Geradores de homologia persistente e aplicações

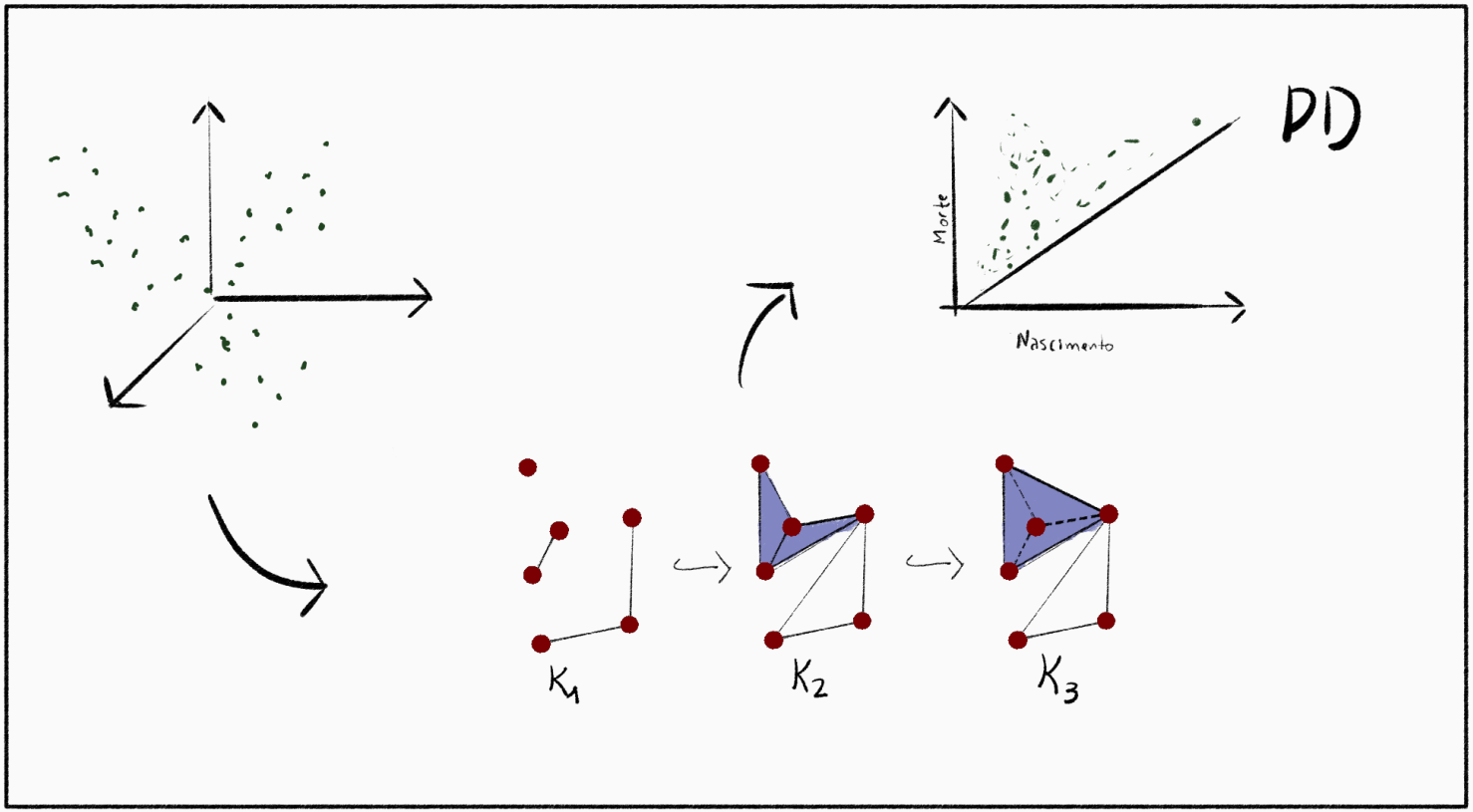
Carlos Ronchi
Marcio Gameiro

13 de Novembro de 2019

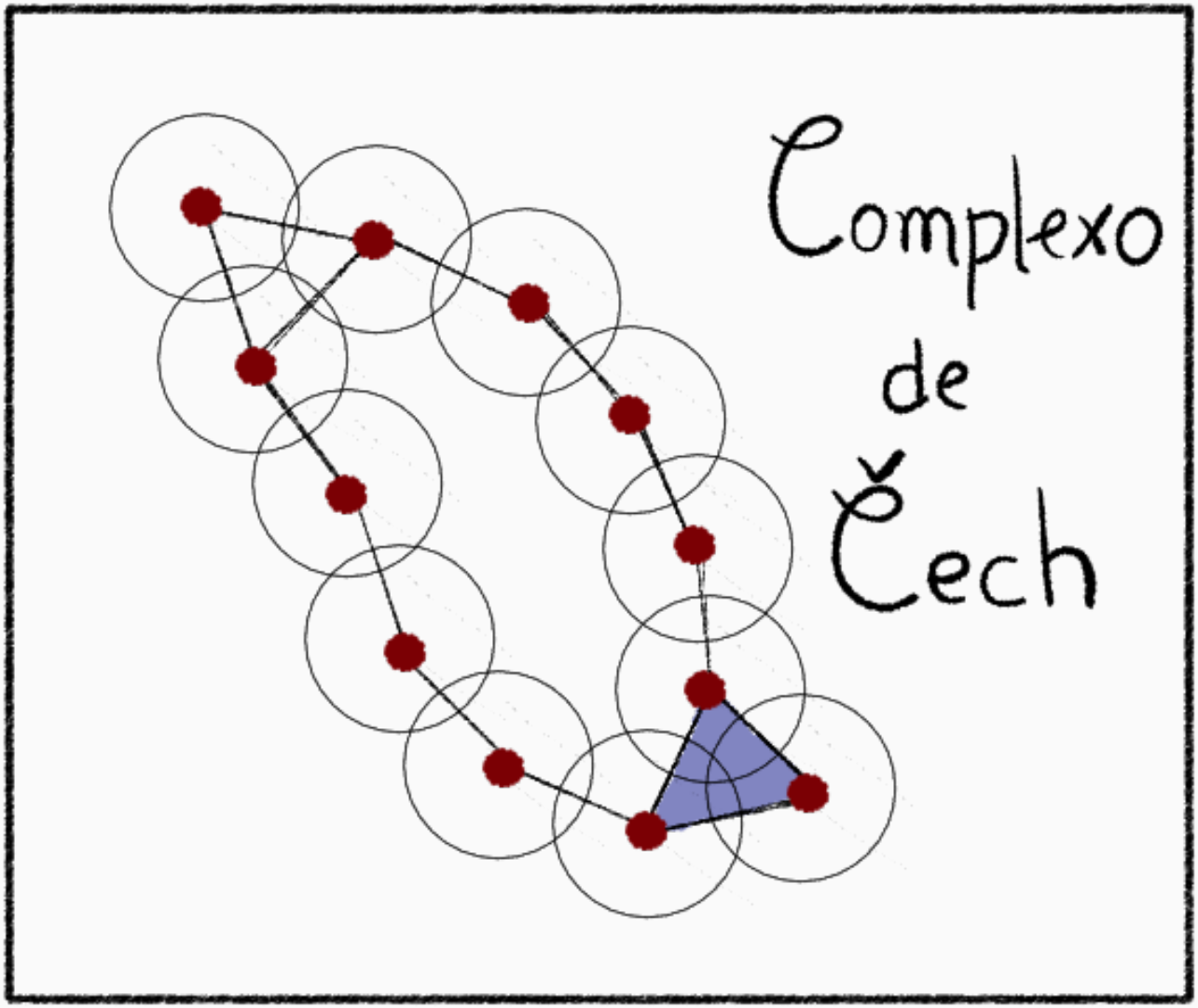
Universidade de São Paulo

Introdução à homologia persistente

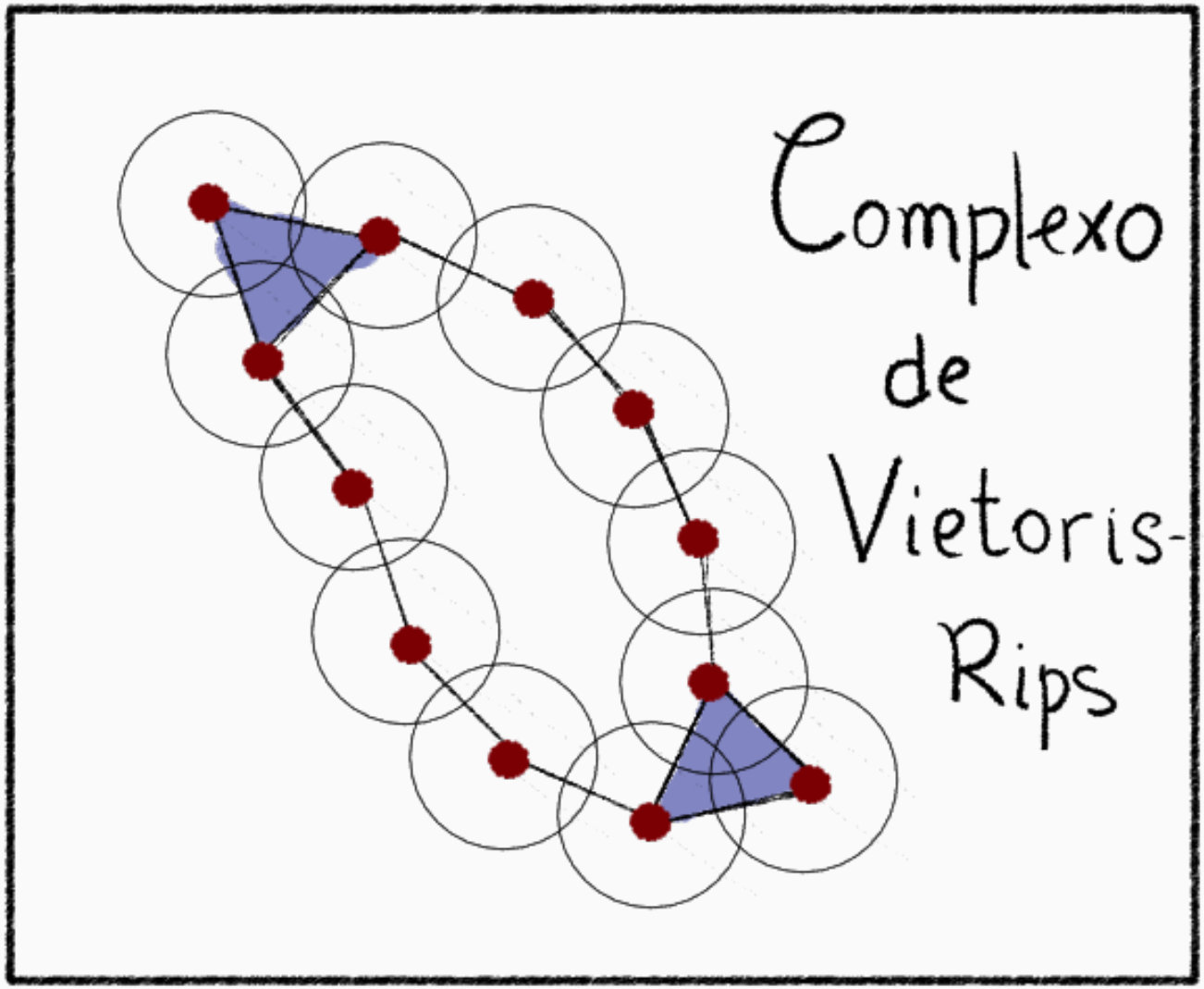
Dos dados à homologia persistente



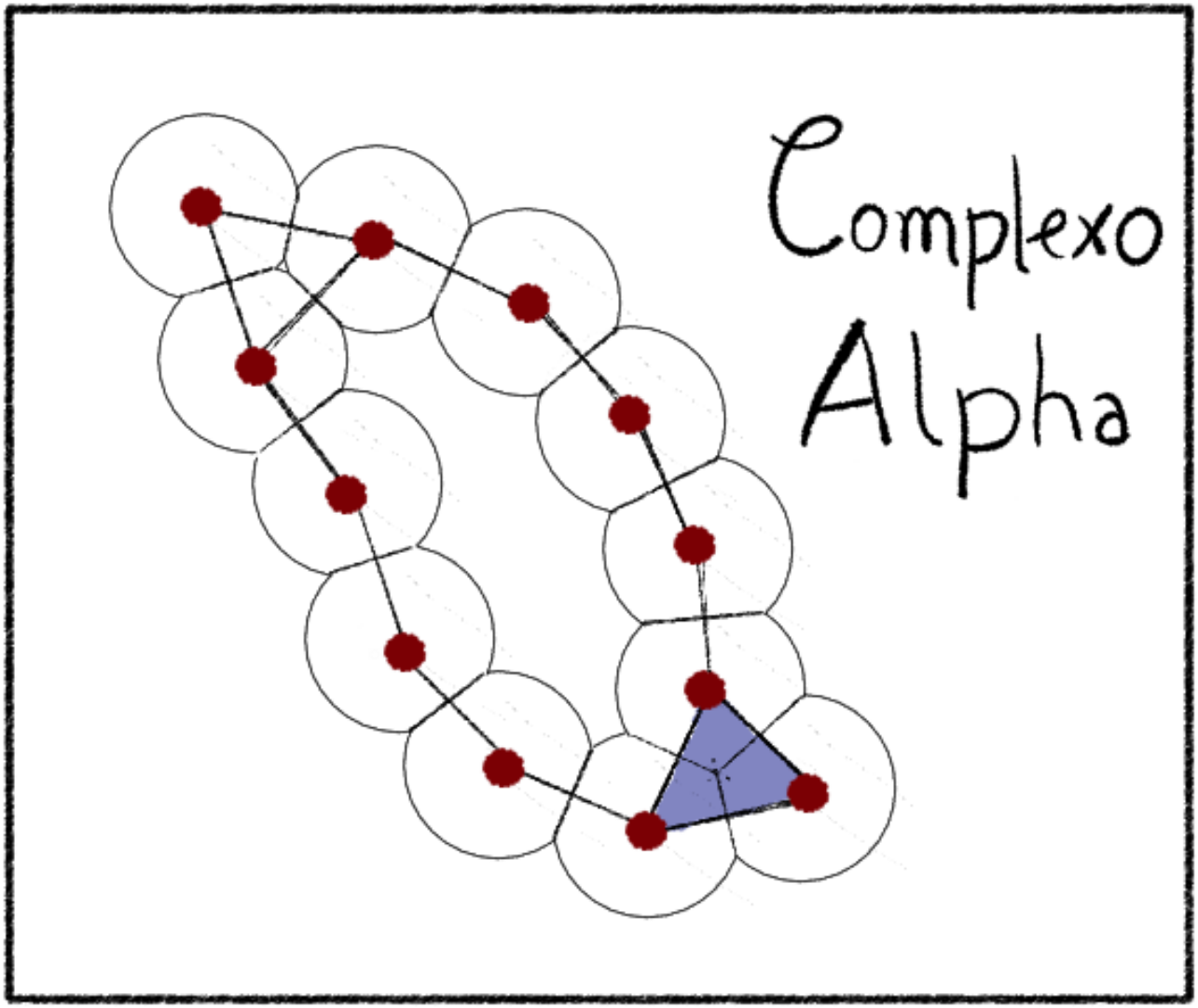
Primeira etapa - o complexo de Čech



Primeira etapa - o complexo de Vietoris-Rips



Primeira etapa - o complexo Alpha



Como os complexos se relacionam?

Seja $r > 0$ fixado, então

- $C^r(X) \subset V^r(X) \subset C^{2r}(X)$
- $A^r(X) \subset C^r(X)$

Dos complexos para a filtração

Para um complexo simplicial K fixado, definimos sua filtração como

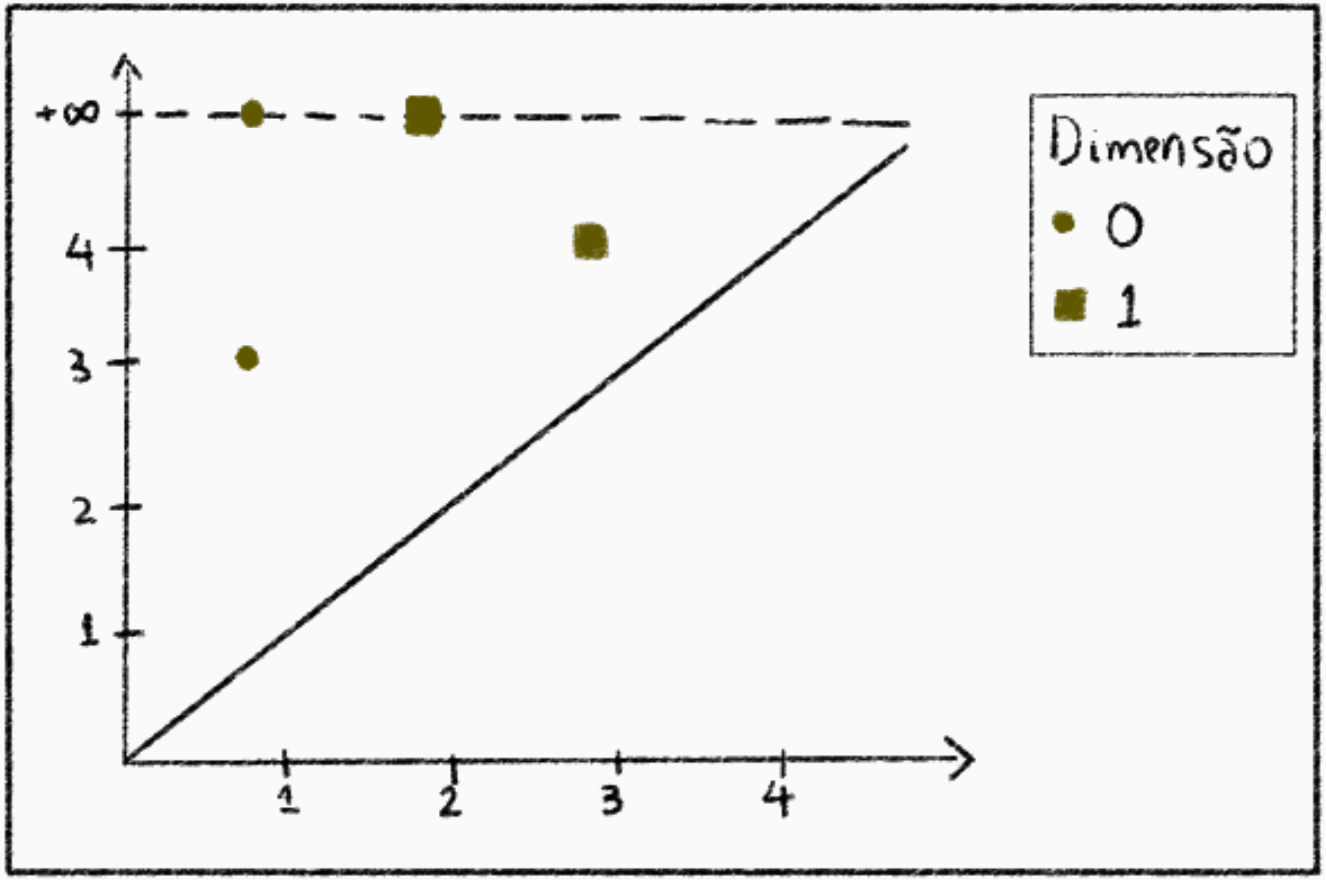
$$K_1 \subset K_2 \subset \cdots \subset K_n = K,$$

em que K_i é um complexo simplicial, para todo $i \in \{1, \ldots, n\}$.

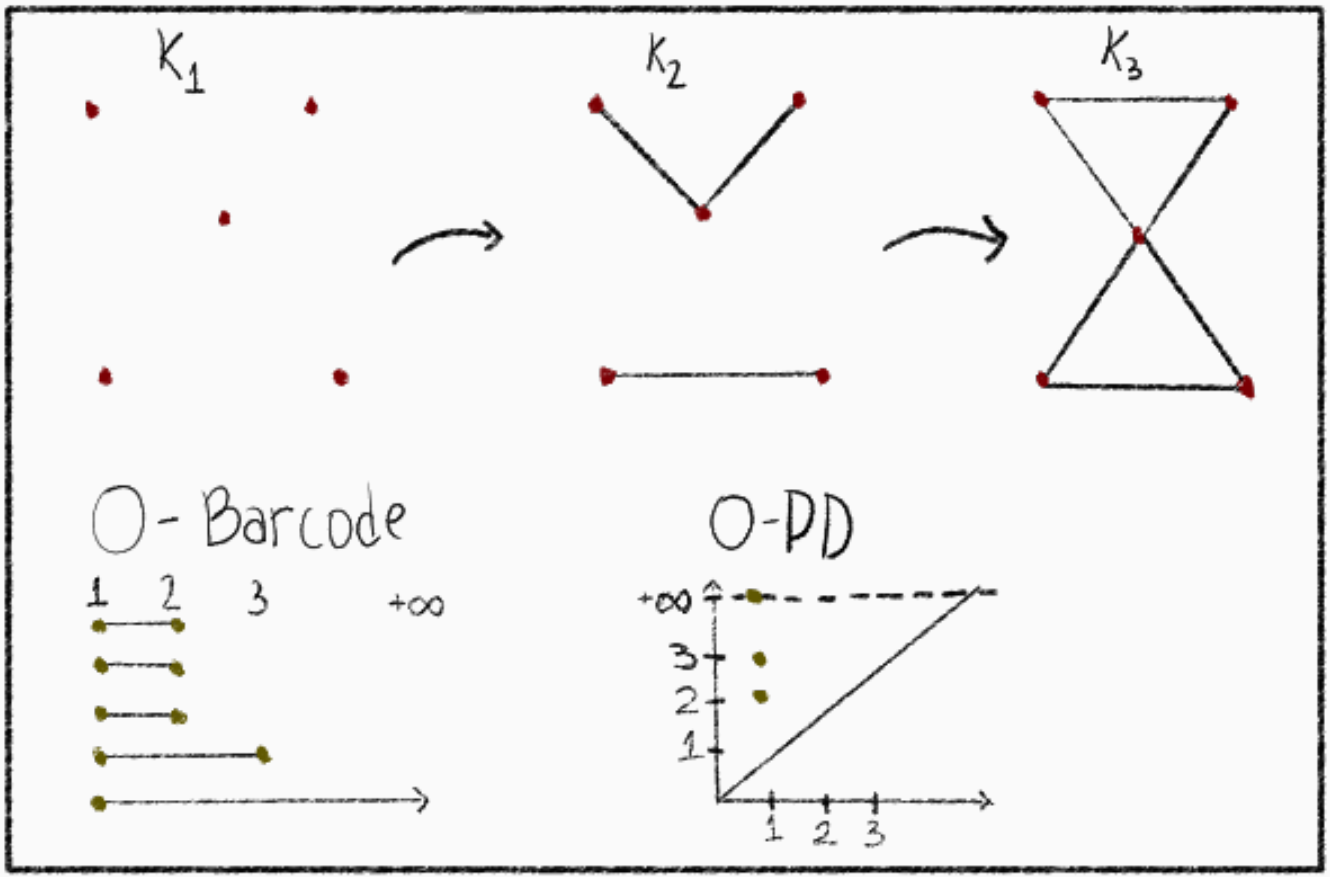
Extraindo informações topológicas da filtração

$$\text{Filtração} \to (H_m(K_1) \to H_m(K_2) \to \cdots \to H_m(K_n))$$

A representação gráfica da homologia persistente



Da filtração às propriedades topológicas



Módulos de persistência

Como podemos formalizar esse processo e garantir a existência do diagrama de persistência?

A formalização - O módulo de persistência

- Seja $T = \mathbb{R}$ um poset e $(V_t)_{t \in T}$ uma sequência de espaços vetoriais
- Sejam $v_t^s \colon V_s \to V_t$, $s, t \in T$ tais que

$$v_r^t \circ v_t^s = v_r^s.$$

A base dos módulos de persistência - Módulos intervalares

Seja $J \subset T$ um intervalo, o módulo intervalar $\mathfrak{I} = \mathbf{k}^J$ é o $\mathbf{T}$-módulo de persistência com os espaços vetoriais

$$I_t = \begin{cases} \mathbf{k} \text{ se } t \in J \\ 0 \text{ caso contrário,} \end{cases}$$

e as aplicações lineares

$$i_t^s = \begin{cases} id \text{ se } s, t \in J \\ 0 \text{ caso contrário.} \end{cases}$$

Sob quais condições temos que o módulo de persistência é decomponível?

Da prática para a teoria - O teorema de Gabriel

Teorema

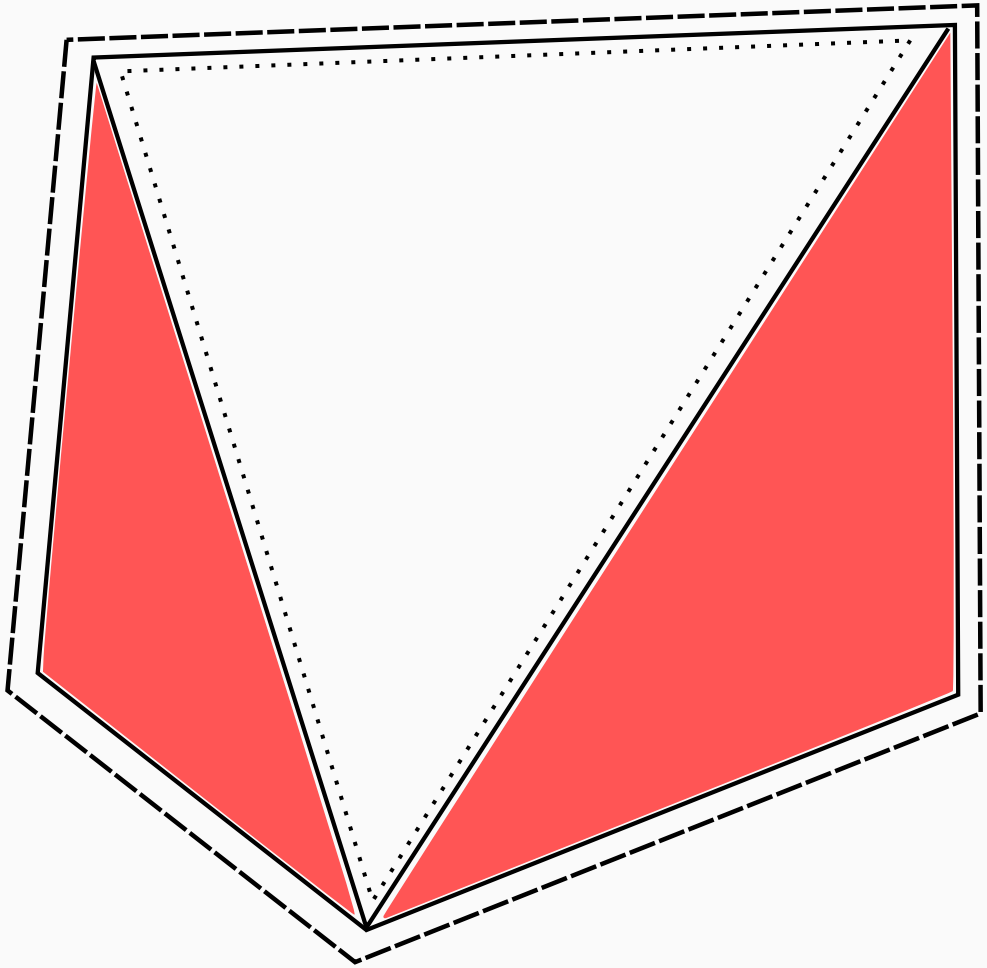
(Gabriel, Auslander, Ringel-Tachikawa, Webb, Crawley-Boevey) Seja $\mathfrak{V}$ um módulo de persistência sobre $\mathbf{T} \subset \mathbb{R}$. Então $\mathfrak{V}$ pode ser decomposto como um soma direta de módulos intervalares sob as seguintes condições:

- $\mathbf{T}$ *é um conjunto finito;*
- *cada* V_t *é um espaço vetorial de dimensão finita.*

Por outro lado, existe um módulo de persistência sob $\mathbb{Z}$ *que não admite uma decomposição intervalar.*

Geradores ótimos

Quando os ciclos são ótimos?



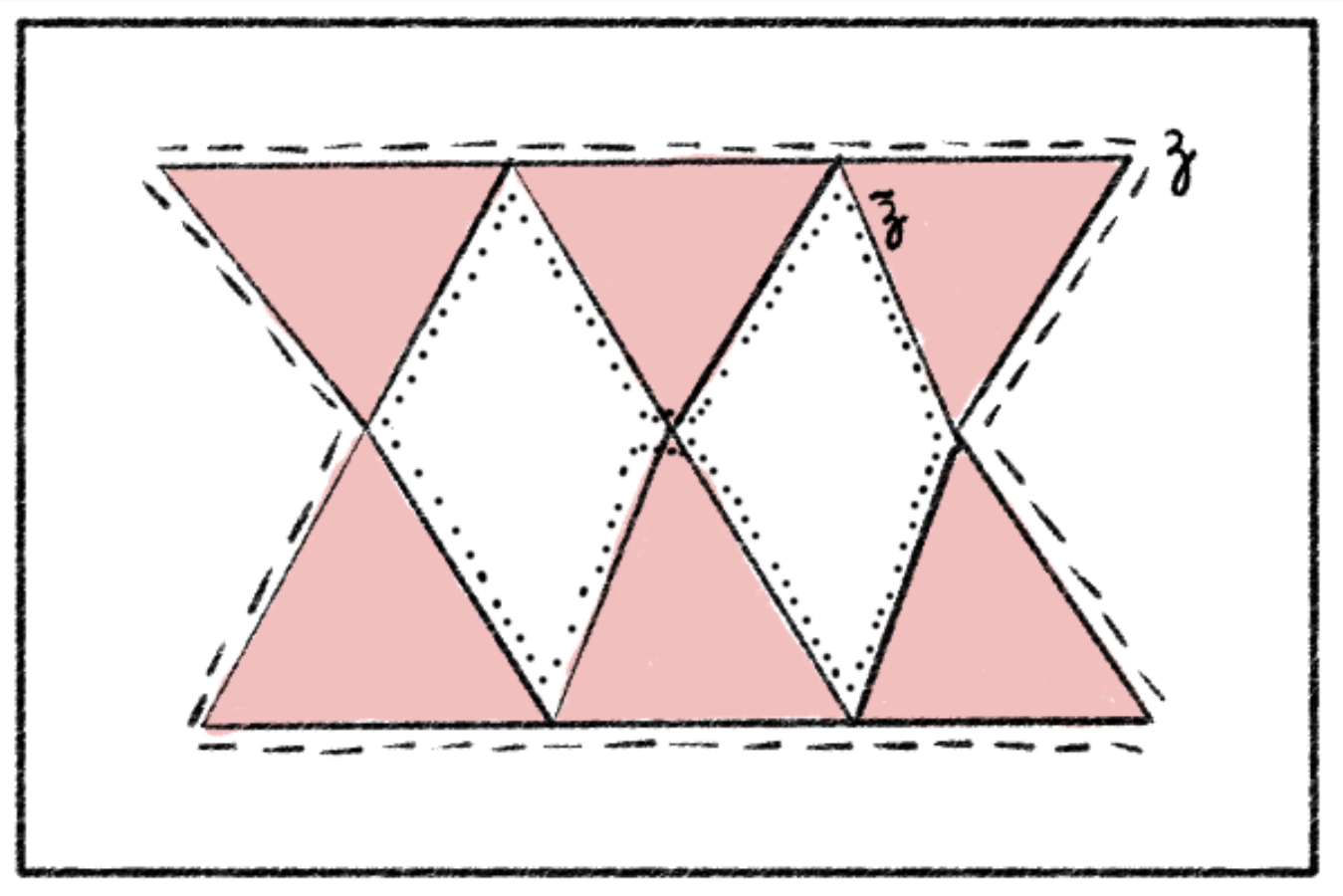
O algoritmo de ciclos ótimos

1: **procedimento** OPTIMIZE_CYCLE($z_j = g_j$)
2: Encontre uma solucao otima $\tilde{z}_j$ para

$$\begin{aligned} &\text{minimize } \|x\|_1 \\ &\text{sujeito a } x + By + \sum_{i \in \mathcal{L}_q(j), i<j} a_i \tilde{z}_i = z_j \end{aligned}$$

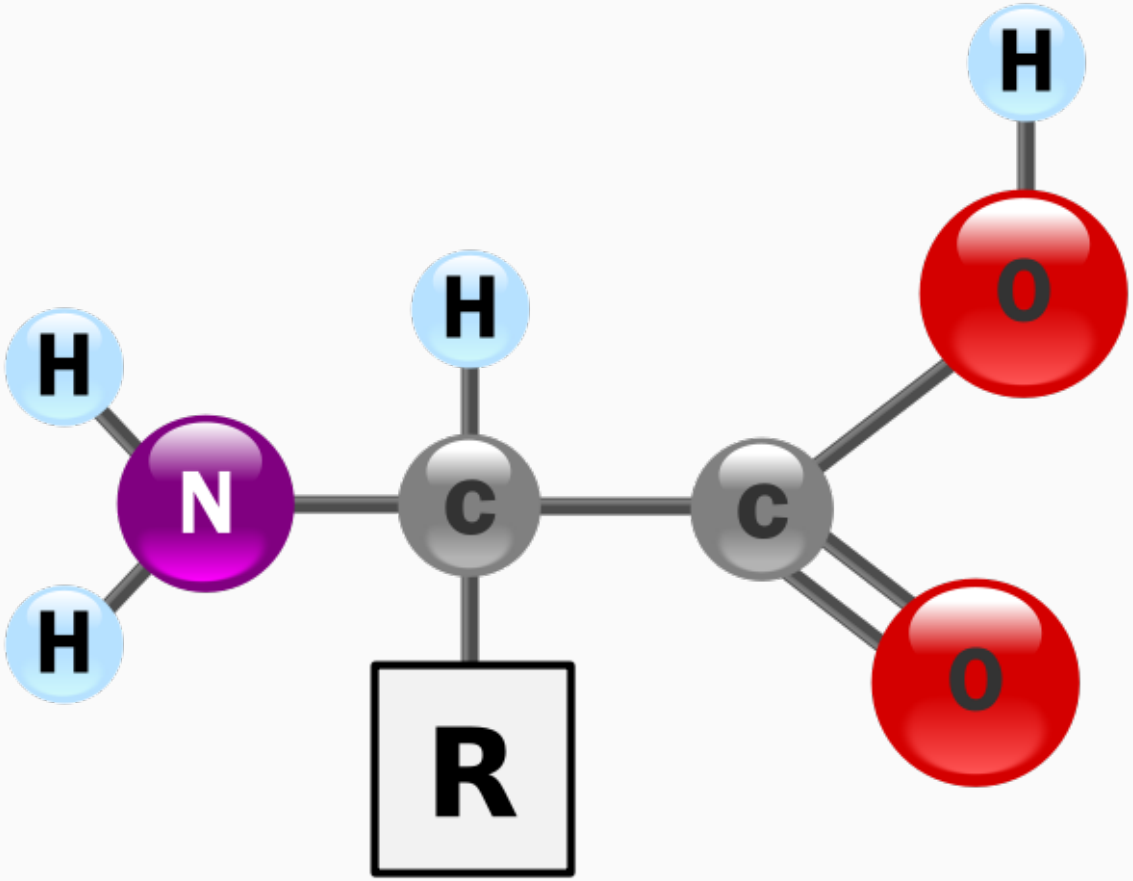
3: **retorna** $\tilde{z}_j$
4: **fim procedimento**

Exemplo de geradores ótimos



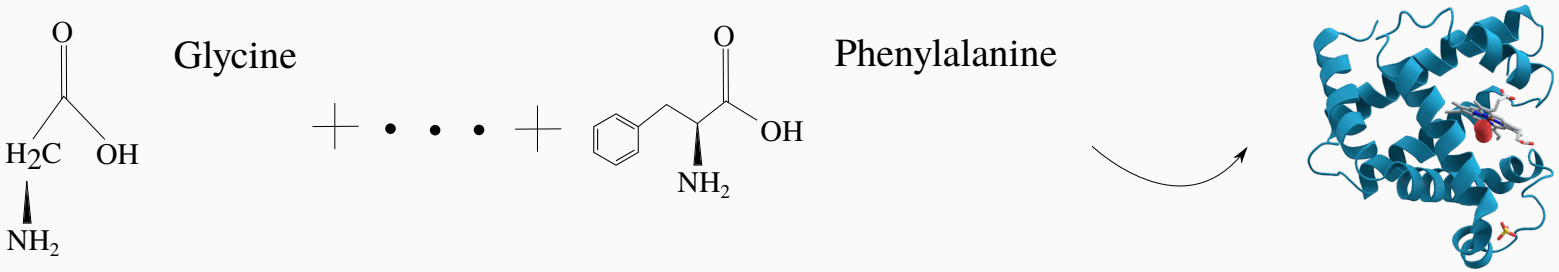
Desenvolvimento computacional de proteínas

Aminoácidos

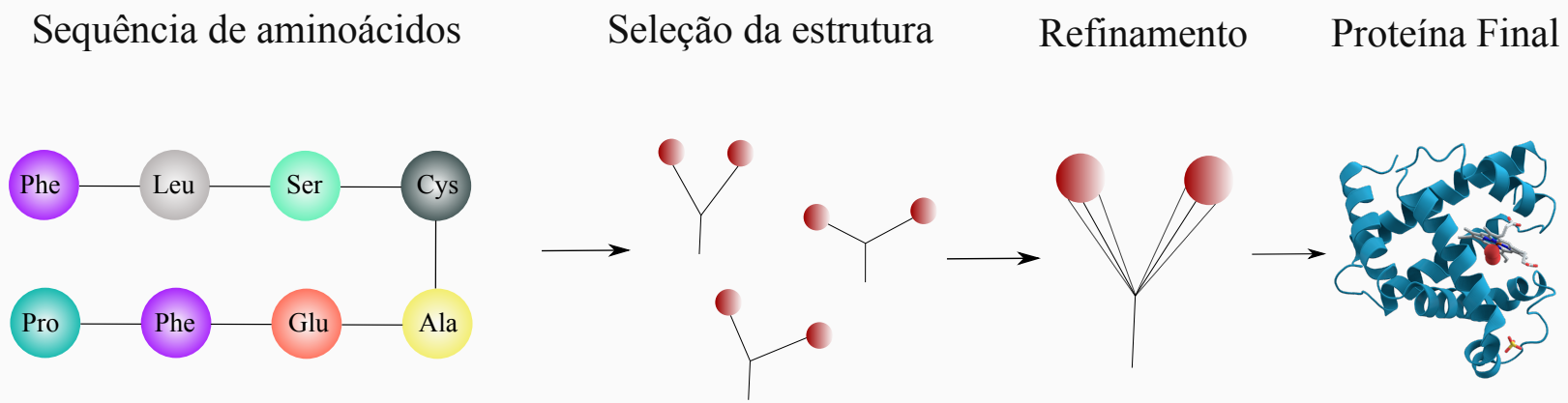


Fonte: Wikimedia Commons

Dos aminoácidos para as proteínas - O processo

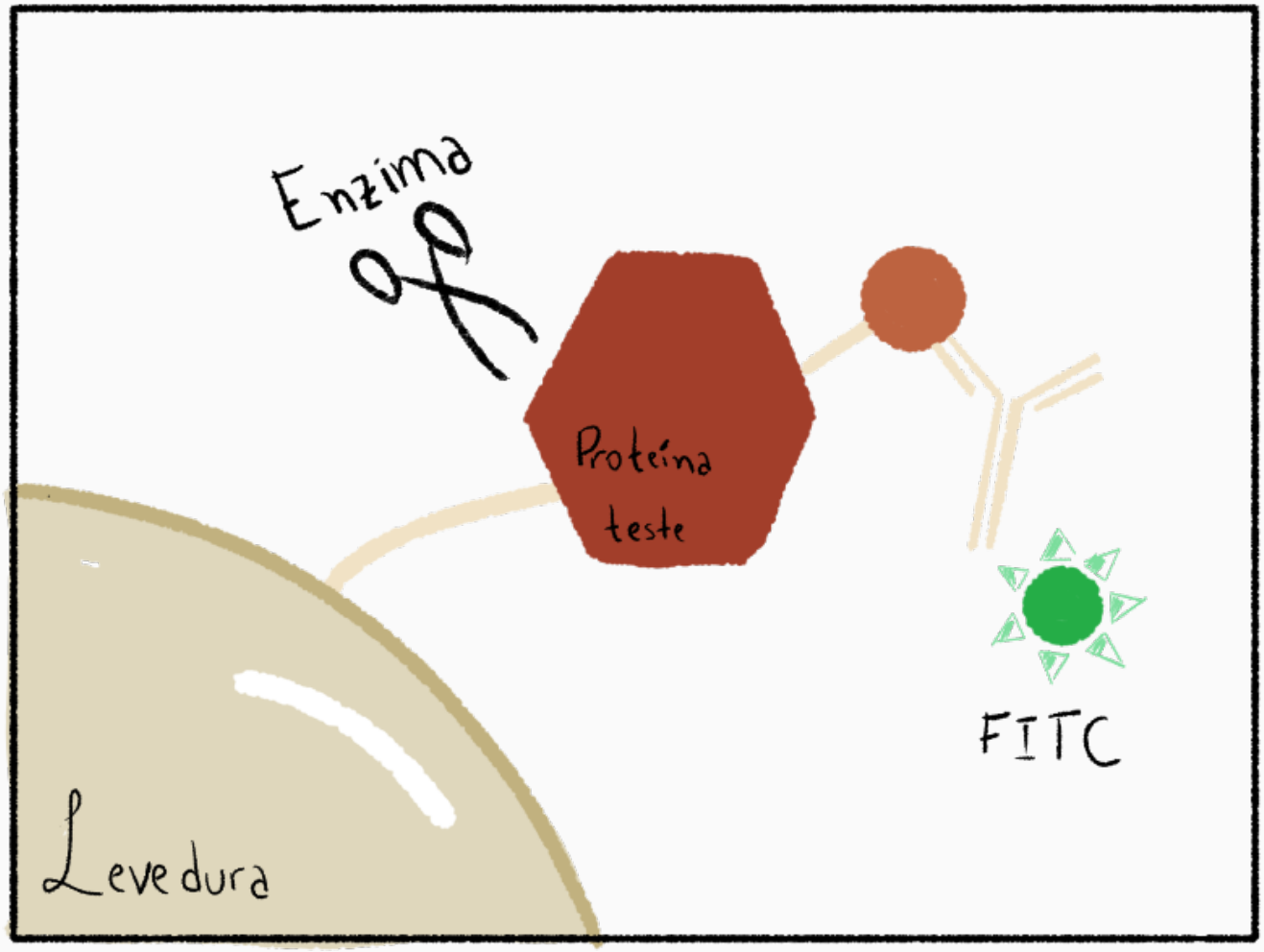


O desenvolvimento de proteínas através de softwares



Experimentos - Rocklin

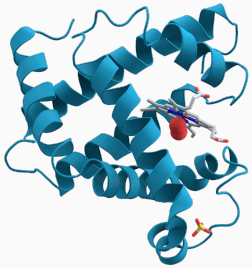
O que é a estabilidade da proteína?



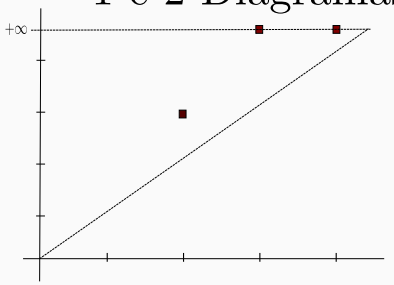
Prevendo a estabilidade - uma abordagem computacional

Modelo	RMSE	Erro percentual (%)
Random Forest	0,419	11,381

Metodologia

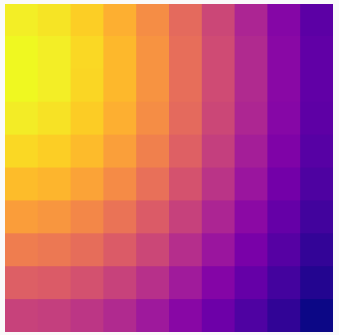


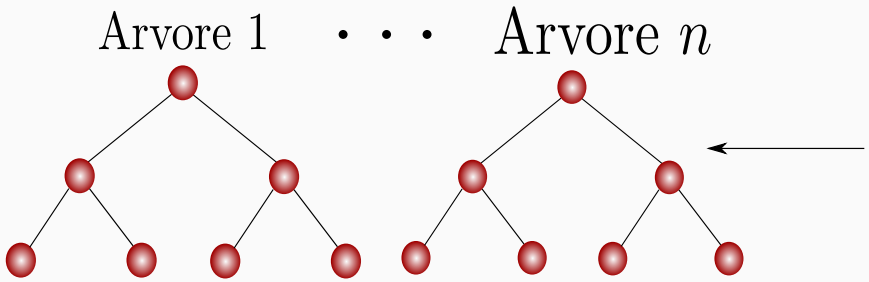
$\simeq 16000$



para cada conjunto de atomos em $\{\{C\}, \{O\}, \{N\}, \{C, O\}, \{C, N\}, \{N, O\}, \{C, N, O\}\}$

Imagens de persistencia





Homologia persistente prevê a estabilidade

Modelo	RMSE	Erro Percentual (%)
Regressão linear	0,5046	13,69
Random Forest I	0,4877	13,24
Random Forest II	0,4874	13,23
GBoost ótimo	0,4770	12,95
Modelo Rocklin [1]	0,419	11,381

Tabela 1: Variância: $0,7$, Grid: 5x4.

[1] Science 14 Jul 2017:
Vol. 357, Issue 6347, pp. 168-175
DOI: 10.1126/science.aan0693

Ciclos unidimensionais de baixa persistência são importantes

Figura 1: Dimensão 1

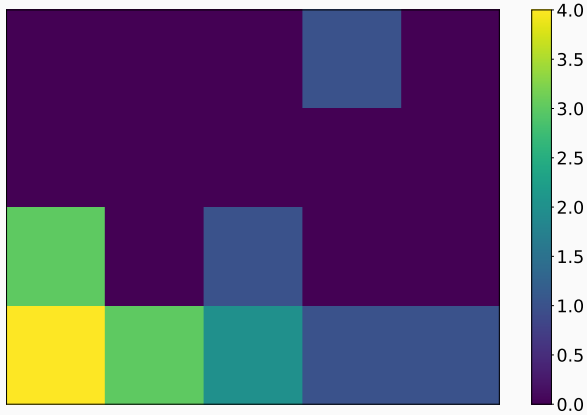
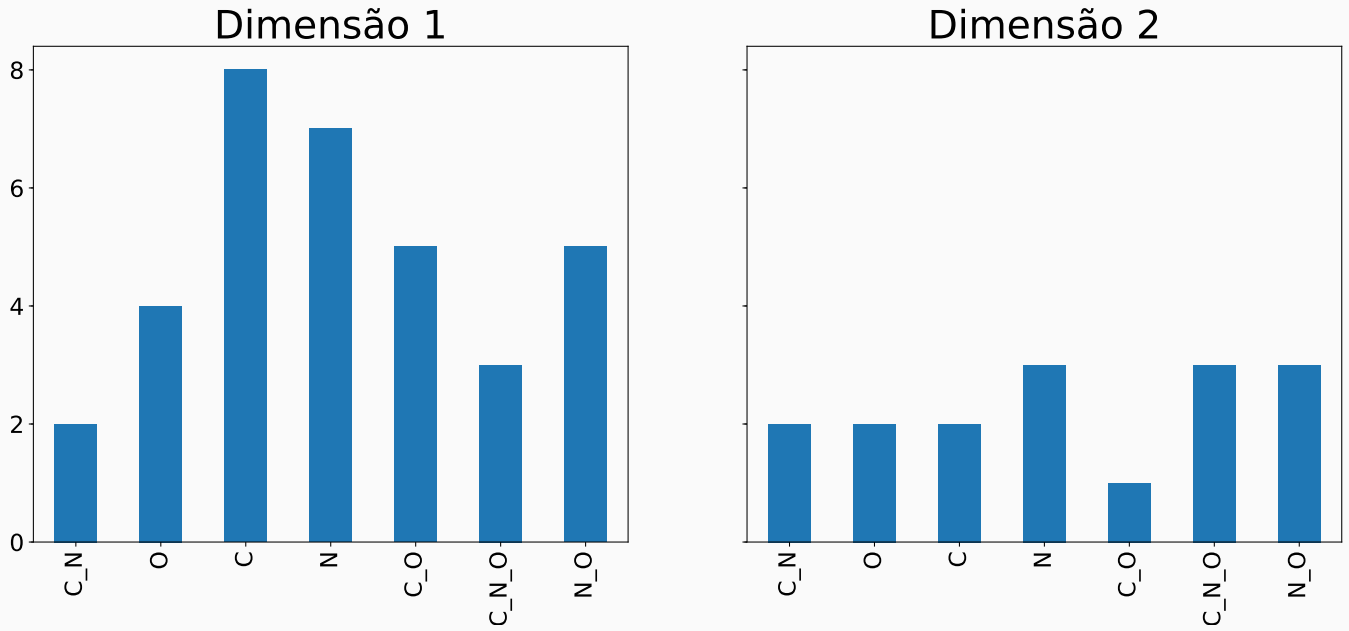


Figura 2: Dimensão 2

Átomos de carbono e nitrogênio são os que mais aparecem no top 50



Experimentos - Sagar

As propriedades das proteínas - Software Rosetta

Rosetta

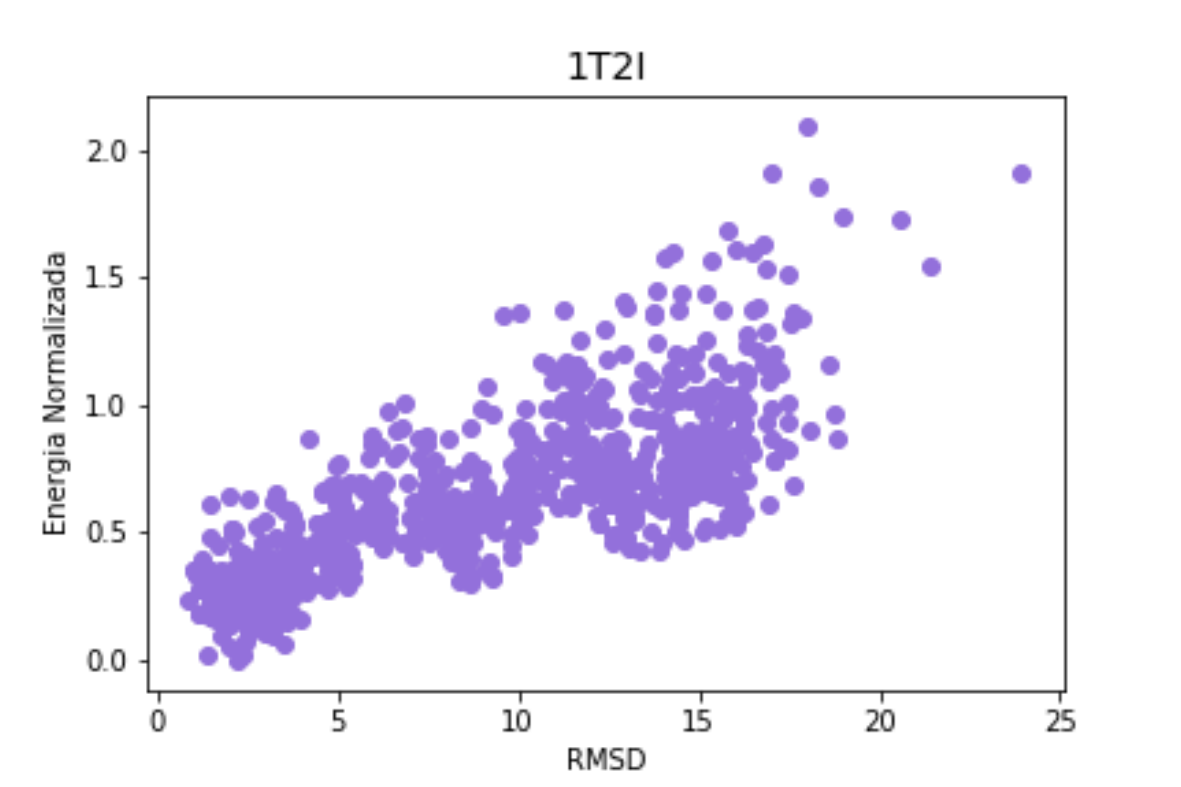
- gera proteínas
- minimiza função de energia

fa_dun, fa_elec, fa_intra_rep, hbond_sc,

fa_rep, fa_sol, hbond_bb_sc, hbond_lr_bb,

hbond_sr_bb, omega, p_aa_pp, pro_close, rama.

O ranking de energia de proteínas simuladas



$$E_{i(norm)} = \frac{E_i - E_{\min}}{E_{95th} - E_{5th}},$$

Como classificar as proteínas em um ranking

Tabela 2: Rank mostrando as top 5 decoys dos experimentos com a proteína 1T2I.

Rank	Energia Normalizada	RMSD
1	0.000	2.233
2	0.023	1.37
3	0.025	2.395
4	0.057	2.004
5	0.061	2.356

Ciclos ótimos e a análise de falsos mínimos

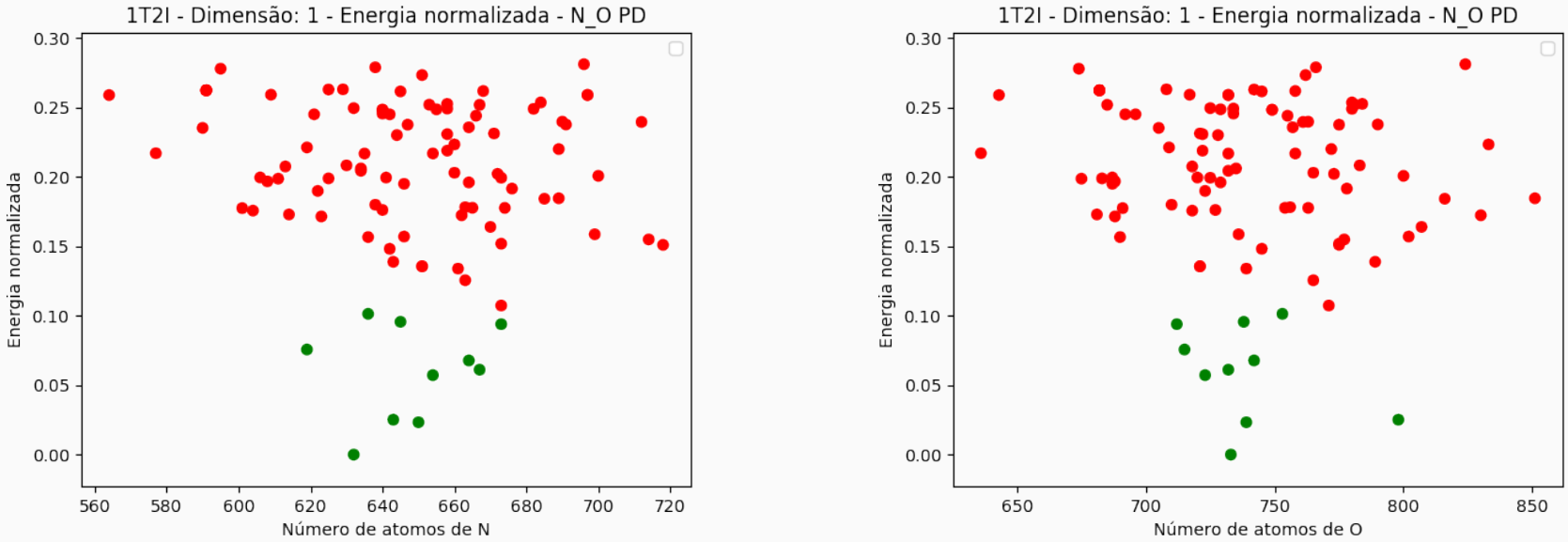


Figura 3: Soma dos átomos de nitrogênio (esquerda) e oxigênio (direita) que compõe os ciclos do 1° diagrama de persistência das decoys da proteína 1T2I.

Ciclos ótimos e a análise de falsos mínimos

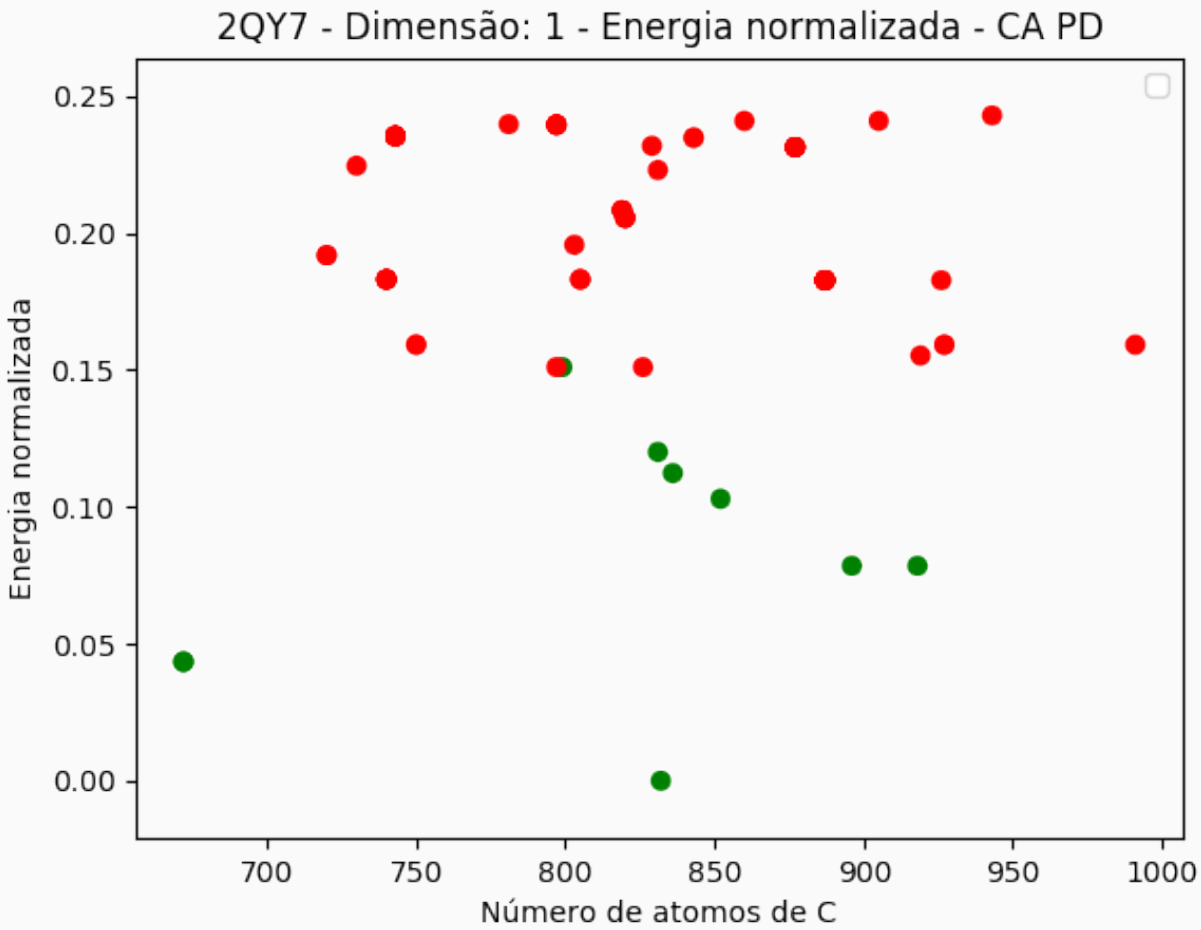


Figura 4: Soma dos átomos de carbono que compõem os ciclos do 1° diagrama de persistência das decoys da 2QY7.

Prevendo o RMSD

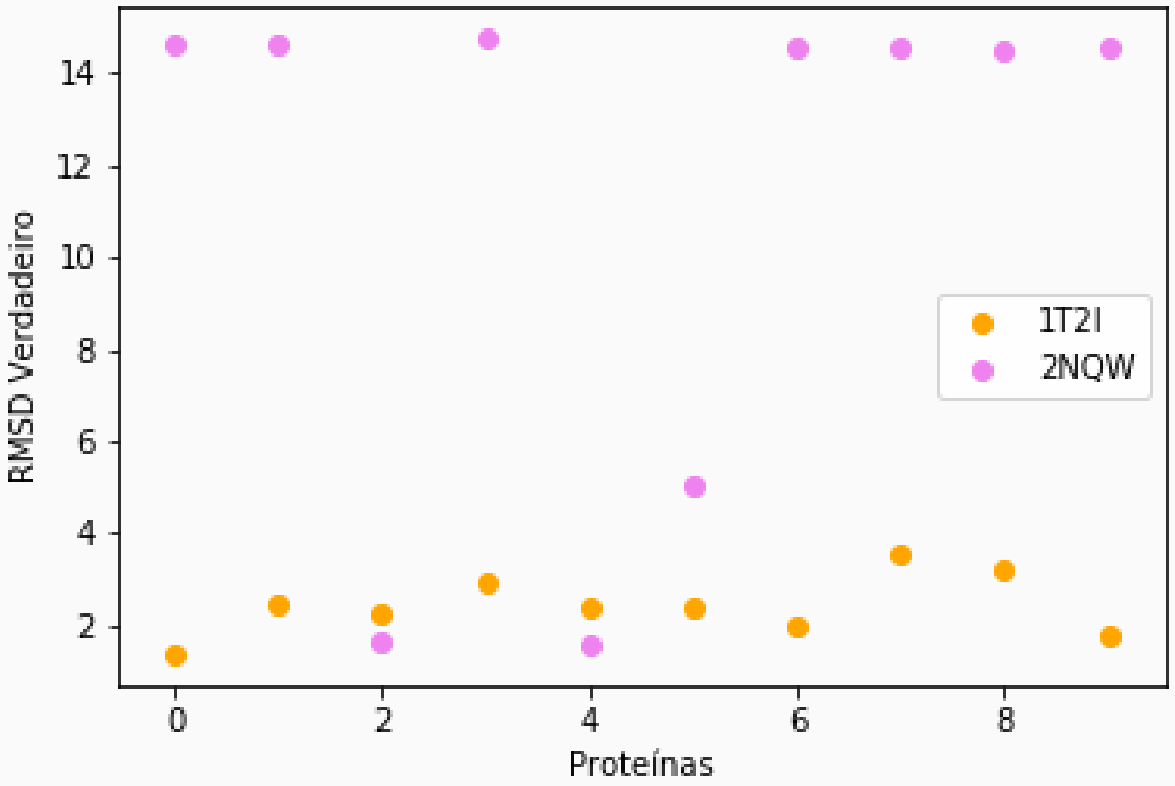


Figura 5: Valor do RMSD para cada decoy no top 10. Não existem falsos mínimos para a proteína 1T2I, enquanto isso existem 7 falsos mínimos para a proteína 2NQW.

Homologia persistente na predição do RMSD

Tabela 3: Melhores parâmetros para cada métrica para os regressores treinados nas imagens de persistência.

Métrica	**Regressor**	**Pixel**	**Var.**	**Átomos**[2]	Score médio
R^2	Redes neurais	100	1,0	C	$-5,780$
MSE	Redes neurais	100	1,0	C	$8,299$
RMSE	Reg. lin c/ reg.	10	1,2	todo	$2,599$
Acur. Bin.	GBoost	10	0,6	N,O	$0,657$

[2] Átomos utilizados para calcular os diagramas de persistência. "todo"significa que todos os átomos menos os de hidrogênio foram usados para os PD's.

As propriedades das proteínas na predição do RMSD

Tabela 4: Melhores regressores treinados com as propriedades das proteínas

Métrica	Regressor	Score médio
R^2	Random Forest II	-13,706
MSE	Random Forest II	10,113
RMSE	Random Forest II	2,707
Acurácia binária	Regressão lin. c/ reg.	0,586

Um comparativo entre homologia persistente e as propriedades de proteínas

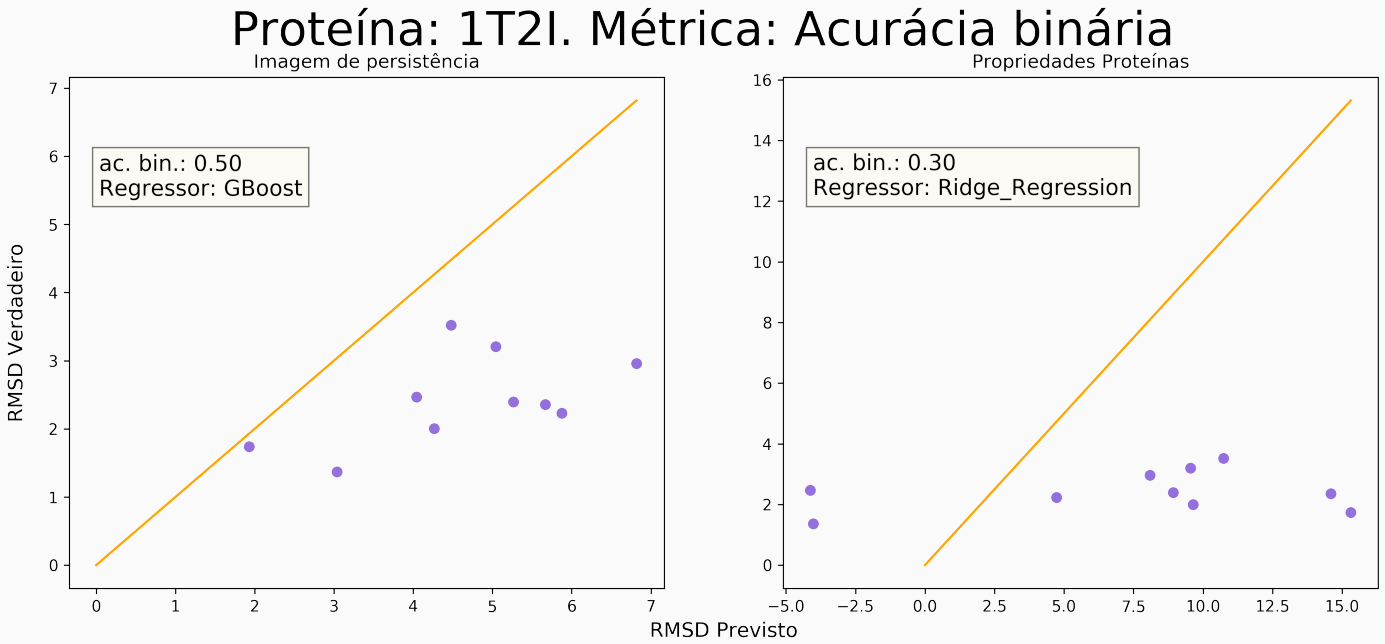


Figura 6: RMSD previsto x RMSD verdadeiro para o top 10 decoys da proteína 1T2I dados os regressores com a melhor acurácia binária no conjunto de validação. Na esquerda os valores para regressores treinados com PI's, já na esquerda com propriedades das proteínas.

Conclusão

Conclusão

- Relação entre hidrofobicidade e as propriedades mais importantes do modelo;
- Modelo proposto nos dá informações geométricas da proteína;
- O modelo pode ser expandido para outras condições da proteína, como outros conjuntos de átomos;
- É possível alterar a definição de ciclos ótimos para outros tipos de problemas.

Agradecimentos

- Marcio Gameiro
- Konstantin Mischaikow
- Lun Zhang
- Priscila Cavassin
- E a todos os amigos

Referências

Referências

[1] Zixuan Cang and Guo-Wei Wei. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLOS Computational Biology*, 13(7):e1005690, July 2017. doi: 10.1371/journal.pcbi.1005690. URL https://doi.org/10.1371/journal.pcbi.1005690.

[2] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. Springer International Publishing, 2016. doi: 10.1007/978-3-319-42545-0. URL https://doi.org/10.1007/978-3-319-42545-0.

Referências II

[3] Emerson G. Escolar and Yasuaki Hiraoka. Optimal cycles for persistent homology via linear programming. In *Optimization in the Real World*, pages 79–96. Springer Japan, sep 2015. doi: 10.1007/978-4-431-55420-2_5. URL https://doi.org/10.1007/978-4-431-55420-2_5.

[4] Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347): 168–175, jul 2017. doi: 10.1126/science.aan0693. URL https://doi.org/10.1126/science.aan0693.

Referências III

[5] Aliza Rubenstein, Kristin Blacklock, Hai Nguyen, David Case, and Sagar Khare. Systematic comparison of amber and rosetta energy functions for protein structure evaluation. 2018. doi: 10.26434/chemrxiv.5314828.v2.