

Anotações de Funções de Proteínas Utilizando Aprendizado de Máquina e Alinhamento Local

Gabriel Bianchin de Oliveira

Orientador: Prof. Dr. Zanoni Dias

Coorientador: Prof. Dr. Hélio Pedrini

22 de Julho de 2025

Instituto de Computação

Universidade Estadual de Campinas

Introdução

Trabalhos Relacionados

Materiais

Métodos

Resultados

Servidor *Web*

Conclusões

Introdução

Trabalhos Relacionados

Materiais

Métodos

Resultados

Servidor *Web*

Conclusões

- Proteínas, aminoácidos e funções
- Sequenciamento de proteínas nos últimos anos
- Custo para definição de funções

Introdução

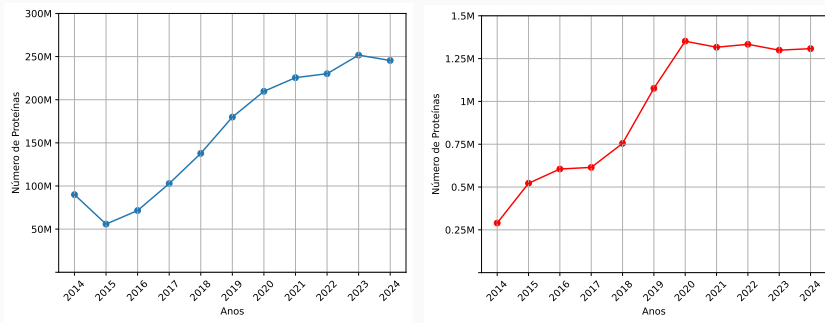


Figura 1: Quantidade de proteínas sequenciadas na base UniProt e proteínas analisadas manualmente na base GOA.

- Anotação via Ontologia Genética (GO) [1]
- Ontologia de Componente Celular (CC), Função Molecular (FM) e Processo Biológico (PB)
- Grafo acíclico direcionado
- Classificação multirrótulo

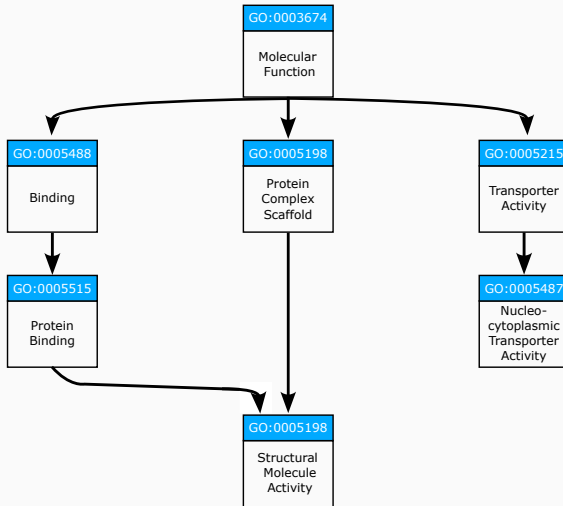


Figura 2: Regra do caminho verdadeiro.

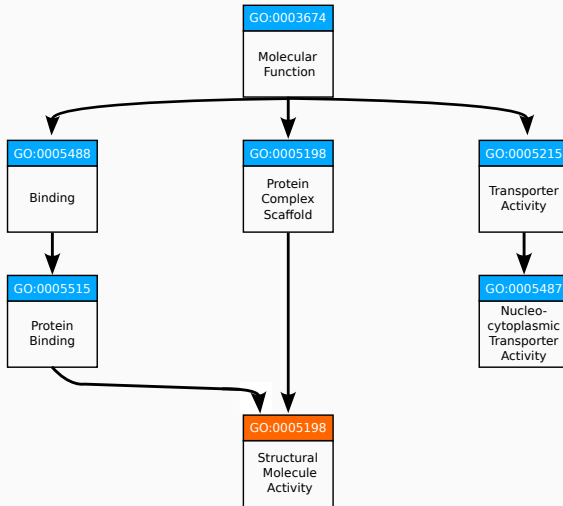


Figura 2: Regra do caminho verdadeiro.

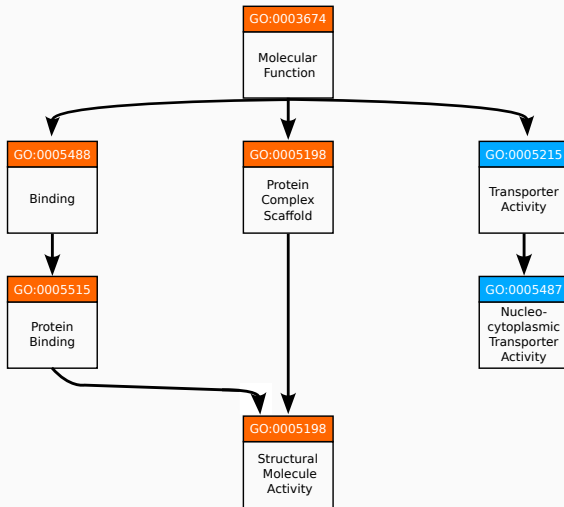


Figura 2: Regra do caminho verdadeiro.

- Métodos computacionais para classificação utilizando a sequência
- Aprendizado de Máquina e Processamento de Linguagem Natural
- Alinhamento local de proteínas
- Continua sendo um problema desafiante

Questões de Pesquisa

1. A utilização de métodos baseados em *ensemble* de aprendizado de máquina com alinhamento local de proteínas apresenta melhorias em relação a cada uma das abordagens separadamente?
2. Qual é a melhor técnica para agregar as classificações de cada método individualmente (aprendizado de máquina e alinhamento local), considerando ambas as predições em um modelo de *ensemble*?
3. Qual é o impacto de utilizar os resultados do alinhamento local para alimentar o modelo de aprendizado de máquina?
4. A transformação das características extraídas pelo modelo Transformer pode ser manipulada para uma representação de imagem, auxiliando na classificação?
5. É possível criar uma abordagem que utilize menos recursos computacionais e mantenha o resultado de um modelo mais complexo para a tarefa de classificação de funções de proteínas?

Introdução

Trabalhos Relacionados

Materiais

Métodos

Resultados

Servidor *Web*

Conclusões

- Métodos baseados em sequência de aminoácidos
 - Aprendizado de Máquina
 - Alinhamento Local
 - *Ensemble* de Aprendizado de Máquina com Alinhamento Local

Tabela 1: Métodos de aprendizado de máquina comparados com a metodologia proposta.

Método	Transformers	Sequência Inteira	Limite da Sequência Analisada
PFmulDL [2]	×	×	2.000
DeepGO [3]	×	×	1.000
DeepGOCNN [4]	×	×	2.000
TALE [5]	✓	×	1.000
ATGO [6]	✓	×	1.022
PU-GO [7]	✓	×	1.022
PROTGOAT [8]	✓	×	1.022
InterLabelGO [9]	✓	×	1.022
TEMPROT [10]	✓	✓	—
MAGO [11]	✓	✓	—

Tabela 2: Métodos de *ensemble* de aprendizado de máquina com alinhamento local comparados com a metodologia proposta.

Método	Transformers	Sequência Inteira	Alinhamento Local
DIAMOND [12]	—	—	—
BLASTp [13]	—	—	—
DeepGOPlus [4]	×	×	DIAMOND
TALE+ [5]	✓	×	DIAMOND
ATGO+ [6]	✓	×	BLASTp
PU-GO+Diamond [7]	✓	×	DIAMOND
InterLabelGO+ [9]	✓	×	DIAMOND
TEMPROT+ [10]	✓	✓	BLASTp
MAGO+ [11]	✓	✓	BLASTp

Introdução

Trabalhos Relacionados

Materiais

Métodos

Resultados

Servidor *Web*

Conclusões

- Conjunto de dados baseado no CAFA5 [14]
- Divisão dos dados do CAFA5 em 80% para treinamento, 10% para validação e 10% para teste
- Número de termos em cada ontologia:
 - Componente Celular (CC): 498
 - Função Molecular (FM): 499
 - Processo Biológico (PB): 500

Tabela 3: Número de proteínas e termos na base CAFA5.

	CC	FM	PB
Treinamento	74.328	62.909	73.768
Validação	9.292	7.864	9.221
Teste	9.292	7.864	9.221
Termos	498	499	500

- $F_{\max}$
- $F_{\max}^*$
- $wF_{\max}$
- $S_{\min}$
- AuPRC
- IAuPRC
- Testes estatísticos:
 - Teste de Ivan-Davenport
 - Teste de Nemenyi

Introdução

Trabalhos Relacionados

Materiais

Métodos

Resultados

Servidor *Web*

Conclusões

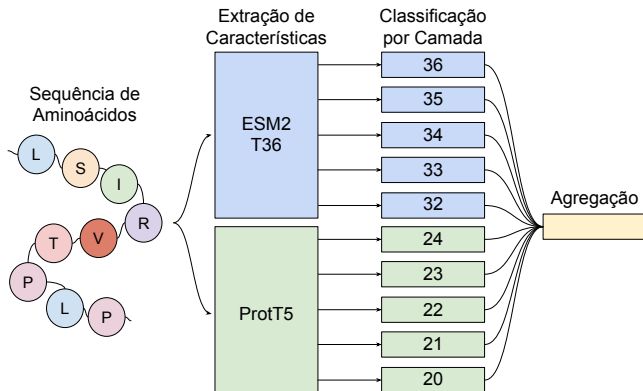


Figura 3: Visão geral do SUPERMAGO.

- Sem pré-processamento para sequências de até 1.022 aminoácidos
- Pré-processamento para sequências com mais do que 1.022 aminoácidos:
 - Divisão em subsequências de até 1.022 aminoácidos
 - Janela deslizante sem sobreposição
- Últimas cinco camadas do ESM2 T36 [15] e ProtT5 [16]
- Média das representações de cada *token*
- Pós-processamento para sequências que foram divididas:
 - Média por posição do vetor de características

SUPERMAGO - Extração de Características

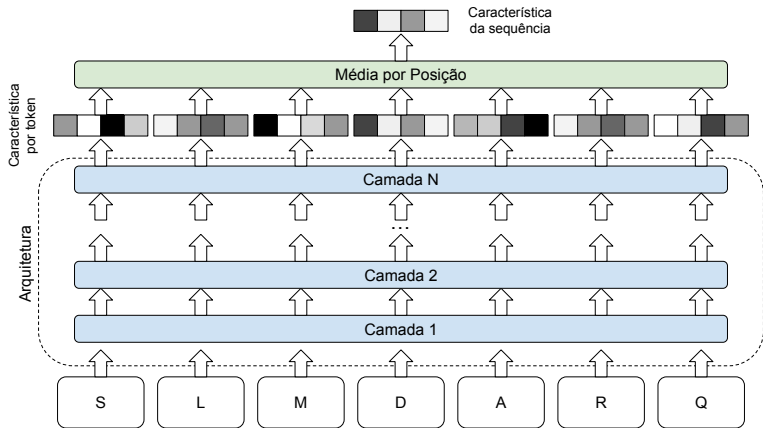


Figura 4: Processo de extração de características.

- Dez modelos *multilayer perceptron*
- Cada modelo especializado em cada representação do ESM2 T36 ou ProtT5
- Modelo com 2 camadas com 4.096 neurônios com função de ativação ReLU

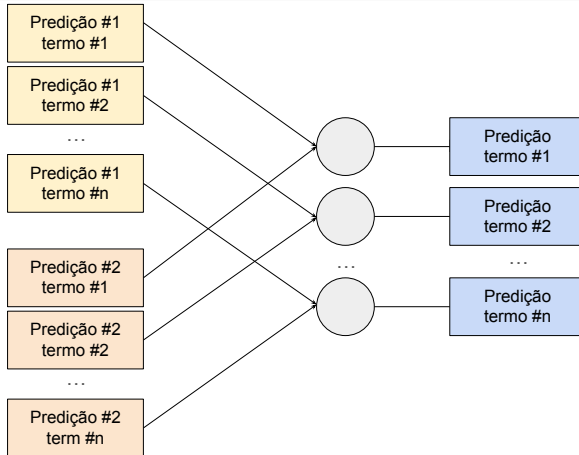


Figura 5: Agregação por *stacking*.

- *Ensemble* do SUPERMAGO com o DIAMOND
- DIAMOND baseado no alinhamento com o treinamento
- Parâmetro *e-value* com valor máximo de 0,001

$$P(p, t) = \frac{\sum_{s \in E} T_s \times \text{bitscore}(p, s)}{\sum_{s \in E} \text{bitscore}(p, s)}. \quad (1)$$

- Adaptação do *stacking* para o SUPERMAGO+

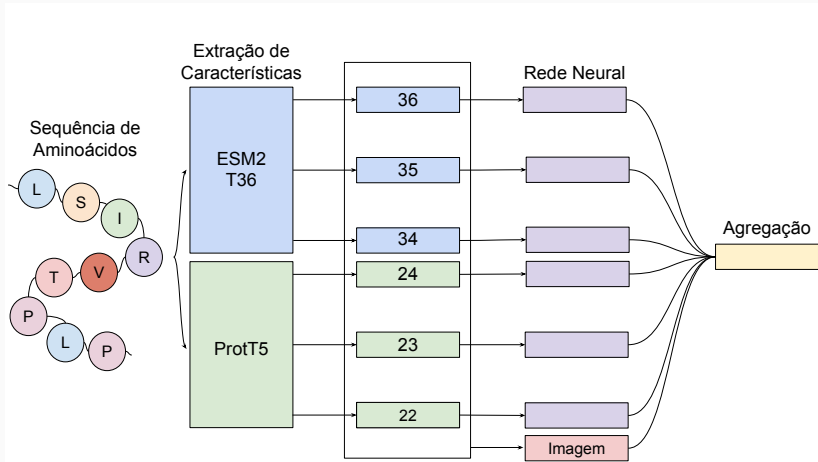


Figura 6: Visão geral do SUPERMAGOv2.

SUPERMAGOv2 - Extração de Características

- Mesma extração do SUPERMAGO
- Sem pré-processamento para sequências de até 1.022 aminoácidos
- Pré-processamento para sequências com mais do que 1.022 aminoácidos:
 - Divisão em subsequências de até 1.022 aminoácidos
 - Janela deslizante sem sobreposição
- Últimas três camadas do ESM2 T36 [15] e ProtT5 [16]
- Média das representações de cada *token*
- Pós-processamento para sequências que foram divididas:
 - Média por posição do vetor de características

SUPERMAGOv2 - Classificação por Rede Neural

- Modelo *multilayer perceptron* com duas entradas:
 - Vetor de características da proteína analisada
 - Vetor de características da ponderação de alinhamentos
- Vetor de características ponderadas por alinhamento local com DIAMOND considerando o conjunto de treinamento
- Parâmetro *e-value* com valor máximo de 0,001

$$W(p, s) = \frac{\text{bitscore}(p, s) \times \text{ID}(p, s)}{\sum_{s \in E} \text{bitscore}(p, s) \times \text{ID}(p, s)}. \quad (2)$$

$$\text{ID}(p, s) = \frac{nident(p, s)}{\max(tam_p, tam_s)}. \quad (3)$$

- Vetor de características da proteína analisada usada como vetor de características da ponderação de alinhamentos caso a proteína não tenha alinhamentos com o treinamento

SUPERMAGOv2 - Classificação por Rede Neural

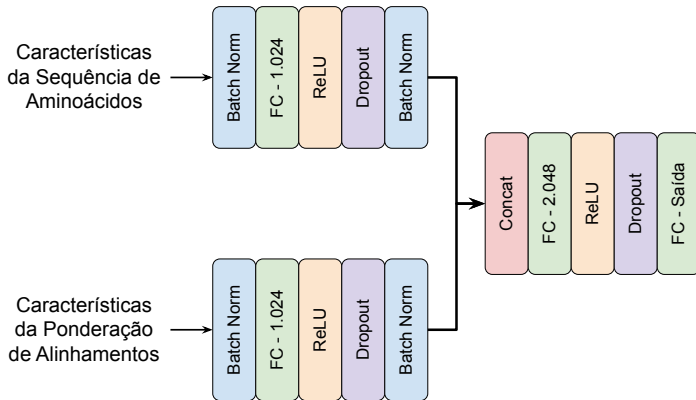


Figura 7: Rede neural do SUPERMAGOv2.

SUPERMAGOv2 - Classificação por Imagem

- Concatenação dos vetores da proteína analisada das camadas correspondentes do ESM2 T36 e ProtT5:
 - 36 (ESM2 T36) com 24 (ProtT5); 35 (ESM2 T36) com 23 (ProtT5); e 34 (ESM2 T36) com 22 (ProtT5)
- Aplicação da função sigmoide para normalizar os valores dos vetores
- Redimensionamento para matrizes de 64x56 e geração da imagem

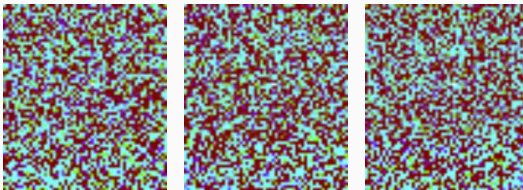


Figura 8: Exemplos de imagens geradas.

- Classificação com ResNet50 [17]

- Mesma agregação do SUPERMAGO
- Utilização do *stacking*

- *Ensemble* do SUPERMAGOv2 com o DIAMOND
- DIAMOND baseado no alinhamento com o treinamento
- Parâmetro *e-value* com valor máximo de 0,001

$$P(p, t) = \frac{\sum_{s \in E} T_s \times \text{bitscore}(p, s) \times \text{ID}(p, s)}{\sum_{s \in E} \text{bitscore}(p, s) \times \text{ID}(p, s)}. \quad (4)$$

- Adaptação do *stacking* para o SUPERMAGOv2+

Introdução

Trabalhos Relacionados

Materiais

Métodos

Resultados

Servidor *Web*

Conclusões

Resultados - Aprendizado de Máquina

Tabela 4: Resultados para a ontologia de **Componente Celular** considerando os métodos de aprendizado de máquina. O melhor resultado para cada métrica está em negrito e o segundo melhor valor está sublinhado.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
PFmulDL	0,707	0,672	0,529	11,320	0,770	0,766
DeepGO	0,655	0,612	0,416	12,923	0,580	0,690
DeepGOCNN	0,678	0,641	0,475	11,816	0,731	0,726
TALE	0,693	0,656	0,506	11,647	0,657	0,753
ATGO	0,724	0,691	0,562	10,829	0,696	0,790
PU-GO	0,692	0,655	0,510	11,924	0,637	0,710
PROTGOAT	0,763	0,736	0,630	9,463	0,775	<u>0,841</u>
InterLabelGO	<u>0,778</u>	<u>0,754</u>	0,620	9,201	<u>0,818</u>	0,785
TEMPROT	<u>0,731</u>	0,699	0,569	10,622	<u>0,731</u>	0,790
MAGO	0,755	0,727	0,615	9,661	0,807	0,829
SUPERMAGO	0,769	0,743	<u>0,641</u>	<u>9,211</u>	0,804	0,848
SUPERMAGOV2	0,786	0,764	0,678	9,305	0,862	0,848

Resultados - Aprendizizado de Máquina

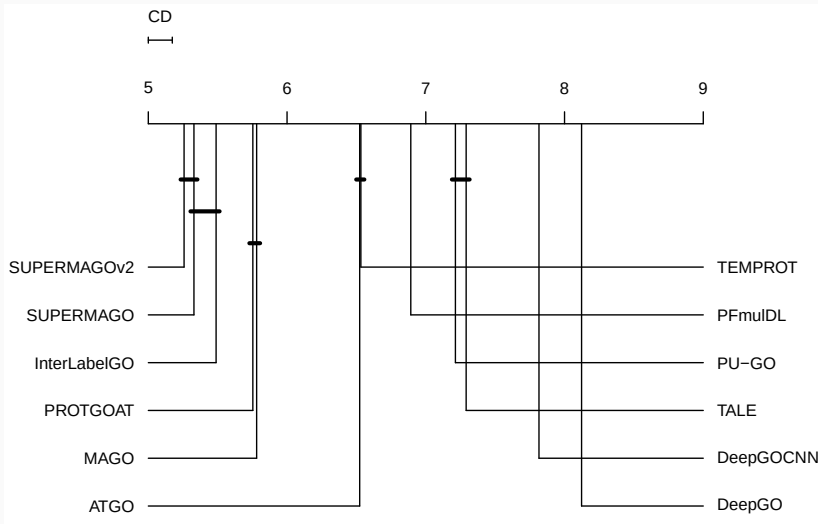


Figura 9: Teste de Nemenyi para a ontologia de **Componente Celular** considerando os métodos de aprendizado de máquina (métrica F_{max}).

Resultados - Aprendizado de Máquina

Tabela 5: Resultados para a ontologia de **Função Molecular** considerando os métodos de aprendizado de máquina. O melhor resultado para cada métrica está em negrito e o segundo melhor valor está sublinhado.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
PFmulDL	0,693	0,628	0,568	7,494	0,622	0,717
DeepGO	0,465	0,387	0,295	10,800	0,149	0,295
DeepGOCNN	0,688	0,622	0,565	7,499	0,706	0,697
TALE	0,685	0,618	0,556	7,492	0,584	0,717
ATGO	0,749	0,700	0,650	6,416	0,701	0,793
PU-GO	0,703	0,642	0,580	7,368	0,547	0,720
PROTGOAT	0,794	0,756	0,716	5,318	0,732	<u>0,855</u>
InterLabelGO	<u>0,807</u>	<u>0,773</u>	<u>0,737</u>	<u>5,203</u>	0,841	0,787
TEMPROT	0,762	0,716	0,671	6,117	0,662	0,806
MAGO	0,793	0,755	0,716	5,405	0,793	0,851
SUPERMAGO	0,802	0,766	0,727	5,257	<u>0,827</u>	0,865
SUPERMAGOV2	0,814	0,782	0,747	5,140	0,642	0,846

Resultados - Aprendizizado de Máquina

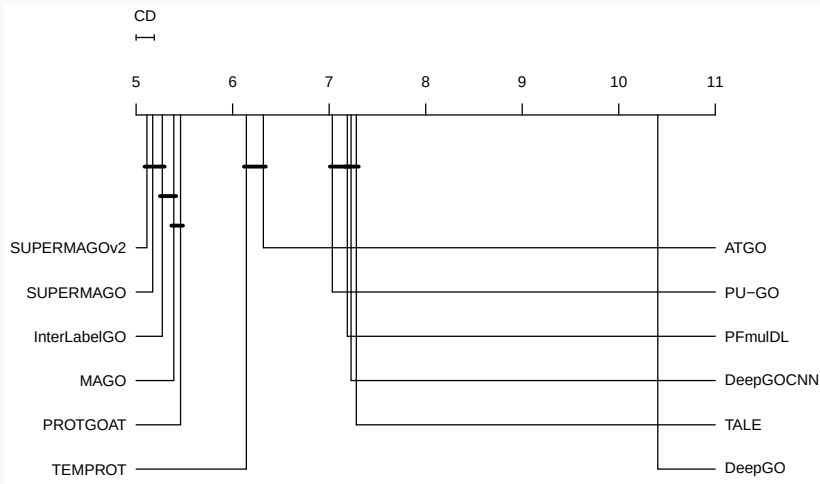


Figura 10: Teste de Nemenyi para a ontologia de **Função Molecular** considerando os métodos de aprendizado de máquina (métrica $F_{\max}$).

Resultados - Aprendizado de Máquina

Tabela 6: Resultados para a ontologia de **Processo Biológico** considerando os métodos de aprendizado de máquina. O melhor resultado para cada métrica está em negrito e o segundo melhor valor está sublinhado.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
PFmulDL	0,460	0,416	0,367	22,625	0,489	0,483
DeepGO	0,372	0,328	0,280	25,304	0,293	0,365
DeepGOCNN	0,497	0,448	0,413	21,044	0,521	0,515
TALE	0,427	0,387	0,343	24,023	0,381	0,436
ATGO	0,505	0,467	0,427	22,003	0,497	0,529
PU-GO	0,463	0,426	0,374	23,490	0,413	0,371
PROTGOAT	0,572	0,538	0,503	19,362	0,567	0,622
InterLabelGO	<u>0,638</u>	<u>0,609</u>	<u>0,586</u>	17,546	<u>0,614</u>	<u>0,664</u>
TEMPROT	0,514	0,472	0,427	21,355	0,483	0,537
MAGO	0,555	0,520	0,483	20,045	0,574	0,593
SUPERMAGO	0,582	0,550	0,516	19,103	0,609	0,630
SUPERMAGOV2	0,644	0,615	0,594	<u>17,874</u>	0,634	0,691

Resultados - Aprendizizado de Máquina

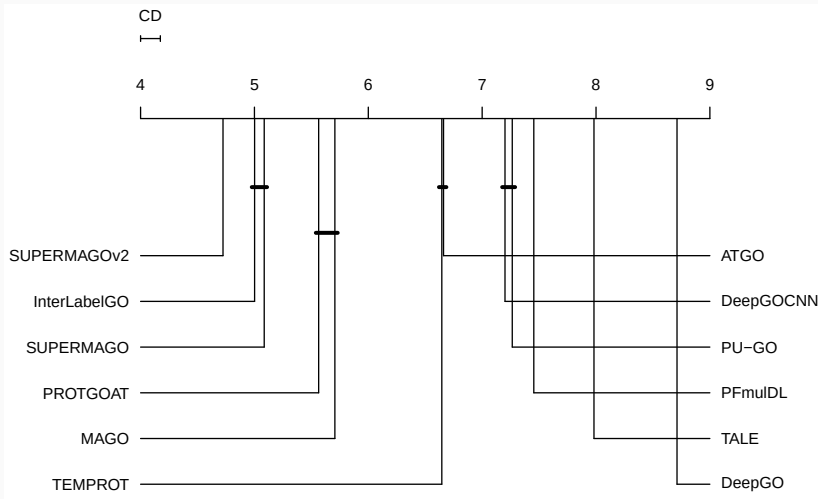


Figura 11: Teste de Nemenyi para a ontologia de **Processo Biológico** considerando os métodos de aprendizado de máquina (métrica F_{max}).

Tabela 7: Resultados para a ontologia de **Componente Celular** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local. O melhor resultado para cada métrica está em negrito e o segundo melhor valor está sublinhado.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
DIAMOND	0,699	0,677	0,586	10,031	0,441	0,673
BLASTp	0,712	0,686	0,582	10,349	0,528	0,729
DeepGOPlus	0,738	0,709	0,594	10,006	0,802	0,799
TALE+	0,753	0,726	0,620	9,644	0,737	0,822
ATGO+	0,743	0,714	0,603	10,082	0,722	0,819
PU-GO+Diamond	0,751	0,723	0,622	9,890	0,729	0,801
InterLabelGO+	0,775	0,752	<u>0,661</u>	9,361	<u>0,835</u>	0,826
TEMPROT+	0,752	0,724	<u>0,614</u>	9,868	<u>0,752</u>	0,822
MAGO+	0,774	0,748	0,649	9,157	0,815	0,844
SUPERMAGO+	<u>0,777</u>	<u>0,753</u>	0,658	8,892	0,813	0,857
SUPERMAGOV2+	0,789	0,766	0,678	<u>8,929</u>	0,856	<u>0,853</u>

Resultados - Ensemble

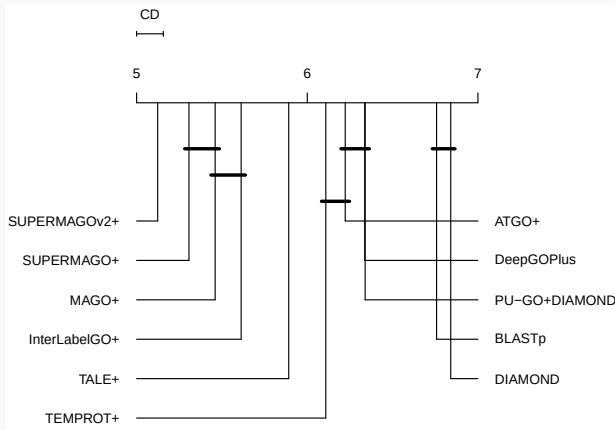


Figura 12: Teste de Nemenyi para a ontologia de **Componente Celular** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local (métrica $F_{\max}$).

Tabela 8: Resultados para a ontologia de **Função Molecular** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local. O melhor resultado para cada métrica está em negrito e o segundo melhor valor está sublinhado.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
DIAMOND	0,742	0,712	0,679	5,579	0,454	0,716
BLASTp	0,763	0,728	0,692	5,613	0,553	0,774
DeepGOPlus	0,784	0,744	0,709	5,573	0,832	0,825
TALE+	0,770	0,727	0,685	5,777	0,707	0,820
ATGO+	0,787	0,749	0,710	5,557	0,749	0,846
PU-GO+Diamond	0,793	0,753	0,717	5,453	0,689	0,829
InterLabelGO+	0,806	0,773	0,736	5,232	<u>0,859</u>	0,847
TEMPROT+	0,794	0,756	0,717	5,437	<u>0,734</u>	0,849
MAGO+	0,801	0,765	0,728	5,285	0,806	0,860
SUPERMAGO+	<u>0,809</u>	<u>0,775</u>	<u>0,739</u>	<u>5,100</u>	0,839	0,874
SUPERMAGOV2+	0,815	0,782	0,747	5,057	0,865	<u>0,863</u>

Resultados - Ensemble

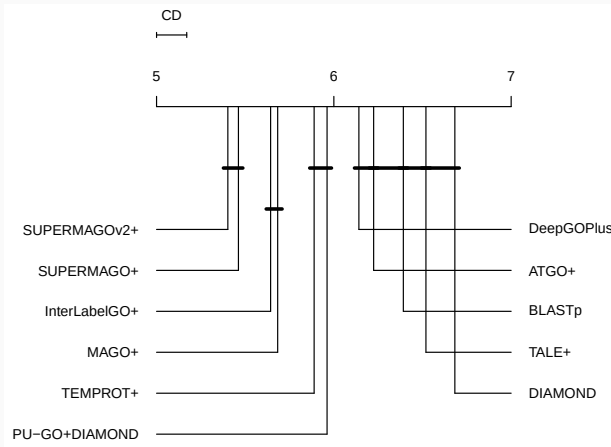


Figura 13: Teste de Nemenyi para a ontologia de **Função Molecular** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local (métrica $F_{\max}$).

Tabela 9: Resultados para a ontologia de **Processo Biológico** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local. O melhor resultado para cada métrica está em negrito e o segundo melhor valor está sublinhado.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
DIAMOND	0,588	0,566	0,541	18,524	0,455	0,571
BLASTp	0,576	0,550	0,521	19,470	0,495	0,590
DeepGOPlus	0,601	0,568	0,542	18,121	0,636	0,631
TALE+	0,580	0,549	0,524	18,792	0,562	0,608
ATGO+	0,569	0,538	0,509	19,669	0,572	0,605
PU-GO+Diamond	0,620	0,566	0,542	18,488	0,595	0,622
InterLabelGO+	<u>0,647</u>	<u>0,617</u>	<u>0,594</u>	<u>17,089</u>	0,680	<u>0,676</u>
TEMPROT+	<u>0,588</u>	0,558	0,528	18,987	0,578	0,624
MAGO+	0,615	0,585	0,559	18,092	0,635	0,660
SUPERMAGO+	0,619	0,591	0,566	17,702	<u>0,645</u>	0,670
SUPERMAGOV2+	0,652	0,625	0,603	16,996	0,633	0,696

Resultados - Ensemble

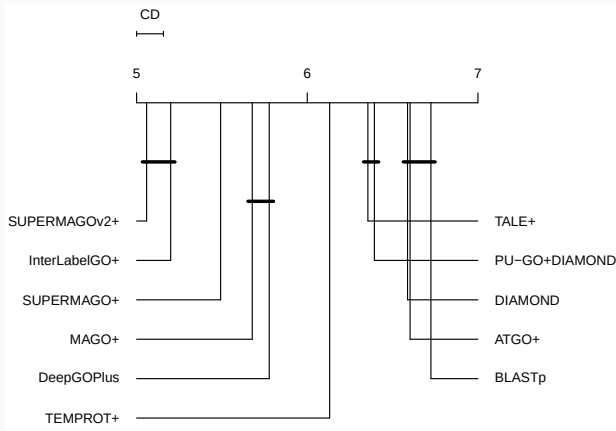


Figura 14: Teste de Nemenyi para a ontologia de **Processo Biológico** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local (métrica $F_{\max}$).

Introdução

Trabalhos Relacionados

Materiais

Métodos

Resultados

Servidor *Web*

Conclusões

- Substituição do ESM2 T36 pelo ESM2 T12 [15]
- Aplicação do LoRA [18] com `int4` [19] e precisão `bf16`
- Adaptações necessárias pela troca de arquitetura
- Servidor *web* disponível em `supermago.ic.unicamp.br`

Tabela 10: Resultados para a ontologia de **Componente Celular** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local e suas versões eficientes em memória.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
SUPERMAGO+	0,777	0,753	0,658	8,892	0,813	0,857
SUPERMAGO+Web	0,775	0,750	0,653	9,015	0,813	0,853
SUPERMAGOV2+	0,789	0,766	0,678	8,929	0,856	0,853
SUPERMAGOV2+Web	0,785	0,762	0,673	9,012	0,852	0,853

Resultados

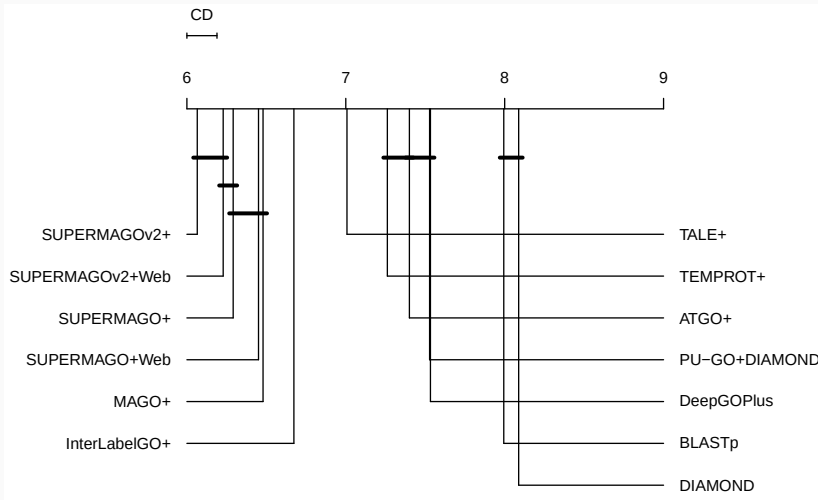


Figura 15: Teste de Nemenyi para a ontologia de **Componente Celular** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local e suas versões eficientes em memória (métrica $F_{\max}$).

Tabela 11: Resultados para a ontologia de **Função Molecular** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local e suas versões eficientes em memória.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
SUPERMAGO+	0,809	0,775	0,739	5,100	0,839	0,874
SUPERMAGO+Web	0,809	0,775	0,739	5,128	0,838	0,871
SUPERMAGOV2+	0,815	0,782	0,747	5,057	0,865	0,863
SUPERMAGOV2+Web	0,812	0,780	0,745	5,110	0,698	0,865

Resultados

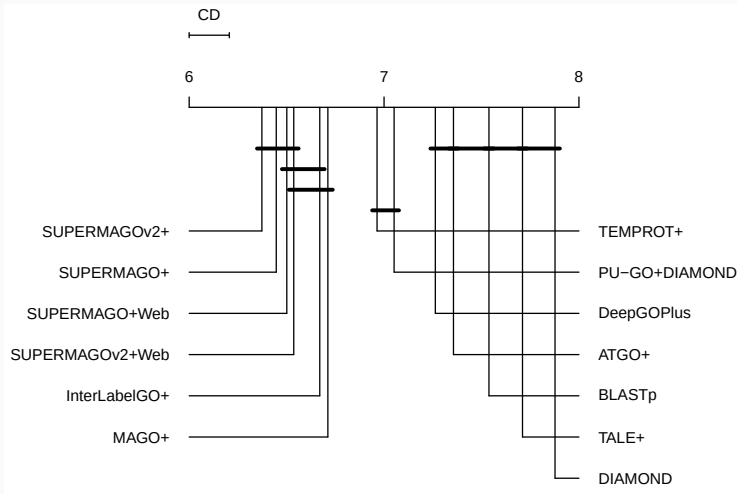


Figura 16: Teste de Nemenyi para a ontologia de **Função Molecular** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local e suas versões eficientes em memória (métrica $F_{\max}$).

Tabela 12: Resultados para a ontologia de **Processo Biológico** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local e suas versões eficientes em memória.

Método	$F_{\max}$	$F_{\max}^*$	$wF_{\max}$	$S_{\min}$	AuPRC	IAuPRC
SUPERMAGO+	0,619	0,591	0,566	17,702	0,645	0,670
SUPERMAGO+Web	0,617	0,589	0,563	17,785	0,636	0,664
SUPERMAGOV2+	0,652	0,625	0,603	16,996	0,633	0,696
SUPERMAGOV2+Web	0,647	0,619	0,597	17,247	0,643	0,693

Resultados

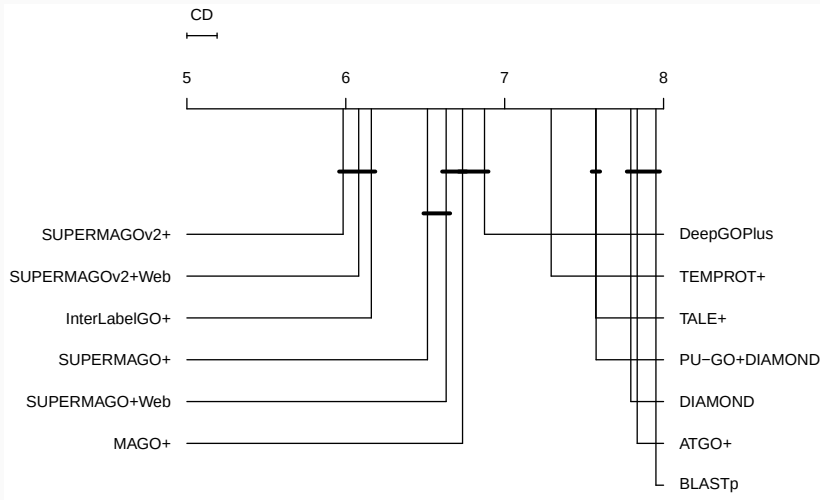


Figura 17: Teste de Nemenyi para a ontologia de **Processo Biológico** considerando os métodos de *ensemble* de aprendizado de máquina com alinhamento local e suas versões eficientes em memória (métrica $F_{\max}$).

Introdução

Trabalhos Relacionados

Materiais

Métodos

Resultados

Servidor *Web*

Conclusões

- Métodos propostos:
 - SUPERMAGO e SUPERMAGOV2
 - SUPERMAGO+ e SUPERMAGOV2+
 - SUPERMAGO+Web e SUPERMAGOV2+Web

- Nova métrica chamada IAuPRC
- Avaliação de técnicas de agregação de classificadores
- Alinhamentos locais como características para aprendizado de máquina
- Tratamento da tarefa como classificação de imagens

1. A utilização de métodos baseados em *ensemble* de aprendizado de máquina com alinhamento local de proteínas apresenta melhorias em relação a cada uma das abordagens separadamente?
- R:** Os resultados da pesquisa indicam que a combinação via *ensemble* supera as abordagens individualmente, conforme outros estudos da literatura.

2. Qual é a melhor técnica para agregar as classificações de cada método individualmente (aprendizado de máquina e alinhamento local), considerando ambas as previsões em um modelo de *ensemble*?

R: Comparando as estratégias de combinação linear simples, amplamente utilizada nos métodos estado da arte, com o *stacking*, esta última abordagem se mostrou superior. Isto ocorre devido a maior complexidade que esse modelo consegue capturar.

3. Qual é o impacto de utilizar os resultados do alinhamento local para alimentar o modelo de aprendizado de máquina?

R: A utilização de resultados do alinhamento local como característica de entrada de modelos de aprendizado de máquina teve um impacto positivo, contribuindo para a melhoria dos resultados quando comparado com a não utilizando dessa informação.

4. A transformação das características extraídas pelo modelo Transformer pode ser manipulada para uma representação de imagem, auxiliando na classificação?

R: A conversão de vetores de características extraídos de modelos Transformers em imagens contribuiu positivamente para a classificação, agregando informações complementares às demais estratégias de classificação e tratando o problema de forma multimodal.

5. É possível criar uma abordagem que utilize menos recursos computacionais e mantenha o resultado de um modelo mais complexo para a tarefa de classificação de funções de proteínas?

R: A combinação de técnicas como quantização com `int4`, adaptação de pesos com LoRA e uso de precisão computacional `bf16`, aliada à substituição de arquiteturas mais pesadas por modelos mais leves para a extração de características, permitiu o desenvolvimento de métodos com desempenho estatisticamente equivalente aos modelos mais complexos. Sendo assim, os modelos menores reduzem significativamente a demanda computacional e mantêm resultados competitivos.

- Melhoria dos resultados em PB
- Integração de outros dados de proteínas
- Entender o porquê das classificações dos modelos computacionais

- Integração de outros tipos de dados:
 - Estruturas tridimensionais
 - Redes de interações de proteínas
- Interpretabilidade

- “SUPERMAGOV2: Protein Function Prediction via Transformer Embeddings and Bitscore-Weighted Features” (Gabriel Bianchin de Oliveira, Hélio Pedrini e Zanoni Dias), em avaliação em uma revista internacional;
- “SUPERMAGO: Protein Function Prediction based on Transformer Embeddings” (Gabriel Bianchin de Oliveira, Hélio Pedrini e Zanoni Dias) [20], publicado na revista *PROTEINS: Structure, Function, and Bioinformatics*, 2025;
- “Integrating Transformers and AutoML for Protein Function Prediction” (Gabriel Bianchin de Oliveira, Hélio Pedrini e Zanoni Dias) [11], apresentado na *46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2024)*;

- “Scaling Up ESM2 Architectures for Long Protein Sequences Analysis: Long and Quantized Approaches” (Gabriel Bianchin de Oliveira, Hélio Pedrini e Zandoni Dias) [21], apresentado na *17th Brazilian Symposium on Bioinformatics* (BSB 2024);
- “TEMPROT: Protein Function Annotation using Transformers Embeddings and Homology Search” (Gabriel Bianchin de Oliveira, Hélio Pedrini e Zandoni Dias) [10], publicado na revista *BMC Bioinformatics*, 2023;
- “Protein Molecular Function Annotation Based on Transformer Embeddings” (Gabriel Bianchin de Oliveira, Hélio Pedrini e Zandoni Dias) [22], apresentado na *11th Brazilian Conference on Intelligent Systems* (BRACIS 2022);

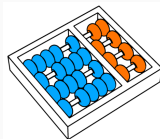
- “Prediction of Protein Molecular Functions Using Transformers” (Felipe Lopes de Mello, Gabriel Bianchin de Oliveira, Hélio Pedrini e Zanoni Dias) [23], apresentado na *21st International Conference on Artificial Intelligence and Soft Computing* (ICAISC 2022);
- “Predição de Funções Moleculares de Proteínas utilizando Aprendizado de Máquina” (Felipe Lopes de Mello, Gabriel Bianchin de Oliveira e Zanoni Dias) [24], relatório técnico relativo ao Projeto Final de Graduação do aluno Felipe Lopes de Mello (2021).

- “Fair Age Prediction through ILP-Based Multi-Attribute Training Subset Selection” (Ana Paula dos Santos Dantas, Gabriel Bianchin de Oliveira, Hélio Pedrini, Cid Carvalho de Souza e Zanoni Dias), em avaliação em uma revista internacional;
- “The Multi-attribute Fairer Cover Problem” (Ana Paula dos Santos Dantas, Gabriel Bianchin de Oliveira, Hélio Pedrini, Cid Carvalho de Souza e Zanoni Dias) [25], apresentado na *12th Brazilian Conference on Intelligent Systems* (BRACIS 2023);
- “Algorithmic Fairness Applied to the Multi-Label Classification Problem” (Ana Paula dos Santos Dantas, Gabriel Bianchin de Oliveira, Daiane Mendes de Oliveira, Hélio Pedrini, Cid Carvalho de Souza e Zanoni Dias) [26], apresentado na *18th International Conference on Computer Vision Theory and Applications* (VISAPP 2023);

Outras Contribuições

- “Bias Assessment in Medical Imaging Analysis: a Case Study on Retinal OCT Image Classification” (Gabriel Oliveira, Lucas David, Rafael Padilha, Ana Paula da Silva, Francine de Paula, Lucas Infante, Lucio Jorge, Patricia Xavier e Zanoni Dias) [27], apresentado na *14th International Conference on Agents and Artificial Intelligence* (ICAART 2022);
- “Ensemble of Patches for COVID-19 X-ray Image Classification” (Thiago Dong Chen, Gabriel Bianchin de Oliveira e Zanoni Dias) [28], apresentado na *14th International Conference on Agents and Artificial Intelligence* (ICAART 2022);
- “Deep Learning-based COVID-19 diagnostics of low-quality CT images” (Daniel Ferber, Felipe Vieira, João Dalben, Mariana Ferraz, Nicholas Sato, Gabriel Oliveira, Rafael Padilha e Zanoni Dias) [29], apresentado na *14th Brazilian Symposium on Bioinformatics* (BSB 2021).

Agradecimentos



- [1] Gene Ontology Consortium.
The Gene Ontology (GO) database and informatics resource.
Nucleic Acids Research, 32(suppl_1):D258–D261, 2004.
- [2] Weiqi Xia, Lingyan Zheng, Jiebin Fang, Fengcheng Li, Ying Zhou, Zhenyu Zeng, Bing Zhang, Zhaorong Li, Honglin Li, and Feng Zhu.
PFmulDL: A Novel Strategy Enabling Multi-Class and Multi-Label Protein Function Annotation by Integrating Diverse Deep Learning Methods.
Computers in Biology and Medicine, 145:105465, 2022.
- [3] Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf.
DeepGO: Predicting Protein Functions from Sequence and Interactions using a Deep Ontology-Aware Classifier.
Bioinformatics, 34(4):660–668, 2018.

- [4] Maxat Kulmanov and Robert Hoehndorf.
DeepGOPlus: improved protein function prediction from sequence.
Bioinformatics, 36(2):422–429, 2019.
- [5] Yue Cao and Yang Shen.
TALE: Transformer-based protein function Annotation with joint sequence–Label Embedding.
Bioinformatics, 37(18):2825–2833, 2021.
- [6] Yi-Heng Zhu, Chengxin Zhang, Dong-Jun Yu, and Yang Zhang.
Integrating Unsupervised Language Model with Triplet Neural Networks for Protein Gene Ontology Prediction.
PLoS Computational Biology, 18(12):e1010793, 2022.

- [7] Fernando Zhapa-Camacho, Zhenwei Tang, Maxat Kulmanov, and Robert Hoehndorf.
Predicting protein functions using positive-unlabeled ranking with ontology-based priors.
bioRxiv, pages 1–9, 2024.
- [8] Zong M. Chua, Adarsh Rajesh, Sanju Sinha, and Peter D. Adams.
PROTGOAT: Improved Automated Protein Function Predictions Using Protein Language Models.
bioRxiv, pages 1–15, 2024.
- [9] Quancheng Liu, Chengxin Zhang, and Lydia Freddolino.
InterLabelGO+: unraveling label correlations in protein function prediction.
Bioinformatics, 40(11):btae655, 2024.

- [10] Gabriel B. Oliveira, Helio Pedrini, and Zanoni Dias.
TEMPROT: Protein Function Annotation using Transformers Embeddings and Homology Search.
BMC Bioinformatics, 24(1):1–16, 2023.
- [11] Gabriel B. Oliveira, Helio Pedrini, and Zanoni Dias.
Integrating Transformers and AutoML for Protein Function Prediction.
In *46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, 2024.
- [12] Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost.
Sensitive protein alignments at tree-of-life scale using DIAMOND.
Nature Methods, 18(4):366–368, 2021.

- [13] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman.
Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.
Nucleic Acids Research, 25(17):3389–3402, 1997.
- [14] Iddo Friedberg, Predrag Radivojac, Clara D. Paolis, Damiano Piovesan, Parnal Joshi, Walter Reade, and Addison Howard.
CAFA 5 Protein Function Prediction, 2023.
[https://kaggle.com/competitions/
cafa-5-protein-function-prediction.](https://kaggle.com/competitions/cafa-5-protein-function-prediction)

- [15] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos S. Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Aleksander Rives.

Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model.

Science, 379(6637):1123–1130, 2023.

- [16] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost.

ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10):7112–7127, 2021.

- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.

Deep residual learning for image recognition.

In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

- [18] Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, Tuan Dinh, Ankur Gandhe, Denis Filimonov, Shalini Ghosh, Andreas Stolcke, Ariya Rastow, and Ivan Bulyko.

Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition.

In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

- [19] Tim Dettmers and Luke Zettlemoyer.
The case for 4-bit precision: k -bit Inference Scaling Laws.
In *40th International Conference on Machine Learning (ICML)*,
pages 7750–7774, 2023.
- [20] Gabriel B. Oliveira, Helio Pedrini, and Zanoni Dias.
**SUPERMAGO: Protein Function Prediction Based on
Transformer Embeddings.**
PROTEINS: Structure, Function, and Bioinformatics,
93(5):981–996, 2025.

- [21] Gabriel B. Oliveira, Helio Pedrini, and Zanoni Dias.
Scaling Up ESM2 Architectures for Long Protein Sequences Analysis: Long and Quantized Approaches.
In *Brazilian Symposium on Bioinformatics (BSB)*, pages 1–11. SBC, 2024.
- [22] Gabriel B. Oliveira, Helio Pedrini, and Zanoni Dias.
Protein Molecular Function Annotation Based on Transformer Embeddings.
In *11th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 210–220. Springer, 2022.

- [23] Felipe L. Mello, Gabriel B. Oliveira, Helio Pedrini, and Zanoni Dias.
Prediction of Protein Molecular Functions Using Transformers.

In 21st International Conference on Artificial Intelligence and Soft Computing (ICAISC), pages 379–387. Springer, 2022.

- [24] Felipe L. Mello, Gabriel B. Oliveira, and Zanoni Dias.
Predição de Funções Moleculares de Proteínas utilizando Aprendizado de Máquina.

Technical Report IC-PFG-21-44, Institute of Computing, University of Campinas, December 2021.

- [25] Ana P. S. Dantas, Gabriel B. Oliveira, Helio Pedrini, Cid C. Souza, and Zanoni Dias.

The Multi-attribute Fairer Cover Problem.

In *12th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 163–177. Springer, 2023.

- [26] Ana P. S. Dantas, Gabriel B. Oliveira, Daiane M. Oliveira, Helio Pedrini, Cid C. Souza, and Zanoni Dias.

Algorithmic Fairness Applied to the Multi-Label Classification Problem.

In *18th International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 737–744. SciTePress, 2023.

- [27] Gabriel Oliveira, Lucas David, Rafael Padilha, Ana P. Silva, Francine Paula, Lucas Infante, Lucio Jorge, Patricia Xavier, and Zanoni Dias.
Bias Assessment in Medical Imaging Analysis: a Case Study on Retinal OCT Image Classification.
In 14th International Conference on Agents and Artificial Intelligence (ICAART), pages 574–580. SciTePress, 2022.
- [28] Thiago D. Chen, Gabriel B. Oliveira, and Zanoni Dias.
Ensemble of Patches for COVID-19 X-ray Image Classification.
In 14th International Conference on Agents and Artificial Intelligence (ICAART), pages 561–567. SciTePress, 2022.

- [29] Daniel Ferber, Felipe Vieira, João Dalben, Mariana Ferraz, Nicholas Sato, Gabriel Oliveira, Rafael Padilha, and Zanoni Dias.
Deep Learning-based COVID-19 diagnostics of low-quality CT images.
In *Brazilian Symposium on Bioinformatics (BSB)*, pages 69–80.
Springer, 2021.

Anotações de Funções de Proteínas Utilizando Aprendizado de Máquina e Alinhamento Local

Gabriel Bianchin de Oliveira

Orientador: Prof. Dr. Zanoni Dias

Coorientador: Prof. Dr. Hélio Pedrini

22 de Julho de 2025

Instituto de Computação

Universidade Estadual de Campinas