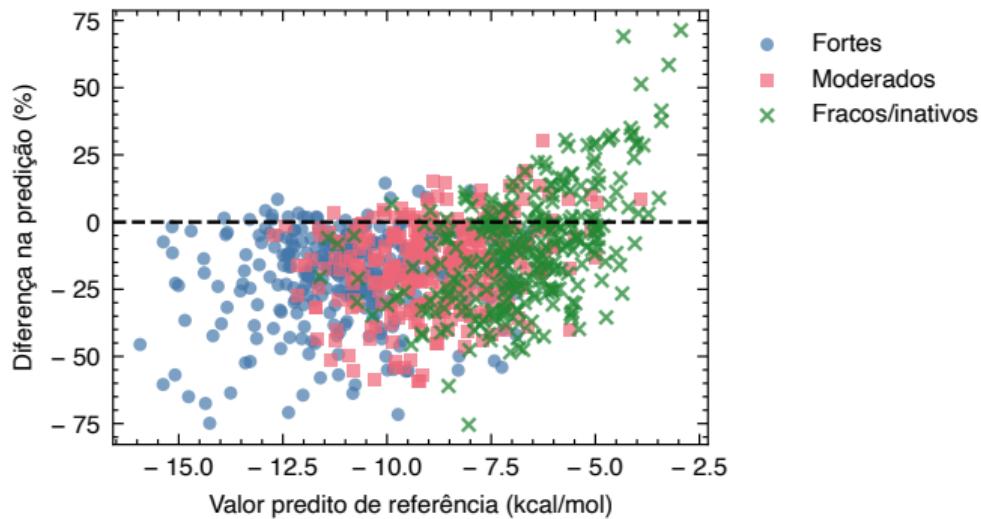


Figura: RMSE before (blue) and after atom removal (red).

# Results: learning of interactions

DockTDeep



**Figura:** Scatter plot comparing affinity predictions for crystallographic structures (reference) with the percentage variation in predictions for the worst poses obtained in the redocking experiment.

# Results: activity ranges

DockTDeep

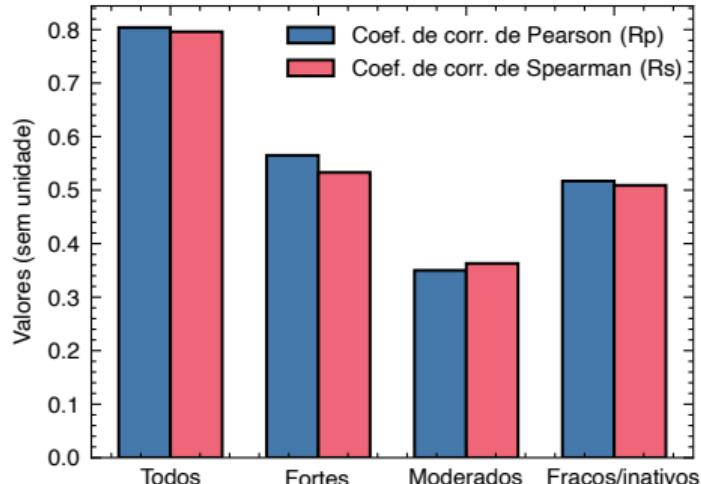


Figura: Comparison of  $R_p$  and  $R_s$  values for the entire validation set and affinity ranges.

		Predito		
		Fortes	Moderados	Fracos/inativos
Verdadeiro	Fortes	225 (74.0%)	66 (21.7%)	13 (4.3%)
	Moderados	61 (19.8%)	186 (60.4%)	61 (19.8%)
Fracos/inativos		8 (2.7%)	80 (27.5%)	203 (69.8%)

Figura: Confusion matrix for the three affinity classes: strong, moderate, and weak/inactive.

# Results: evaluation on external sets

DockTDeep

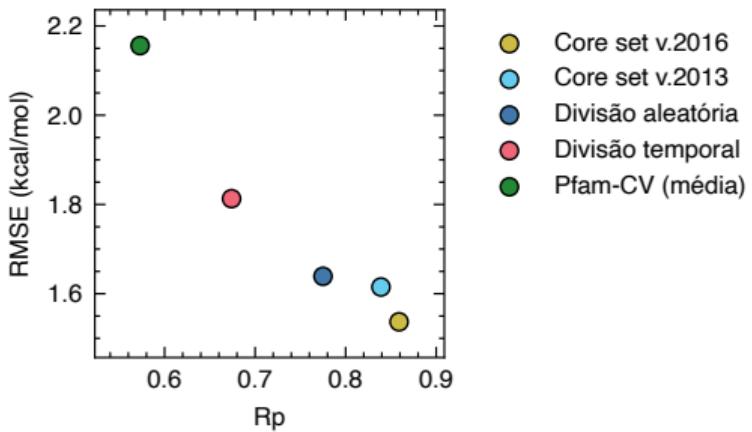


Figura: Comparison of RMSE (lower is better) and Rp (higher is better) values in all evaluated test sets.

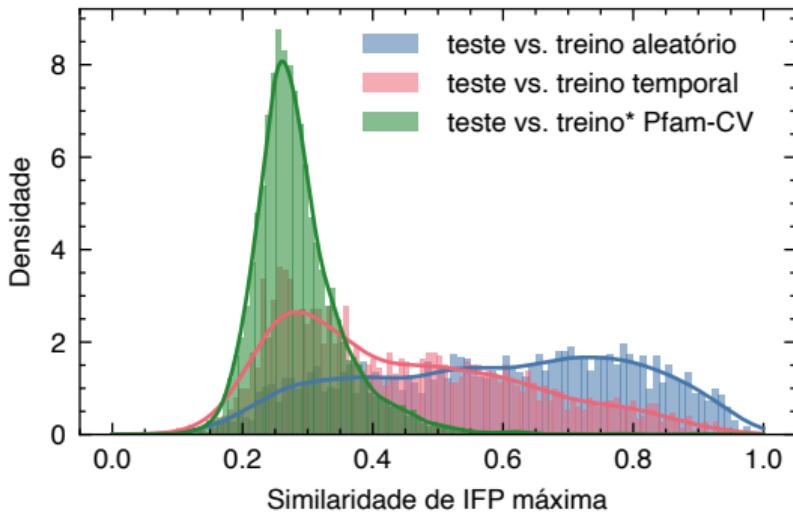


Figura: Density plot of maximum IFP similarity between sets.

# Results: PDBbind coresets Pfam-CV

DockTDeep

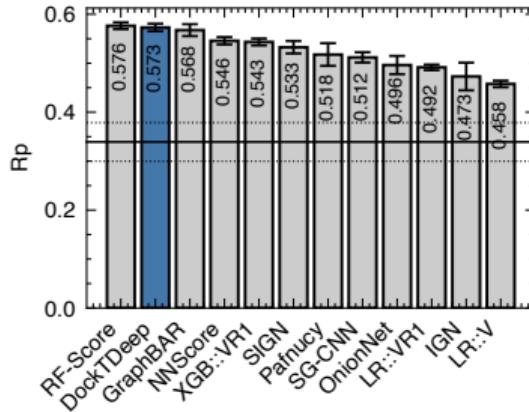
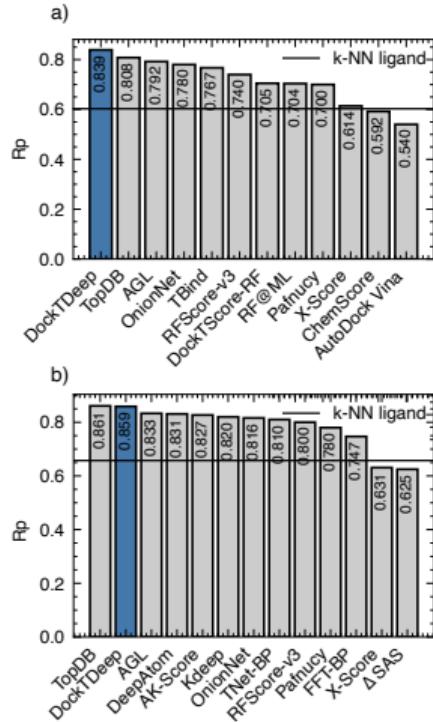


Figura: Rp in the Pfam-CV split scheme.

Figura: Rp in PDBbind v. (a) 2013 and (b) 2016.

- DockTDeep: simple 3D CNN model, trained in a systematic and grounded manner, shows robust generalization and is competitive with the state of the art.
- Random rotation surpasses 90° rotation as a data augmentation technique, improving stability and performance.
- Molecular dropout reduces protein/ligand bias, favoring learning of significant interactions.
- Model is more robust in **categorization** tasks (strong/moderate/weak) than in fine ranking.
- Performance drops in rigorous scenarios (Pfam-CV), highlighting limitations imposed by the diversity of training and test data.

# DockTDesign

# Generative chemistry model

DockTDesign

- HierVAE model based on graphs for generation of molecules with **100% validity** [8].
- Uses **structural motifs** as building blocks, extracted from recurring patterns in ChEMBL.
- Representation in **3 levels**: motifs, bonds between motifs, and atomic graph.
- Regular latent space ( $z \in \mathbb{R}^{32}$ ).
- Pre-trained on **1.8M molecules** from ChEMBL.

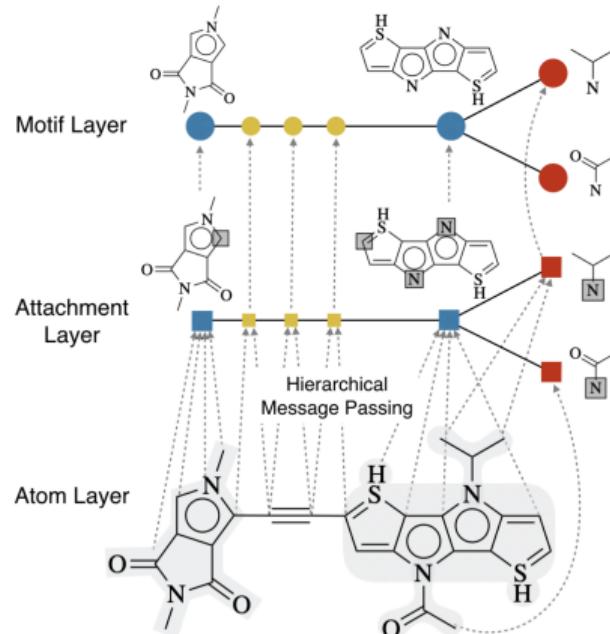


Figura: Hierarchical encoder of the HierVAE model [8].

## Compared algorithms:

- NSGA-II (multi-objective).
- NSGA-III (*many*-objective).
- Single-objective GA with aggregation.

## General execution:

- 100 generations.
- Population: 800 individuals.
- 11 independent runs per configuration  
(only 1 run when using molecular docking).

## Genetic operators:

- 2-point crossover,  $p_c = 0,9$ .
- Gaussian mutation,  $p_m = 0,1$ ,  
 $\sigma = 0,01$ .
- Selection: Binary tournament.

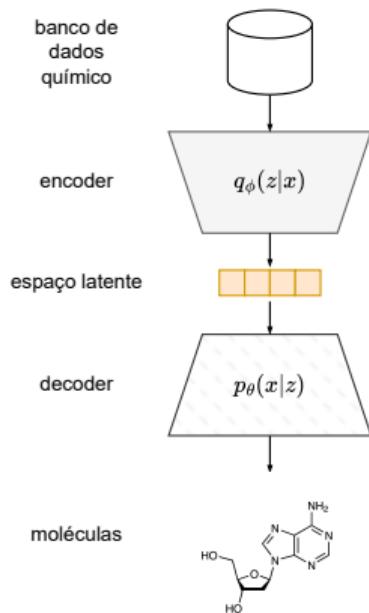
## Specific configurations:

- NSGA-III: reference directions via Das-Dennis method (num. equal to population size).
- Single-objective GA: 5 weight combinations (0.6, 0.1, 0.1, 0.1, 0.1) to (0.1, 0.1, 0.1, 0.1, 0.6).

# Proposed generative platform

DockTDesign

a)



b)

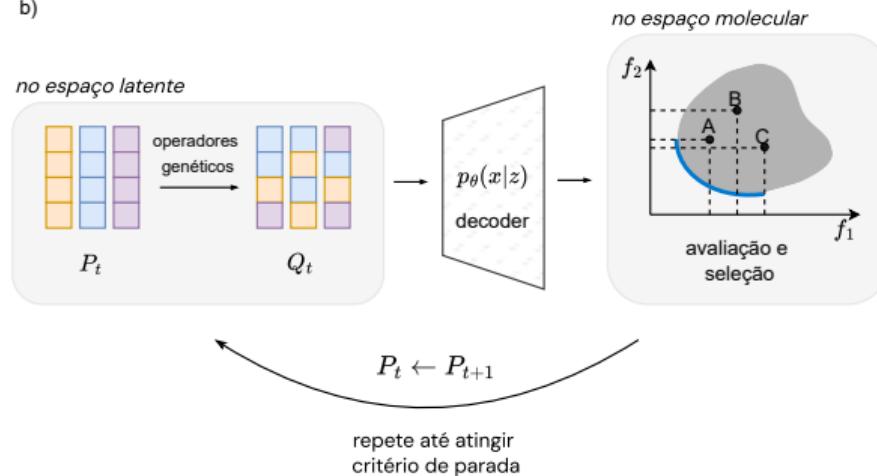


Figura: DockTDesign generative platform.

# Pharmacological targets

DockTDesign

A **target** and an **anti-target** (*off-target*) were selected as a case study for generating selective inhibitors:

- LpxC<sup>a</sup> (**target** antimicrobial involved in LPS synthesis).
- MMP-9<sup>b</sup> (**anti-target** human important for extracellular matrix degradation).

The choice of targets was made based on the identification of a potent LpxC inhibitor that also inhibits MMP-9 (BDBM50478376).

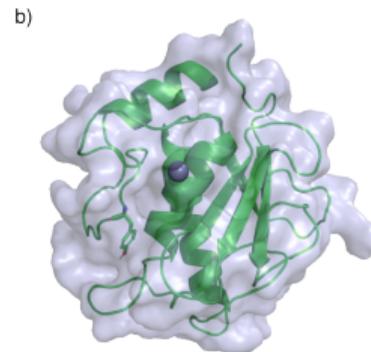
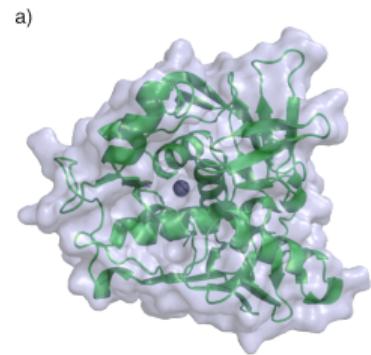


Figura: Receptors (a) LpxC and (b) MMP-9.

<sup>a</sup>UDP-3-O-acyl-N-acetylglucosamine deacetylase

<sup>b</sup>matrix metalloproteinase 9

## Drug-likeness (QED)

Maximize similarity to the profile of approved drugs.

$$f_{\text{QED}}(m) = 1 - \text{QED}(m)$$

QED values  $\in [0,1]$ .

## Synthetic Accessibility (SA)

Penalize molecules that are difficult to synthesize.

$$f_{\text{SA}}(m) = (\text{SA}(m) - 1)^2$$

SA(m)  $\in [1,10]$ .

## Molecular Weight (MW)

Optimize desired target value.

$$f_{\text{MW}}(m, v) = (\text{MW}(m) - v)^2$$

MW(m): mass in Da; v: target value.

## Molecular Complexity (Cx)

Minimize structural complexity.

$$f_{\text{Cx}}(m) = \text{Cx}(m)^2$$

Cx  $\in [0, \infty]$ .

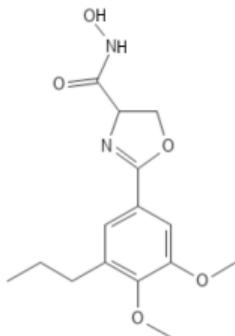
# Objectives: Tanimoto similarity

DockTDesign

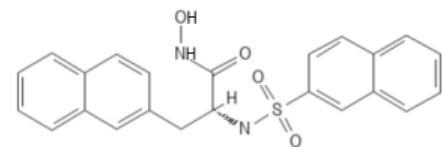
1. Tanimoto similarity against molecule BDBM50074960 ( $K_i = 0.053 \text{ nM LpxC}$ ):

$$f_{\text{sim}}(m, m_{\text{ref}}) \stackrel{\text{def}}{=} 1 - \text{TS}(m, m_{\text{ref}}).$$

a)



b)



2. Tanimoto dissimilarity against molecule BDBM50478376 ( $K_i = 0.069 \text{ nM LpxC}$ ;  $\text{IC}_{50} = 97 \text{ nM MMP-9}$ ):

$$f_{\text{dissim}}(m, m'_{\text{ref}}) \stackrel{\text{def}}{=} \text{TS}(m, m'_{\text{ref}}),$$

Figura: Compound (a) BDBM50074960 and (b) BDBM50478376

Integration DockTDesign (generative), DockThor (pose) and DockTDeep (binding affinity):

$$f_{\text{docking}}(m, p, v) = (v - \hat{y}(m, p))^2,$$

where  $m$ : molecule,  $p$ : protein target,  $v$ : target value,  $\hat{y}$ : affinity prediction (DockTDeep).

- **Target:**  $v = -25$  kcal/mol (maximize activity).
- **Anti-target:**  $v = 0$  kcal/mol (minimize activity).

# **Resultados**

## Experimental design

**Algorithms:** Single-objective GA, NSGA-II and NSGA-III.

### Objectives:

1. maximize QED;
2. minimize SA;
3. minimize complexity;
4. maximize similarity against molecule BDBM50074960 (active LpxC);
5. minimize similarity against molecule BDBM50478376 (active LpxC, active MMP-9).

Work resulted in a publication in an international conference [9].

# Results: Tanimoto similarity

DockTDesign

Tabela: Multi-objective performance indicators.

Algo.	HV <sup>1</sup>	IGD <sup>2</sup>
Single-obj. GA	0.783	0.073
NSGA-II	0.805	0.040
NSGA-III	<b>0.880</b>	<b>0.031</b>

<sup>a</sup>Hypervolume.

<sup>b</sup>Inverted generational distance.

Tabela: Generative chemistry metrics.

Algo.	Valid. <sup>3</sup>	Unic. <sup>4</sup>	DivInt <sup>5</sup>	Nov. <sup>6</sup>
NSGA-II	<b>1.00±0.0</b>	<b>1.00±0.0</b>	<b>0.70±0.01</b>	0.87±0.01
NSGA-III	<b>1.00±0.0</b>	0.75±0.03	0.65±0.01	<b>0.93±0.01</b>

<sup>c</sup>Validity.

<sup>d</sup>Uniqueness@FP.

<sup>e</sup>Internal diversity.

<sup>f</sup>Novelty.

# Results: Tanimoto similarity

DockTDesign

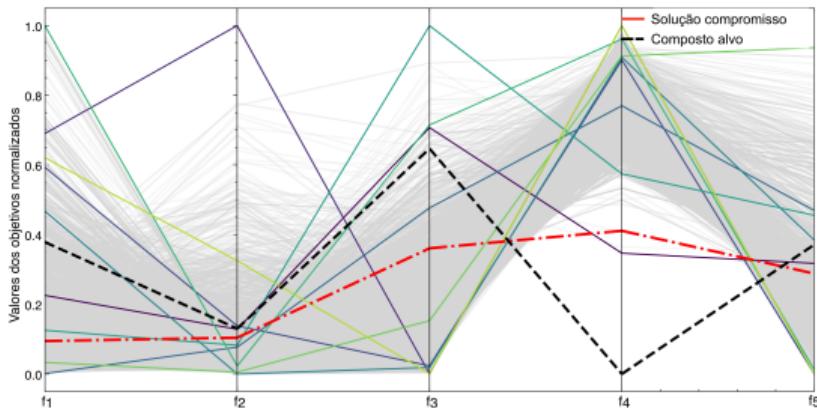


Figura: Non-dominated solutions generated by the NSGA-III algorithm. Order:  
 $f_{\text{QED}}$ ,  $f_{\text{SA}}$ ,  $f_{\text{Cx}}$ ,  $f_{\text{sim}}$ ,  $f_{\text{dissim}}$ .

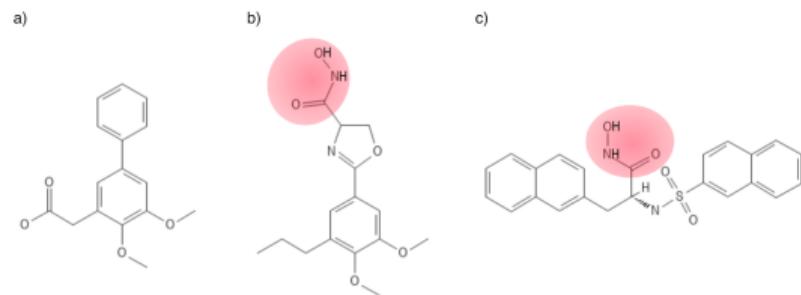


Figura: (a) compromise solution, (b) similarity objective and (c) dissimilarity objective.

# Results: molecular docking

DockTDesign

## Protocol I

### Objectives:

1. maximize QED;
2. minimize SA;
3. minimize complexity;
4. minimize binding affinity for LpxC;
5. maximize binding affinity for MMP-9.

## Protocol II

### Objectives:

1. maximize QED;
2. minimize binding affinity for LpxC;
3. maximize binding affinity for MMP-9.

### Constraints:

1. SA less than or equal to 4;
2. complexity between 150 and 600;
3. molecular weight above 150 Da;
4. QED above 0.5.

# Results: molecular docking (protocol I)

DockTDesign

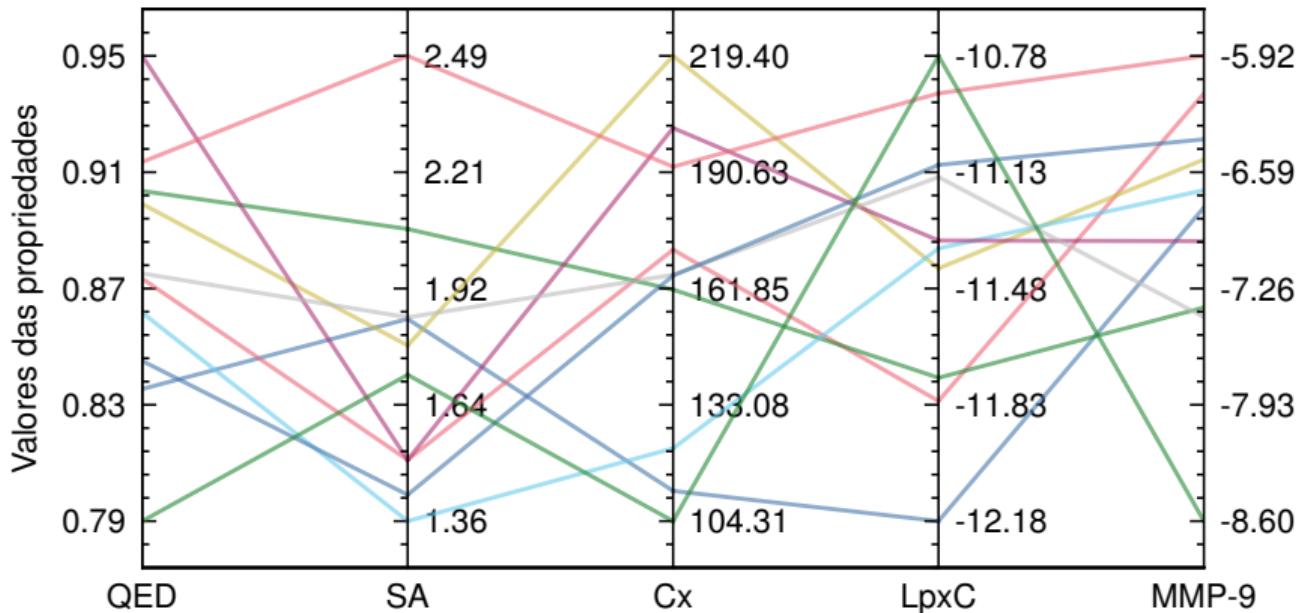


Figura: Parallel coordinates plot for the solutions obtained with protocol I.

# Results: molecular docking (protocol II)

DockTDesign

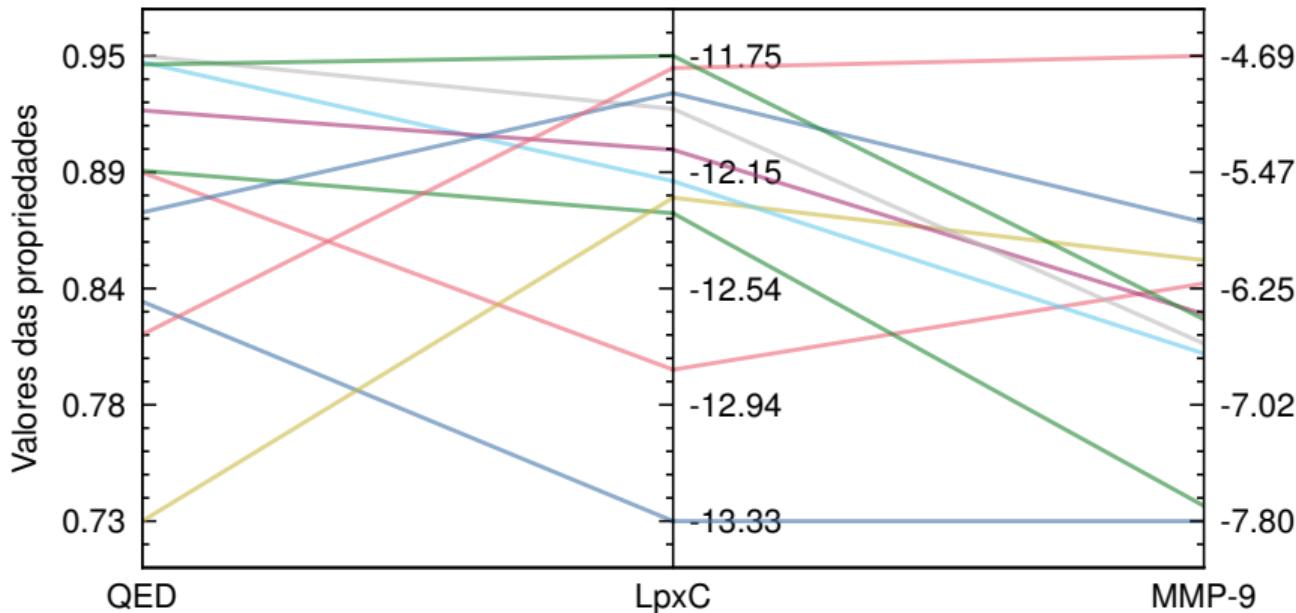


Figura: Parallel coordinates plot for the solutions obtained with protocol II.

# Results: molecular docking

DockTDesign

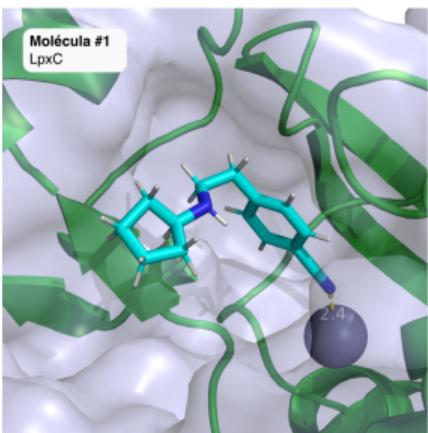
Tabela: Uniqueness metrics on the Pareto front, internal diversity (DivInt) and novelty considering the molecules generated by experimental protocols I and II.

Protocol	Uniqueness@FP	DivInt	Novelty
I	<b>0,920</b>	<b>0,745</b>	0,653
II	0,825	0,451	<b>0,915</b>

# Results: molecular docking (protocol I)

DockTDesign

a)



b)

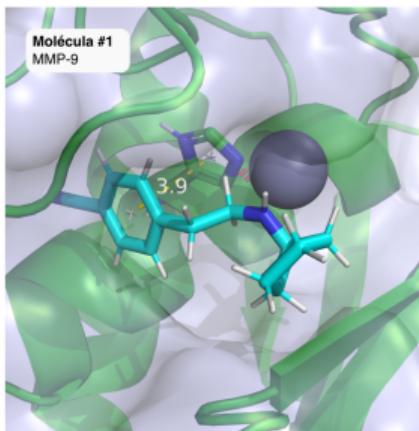
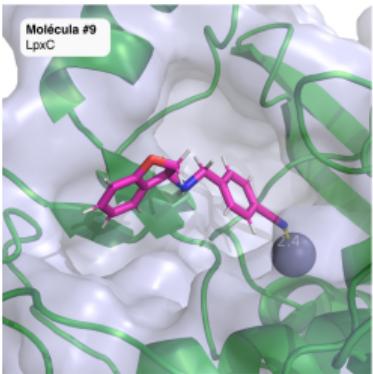


Figura: Compound 1 from protocol I in the active sites of (a) LpxC and (b) MMP-9. LpxC: -12,176 kcal/mol; MMP-9: -6,795 kcal/mol; QED: 0,836; SA: 1,85; Cx: 111,781.

# Results: molecular docking (protocol I)

DockTDesign

a)



b)

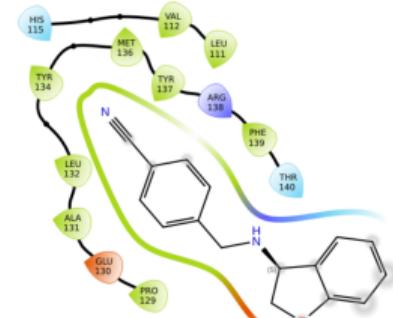
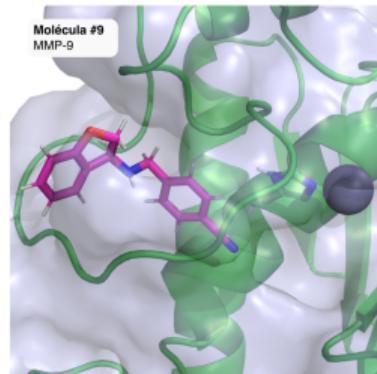


Figura: Compound 9 from protocol I in the active sites of (a) LpxC and (b) MMP-9. LpxC: -10,897 kcal/mol; MMP-9: -5,924 kcal/mol; QED: 0,911; SA: 2,488; Cx: 191,971.

# Results: molecular docking (protocol II)

DockTDesign

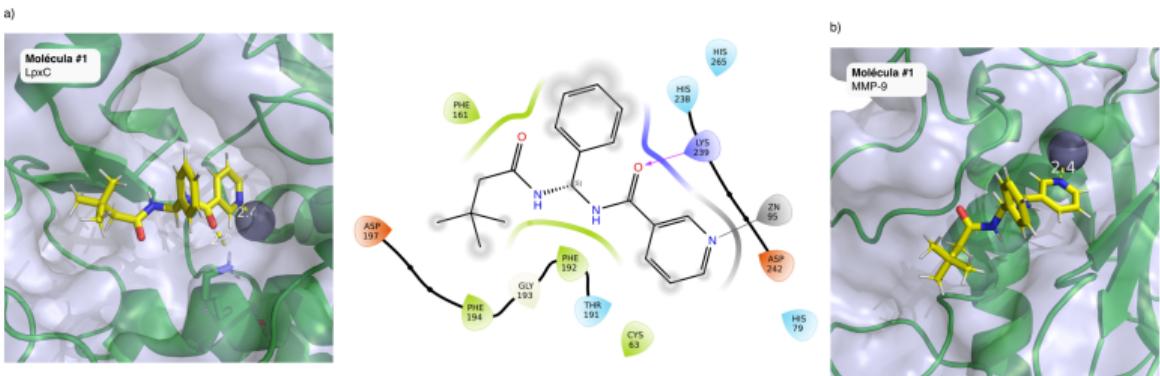
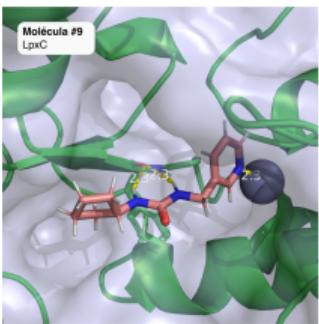


Figura: Compound 1 from protocol II in the active sites of (a) LpxC and (b) MMP-9. LpxC: -13,331 kcal/mol; MMP-9: -7,801 kcal/mol; QED: 0,83; SA: 2,66; Cx: 247,997.

# Results: molecular docking (protocol II)

DockTDesign

a)



b)

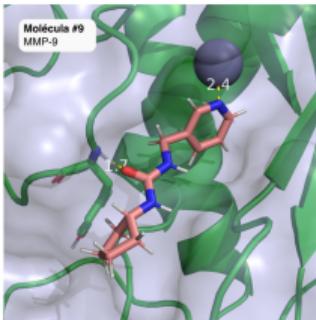


Figura: Compound 9 from protocol II in the active sites of (a) LpxC and (b) MMP-9. LpxC: -11,794 kcal/mol; MMP-9: -4,694 kcal/mol; QED: 0,814; SA: 1,813; Cx: 178,732.

- Integration of **HierVAE** with *many*-objective evolutionary algorithms presents itself as an effective and flexible approach.
- NSGA-III outperforms NSGA-II and GA with aggregation in convergence and coverage of the Pareto front.
- NSGA-II generates molecules more diverse among themselves; NSGA-III explores innovative regions of chemical space. Adjustments in hyperparameters can improve both.
- Predicted affinity and selectivity were optimized without dependence on structural similarity.
- Treating SA and Cx as **constraints** (not objectives) improves predicted activity and selectivity.

## General conclusions

---

- **DockTDeep** is a scoring function that is easy to train, robust to ligand/protein biases and rotational variance, and competitive with the state of the art in various evaluation scenarios.
- **DockTDesign**, together with **DockThor** and **DockTDeep**, presents itself as a promising and flexible platform for hit identification, capable of suggesting to the specialist a set of molecules that simultaneously meet multiple objectives.

## Perspectives

---

- Incorporate **DockTDeep** into the DockThor portal (<https://dockthor.lncc.br>).
- Include multiple conformations (docking/molecular dynamics) to model **dynamic** aspects of binding.
- Evaluate **DockTDeep** in *lead* optimization tasks with experimental data and compare with FEP.
- Use **metamodels** to reduce computational cost of docking in optimization.
- Expand the approach to other therapeutic targets (including the multi-target scenario).
- Collaborate with experimental groups to define realistic objectives and validate generated molecules.
- Develop a **web portal** for the *DockTDesign* platform, integrated with the Santos Dumont supercomputer.

# Referências I

---



Hongjian Li, Gang Lu, Kam-Heung Sze, Xianwei Su, Wai-Yee Chan, and Kwong-Sak Leung.

Machine-learning scoring functions trained on complexes dissimilar to the test set already outperform classical counterparts on a blind benchmark.

*Briefings in bioinformatics*, 22(6):bbab225, 2021.



Jincai Yang, Cheng Shen, and Niu Huang.

Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets.

*Frontiers in pharmacology*, 11:69, 2020.



Hui Zhu, Jincai Yang, and Niu Huang.

Assessment of the generalization abilities of machine-learning scoring functions for structure-based virtual screening.

*Journal of chemical information and modeling*, 62(22):5485–5502, 2022.

## Referências II

---

-  Matheus da Silva, Lincon Vidal, Isabella Guedes, Camila de Magalhães, Fábio Custódio, and Laurent Dardenne.  
Data-centric training enables meaningful interaction learning in protein–ligand binding affinity prediction.  
2025.
-  Jaqueline S Angelo, Isabella A Guedes, Helio JC Barbosa, and Laurent E Dardenne.  
Multi-and many-objective optimization: present and future in de novo drug design.  
*Frontiers in Chemistry*, 11:1288626, 2023.
-  Rıza Özçelik, Helena Brinkmann, Emanuele Criscuolo, and Francesca Grisoni.  
Generative deep learning for de novo drug design—a chemical space odyssey.  
*Journal of Chemical Information and Modeling*, 65(14):7352–7372, 2025.
-  Matheus Müller Pereira da Silva, Isabella Alvim Guedes, and Fábio Lima.  
Docktgrid: A python package for generating deep learning-ready voxel grids of molecular complexes.

## Referências III

---

-  Wengong Jin, Regina Barzilay, and Tommi Jaakkola.  
Hierarchical generation of molecular graphs using structural motifs.  
In *International conference on machine learning*, pages 4839–4848. PMLR, 2020.
-  Matheus Muller Pereira Da Silva, Jaqueline Silva Angelo, Isabella Alvim Guedes, and Laurent Emmanuel Dardenne.  
A generative evolutionary many-objective framework: A case study in antimicrobial agent design.  
In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1623–1630, 2024.
-  Madura KP Jayatunga, Margaret Ayers, Lotte Bruens, Dhruv Jayanth, and Christoph Meier.  
How successful are ai-discovered drugs in clinical trials? a first analysis and emerging lessons.  
*Drug discovery today*, 29(6):104009, 2024.

# Thank you!



# NSGA-II

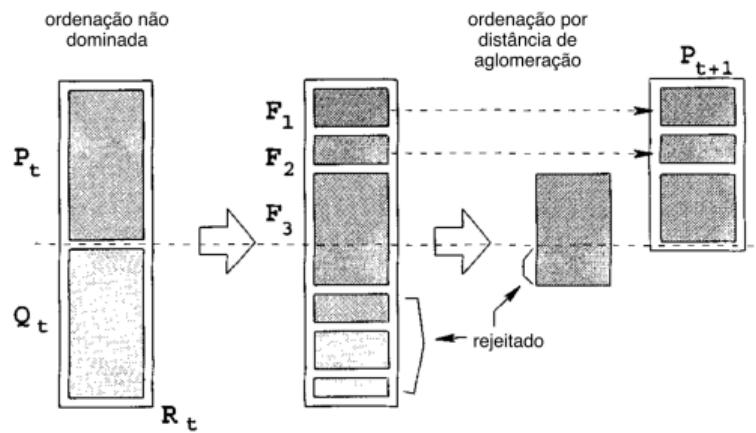


Figura: NSGA-II algorithm.

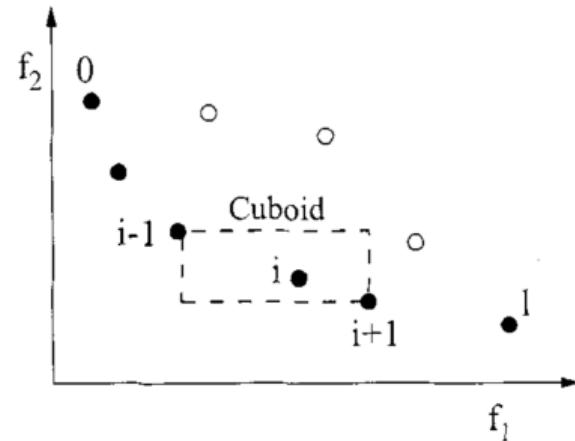


Figura: Crowding distance in NSGA-II.

## NSGA-III: reference directions

---

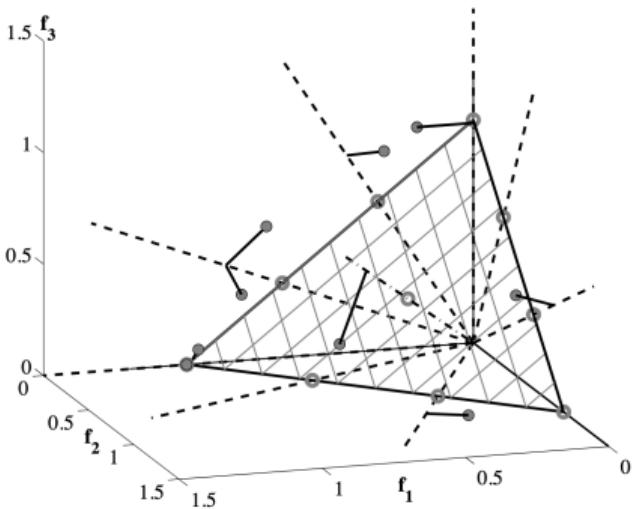


Figura: Reference directions in NSGA-III.

# *de novo* Drug Design | DockTDesign

## Protocol I (NSGA-III) – Multi Target Scenario

Objectives :

- 1) QED (drug-likeness) ↑
- 2) SA (synthetic accessibility). ↓
- 3) LE (ligand efficiency) for LpxC ↑
- 4) LE (ligand efficiency) for GshA ↑
- 5) Affinity Prediction LpxC (DockThor + DockTDeep) ↑
- 6) Affinity Prediction gshA (DockThor + DockTDeep) ↑

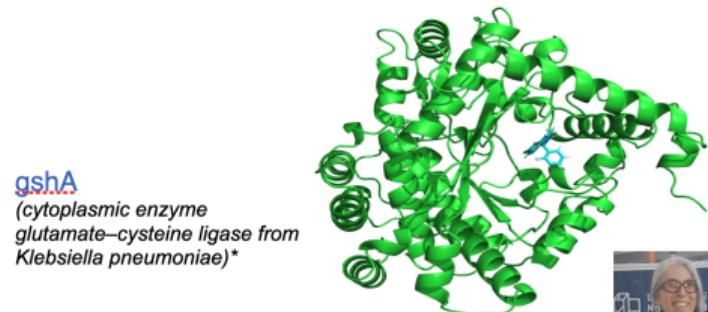
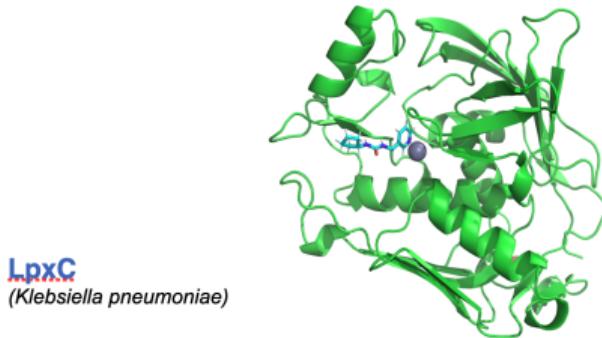
## Protocol II (NSGA-III) – Multi Target Scenario

Objectives :

- 1) QED (drug-likeness) ↑
- 2) Affinity Prediction LpxC (DockThor + DockTDeep) ↑
- 3) Affinity Prediction gshA (DockThor + DockTDeep) ↑

Restrictions :

- 1) SA (synthetic accessibility) ≤ 6
- 2) LE (ligand efficiency) for LpxC ≥ 0.30 kcal·mol<sup>-1</sup>·HA<sup>-1</sup>
- 3) LE (ligand efficiency) for GshA ≥ 0.30 kcal·mol<sup>-1</sup>·HA<sup>-1</sup>

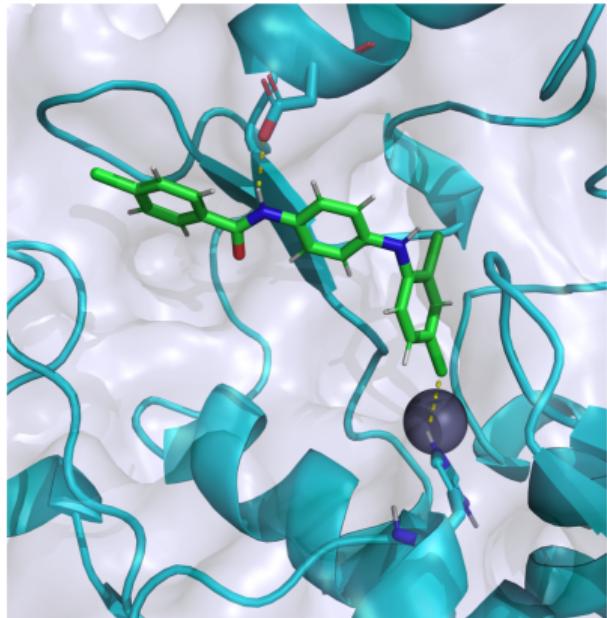


\* An integrative, multi-omics approach towards the prioritization of *Klebsiella pneumoniae* drug targets  
Scientific Reports 2018 Jul 17;8(1):10755. doi: 10.1038/s41598-018-28916-7.

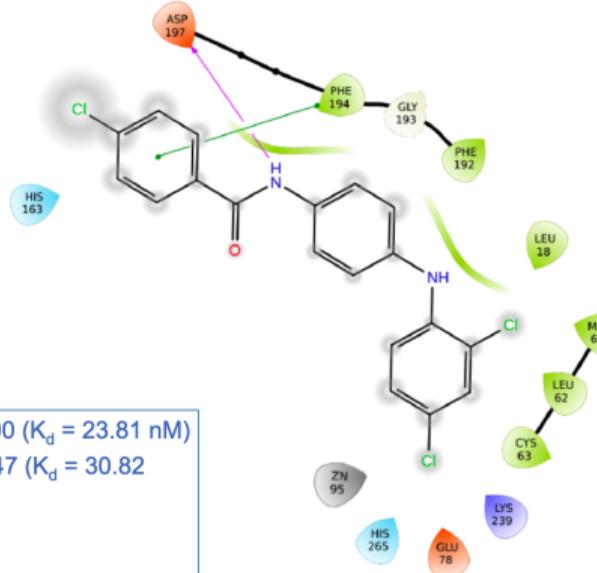


Marisa Nicolás  
Bioinformatics

## Protocol 1 – LpxC - molecule 4 – $\Delta G = -10.400$ kcal/mol LE = 0.416

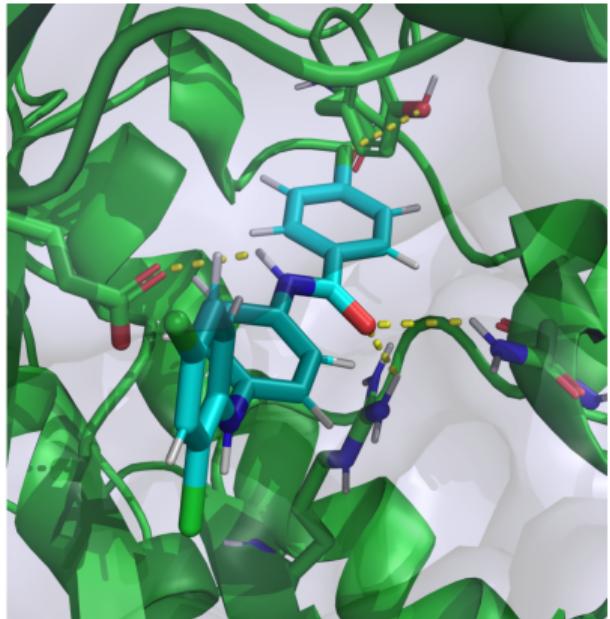


LpxC ( $\Delta G$ ) = -10.400 ( $K_d = 23.81$  nM)  
GshA ( $\Delta G$ ) = -10.247 ( $K_d = 30.82$  nM)  
QED = 0.523  
SA = 1.661  
LE (LpxC) = 0.416  
LE (GshA)) = 0.410



- Charged (negative)
- Charged (positive)
- Glycine
- Hydrophobic
- Metal
- Polar
- Unspecified residue
- Water
- Hydration site
- Hydration site (displaced)
- Distance
- ↔ H-bond
- Halogen bond
- Metal coordination
- ✖ Hydration site (displaced)
- Pi-Pi stacking

## Protocol 1 – GshA - molecule 4 – $\Delta G = -10.247$ kcal/mol LE = 0.410



gshA

