

Anotação de Funções Proteicas Utilizando Redes de Interações Proteína-Proteína

Exame de Qualificação de Mestrado

Candidato: Yeonatan Mauhnoom

Orientador: Prof. Dr. Zanoni Dias

Coorientador: Prof. Dr. Hélio Pedrini

23 de Abril de 2025

Instituto de Computação (IC)

Universidade Estadual de Campinas (Unicamp)

Introdução

Fundamentação Teórica

Trabalhos Correlatos

Material e Métodos

Resultados Iniciais

Próximos Passos

Introdução

- Proteínas são macromoléculas essenciais para a vida: catálise, estrutura, transporte, sinalização, etc.
- Grande volume de proteínas sequenciadas, entretanto, a anotação funcional ainda é lenta e custosa.
- Necessidade de métodos computacionais baseados em **Aprendizado de Máquina** para acompanhar essa expansão.
- Redes de Interações Proteína-Proteína (**PPI**) podem fornecer informações valiosas ao considerar que proteínas que interagem tendem a ter funções semelhantes ou em comum.

Desenvolver abordagens de aprendizado de máquina que utilizem redes de interação PPI, explorando a estrutura topológica dos grafos para aprimorar a predição de funções proteicas.

Objetivos Específicos

1. Realizar revisão bibliográfica e estudo das principais técnicas de predição funcional que utilizam redes PPI;
2. Desenvolver e testar modelos iniciais, investigando configurações mais simples e estabelecendo uma referência de desempenho;
3. Explorar diferentes representações das redes de interações, como dados tabulares, representações visuais e abordagens baseadas em grafos;
4. Investigar a combinação de técnicas de aprendizado autossupervisionado;
5. Comparar e validar as abordagens propostas em relação a métodos de referência que utilizam redes PPI;
6. Integrar o melhor modelo obtido ao servidor do laboratório, a fim de complementar, o método baseado em sequência desenvolvido pelo grupo;
7. Documentar e publicar os resultados obtidos.

Questões de Pesquisa

- Modelos baseados em grafos podem apresentar vantagens consideráveis em relação a outras representações, ao explorar as relações topológicas inerentes às redes PPI?
- A aplicação de técnicas de aprendizado autossupervisionado (como reconstrução e/ou estratégias contrastivas em *autoencoders*) pode produzir *embeddings* mais robustos, resultando em ganhos de desempenho?
- O uso de *transfer learning* em modelos pré-treinados, adaptados às imagens ou representações derivadas das PPI, é efetivo ou acarreta perdas decorrentes da geração de tais representações visuais das redes?
- A integração dos modelos gerados com base em *embeddings* de sequências pode aprimorar a predição de anotações funcionais ao combinar informações de interações proteína-proteína e dados de sequência?

Fundamentação Teórica

Proteínas

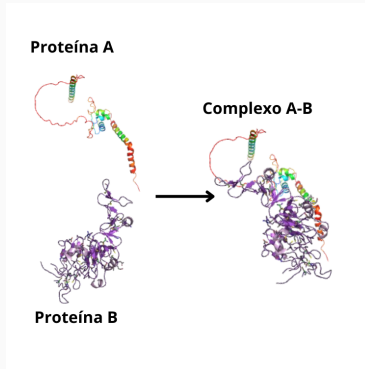


Figura 1: Complexo proteico formado pela interação de duas proteínas.

- Macromoléculas de aminoácidos (ligações peptídicas).
- Estrutura e função dependem da sequência e das interações químicas [1].
- Papéis fundamentais: catálise, transporte, regulação e suporte estrutural [2].

Gene Ontology (GO)

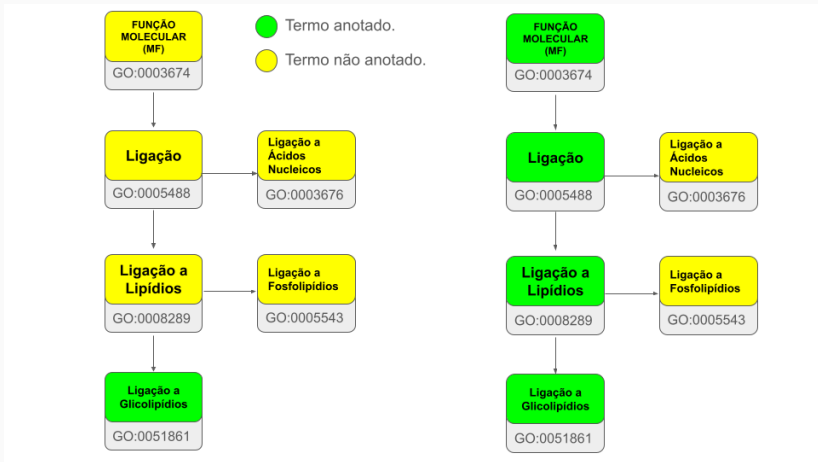


Figura 2: Exemplo de DAG da GO, em que termos específicos (filhos) herdam todos os termos ancestrais.

Gene Ontology (GO)

- **GO**: vocabulário controlado para anotar funções de genes/proteínas [3].
- Três domínios: Função Molecular (MF), Processo Biológico (BP) e Componente Celular (CC).
- Estrutura em grafo acíclico direcionado (DAG).
- *True Path Rule*: anotar um termo filho implica anotar os ancestrais.

Redes de Interação Proteína-Proteína (PPI)

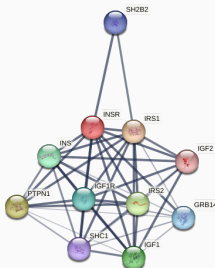


Figura 3: Rede PPI gerada na STRING para a proteína INSR.

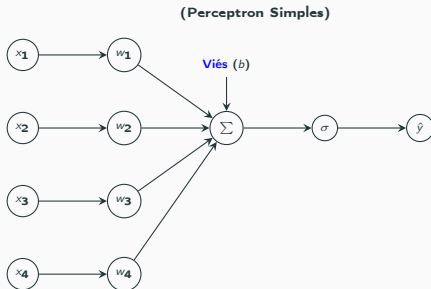
- Interações podem ser físicas, funcionais ou preditas.
- Bases: *BioGRID*, *IntAct*, *STRING*¹.
- Nós representam proteínas, arestas representam relações (coexpressão, experimentos, literatura, etc.) [4, 5].

¹<https://string-db.org>

- **Aprendizado Supervisionado:** modelos treinados a partir de exemplos rotulados.
 - Classificação ou Regressão.
 - Função de perda minimizada (via *backpropagation*), visando boa generalização.
- **Classificação Multirrótulo:** cada amostra pode ter múltiplos rótulos [6, 7]. Exemplo: várias anotações GO numa mesma proteína.

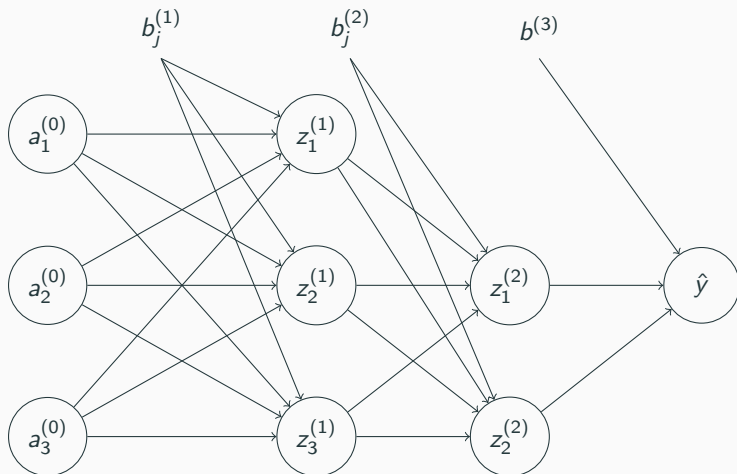
Perceptron Simples

- Combinação linear + ativação.
- Unidade básica em RNAs.
- Separa dados que podem ser distinguidos por uma fronteira linear.
- $\hat{y} = \sigma(w^T x + b)$.



Multi-Layer Perceptrons (MLP)

- Extensão: múltiplas camadas densas (ReLU, Sigmoid, etc.).
- Treino via *backpropagation* e otimização por gradiente [8].
- Capaz de modelar funções complexas, com risco de *overfitting*.



Redes Neurais Convolucionais (CNNs)

- Convolução e *pooling* para extrair padrões hierárquicos [9].
- *Transfer learning*: redes pré-treinadas (por exemplo, EfficientNet) ajustadas em dados específicos.

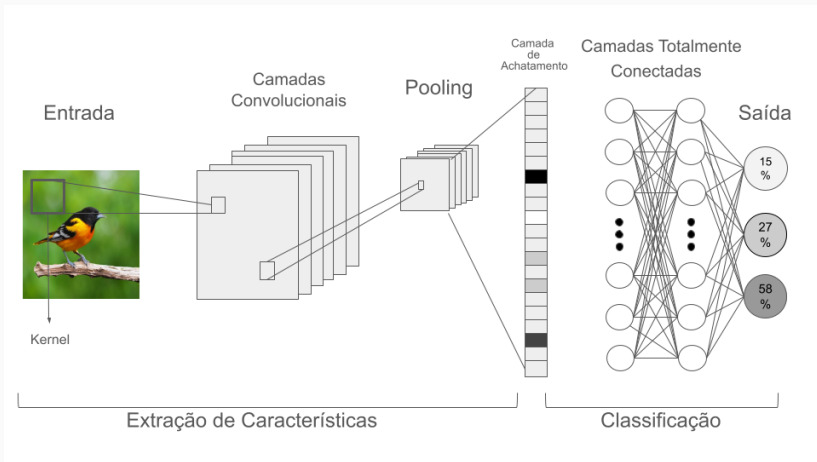


Figura 4: Estrutura genérica de CNN.

Aprendizado Autossupervisionado

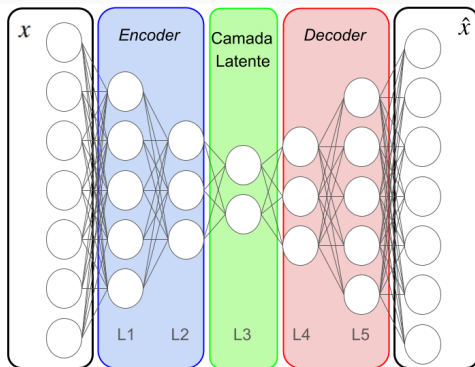


Figura 5: *Encoder* gera representação latente, *decoder* reconstrói.

- Extrai representações ricas de dados não rotulados definindo tarefas auxiliares (*pretext tasks*) [10].
- Exemplos:
 - **Autoencoders:** aprendizado de compressão/reconstrução.
 - **Contrastivo:** maximizar similaridade de pares “positivos” e minimizar de pares “negativos” [11].

Redes Neurais Convolucionais Baseadas em Grafos (GCNNs)

- Convolução em grafos: cada nó agrega info dos vizinhos [12].
- Em PPI: nós representam proteínas e arestas representam interações.
- Captura topologia do grafo (adjacências, caminhos, etc.).

$$h_v^{(l+1)} = \sigma \left(W^{(l)} \cdot \text{AGG} \{ h_u^{(l)} : u \in N(v) \cup \{v\} \} \right).$$

Trabalhos Correlatos

Principais Trabalhos Correlatos

- DeepGraphGO [13]:
 - Usa sequência e PPI em um GCN treinado em múltiplas espécies.
- Graph2GO [14]:
 - Usa PPI e similaridade de sequência via autoencoders variacionais.
 - Concatena as representações para um classificador final de GO.
- DualNetGO [15]:
 - Combina múltiplas PPIs e características de sequência em MLPs.
- deepNF [16]:
 - Funde diversas redes de interação com *autoencoders* profundos.
 - Gera *embeddings* unificados que alimentam classificadores de função.
- CSGNN [17]:
 - Aprendizado contrastivo autossupervisionado em PPI.
 - Emprega um agregador *multihop* para dependências de alta ordem.
- SSLpheno [18]:
 - Integra rede PPI com conhecimento ontológico.
 - Pré-treinamento autossupervisionado contrastivo em PPI.

Material e Métodos

Metodologia Proposta

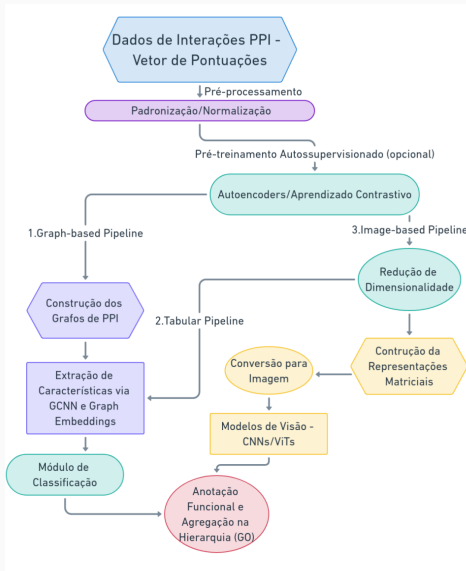


Figura 2: Fluxograma das abordagens (tabular, imagem, grafo).

1. Tabular:

- Vetor $v \in \mathbb{R}^D$.
- Pode usar reduções (*max/avg pooling*) ou *autoencoders* para compactar.

2. Imagem:

- Matriz $A_{ij} = |v_i - v_j|$, interpretada como imagem (1 ou 3 canais).

3. Grafo:

- Nós representam proteínas, arestas representam pontuações da STRING.
- GCNN, Node2Vec e Graph2Vec.

Abordagens Supervisionadas e Autossupervisionadas

Supervisionadas:

- **Tabular:** MLP, Random Forest.
- **Imagem:** CNN, Vision Transformers (e.g., ResNet50 e EfficientNetB0).
- **Grafo:** GCN, GNN, entre outras.

Autossupervisionadas:

- *Autoencoders* para reconstruir vetores de interação.
- Aprendizado contrastivo em grafos.
- *Embeddings* (Node2Vec e Graph2Vec) via caminhadas aleatórias.

Fluxo: Abordagem em Grafos

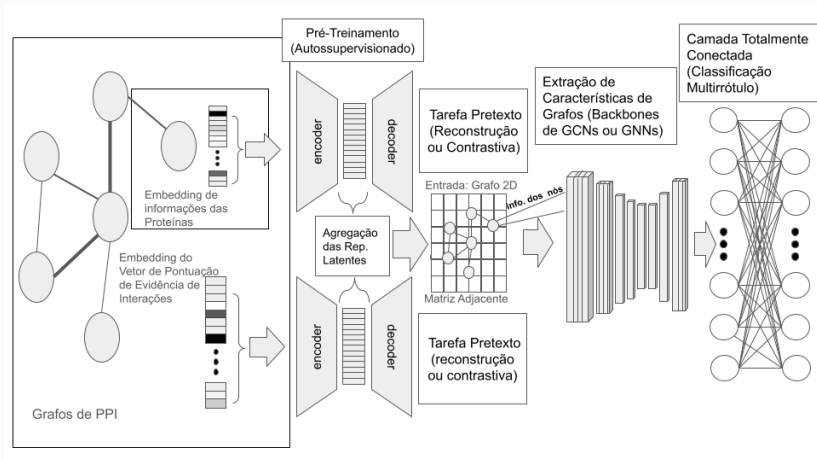


Figura 7: Construção do grafo PPI, aprendido de embeddings com GCNN e classificação final.

- F_{max} : melhor harmonia de precisão e revocação ao longo de limiares.
- S_{min} : custo semântico combinando falsos negativos (RU) e falsos positivos (MI).
- AuPRC e IAuPRC: área sob a curva Precisão-Revocação, com interpolação no caso de IAuPRC.

- **CAFA5²** [19]:

Tabela 1: Número de proteínas e termos.

Ontologia	Treinamento	Validação	Teste	Termos
<i>BP</i>	73.768	9.221	9.221	500
<i>CC</i>	74.328	9.292	9.292	498
<i>MF</i>	62.909	7.864	7.864	499

- Voltada à predição de função proteica.
 - Seleccionados 500 termos BP, 498 termos CC e 499 termos MF.
- **STRING³** [4]:
 - Integra múltiplas fontes de interações PPI.
 - Atribui pontuações de confiança (0 a 1).

²<https://biofunctionprediction.org/cafa>

³<https://string-db.org>

- **Linguagem:** Python.
- **Bibliotecas:** NumPy, PyTorch, scikit-learn e Matplotlib.
- **Infraestrutura:**
 - Google Colab (GPUs T4) para protótipos.
 - CENAPAD⁴ para execuções em larga escala.

⁴<https://www.cenapad.unicamp.br>

Resultados Iniciais

- Teste com duas representações:
 1. Tabular ($v \in \mathbb{R}^D$).
 2. Imagem (matriz de diferenças, $d \times d$).
- **Baseline**: média ponderada simples das anotações dos vizinhos.
- Modelos: MLP (2 camadas) e CNNs (ResNet50, EfficientNetB0).
- Avaliado em BP, CC e MF.

Abordagem Tabular: MLP vs. Baseline

Tabela 2: Resultados com vetor tabular (MLP) vs. baseline. Em **negrito** o melhor valor e sublinhado o segundo melhor.

Método	Ontologia	Fmax	Smin	AuPRC	IAuPRC
Baseline	BP	0,4368	21,9111	0,3977	0,3651
MLP (raw)	BP	<u>0,3969</u>	23,8144	<u>0,3935</u>	<u>0,3921</u>
MLP (max pool)	BP	<u>0,3944</u>	<u>23,4926</u>	0,3773	0,3951
MLP (avg pool)	BP	0,3558	24,7065	0,2775	0,3401
Baseline	CC	0,5759	10,9067	0,5550	0,5029
MLP (raw)	CC	0,5745	<u>11,6893</u>	0,5812	0,6204
MLP (max pool)	CC	0,6620	11,7706	0,6700	0,7056
MLP (avg pool)	CC	<u>0,5997</u>	12,0358	<u>0,5903</u>	<u>0,6352</u>
Baseline	MF	0,5324	8,4366	0,4979	0,4656
MLP (raw)	MF	0,5631	10,4853	<u>0,4789</u>	<u>0,5126</u>
MLP (max pool)	MF	0,6017	<u>8,9388</u>	0,5106	0,5018
MLP (avg pool)	MF	<u>0,5834</u>	9,3573	0,4230	0,5424

Abordagem por Imagem (Diferenças Absolutas)

- Vetor $v \in \mathbb{R}^D$ reduzido a d , gerando matriz $A_{ij} = |v_i - v_j|$.
- CNNs (ResNet50 e EfficientNetB0) com *transfer learning*.

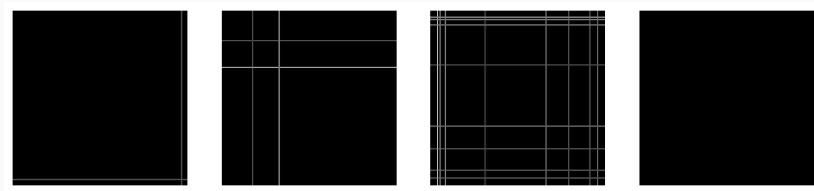


Figura 9: Exemplos de imagens a partir de v .

Resultados: Representação em Imagem

Tabela 3: Resultados com CNNs em imagens. Em **negrito** o melhor valor e sublinhado o segundo melhor.

Método	Ontologia	Fmax	Smin	AuPRC	IAuPRC
Baseline	BP	0,4368	21,9111	0,3977	0,3651
ResNet50	BP	<u>0,3582</u>	<u>24,7360</u>	<u>0,3538</u>	<u>0,3453</u>
EfficientNetB0	BP	0,3562	24,7878	0,3519	0,3430
Baseline	CC	0,5759	10,9067	0,5550	0,5029
ResNet50	CC	0,6360	<u>12,6971</u>	0,6476	0,6679
EfficientNetB0	CC	<u>0,6351</u>	12,7187	<u>0,5753</u>	<u>0,6657</u>
Baseline	MF	0,5324	8,4366	0,4979	0,4656
ResNet50	MF	0,5903	<u>9,2395</u>	<u>0,5248</u>	0,5604
EfficientNetB0	MF	<u>0,5873</u>	9,3204	0,5583	<u>0,5562</u>

Conclusões Parciais

- Anotação funcional usando PPI é tarefa multirrótulo complexa.
- *Baseline* simples produziu os melhores resultados em BP.
- MLP (max) foi destaque em CC e MF; CNNs (em imagens) não superaram esse caso de MLP.
- Próximos passos incluem arquiteturas de grafo (GCN/GNNs) e métodos autossupervisionados para representações mais robustas.

Cronograma de Atividades

	1º Ano (bimestres)						2º Ano (bimestres)					
	1	2	3	4	5	6	1	2	3	4	5	6
Etapa 1 - Preparação												
Estudo e análise das técnicas	•	•	•	•	•	•						
Preparação da base de dados		•	•	•	•	•						
Etapa 2 - Abordagens supervisionadas												
Desenvolvimento dos modelos				•	•	•						
Realização de testes					•	•	•					
Etapa 3 - Pré-Treinamento Autossupervisionado												
Desenvolvimento dos modelos						•	•	•				
Realização de testes								•	•	•		
Etapa 4 - Comparação e Análise												
Comparação dos resultados									•	•		
Etapa 5 - Conclusão												
Documentação e publicação									•	•	•	
Escrita da dissertação										•	•	•
Defesa da dissertação												•

Próximos Passos

Próximos Passos

- Explorar GCN/GNNs em redes PPI diretamente ou a partir de representações obtidas por abordagens como caminhada aleatória (por exemplo, Node2Vec e Graph2Vec).
- Aprendizado autossupervisionado (autoencoders, aprendizado contrastivo, entre outras) para vetores e grafos.
- Experimentação com integração de *embeddings* de sequência ou outras informações proteicas para abordagem de grafos.
- Melhorias no pré-processamento, principalmente para abordagem de imagens.
- Ajuste mais fino (*fine-tuning*) das CNNs.
- Avaliação dos modelos para integração final e documentação dos resultados.

- [1] U. Satyanarayana.
Biochemistry.
2013.
- [2] Christian B. Anfinsen.
Principles That Govern the Folding of Protein Chains.
Science, 181(4096):223–230, 1973.
- [3] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock.
Gene Ontology: Tool for the Unification of Biology.
Nature Genetics, 25(1):25–29, 2000.

- [4] Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J. Jensen, and Christian von Mering.

The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets.

Nucleic Acids Research, 49(D1):D605–D612, 2021.

- [5] Nahid Safari-Alighiarloo, Mohammad Taghizadeh, Mostafa Rezaei-Tavirani, Bahram Goliaei, and Ali Asghar Peyvandi.

Protein-Protein Interaction Networks (PPI) and Complex Diseases.

Gastroenterology and Hepatology from Bed to Bench, 7(1):17–31, 2014.

- [6] Christopher M. Bishop and Nasser M. Nasrabadi.
***Pattern Recognition and Machine Learning*, volume 4.**
2006.
- [7] Grigorios Tsoumakas and Ioannis Katakis.
Multi-Label Classification: An Overview.
Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications, 3(3):64–74, 2008.
- [8] David E. Rumelhart, Geoffrey E. Hinton, and R. J. Williams.
Learning Representations by Back-Propagating Errors.
Nature, 323(6088):533–536, 1986.
- [9] Waseem Rawat and Zenghui Wang.
Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review.
Neural Computation, 29(9):2352–2449, 2017.

- [10] Longlong Jing and Yingli Tian.
Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11):4037–4058, 2020.
- [11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton.
A Simple Framework for Contrastive Learning of Visual Representations.
In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

- [12] Thomas N. Kipf and Max Welling.
Semi-Supervised Classification with Graph Convolutional Networks.
arXiv, pages 1–14, 2016.
- [13] Ronghui You, Shuwei Yao, Hiroshi Mamitsuka, and Shanfeng Zhu.
DeepGraphGO: Graph Neural Network for Large-Scale, Multispecies Protein Function Prediction.
Bioinformatics, 37:i262–i271, 2021.
- [14] Kunjie Fan, Yuanfang Guan, and Yan Zhang.
Graph2go: A multi-modal attributed network embedding method for inferring protein functions.
GigaScience, 9(8):giaa081, 2020.

- [15] Zhuoyang Chen and Qiong Luo.
DualNetGO: A Dual Network Model for Protein Function Prediction via Effective Feature Selection.
Bioinformatics, 40(7):btae437, 2024.
- [16] Vladimir Gligorijevic, Meet Barot, and Richard Bonneau.
deepNF: Deep Network Fusion for Protein Function Prediction.
Bioinformatics, 34(22):3873–3881, 2018.
- [17] Chengshuai Zhao, Shuai Liu, Feng Huang, Shichao Liu, and Wen Zhang.
CSGNN: Contrastive Self-Supervised Graph Neural Network for Molecular Interaction Prediction.
In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3756–3763, 2021.

- [18] Xuehua Bi, Weiyang Liang, Qichang Zhao, and Jianxin Wang.
SSLpheno: Self-Supervised Learning for Gene-Phenotype Association Prediction.
Bioinformatics, 39(11):btad662, 2023.
- [19] Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, et al.
The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes Through Experimental Screens.
Genome Biology, 20(1):1–23, 2019.

Anotação de Funções Proteicas Utilizando Redes de Interações Proteína-Proteína

Exame de Qualificação de Mestrado

Candidato: Yeonatan Mauhnoom

Orientador: Prof. Dr. Zanoni Dias

Coorientador: Prof. Dr. Hélio Pedrini

23 de Abril de 2025

Instituto de Computação (IC)

Universidade Estadual de Campinas (Unicamp)

Métricas de Avaliação: $F_{\max}$, $S_{\min}$, AuPRC e IAuPRC

$F_{\max}$:

$$F_{\max} = \max_{\tau} \left\{ \frac{2 \cdot pr(\tau) \cdot rc(\tau)}{pr(\tau) + rc(\tau)} \right\},$$

onde

$$pr(\tau) = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{|P_i(\tau) \cap T_i|}{|P_i(\tau)|}, \quad rc(\tau) = \frac{1}{n} \sum_{i=1}^n \frac{|P_i(\tau) \cap T_i|}{|T_i|}.$$

$S_{\min}$:

$$S(\tau) = \sqrt{\frac{RU(\tau)^2 + MI(\tau)^2}{2}}, \quad S_{\min} = \min_{\tau} S(\tau),$$

AuPRC:

$$AuPRC = \sum_{k=1}^{K-1} (r_{k+1} - r_k) \cdot p_{k+1},$$

com r_k e p_k sendo revocação e precisão no ponto k .

IAuPRC:

$$p_{\text{interp}}(r) = \max_{r_k \geq r} p_k, \quad IAuPRC = \int (r, p_{\text{interp}}(r)) dr.$$