DockTDesign

Deep Generative Models for *de novo* Drug Design

Matheus M. P. da Silva

PhD in Computational Modeling
Laboratório Nacional de Computação Científica (LNCC/MCTI)

matheusp@posgrad.lncc.br github.com/mpds

November 6, 2025

Summary

Motivation

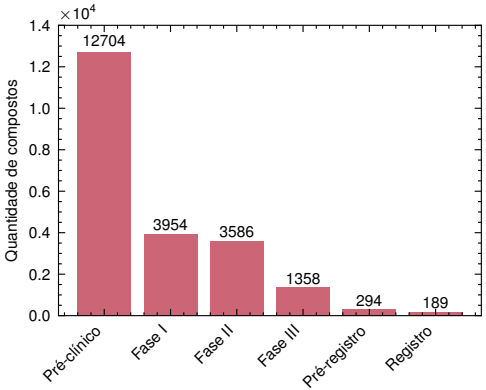


Figure: Number of compounds per development phase in 2025. Source: Pharmaprojects.

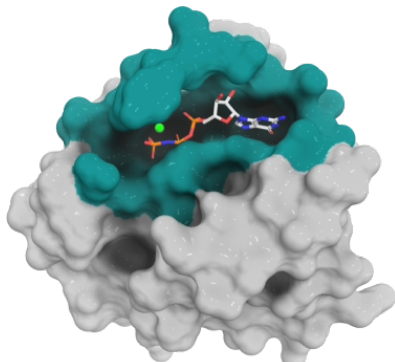
- Drug development: long, costly, and with low success rates ($\leq 5\%$ in the preclinical phase).
- Increasing the success rate in early phases can have a large economic and public-health impact.
- Computational methods are essential for the early identification of promising molecules (*hits*).

Objectives

- Pose prediction
- Binding affinity prediction

Components

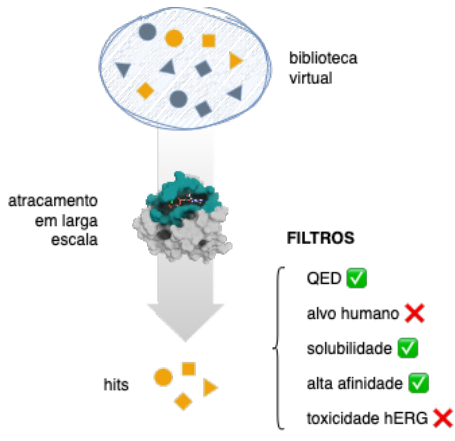
- Search algorithm
- Scoring function:
  - pose
  - binding affinity



Hit identification

Virtual screening

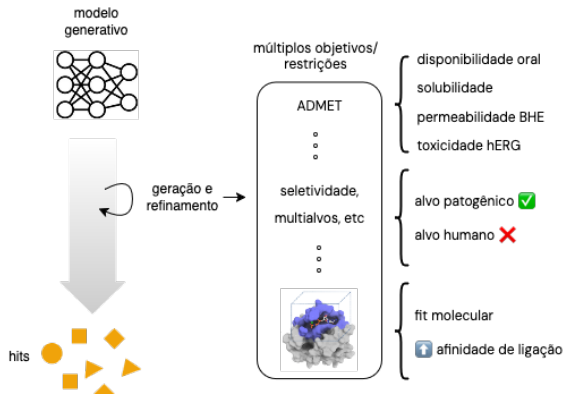
- Identify *hits* with high binding affinity from large virtual compound libraries.



Hit identification

De novo molecule design

- Generate molecules with desirable properties without the need for pre-defined structures and fixed databases.
- Essentially multi-objective problem.



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:

$$\mathcal{D}' = \{x_i\}_{i=1}^{n},$$
$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{x\sim P}[\ell(h(x))].$$

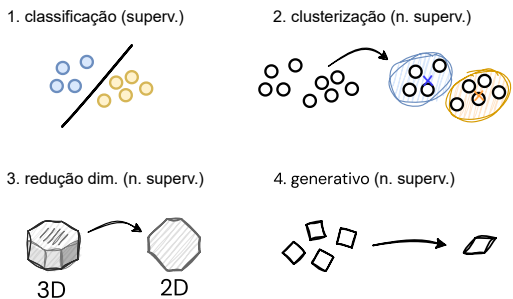


Figure: Learning tasks.

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.

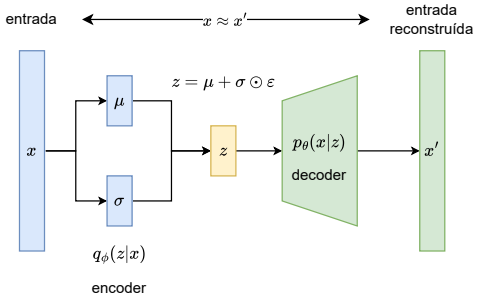


Figure: Variational Autoencoder (VAE).

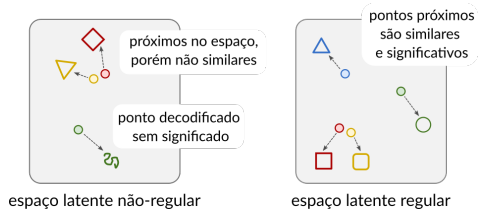


Figure: Properties of the latent space.

Multi-objective optimization

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

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

Intuitive resolution

Convert to single-objective problem via weighted sum:

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

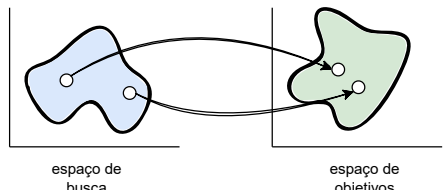


Figure: Different spaces in multi-objective optimization.

Objective space:
$F(x) = z = (z_1, z_2, \ldots, z_k)^T$, 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).

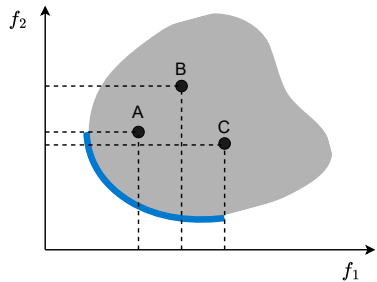


Figure: Pareto Front.

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: *de novo* design

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].

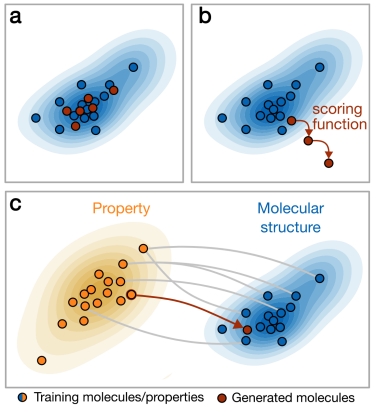


Figure: Generative approaches: (a) distribution learning, (b) objective-guided learning, and (c) conditional generation [6].

Generative chemistry model

- **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.

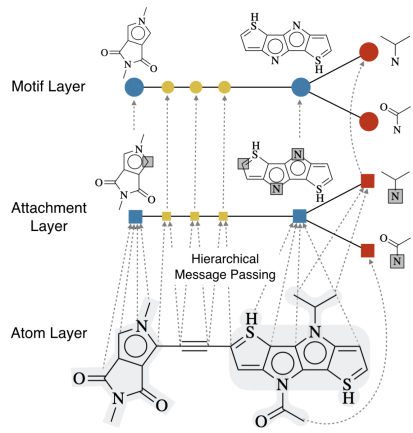


Figure: Hierarchical encoder of the HierVAE model [8].

Multi-objective evolutionary algorithms

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

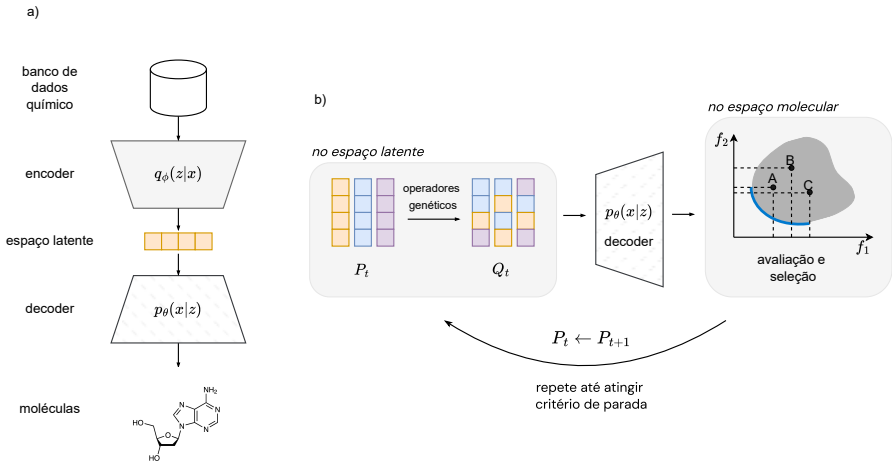


Figure: DockTDesign generative platform.

Pharmacological targets

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

- LpxC[a] (**target** antimicrobial involved in LPS synthesis).
- MMP-9[b] (**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).

[a] UDP-3-O-acyl-N-acetylglucosamine deacetylase
[b] matrix metalloproteinase 9

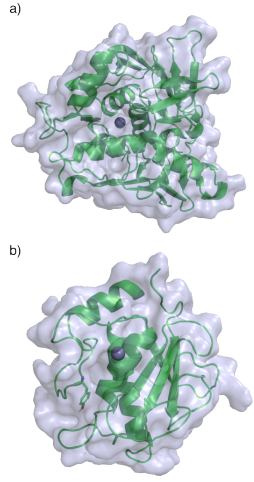


Figure: Receptors (a) LpxC and (b) MMP-9.

Optimization objectives and constraints

Drug-likeness (QED)

Maximize similarity to the profile of approved drugs.

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

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

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.

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 Complexity (Cx)

Minimize structural complexity.

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

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

Objectives: Tanimoto similarity

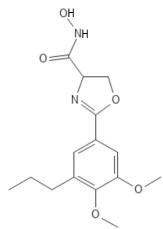
1. Tanimoto similarity against molecule BDBM50074960 ($K_i = 0.053$ nM LpxC):

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

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

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

a)



b)

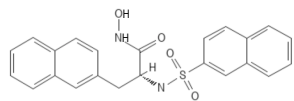


Figure: Compound (a) BDBM50074960 and (b) BDBM50478376

Objectives: molecular docking

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

Results: Tanimoto similarity

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

Table: Multi-objective performance indicators.

Algo.	HV[1]	IGD[2]
Single-obj. GA	0.783	0.073
NSGA-II	0.805	0.040
NSGA-III	**0.880**	**0.031**

[a] Hypervolume.
[b] Inverted generational distance.

Table: Generative chemistry metrics.

Algo.	Valid.[3]	Unic.[4]	DivInt[5]	Nov.[6]
NSGA-II	**1.00±0.0**	**1.00±0.0**	**0.70±0.01**	0.87±0.01
NSGA-III	**1.00±0.0**	0.75±0.03	0.65±0.01	**0.93±0.01**

[c] Validity.
[d] Uniqueness@FP.
[e] Internal diversity.
[f] Novelty.

Results: Tanimoto similarity

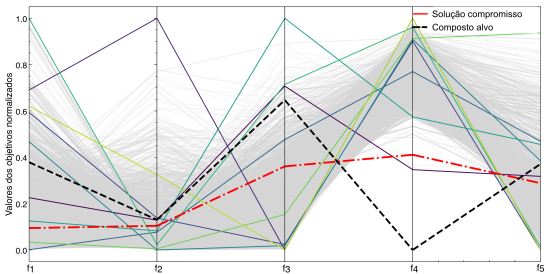


Figure: Non-dominated solutions generated by the NSGA-III algorithm. Order: f_{QED}, f_{SA}, f_{Cx}, f_{sim}, f_{dissim}.

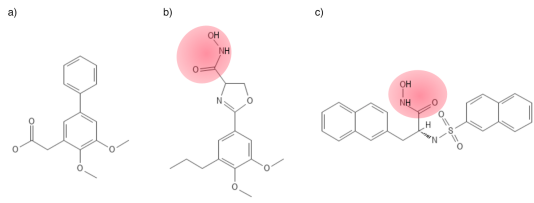


Figure: (a) compromise solution, (b) similarity objective and (c) dissimilarity objective.

Results: molecular docking

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)

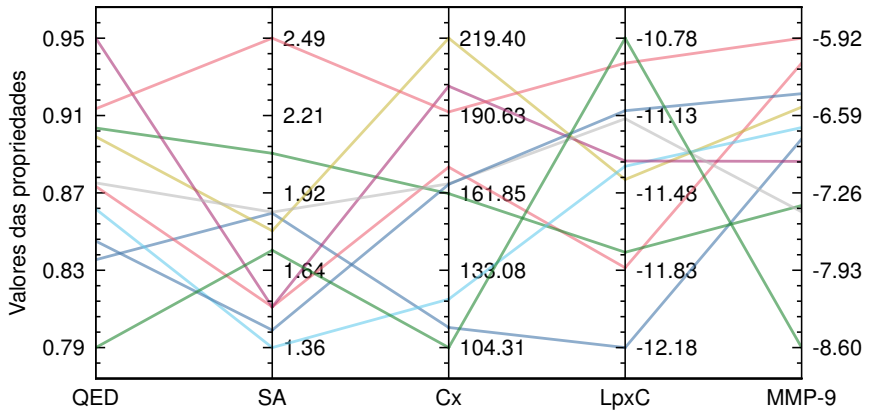


Figure: Parallel coordinates plot for the solutions obtained with protocol I.

Results: molecular docking (protocol II)

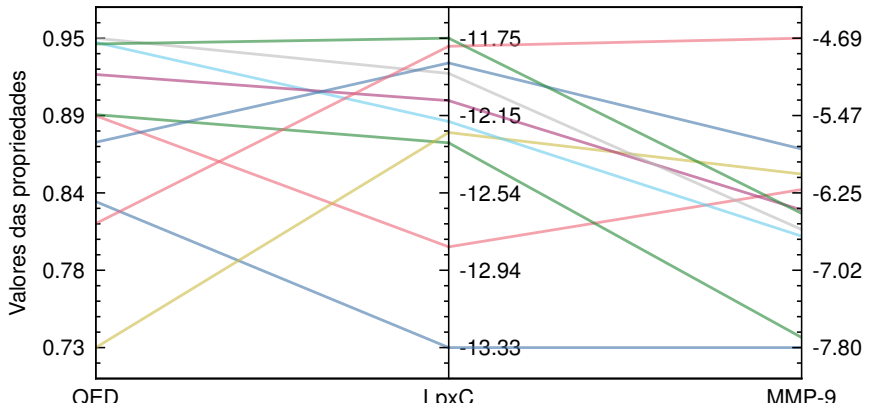


Figure: Parallel coordinates plot for the solutions obtained with protocol II.

Results: molecular docking

Table: 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	**0,920**	**0,745**	0,653
II	0,825	0,451	**0,915**

Results: molecular docking (protocol I)

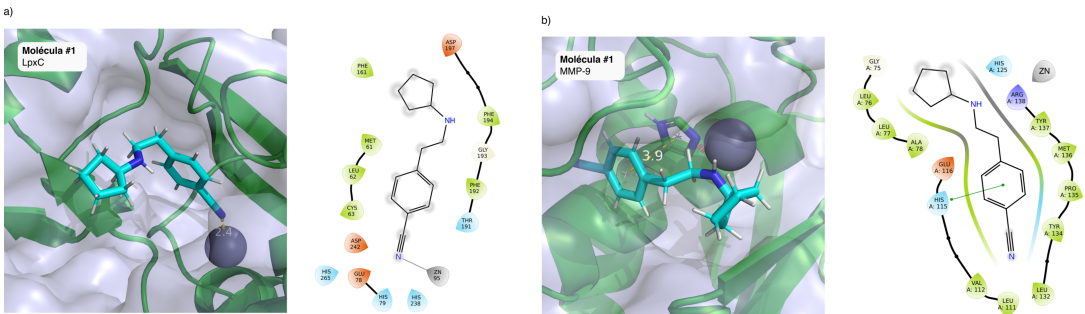


Figure: Compound 1 from protocol I in the active sites of (a) LpxC and (b) MMP-9. LpxC: -12,176 kcal/mol; MMP-9: -6,795 kcal/mol; QED: 0,836; SA: 1,85; Cx: 111,781.

Results: molecular docking (protocol I)

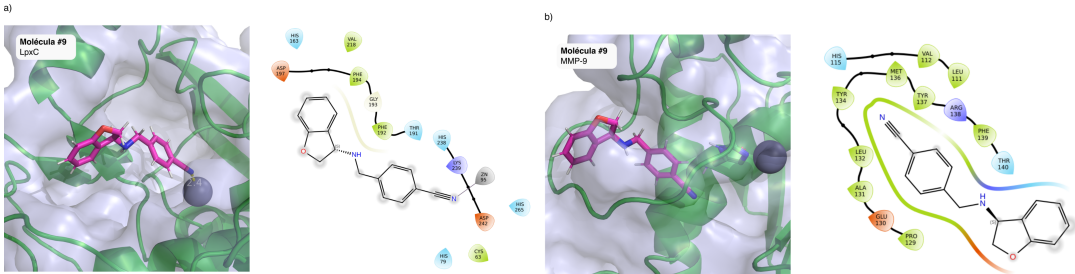


Figure: Compound 9 from protocol I in the active sites of (a) LpxC and (b) MMP-9. LpxC: -10,897 kcal/mol; MMP-9: -5,924 kcal/mol; QED: 0,911; SA: 2,488; Cx: 191,971.

Results: molecular docking (protocol II)

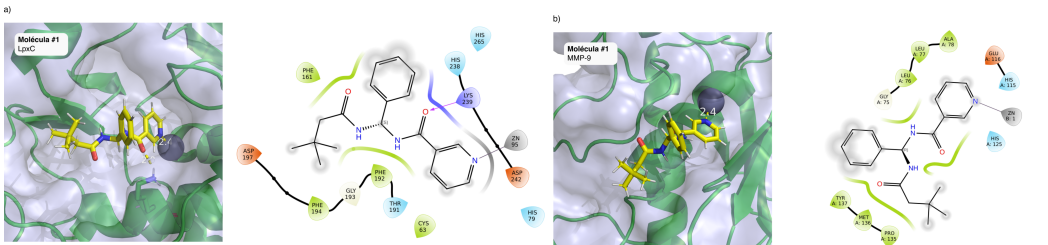


Figure: Compound 1 from protocol II in the active sites of (a) LpxC and (b) MMP-9. LpxC: -13,331 kcal/mol; MMP-9: -7,801 kcal/mol; QED: 0,83; SA: 2,66; Cx: 247,997.

Results: molecular docking (protocol II)

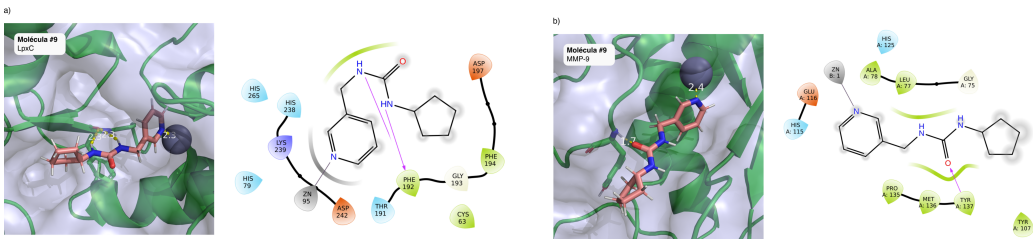


Figure: Compound 9 from protocol II in the active sites of (a) LpxC and (b) MMP-9. LpxC: -11,794 kcal/mol; MMP-9: -4,694 kcal/mol; QED: 0,814; SA: 1,813; Cx: 178,732.

Conclusions

- Integration of **HierVAE** with *many*-objective evolutionary algorithms presents itself as an effective and flexible approach.
- NSGA-II generates molecules more diverse among themselves; NSGA-III explores innovative regions of chemical space. Adjustments in hyperparameters can improve both.
- **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.

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!


LNCC
Laboratório Nacional de
Computação Científica


SINAPAD


CAPES


FAPERJ
Fundação Carlos Chagas Filho de Amparo
à Pesquisa do Estado do Rio de Janeiro

NSGA-II

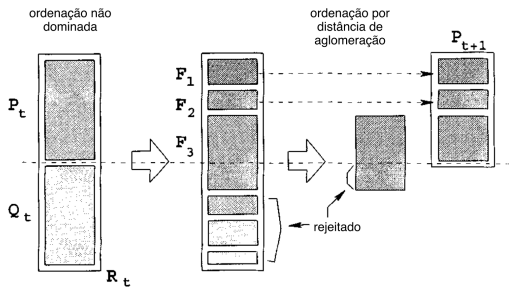


Figure: NSGA-II algorithm.

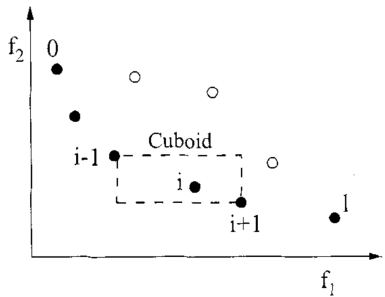


Figure: Crowding distance in NSGA-II.

NSGA-III: reference directions

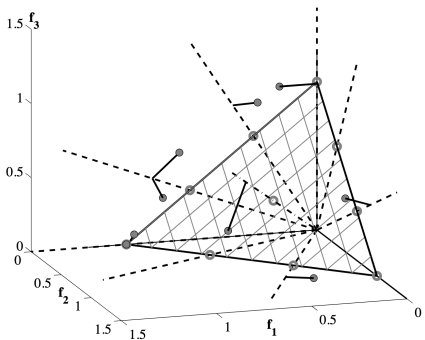


Figure: Reference directions in NSGA-III.

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) ↑

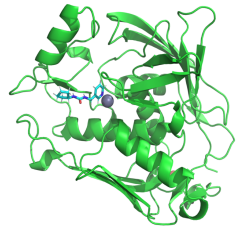
Protocol II (NSGA-III) – Multi Target Scenario

Objectives :

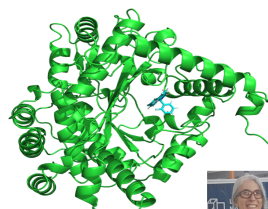
1) QED (drug-likeness) ↑
2) Affinity Prediction LpxC (DockThor + DockTDeep) ↑
3) Affinity Prediction gshA (DockThor + DockTDeep) ↑

Restrictions :

1) SA (synthetic accessibility) ≤ 6
2) LE (ligand efficiency) for LpxC $\geq 0.30\ \text{kcal}\cdot\text{mol}^{-1}\cdot\text{HA}^{-1}$
3) LE (ligand efficiency) for GshA $\geq 0.30\ \text{kcal}\cdot\text{mol}^{-1}\cdot\text{HA}^{-1}$



LpxC
(Klebsiella pneumoniae)

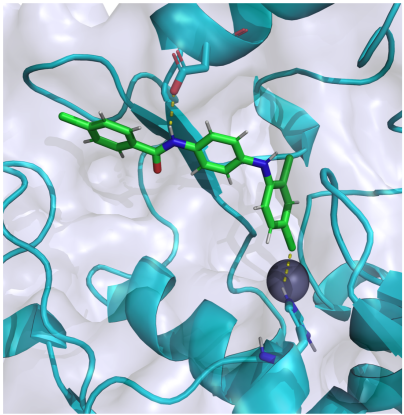


gshA
*(cytoplasmic enzyme glutamate–cysteine ligase from Klebsiella pneumoniae)**

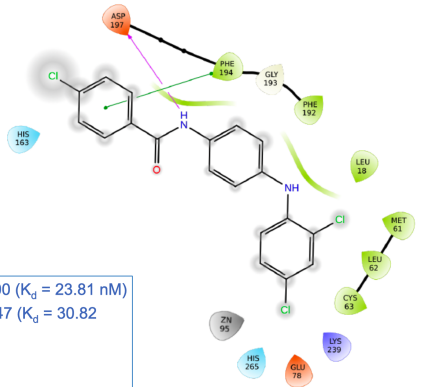
Marisa Nicolás
Bioinformatics

** 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.

Protocol 1 – LpxC - molecule 4 – ΔG = -10.400 kcal/mol LE = 0.416



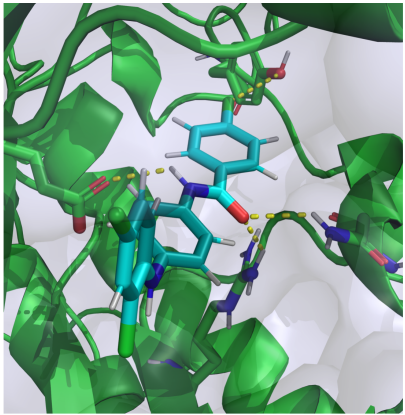
LpxC



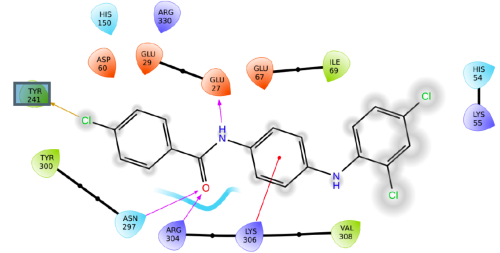
LpxC (ΔG) = -10.400 (K_d = 23.81 nM)
GshA (ΔG) = -10.247 (K_d = 30.82 nM)
QED = 0.523
SA = 1.661
LE (LpxC) = 0.416
LE (GshA)) = 0.410

Charged (negative)
Charged (positive)
Glycine
Hydrophobic
Metal
Polar
Unspecified residue
Water
Hydration site
Hydration site (displaced)
Distance
H-bond
Halogen bond
Metal coordination
Pi-Pi stacking
Pi-cation
Salt bridge
Solvent exposure

Protocol 1 – GshA - molecule 4 – ΔG = -10.247 kcal/mol LE = 0.410



gshA



LpxC (ΔG) = -10.400 (K_d = 23.81 nM)
GshA (ΔG) = -10.247 (K_d = 30.82 nM)
QED = 0.523
SA = 1.661
LE (LpxC) = 0.416
LE (GshA)) = 0.410