

# DockTDesign

**Uma Plataforma de Inteligência Artificial Generativa para Predição de  
Afinidade e Desenho de Fármacos *de novo***

Matheus M. P. da Silva

Defesa de Doutorado em Modelagem Computacional  
Laboratório Nacional de Computação Científica (LNCC/MCTI)

Orientadores: Laurent E. Dardenne & Isabella A. Guedes

10 de outubro de 2025

# Sumário

---

## 1. Introdução e Motivação

1.1 Objetivo geral

## 2. Fundamentação Teórica

2.1 Aprendizado de Máquina e Aprendizado Profundo

2.2 Otimização Multi e *Many*-objetivo

## 3. Revisão da literatura

## 4. DockTDeep

4.1 Metodologia

4.2 Resultados

## 5. DockTDesign

5.1 Metodologia

5.2 Resultados

## 6. Conclusões e Perspectivas

# Motivação

---

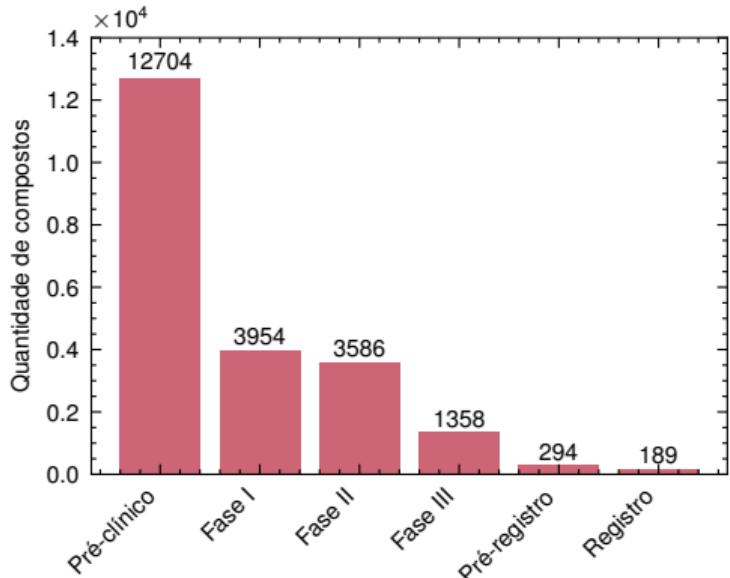


Figura: Número de compostos por fase de desenvolvimento em 2025. Fonte:  
[Pharmaprojects](#).

- Desenvolvimento de fármacos: longo, custoso e com baixas txs. de sucesso ( $\leq 5\%$  na fase pré-clínica).
- Aumentar a taxa de sucesso nas fases iniciais tem o potencial de gerar grande impacto econômico e na saúde pública.
- Métodos computacionais são essenciais para identificar precocemente moléculas promissoras (*hits*).

## Identificação de *hits*

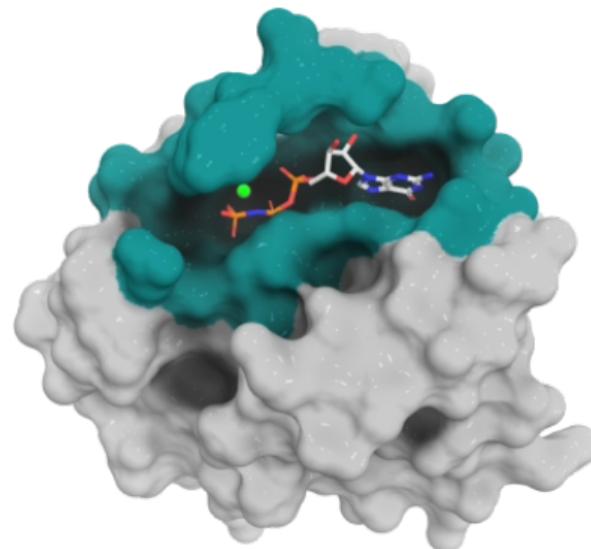
### Objetivos

- Predição da pose
- Predição da afinidade de ligação

### Componentes

- Algoritmo de busca
- Função de pontuação:
  - pose
  - afinidade de ligação

## Atracamento receptor-ligante



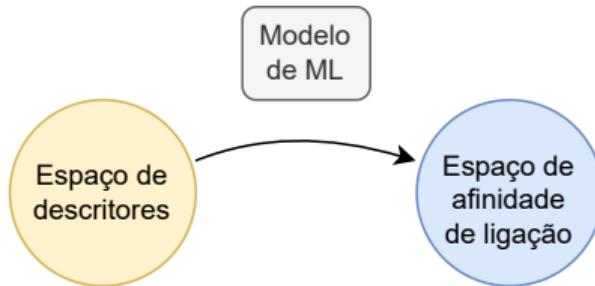
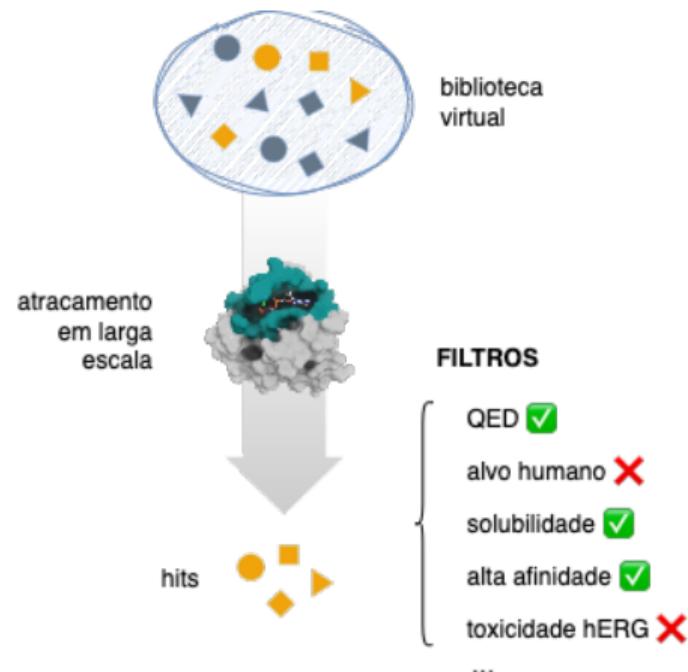


Figura: Funções de pontuação baseadas em técnicas de aprendizado de máquina (MLSFs).

- Buscam reproduzir valores de afinidade de ligação experimentais.
- Predições baseadas em uma única pose do complexo receptor-ligante.
- Utilizam conjuntos de dados com informações estruturais e de afinidade na sua construção.

<sup>1</sup>aprendizado de máquina (*machine learning*).

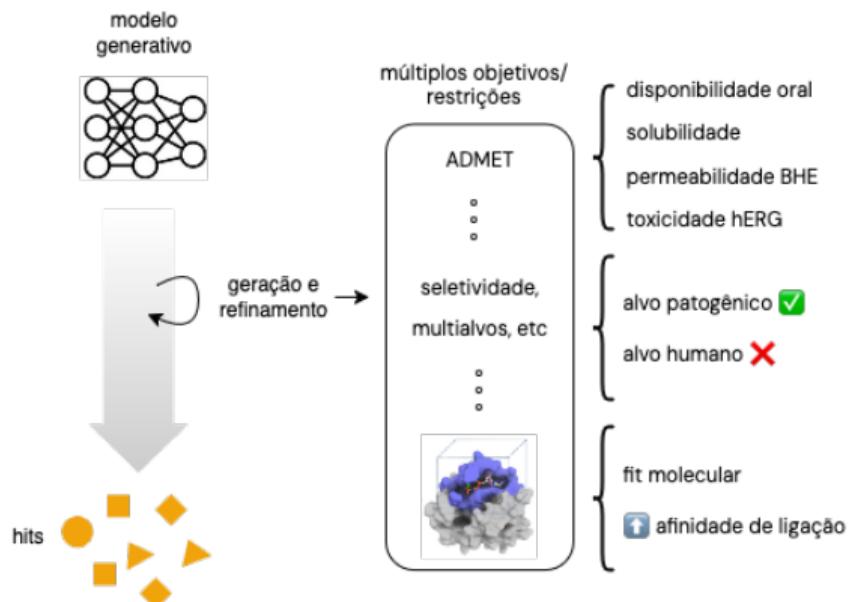
- Identificar *hits* com alta afinidade de ligação a partir de grandes bibliotecas virtuais de compostos.
- Funções de pontuação rápidas e acuradas são essenciais.



## Identificação de *hits*

- Gerar moléculas com propriedades desejáveis sem a necessidade de estruturas pré-definidas e bancos de dados fixos.
- Problema essencialmente multiobjetivo.
- Funções de pontuação rápidas e acuradas são essenciais.

## Desenho de moléculas *de novo*



# Objetivos da tese

---

## Objetivo geral

Desenvolvimento de metodologias baseadas em inteligência artificial (IA) para descoberta de *hits*, com foco em:

1. predição de afinidade receptor-ligante via redes neurais convolucionais 3D;
2. geração *de novo* de moléculas usando modelos de ML generativos e algoritmos evolucionistas *many*-objetivo.

## Fundamentação teórica

# Aprendizado de máquina

O aprendizado de máquina consiste em:

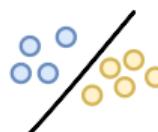
1. obter um conjunto de dados representativo do problema de interesse;
2. construir um modelo estatístico baseado no conjunto de dados (treinamento).

## Aprendizado supervisionado:

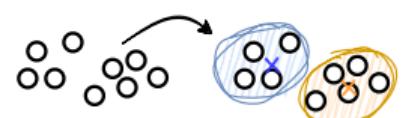
$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n,$$

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

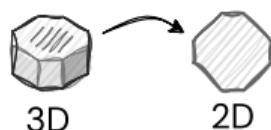
1. classificação (superv.)



2. clusterização (n. superv.)



3. redução dim. (n. superv.)



4. gerativo (n. superv.)



## Aprendizado não supervisionado:

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

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

Figura: Tarefas de aprendizado.

# Aprendizado profundo

Aprendizado profundo refere-se a modelos de ML com múltiplas camadas de transformações:

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

Computação realizada por um neurônio:

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

sendo  $g(\cdot)$  uma função de ativação não linear,  $w \in \mathbb{R}^m$  os pesos,  $b \in \mathbb{R}$  o viés e  $x \in \mathbb{R}^m$  a entrada.

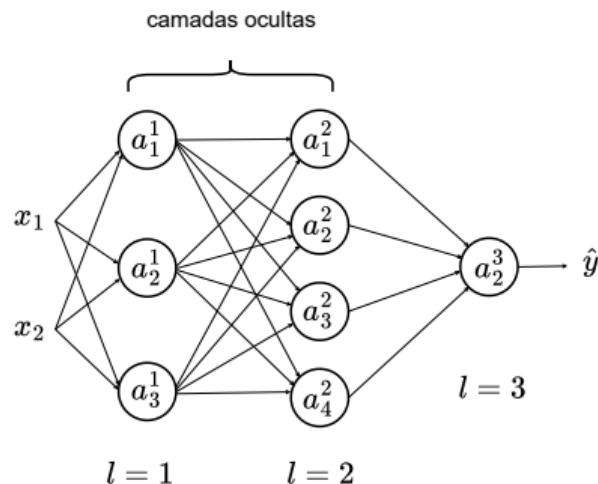


Figura: Arquitetura de uma rede neural artificial.

# Redes neurais convolucionais (CNNs)

- Campos receptivos locais.
- Compartilhamento de pesos.
- Aprendizado de representações hierárquicas.

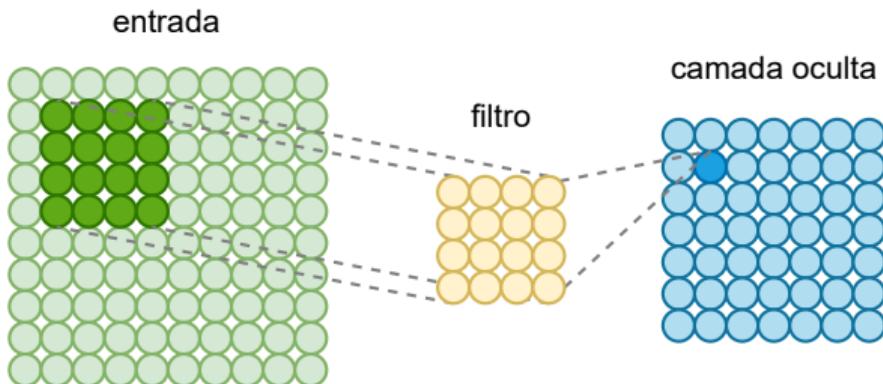
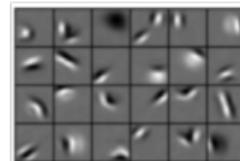
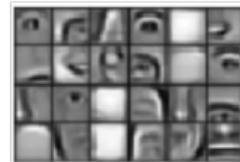


Figura: Operação de convolução ( $X * F = Z$ ).

características baixo nível



características intermediárias



características alto nível



Figura: Representações hierárquicas.

# Autocodificadores variacionais (VAEs)

- VAEs aprendem um espaço latente **contínuo** e **regular** ao modelar variáveis latentes como distribuições probabilísticas.
- Espaço latente adquire “semântica”, favorecendo interpretação, manipulação e robustez.

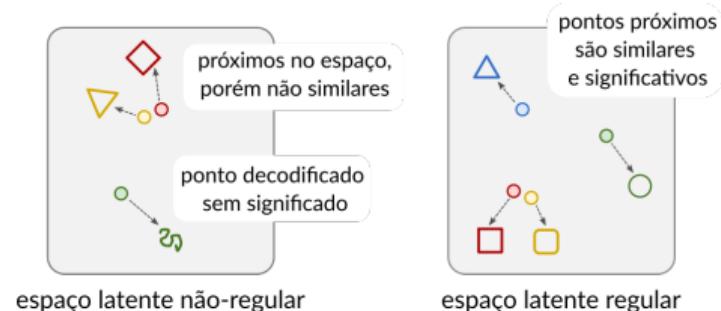
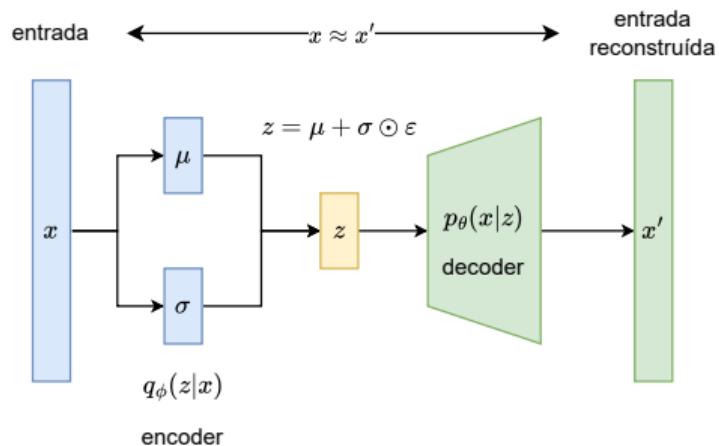


Figura: Propriedades do espaço latente.

Figura: Autocodificador variacional (VAE).

# Otimização multi-objetivo

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

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

## Resolução intuitiva

Transformar em problema mono-objetivo via soma ponderada:

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

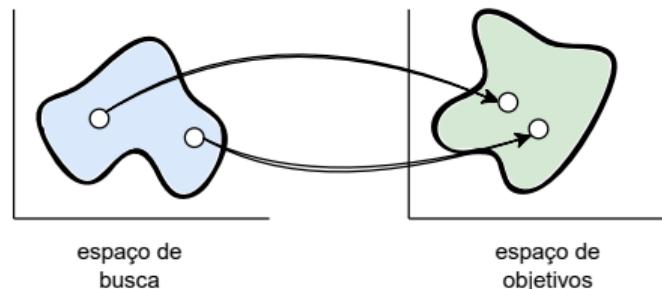


Figura: Diferentes espaços na otimização multi-objetivo.

Espaço dos objetivos:

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

# Dominância e fronteira de Pareto

---

Dizemos que  $x_1$  domina  $x_2$  se:

- $f_i(x_1) \leq f_i(x_2)$  para todo  $i$  (ou seja,  $x_1$  é melhor ou igual em todos os objetivos)
- $f_j(x_1) < f_j(x_2)$  para pelo menos um  $j$  (ou seja,  $x_1$  é estritamente melhor em pelo menos um objetivo)

Na otimização multi-objetivo deseja-se encontrar soluções o mais próximo possível da fronteira de Pareto (**convergência**) e tão diversas quanto possível ao longo dela (**diversidade**).

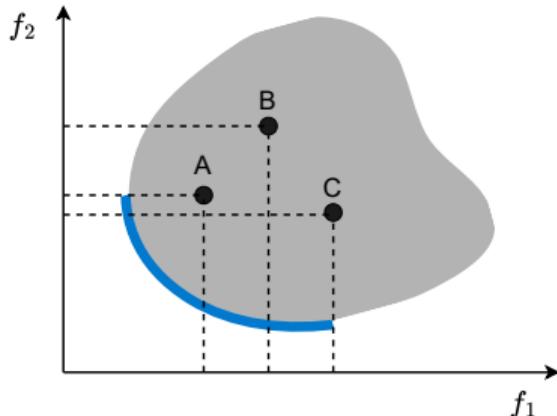


Figura: Fronteira de Pareto.

# Otimização *many*-objetivo

---

Problemas de otimização com  $k > 3$  objetivos são denominados *many*-objetivo e apresentam desafios adicionais:

- Maldição da dimensionalidade: o volume do espaço de busca cresce exponencialmente com  $k$ .
- Resistência à dominância: a maioria das soluções tende a ser não-dominada.
- Visualização das soluções: difícil de representar e interpretar.

## Estado da arte

Técnicas específicas para otimização multi-objetivo mantém as funções objetivo separadas. Exemplos incluem algoritmos evolucionistas, como **NSGA-II** e **NSGA-III**.

## **Revisão da literatura**

MLSFs<sup>a</sup> apresentam desempenho superior às funções clássicas em diversos benchmarks.

<sup>a</sup>funções de pontuação baseadas em aprendizado de máquina (MLSFs).

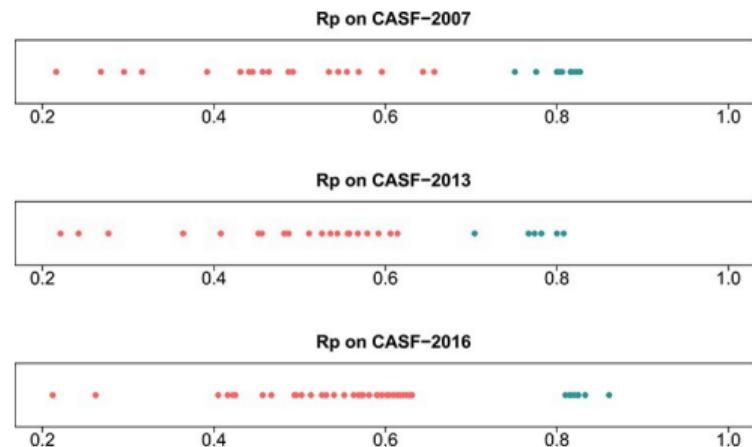


Figura: Desempenho comparativo entre funções clássicas (vermelho) e MLSFs (verde) [1].

# Revisão da literatura: MLSFs

DockTDeep

- Dúvidas sobre o aprendizado de interações intermoleculares e a real **capacidade de generalização** para novos alvos e ligantes.
- Conjuntos de dados “fora do domínio” surgiram como uma avaliação mais desafiadora.
- Estratégias propostas para mitigar esses vieses: *docking*, *decoys* e *crossdocking*.

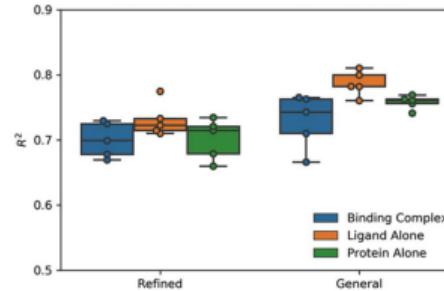


Figura: Viés do ligante e da proteína [2].

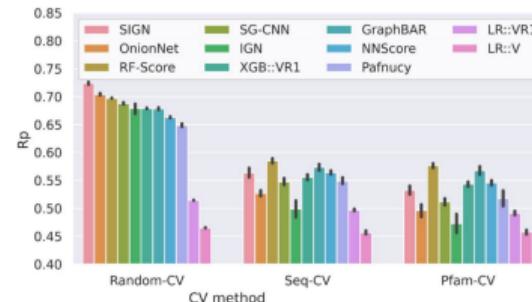


Figura: Desempenho em conjuntos fora do domínio [3].

# Revisão da literatura: MLSFs

DockTDeep

- CNNs não são **invariantes a rotações**.
- Duas abordagens de aumento de dados [4]:
  - rotações de 90º (amplamente utilizada);
  - rotações aleatórias no complexo.

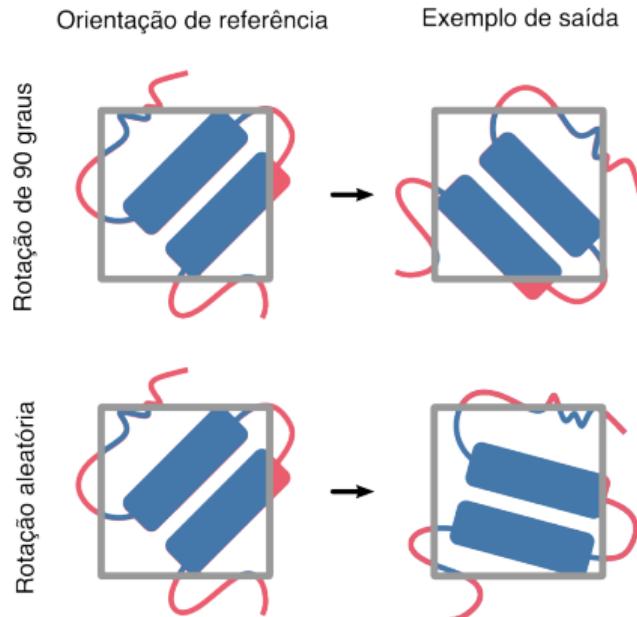


Figura: Diferentes rotações aplicadas ao complexo proteína-ligante.

Três abordagens principais são empregadas para o desenho *de novo* de moléculas usando modelos generativos:

- Aprendizado por distribuição
- Geração condicional
- Aprendizado guiado por objetivos

## Limitações atuais

Metodologias que lidam com  $k \geq 4$  objetivos ainda são pouco exploradas na literatura, sendo raros os estudos que consideram mais de três objetivos, especialmente integrando técnicas apropriadas e modelos generativos [5].

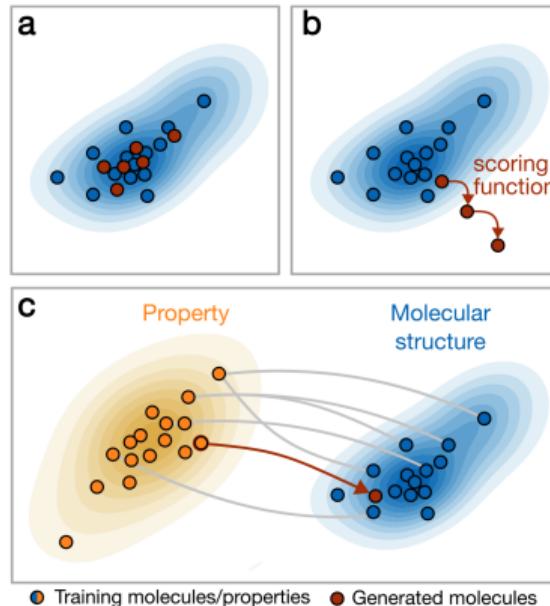


Figura: Abordagens generativas: (a) aprendizado por distribuição, (b) aprendizado guiado por objetivos e (c) geração condicional [6].

DockTDeep

## PDBbind v.2020

General set: **19.443** cpxs

Refined set: **5.316** cpxs

Coreset v.2013: **170** cpxs

Coreset v.2016: **261** cpxs

## Busca de Hiperparâmetros

Utilizando o refined set (divisão aleatória)

- Treino: **3.599** cpxs
- Validação: **904** cpxs

## Faixas de atividade

O dados do conjunto de validação (**904** cpxs) foram divididos em faixas de atividade:

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

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

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

O desempenho em conjuntos teste foi avaliado utilizando o *general set*, em diferentes divisões:

- Aleatória (15.699 treino/3.404 teste).
- Temporal (13.317 treino/1.786 teste).
- Coreset v.2013 (18.933 treino/170 teste).
- Coreset v.2016 (18.842 treino/261 teste).
- Protocolo Pfam-CV (*general set*):
  - Avaliação da generalização fora de domínio.
  - Agrupamento pela similaridade estrutural do sítio de ligação (base de dados Pfam).
  - Validação cruzada repetida 30x.
  - Proporção:  $\frac{2}{3}$  treinamento e  $\frac{1}{3}$  teste.

- Grade de voxels  $24 \text{ \AA}^3$  e discretização  $1 \text{ \AA}$ .
- Canais para os elementos: C, H, O, N, S e X (outros).
- 21 canais: 6 proteína, 6 ligante, 6 cpx e 3 volume total.
- Representações geradas com DockTGrid [7].
- GitHub: [github.com/gmmsb-Incc/docktgrid](https://github.com/gmmsb-Incc/docktgrid).

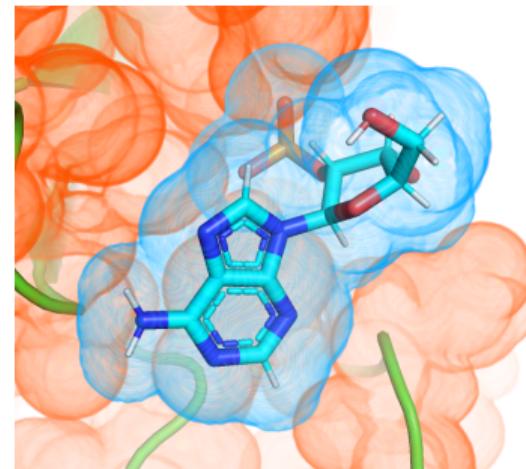


Figura: Exemplo ilustrativo da representação de voxels usada.

- CNN 3D com três camadas convolucionais.
- Uma camada de densa de 1000 neurônios.
- Camada de saída linear (regressão).
- $\sim 2M$  de parâmetros treináveis.
- Taxa de aprendizado:  
 $8,74 \times 10^{-4}$
- Épocas de treinamento: 1500.
- GitHub: [github.com/gmmsb-Incc/docktdeep](https://github.com/gmmsb-Incc/docktdeep) e pré-publicação [4].

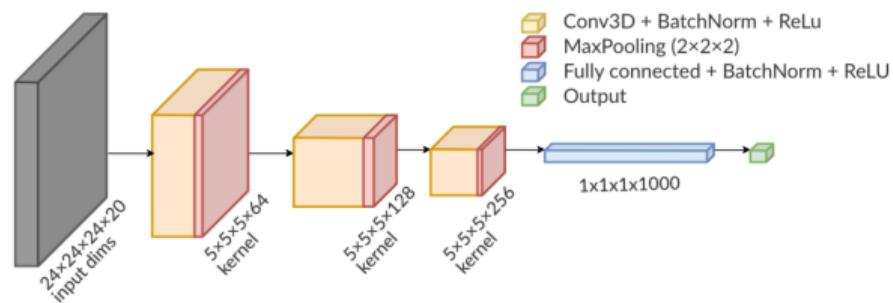


Figura: Arquitetura da rede CNN.

# Estratégias de aumento de dados

DockTDeep

Duas estratégias de aumento de dados foram comparadas:

- rotações de 90°;
- rotações aleatórias no complexo (aplicadas a cada nova época).

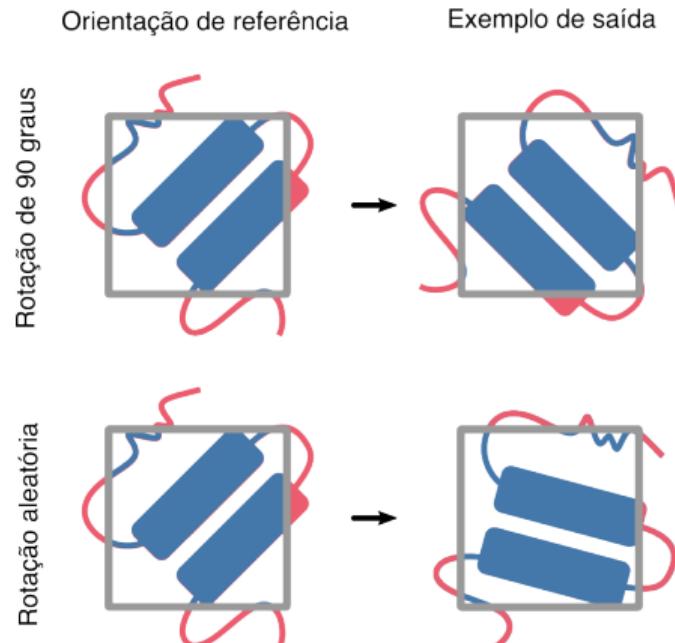
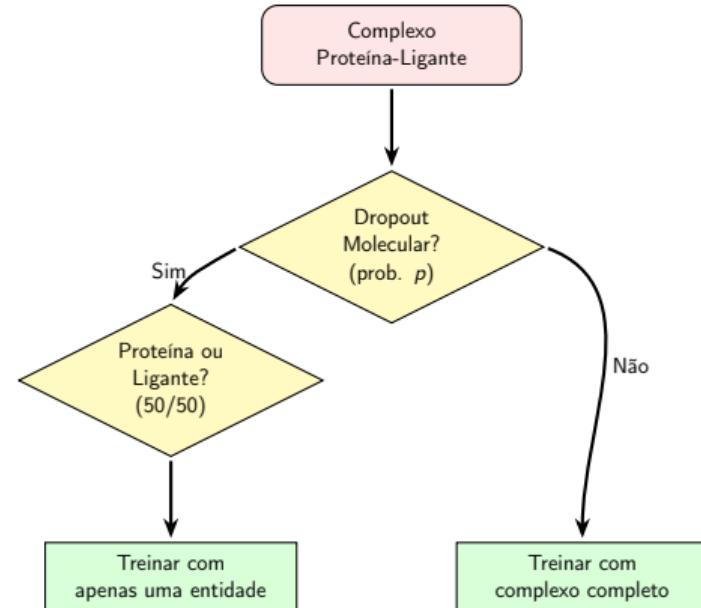


Figura: Diferentes rotações aplicadas ao complexo proteína-ligante.

# Dropout molecular

DockTDeep

- Técnica de regularização que evita superajuste removendo aleatoriamente proteína *ou* ligante durante o treinamento e definindo a afinidade como zero.
- Cada complexo tem chance  $p = 0,06$  de sofrer dropout por época.
- Busca forçar o modelo a aprender interações relevantes, não dependendo de um único componente.



# **Resultados**

# Resultados: variância rotacional

DockTDeep

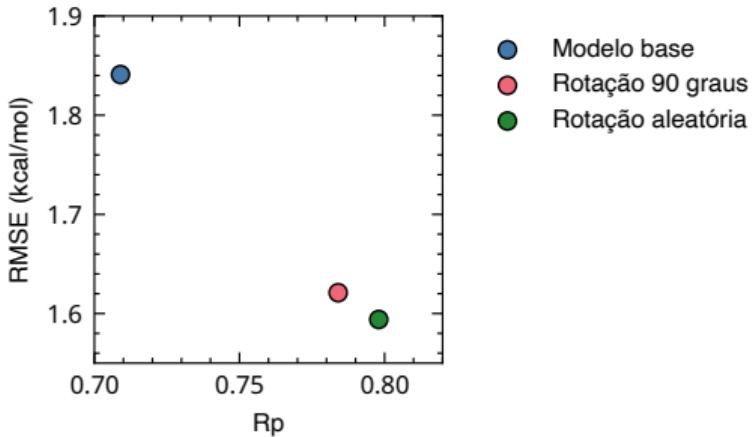


Figura: Comparação do desempenho preditivo usando as métricas RMSE (menor é melhor) e Rp (maior é melhor) para cada uma das estratégias.

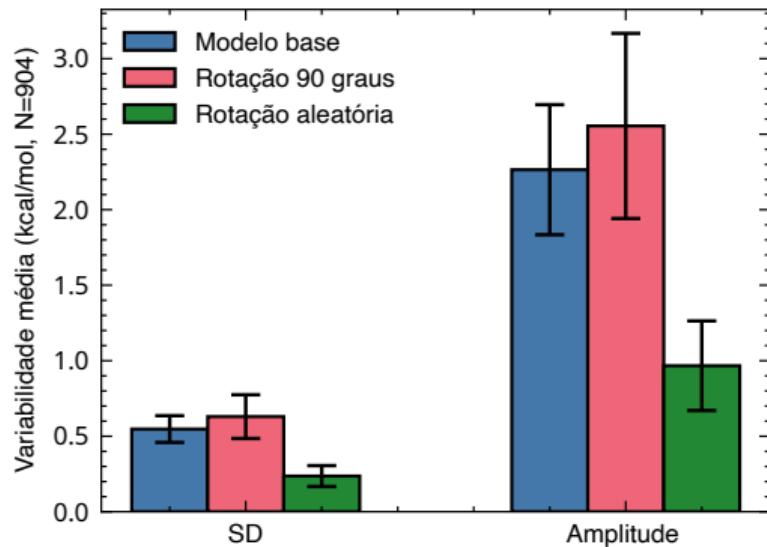


Figura: Média do desvio padrão (SD) e amplitude das previsões (kcal/mol) para o conjunto de validação ( $N=904$ ).

# Resultados: viés proteína/ligante

DockTDeep

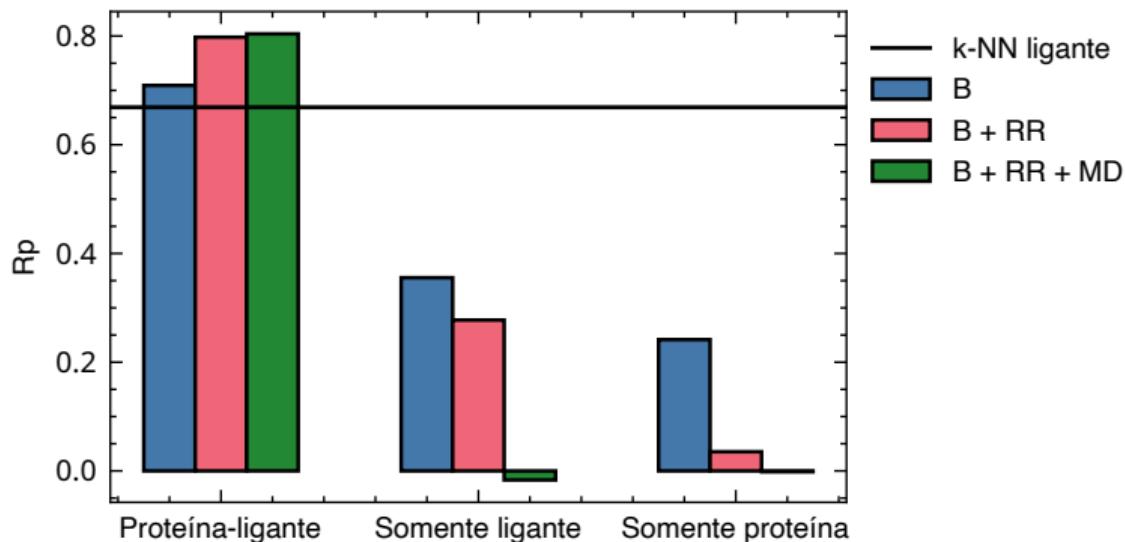


Figura: Comparação dos valores de  $R_p$  entre três modelos: modelo base (B), usando rotação aleatória (B + RR) e usando rotação aleatória e *dropout* molecular (B + RR + MD), além do modelo *k*-NN baseado apenas no ligante. A avaliação foi realizada em três cenários distintos: complexo proteína-ligante completo, apenas o ligante e apenas a proteína; mantendo-se inalterados os rótulos de afinidade.

# Resultados: viés proteína/ligante

DockTDeep

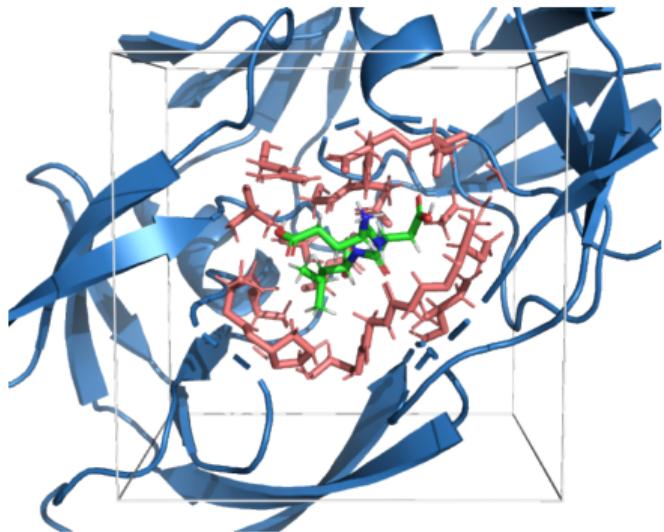


Figura: Visualização dos átomos da proteína removidos (destacados em vermelho), que estão a até 5 Å de distância do ligante.

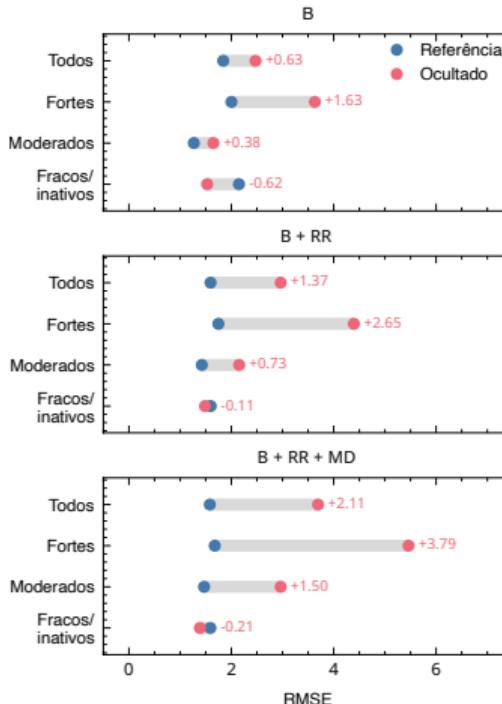
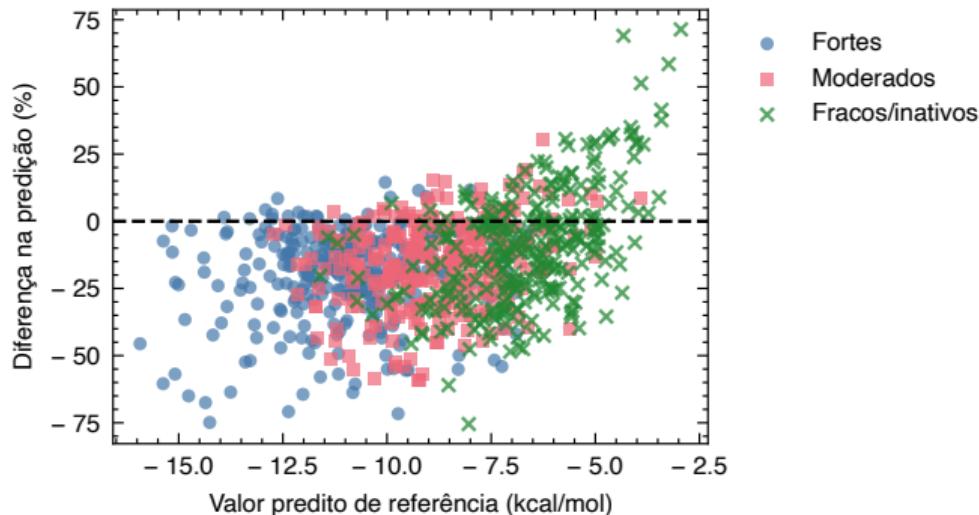


Figura: RMSE antes (azul) e depois da remoção dos átomos (vermelho).

# Resultados: aprendizado das interações

DockTDeep



**Figura:** Gráfico de dispersão comparando as previsões de afinidade para as estruturas cristalográficas (referência) com a variação percentual das previsões referentes às piores poses obtidas no experimento de reatracamento.

# Resultados: faixas de atividade

DockTDeep

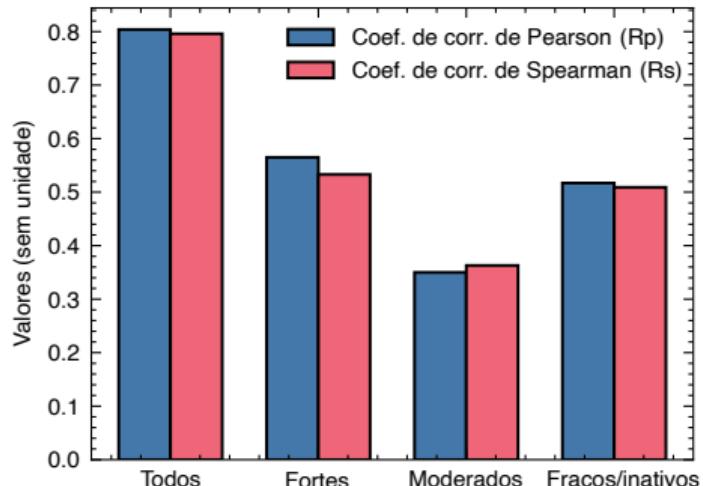


Figura: Comparação dos valores de  $R_p$  e  $R_s$  para todo o conjunto de validação e faixas de afinidade.

	Preditos			
	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: Matriz de confusão para as três classes de afinidade: fortes, moderados e fracos/inativos.

# Resultados: avaliação em conjuntos externos

DockTDeep

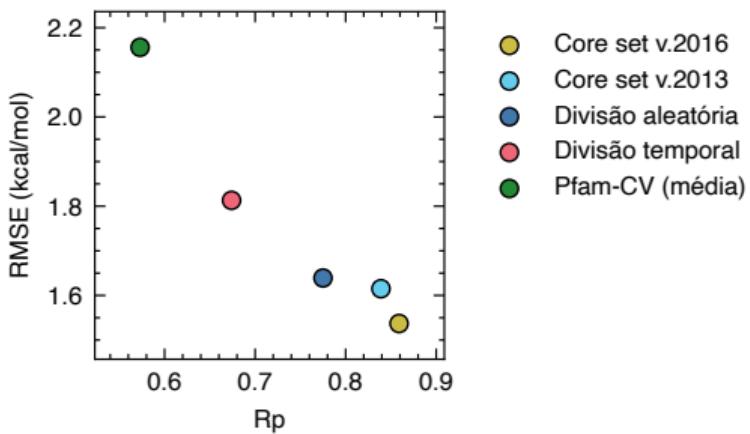


Figura: Comparação dos valores de RMSE (quanto menor, melhor) e  $R_p$  (quanto maior, melhor), em todos os conjuntos teste avaliados.

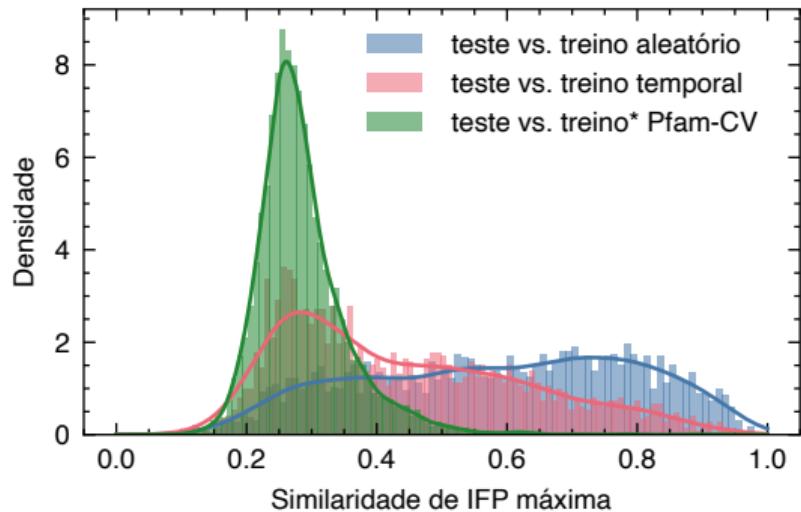


Figura: Gráfico de densidade da similaridade máxima das IFPs entre os conjuntos.

# Resultados: PDBbind coressets Pfam-CV

DockTDeep

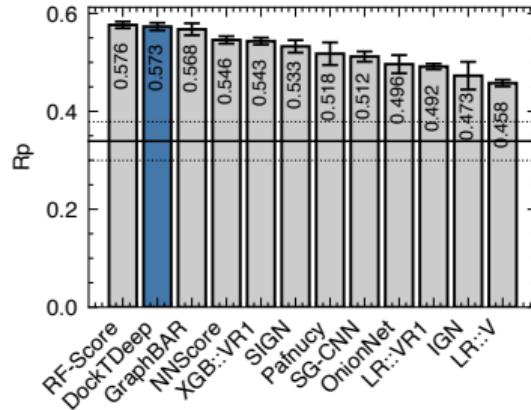
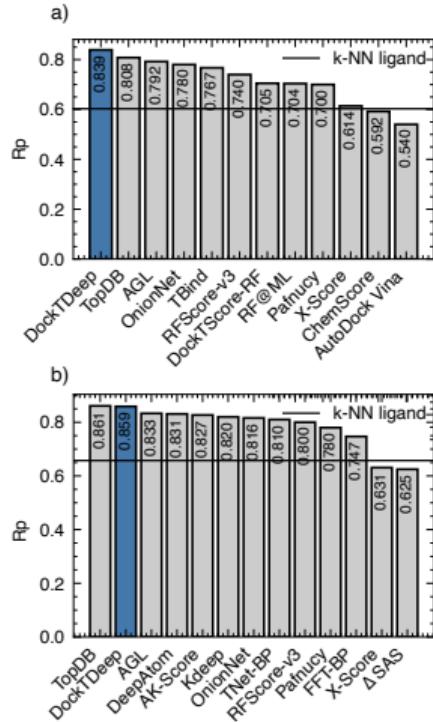


Figura: R<sub>p</sub> no esquema de divisão Pfam-CV

Figura: R<sub>p</sub> no PDBbind v. (a) 2013 e (b) 2016.

- **DockTDeep**: modelo CNN 3D simples, treinado de maneira sistemática e fundamentada, apresenta generalização robusta e é competitivo com o estado da arte.
- Rotação aleatória supera rotação de 90° como técnica de aumento de dados, melhorando estabilidade e desempenho.
- Dropout molecular reduz viés proteína/ligante, favorecendo aprendizado de interações significativas.
- Modelo é mais robusto em tarefas de **categorização** (forte/moderado/fraco) que em ranqueamento fino.
- Desempenho cai em cenários rigorosos (Pfam-CV), destacando limitações impostas pela diversidade dos dados de treino e teste.

# DockTDesign

# Modelo de química generativa

DockTDesign

- Modelo **HierVAE** baseado em grafos para geração de moléculas com **validade de 100%** [8].
- Usa **motivos estruturais** como blocos de construção, extraídos de padrões recorrentes no ChEMBL.
- Representação em **3 níveis**: motivos, ligações entre motivos e grafo atômico.
- Espaço latente regular ( $z \in \mathbb{R}^{32}$ ).
- Pré-treinado em **1,8M moléculas** do ChEMBL.

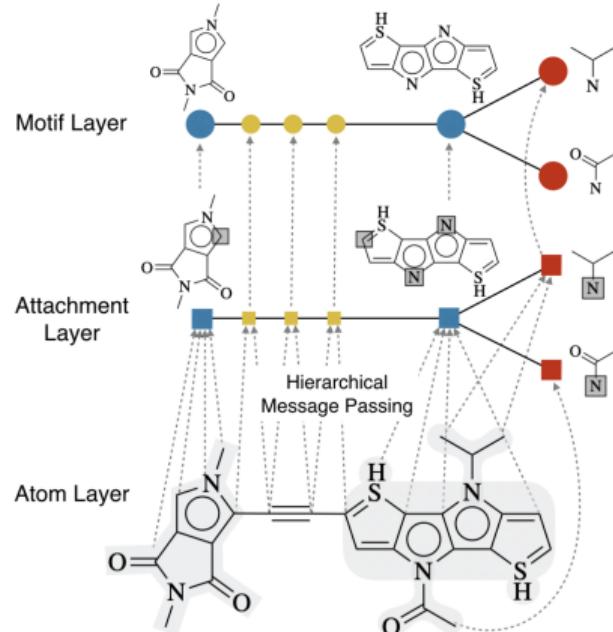


Figura: Codificador hierárquico do modelo HierVAE [8].

## Algoritmos comparados:

- NSGA-II (multiobjetivo).
- NSGA-III (*many*-objetivo).
- AG mono-objetivo com agregação.

## Execução geral:

- 100 gerações.
- População: 800 indivíduos.
- 11 execuções independentes por configuração (apenas 1 execução no caso de usar o atracamento molecular).

## Operadores genéticos:

- Crossover 2 pontos,  $p_c = 0,9$ .
- Mutação Gaussiana,  $p_m = 0,1$ ,  $\sigma = 0,01$ .
- Seleção: Torneio binário.

## Configurações específicas:

- NSGA-III: direções de referência via método Das-Dennis (núm. igual ao tamanho da população).
- AG mono-objetivo: 5 combinações de pesos  $(0.6, 0.1, 0.1, 0.1, 0.1)$  e  $(0.1, 0.1, 0.1, 0.1, 0.6)$ .

# Plataforma generativa proposta

DockTDesign

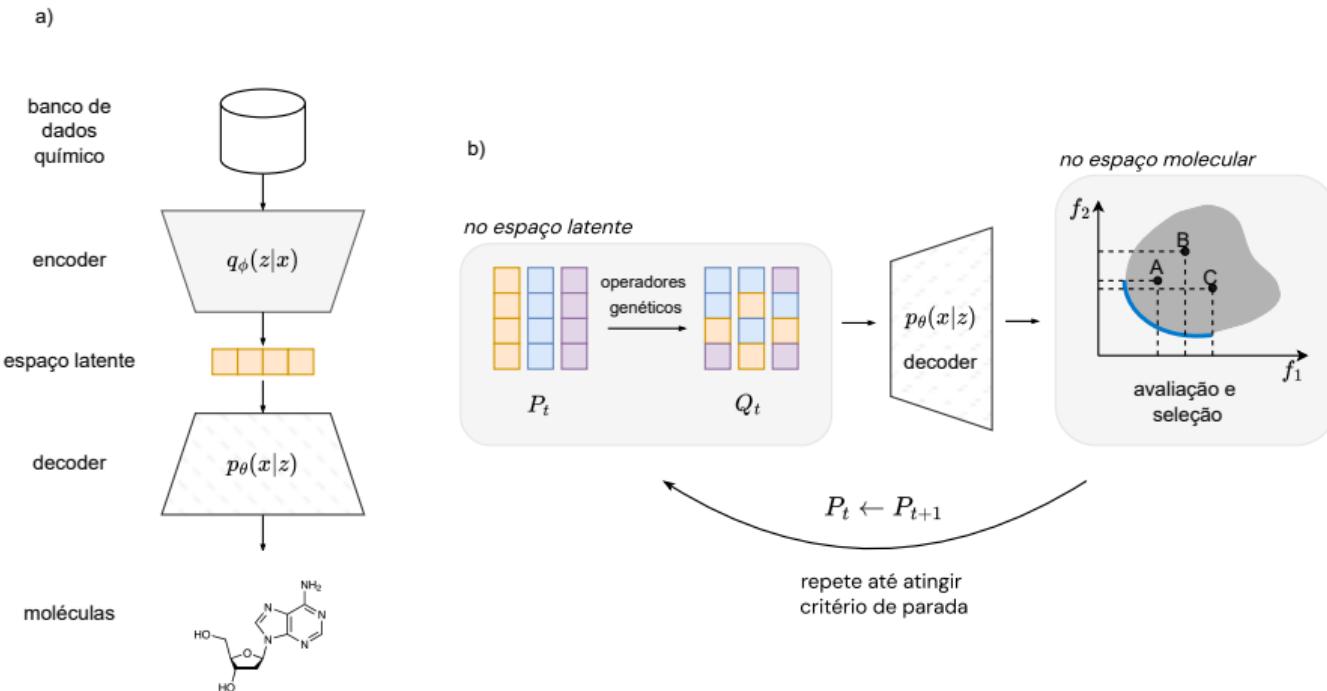


Figura: Plataforma generativa DockTDesign.

# Alvos farmacológicos

DockTDesign

Um **alvo** e um **antialvo** (*off-target*) foram selecionados como estudo de caso para geração de inibidores seletivos:

- LpxC<sup>a</sup> (**alvo** antimicrobiano envolvido na síntese de LPS).
- MMP-9<sup>b</sup> (**antialvo** humano importante para degradação da matriz extracelular).

A escolha dos alvos foi feita a partir da identificação de um inibidor potente contra LpxC que também inibe a MMP-9 (BDBM50478376).

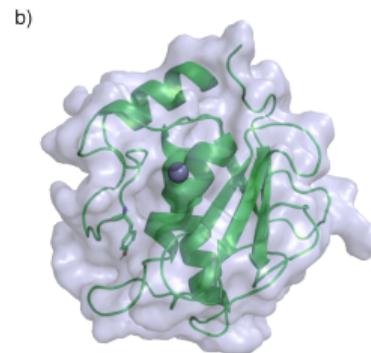
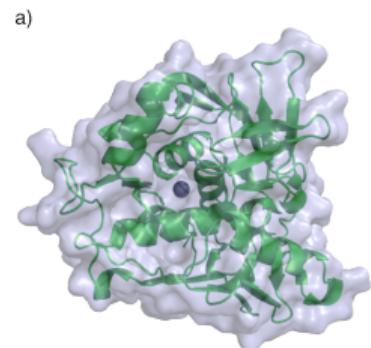


Figura: Receptores (a) LpxC e (b) MMP-9.

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

<sup>b</sup>metaloproteinase de matriz 9

## Drug-likeness (QED)

Maximizar a semelhança ao perfil de fármacos aprovados.

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

Valores de QED  $\in [0,1]$ .

## Acessibilidade Sintética (SA)

Penalizar moléculas difíceis de sintetizar.

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

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

## Peso Molecular (MW)

Otimizar valor-alvo desejado.

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

MW(m): massa em Da; v: valor-alvo.

## Complexidade Molecular (Cx)

Minimizar complexidade estrutural.

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

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

# Objetivos: similaridade de Tanimoto

DockTDesign

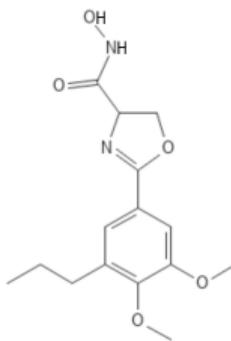
## 1. Similaridade de Tanimoto

contra a molécula

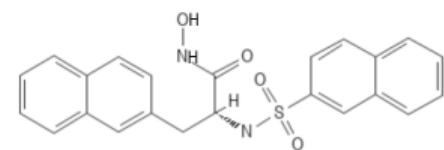
BDBM50074960 ( $K_i = 0.053 \text{ nM}$   
LpxC):

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

a)



b)



## 2. Dissimilaridade de Tanimoto

contra a molécula

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: Composto (a) BDBM50074960 e (b)  
BDBM50478376

Integração DockTDesign (generativo), DockThor (pose) e DockTDeep (afinidade de ligação):

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

onde  $m$ : molécula,  $p$ : alvo proteico,  $v$ : valor-alvo,  $\hat{y}$ : predição de afinidade (DockTDeep).

- **Alvo:**  $v = -25 \text{ kcal/mol}$  (maximizar atividade).
- **Antialvo:**  $v = 0 \text{ kcal/mol}$  (minimizar atividade).

# **Resultados**

## Desenho experimental

**Algoritmos:** AG mono-objetivo, NSGA-II e NSGA-III.

**Objetivos:**

1. maximizar QED;
2. minimizar SA;
3. minimizar complexidade;
4. maximizar a similaridade contra a molécula BDBM50074960 (ativo LpxC);
5. minimizar a similaridade contra a molécula BDBM50478376 (ativo LpxC, ativo MMP-9).

Trabalho resultou em uma publicação em congresso internacional [9].

# Resultados: similaridade de Tanimoto

DockTDesign

Tabela: Indicadores de desempenho multiobjetivo.

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

<sup>a</sup>Hipervolume.

<sup>b</sup>Distância geracional invertida.

Tabela: Métricas de química generativa.

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>Validade.

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

<sup>e</sup>Diversidade interna.

<sup>f</sup>Novidade.

# Resultados: similaridade de Tanimoto

DockTDesign

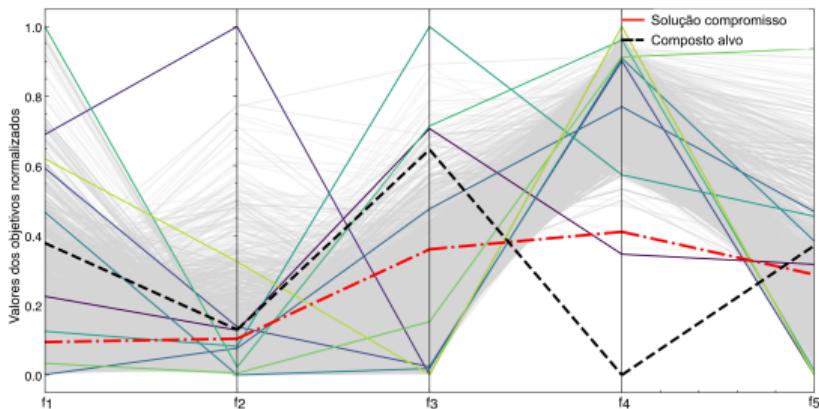


Figura: Soluções não dominadas geradas pelo algoritmo NSGA-III. Ordem:  
 $f_{QED}$ ,  $f_{SA}$ ,  $f_{Cx}$ ,  $f_{sim}$ ,  $f_{dissim}$ .

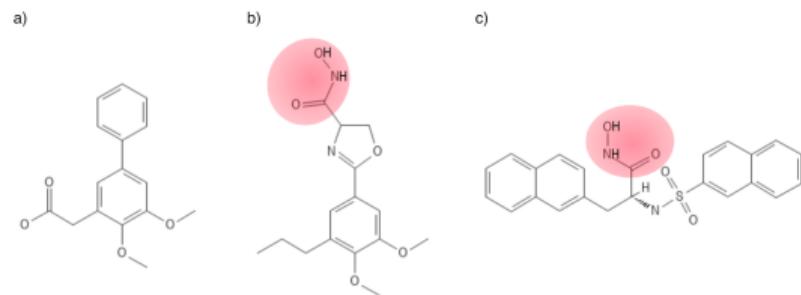


Figura: (a) solução compromisso, (b) objetivo de similaridade e (c) objetivo de dissimilaridade.

## Protocolo I

### Objetivos:

1. maximizar QED;
2. minimizar SA;
3. minimizar complexidade;
4. minimizar a afinidade de ligação para a LpxC;
5. maximizar a afinidade de ligação para a MMP-9.

## Protocolo II

### Objetivos:

1. maximizar QED;
2. minimizar a afinidade de ligação para a LpxC;
3. maximizar a afinidade de ligação para a MMP-9.

### Restrições:

1. SA menor ou igual a 4;
2. complexidade entre 150 e 600;
3. peso molecular acima de 150 Da;
4. QED acima de 0,5.

# Resultados: atracamento molecular (protocolo I)

DockTDesign

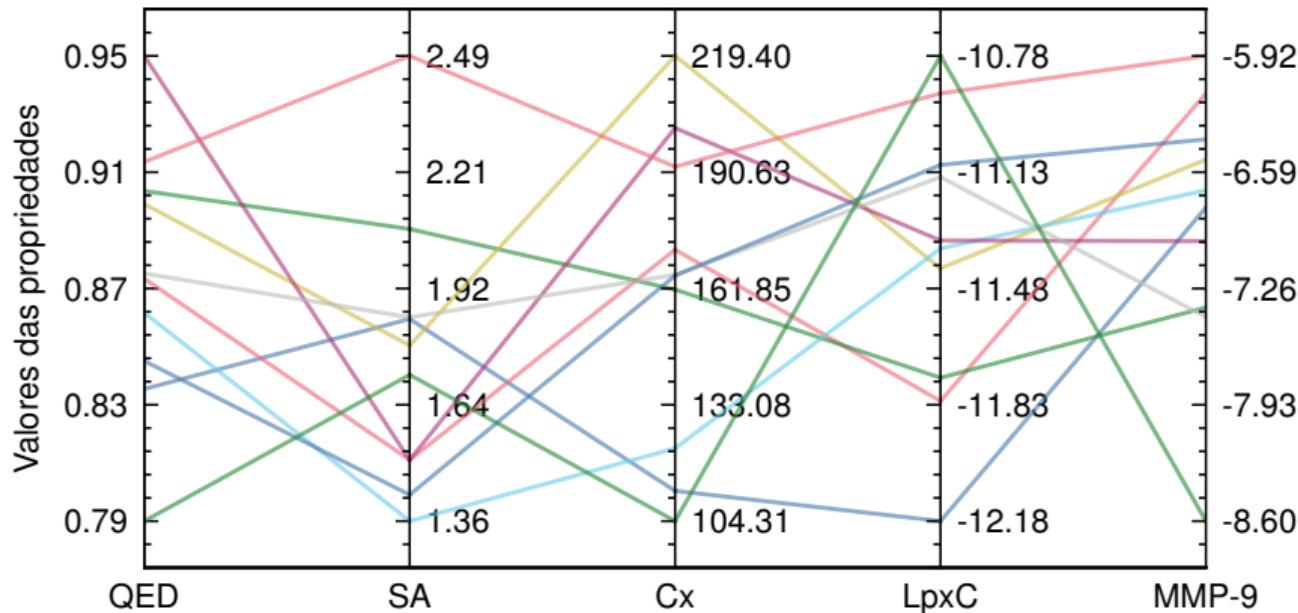


Figura: Gráfico de coordenadas paralelas para as soluções obtidas com o protocolo I.

# Resultados: atracamento molecular (protocolo II)

DockTDesign

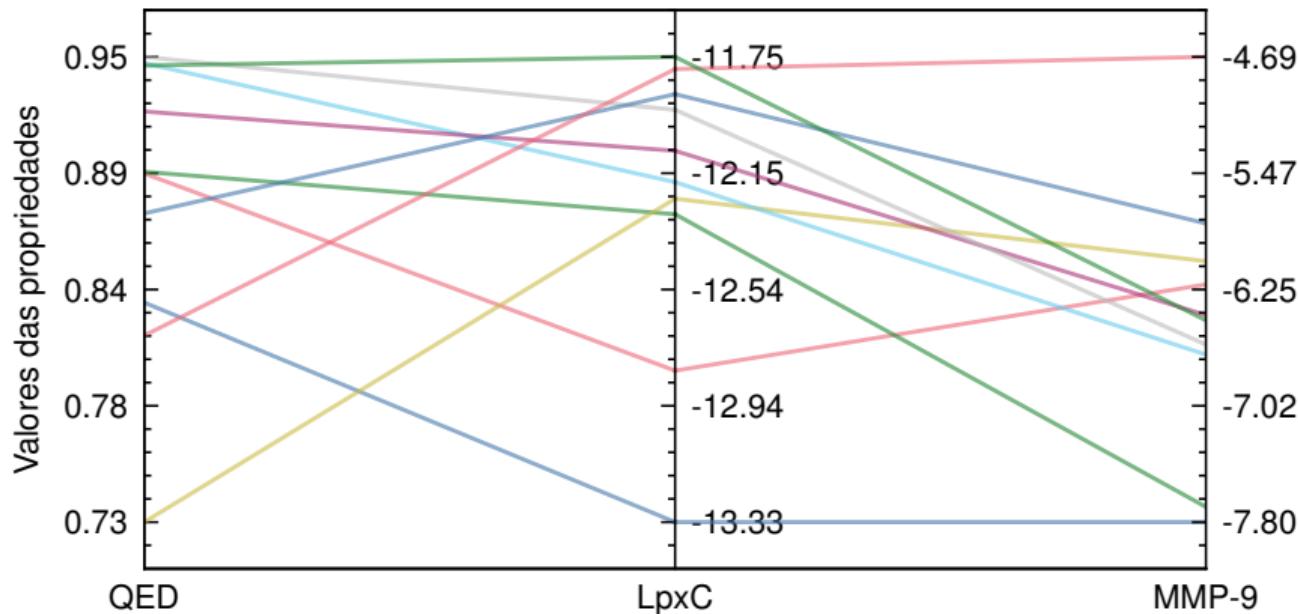


Figura: Gráfico de coordenadas paralelas para as soluções obtidas com o protocolo II.

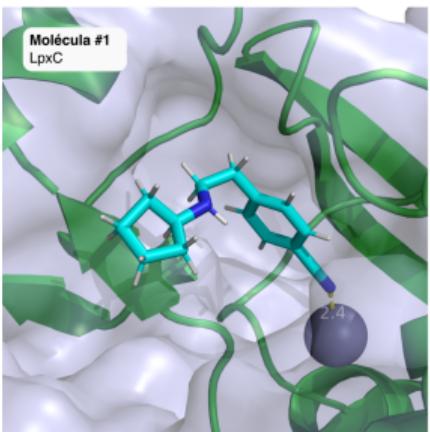
Tabela: Métricas de unicidade na fronteira de Pareto, diversidade interna (DivInt) e novidade considerando as moléculas geradas pelos protocolos experimentais I e II.

Protocolo	Unicidade@FP	DivInt	Novidade
I	<b>0,920</b>	<b>0,745</b>	0,653
II	0,825	0,451	<b>0,915</b>

# Resultados: atracamento molecular (protocolo I)

DockTDesign

a)



b)

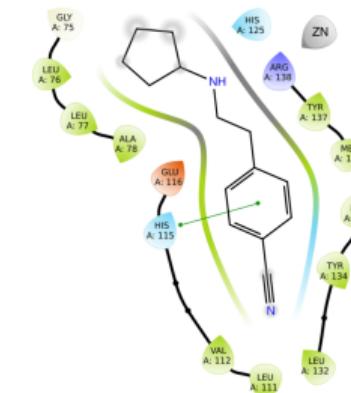
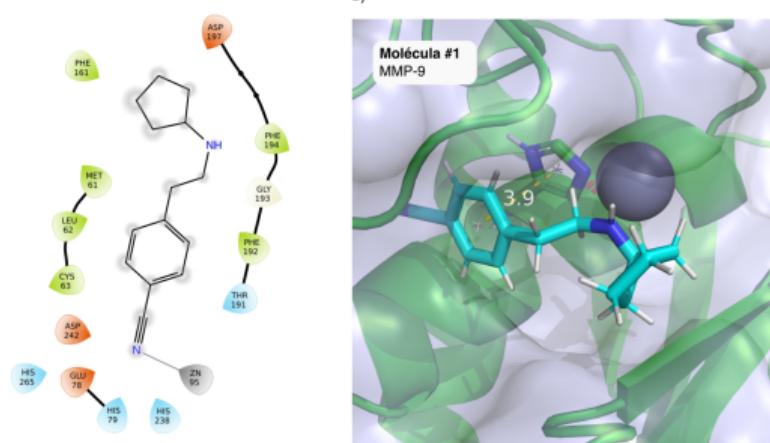
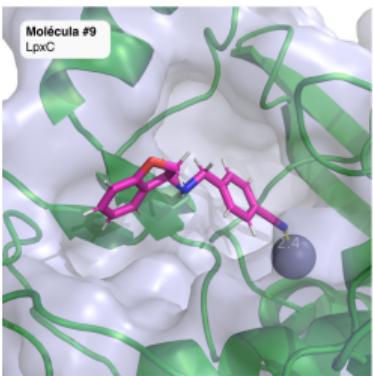


Figura: Composto 1 do protocolo I nos sítios ativos de (a) LpxC e (b) MMP-9. LpxC: -12,176 kcal/mol; MMP-9: -6,795 kcal/mol; QED: 0,836; SA: 1,85; Cx: 111,781.

# Resultados: atracamento molecular (protocolo I)

DockTDesign

a)



b)

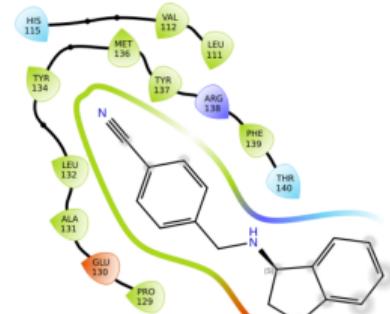
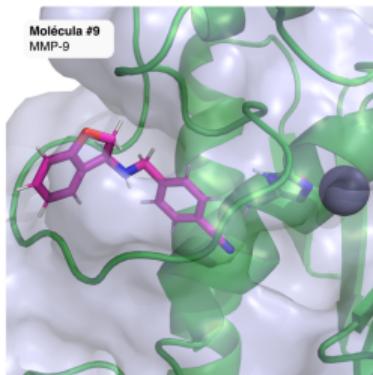
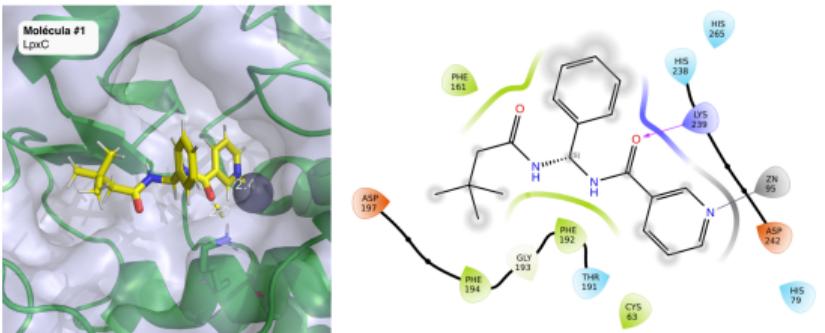


Figura: Composto 9 do protocolo I nos sítios ativos de (a) LpxC e (b) MMP-9. LpxC: -10,897 kcal/mol; MMP-9: -5,924 kcal/mol; QED: 0,911; SA: 2,488; Cx: 191,971.

# Resultados: atracamento molecular (protocolo II)

DockTDesign

a)



b)

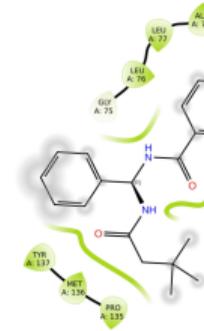
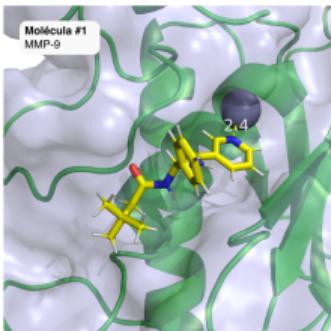
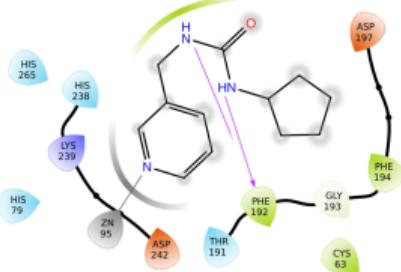
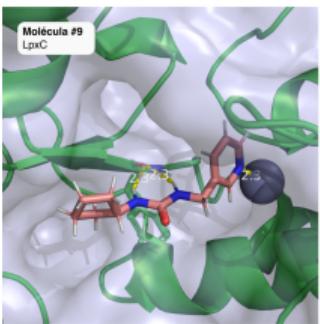


Figura: Composto 1 do protocolo II nos sítios ativos de (a) LpxC e (b) MMP-9. LpxC: -13,331 kcal/mol; MMP-9: -7,801 kcal/mol; QED: 0,83; SA: 2,66; Cx: 247,997.

# Resultados: atracamento molecular (protocolo II)

DockTDesign

a)



b)

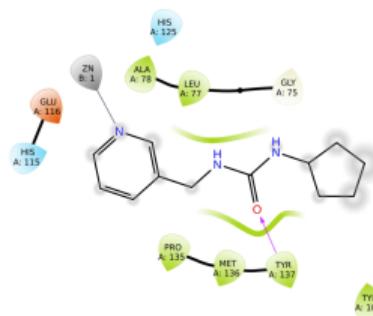
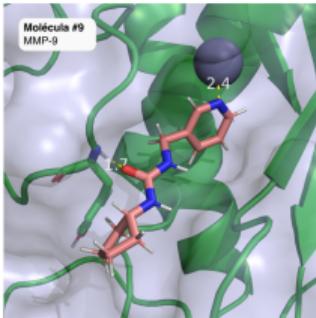


Figura: Composto 9 do protocolo II nos sítios ativos de (a) LpxC e (b) MMP-9. LpxC: -11,794 kcal/mol; MMP-9: -4,694 kcal/mol; QED: 0,814; SA: 1,813; Cx: 178,732.

- Integração de **HierVAE** com algoritmos evolucionistas *many*-objetivo apresenta-se como uma abordagem eficaz e flexível.
- NSGA-III supera NSGA-II e AG com agregação em convergência e cobertura da fronteira de Pareto.
- NSGA-II gera moléculas mais diversas entre si; NSGA-III explora regiões inovadoras do espaço químico. Ajustes nos Hiperparâmetros podem melhorar ambos.
- Afinidade e seletividade preditas foram otimizadas sem dependência de similaridade estrutural.
- Tratar SA e Cx como **restrições** (não objetivos) melhora atividade e seletividade preditas.

## Conclusões gerais

---

- **DockTDeep** é uma função de pontuação de fácil treinamento, robusta a vieses de ligante/proteína e à variância rotacional, além de competitiva com o estado da arte em diversos cenários de avaliação.
- **DockTDesign**, em conjunto com **DockThor** e **DockTDeep**, apresenta-se como uma plataforma promissora e flexível para identificação de *hits*, capaz de sugerir ao especialista um conjunto de moléculas que atendam simultaneamente a múltiplos objetivos.

## Perspectivas

---

- Incorporar **DockTDeep** no portal DockThor (<https://dockthor.lncc.br>).
- Incluir conformações múltiplas (atracamento/dinâmica molecular) para modelar aspectos **dinâmicos** da ligação.
- Avaliar **DockTDeep** em tarefas de otimização de *leads* com dados experimentais e comparar com FEP.
- Usar **metamodelos** para reduzir custo computacional do atracamento na otimização.
- Expandir a abordagem para outros alvos terapêuticos (incluindo o cenário multialvo).
- Colaborar com grupos experimentais para definir objetivos realistas e validar moléculas geradas.
- Desenvolver um **portal web** para a plataforma *DockTDesign*, integrado ao supercomputador Santos Dumont.

# 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.

# Obrigado!



# NSGA-II

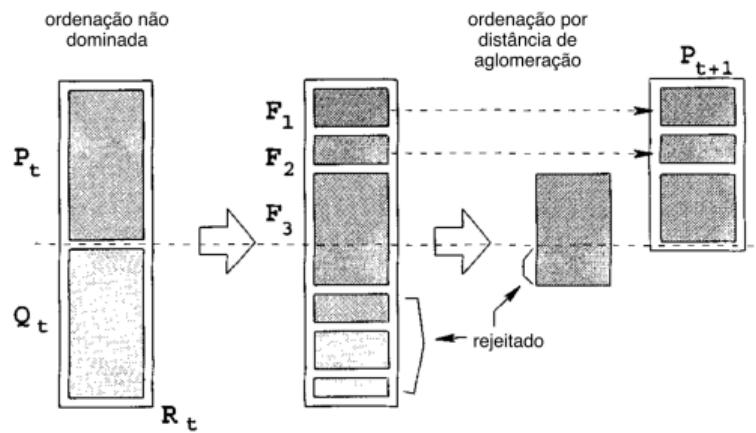


Figura: Algoritmo NSGA-II.

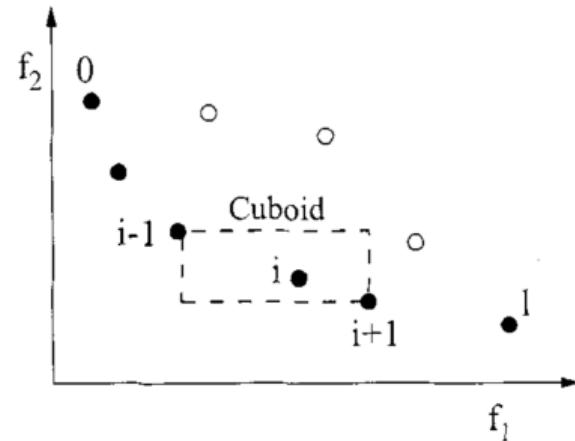


Figura: Distância de agrupamento no NSGA-II.

## NSGA-III: direções de referência

---

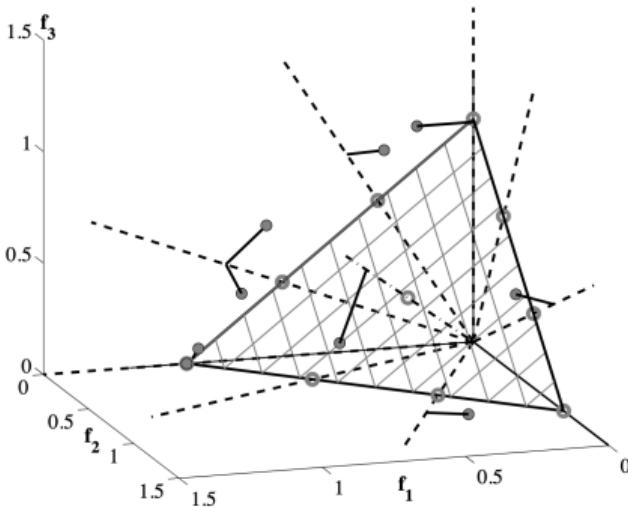


Figura: Direções de referência no NSGA-III.

# *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) ↑

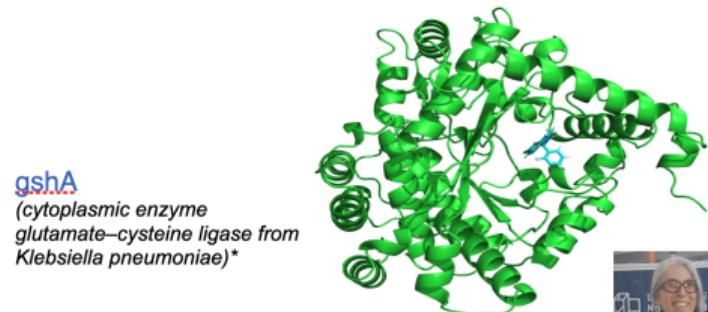
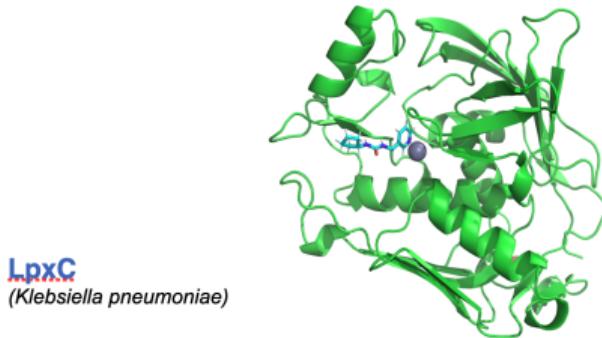
## Protocol II (NSGA-III) – Multi Target Scenario

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>

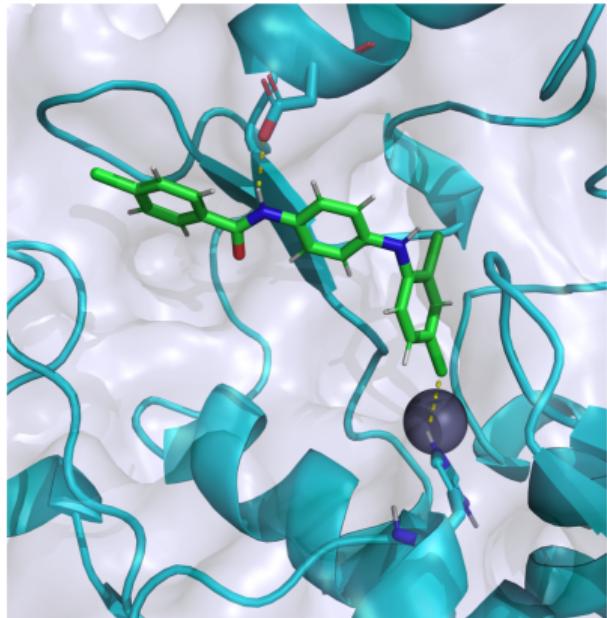


\* An integrative, multi-omics approach towards the prioritization of *Klebsiella pneumoniae* drug targets  
Scientific Reports 2018 Jul 17;8(1):10755. doi: 10.1038/s41598-018-28916-7.

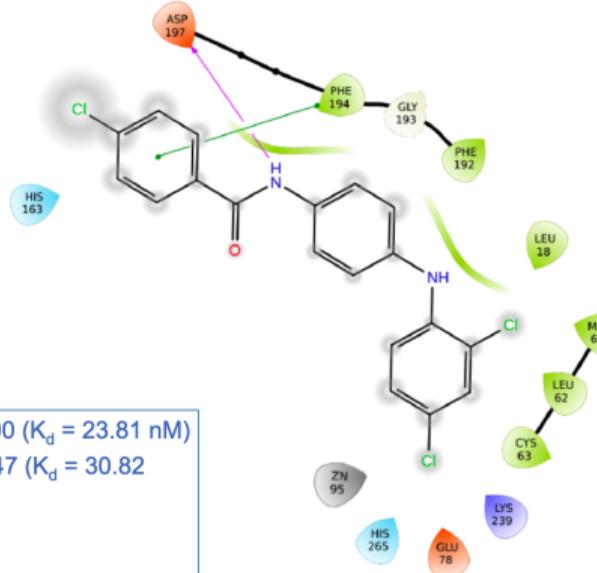


Marisa Nicolás  
Bioinformatics

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

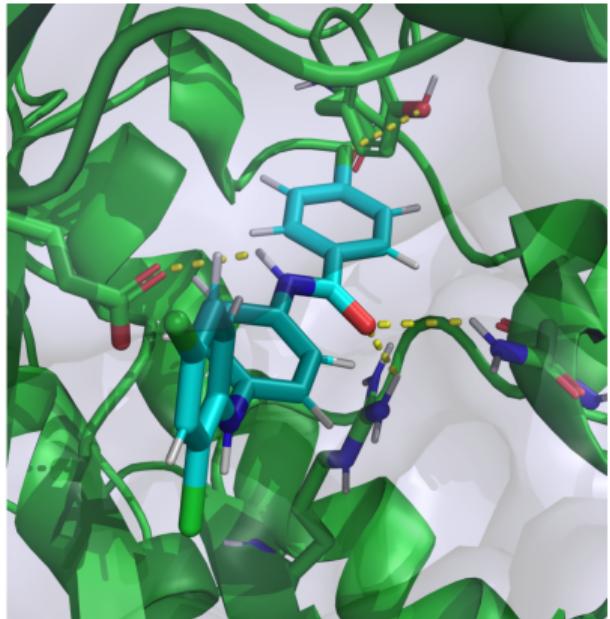


LpxC ( $\Delta G$ ) = -10.400 ( $K_d = 23.81$  nM)  
GshA ( $\Delta G$ ) = -10.247 ( $K_d = 30.82$  nM)  
QED = 0.523  
SA = 1.661  
LE (LpxC) = 0.416  
LE (GshA)) = 0.410



- 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
- Pi-cation
- Salt bridge
- Solvent exposure

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



gshA

