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

Gabriel Bianchin de Oliveira[1], **Hélio Pedrini**[1] **e Zanoni Dias**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (Unicamp)
Av. Albert Einstein, 1251 – 13.083-852 – Campinas – SP – Brasil

{gabriel.oliveira,helio,zanoni}@ic.unicamp.br

Abstract. *Advances in sequencing technologies have led to the determination of millions of protein sequences, while experimental annotation of their functions remains limited. As a result, computational protein function prediction has become a central challenge in bioinformatics, typically formulated as a large-scale hierarchical multi-label classification problem. In this thesis, we introduce two machine learning methods based on Transformer-derived embeddings, along with two ensemble approaches that combine these predictions with local sequence alignment. When evaluated on a dataset derived from CAFA5, the main benchmark in the field, the proposed methods consistently outperformed the leading approaches in the literature, establishing themselves as the new state-of-the-art for the prediction function problem using only amino acid sequences. We also present memory-efficient versions of the models and a publicly available web server for use by the scientific community.*

Resumo. *Com o avanço das tecnologias de sequenciamento, milhões de proteínas tiveram suas sequências determinadas, enquanto a anotação experimental de suas funções permanece limitada. A predição computacional de funções proteicas tornou-se, portanto, um problema central em bioinformática, caracterizado como uma tarefa de classificação multirrótulo hierárquica de larga escala. Nesta tese, propomos dois métodos baseados em aprendizado de máquina utilizando embeddings de modelos Transformers, bem como duas abordagens de ensemble que integram essas predições com alinhamento local de sequências. Avaliados na base derivada do CAFA5, principal conjunto de dados da área, os métodos propostos superaram consistentemente as principais abordagens da literatura, estabelecendo-se como os novos estado da arte para o problema de predição de funções proteicas a partir exclusivamente da sequência de aminoácidos. Além disso, apresentamos versões otimizadas em memória e um servidor Web público para uso da comunidade científica.*

1. Introdução

Proteínas desempenham papéis fundamentais em diversos processos biológicos. Com os avanços tecnológicos, especialmente nas técnicas de sequenciamento, tornou-se simples e relativamente barato determinar, em laboratório, as sequências de aminoácidos que as compõem. No entanto, a análise de características de mais alto nível, como suas funções específicas, ainda exige alto custo financeiro e laboratorial [Radivojac 2013].

A identificação dessas funções impacta diretamente aplicações como desenvolvimento de medicamentos, análises genéticas e estudos em saúde. Por exemplo, doenças

como fibrose cística e certos tipos de enfisema podem ser causadas por alterações estruturais que comprometem o funcionamento das proteínas [Dobson 1999].

Devido ao alto custo experimental, diversas pesquisas buscam predizer funções proteicas por meio de métodos computacionais. Entre eles, destacam-se abordagens baseadas na sequência de aminoácidos, que é a informação mais abundante disponível, incluindo técnicas de alinhamento local e métodos de aprendizado de máquina.

Com o avanço do processamento de linguagem natural, especialmente dos Transformers [Vaswani et al. 2017], foram obtidos progressos significativos na análise de proteínas [Elnaggar et al. 2023, Elnaggar et al. 2021, Lin et al. 2023]. Estudos recentes mostram que a combinação dessas técnicas com alinhamento local pode alcançar alto desempenho [Cao and Shen 2021, Zhapa-Camacho et al. 2024, Zhu et al. 2022]. Ainda assim, os métodos existentes apresentam limitações relevantes: dependem de truncamento da sequência para viabilizar o processamento, utilizam combinações lineares fixas na integração com alinhamento local e não exploram estratégias supervisionadas de agregação entre representações diferentes.

Nesta tese de doutorado, apresentamos uma metodologia eficaz para a anotação de funções de proteínas utilizando exclusivamente a sequência de aminoácidos, combinando aprendizado de máquina e alinhamento local. Os métodos propostos endereçam as limitações da literatura por meio de janelas não sobrepostas para eliminação do truncamento, agregação supervisionada via *stacking* como alternativa a combinações lineares fixas, e integração estrutural do alinhamento local ao modelo de aprendizado. Sob a perspectiva da Ciência da Computação, esta tese propõe avanços metodológicos em aprendizado multirrótulo hierárquico em larga escala, integrando aprendizado de representações, agregação supervisionada e técnicas de eficiência computacional voltadas a aplicações reais. As principais contribuições desta pesquisa são listadas a seguir:

1. Desenvolvimento de uma nova métrica de avaliação, denominada IAuPRC, baseada na área sob a curva precisão-revocação interpolada, com o objetivo de reduzir efeitos indesejados da AuPRC na comparação de métodos, especialmente em regiões de alta revocação;
2. Implementação de dois novos métodos computacionais baseados em aprendizado de máquina, chamados SUPERMAGO e SUPERMAGOv2, e de dois novos métodos computacionais baseados em *ensemble* de aprendizado de máquina com alinhamento local de proteínas, nomeados SUPERMAGO+ e SUPERMAGOv2+;
3. Obtenção de desempenho de estado da arte com os métodos apresentados, tanto no cenário de abordagens baseadas em aprendizado de máquina quanto no de métodos de *ensemble* que combinam aprendizado de máquina com alinhamento local de proteínas;
4. Desenvolvimento de um servidor *web* baseado no SUPERMAGO+ e no SUPERMAGOv2+, capaz de predizer funções de proteínas a partir da sequência de aminoácidos, necessitando de menos recursos computacionais do que as versões originais dos métodos.

O restante deste resumo está organizado da seguinte forma. Na Seção 2, descrevemos os trabalhos relacionados com a tarefa de anotação de funções de proteínas. Na Seção 3, detalhamos os métodos desenvolvidos na pesquisa, a base de dados, as métricas de avaliação e os métodos comparados. Na Seção 4, avaliamos os métodos propostos

comparados com as abordagens da literatura. Na Seção 5, apresentamos as publicações realizadas durante o doutorado. Na Seção 6, apresentamos as conclusões e possíveis caminhos para trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção, apresentamos os principais trabalhos relacionados com o tema investigado. A revisão da literatura completa da área pode ser encontrada no texto da tese.

A anotação automática de funções de proteínas a partir da sequência de aminoácidos tem sido abordada por três linhas metodológicas: (i) modelos baseados exclusivamente em aprendizado profundo sobre sequência; (ii) métodos baseados em alinhamento local; e (iii) abordagens híbridas que combinam ambas as estratégias.

2.1. Métodos Baseados em Sequência de Aminoácidos

Antes do desenvolvimento de modelos baseados em Transformers, a literatura explorou arquiteturas profundas tradicionais, incluindo redes neurais multicamadas, redes convolucionais e redes recorrentes. Essas abordagens geralmente utilizam representações *one-hot encoding* ou fragmentação em k-mers como entrada, seguidas por classificadores multirrótulo [Ranjan et al. 2021, Kulmanov and Hoehndorf 2019, Xia et al. 2022]. Embora tenham apresentado avanços importantes, esses métodos enfrentam limitações relacionadas à capacidade de modelar dependências de longo alcance e à necessidade de truncamento das sequências para viabilizar o treinamento.

Com a introdução de modelos de linguagem pré-treinados em grandes bases proteicas, os métodos baseados em representações (do inglês, *embeddings*) de Transformers tornaram-se o estado da arte [Cao and Shen 2021, Zhu et al. 2022, Chua et al. 2024, Liu et al. 2024, Zhapa-Camacho et al. 2024]. Esses trabalhos utilizam representações extraídas de camadas profundas e treinam meta-classificadores para predição multirrótulo. Em geral, tais métodos ainda dependem de truncamento da sequência e não exploram de forma sistemática estratégias de agregação ou integração com alinhamento local.

2.2. Métodos Baseados em Alinhamento Local

Métodos baseados exclusivamente em homologia utilizam ferramentas como BLASTp [Altschul et al. 1997] e DIAMOND [Buchfink et al. 2021] para transferir anotações de proteínas similares. Embora eficazes quando há similaridade significativa, essas abordagens apresentam limitações em cenários de homologia remota ou quando não existem proteínas representativas na base de dados.

2.3. Abordagens Híbridas

Diversos trabalhos combinam aprendizado profundo com alinhamento local por meio de combinações lineares das predições [Kulmanov and Hoehndorf 2019, Cao and Shen 2021, Zhu et al. 2022, Liu et al. 2024]. Apesar de apresentarem melhorias consistentes em relação às abordagens isoladas, tais métodos normalmente utilizam esquemas simples de combinação e não exploram estratégias supervisionadas de agregação ou integração estrutural entre as fontes de informação.

3. Metodologia

Nesta seção, apresentamos os métodos propostos (SUPERMAGO e SUPERMAGOv2), a base de dados, métricas de avaliação e abordagens estado da arte comparadas.

3.1. SUPERMAGO

O SUPERMAGO é composto por três etapas: (i) extração de características via Transformers; (ii) classificação independente por camada; e (iii) agregação supervisionada.

Extração de Características Utilizamos os modelos pré-treinados ESM2 T36 [Lin et al. 2023] e ProtT5 [Elnaggar et al. 2021] para extrair *embeddings* das cinco últimas camadas de cada arquitetura. Para evitar truncamento, proteínas com mais de 1.022 aminoácidos são divididas em janelas não sobrepostas, e as representações finais são obtidas pela média das subsequências. Para cada camada, calculamos a média da representação de cada um dos *tokens*, obtendo um vetor fixo por proteína. Ao final, cada proteína é representada por dez vetores (cinco por modelo).

Classificação por Camada Para cada vetor extraído, treinamos um classificador do tipo rede neural *multilayer perceptron* independente, resultando em dez modelos por ontologia. Essa estratégia permite explorar a complementaridade informacional entre camadas profundas dos modelos de linguagem.

Agregação As predições dos classificadores são combinadas por meio de um modelo de agregação do tipo *stacking* baseado em redes neurais, com pesos específicos por termo da ontologia. Diferentemente de combinações lineares fixas, que são amplamente adotadas na literatura, os pesos são aprendidos de forma supervisionada, permitindo que cada termo receba contribuições distintas das camadas e modelos.

SUPERMAGO+ Para a criação do *ensemble* de aprendizado de máquina com alinhamento local de proteínas, incorporamos o DIAMOND ao SUPERMAGO. Para obter as predições do DIAMOND, as anotações são transferidas por média ponderada pelo *bitscore*. A combinação com o SUPERMAGO é realizada por meio do mesmo mecanismo de *stacking* supervisionado, com tratamento explícito para casos em que o alinhamento não retorna *hits*. Nessa situação, apenas as predições do SUPERMAGO são consideradas, garantindo robustez em cenários sem homologia detectável.

3.2. SUPERMAGOv2

O SUPERMAGOv2 estende o método anterior ao introduzir duas inovações principais: (i) a incorporação explícita da informação de alinhamento como característica estruturada no modelo e (ii) uma estratégia alternativa de representação baseada na transformação dos *embeddings* em imagens. Além disso, diferentemente do SUPERMAGO, que utiliza as cinco últimas camadas dos modelos de linguagem, o SUPERMAGOv2 emprega apenas as três últimas camadas, priorizando as representações mais semânticas e reduzindo a complexidade do modelo.

Integração de Alinhamento como Característica Além do *embedding* original da proteína na etapa de classificação por camada, introduzimos um vetor ponderado pelas proteínas alinhadas via DIAMOND. Os pesos são calculados considerando *bitscore* e

identidade normalizada dos alinhamentos. O vetor final ponderado é utilizado juntamente com o *embeddings* original como entrada para redes neurais, permitindo incorporar homologia de forma diferenciável e estruturada ao modelo.

Classificação por Imagem Adicionalmente, introduzimos um módulo de classificação por imagem, no qual os *embeddings* das três últimas camadas dos modelos ESM2 T36 e ProtT5 são concatenados, normalizados e reorganizados em matrizes tridimensionais, formando imagens com três canais classificadas via ajuste fino da ResNet50 [He et al. 2016]. A inclusão desse módulo contribui para ganhos de desempenho, conforme evidenciado pelo estudo de ablação disponível no texto da tese.

SUPERMAGOv2+ Para a criação do *ensemble* de aprendizado de máquina com alinhamento local de proteínas baseado no SUPERMAGOv2, adicionamos as predições do DIAMOND a partir de uma média ponderada que considera simultaneamente o *bitscore* e identidade. Assim como no SUPERMAGO+, caso não haja anotações para o DIAMOND, apenas a predição do SUPERMAGOv2 é utilizada.

3.3. Base de Dados

Para comparar os métodos propostos com a literatura, utilizamos um conjunto de dados derivado do quinto e mais recente Desafio de Avaliação Crítica de Anotações de Funções de Proteínas (do inglês, *Critical Assessment of protein Function Annotation*), denominado CAFA Challenge[1].

Em relação às funções exercidas pelas proteínas, elas são organizadas seguindo a Ontologia Genética (do inglês, *Gene Ontology* [Consortium 2004]), dividida em três categorias: ontologia de Componente Celular (CC), ontologia de Função Molecular (FM) e ontologia de Processo Biológico (PB). Essas três ontologias são estruturadas em um formato de grafo acíclico direcionado, de forma que os termos mais próximos do termo raiz são mais gerais e termos mais profundos representam condições mais específicas, formando uma relação hierárquica entre descendentes e ancestrais. Caso uma proteína seja designada para um termo mais profundo do grafo, ela também é representada por todos os termos ancestrais, o que transforma a tarefa em um problema de classificação multirrótulo.

Utilizamos o conjunto de dados do CAFA5, contendo aproximadamente 142 mil proteínas, o qual foi dividido em 80% para treinamento, 10% para validação e 10% para teste. Foram considerados os termos mais frequentes de cada ontologia, com frequência mínima de 0,1%, 0,1% e 1% para CC, FM e PB, respectivamente.

3.4. Métricas de Avaliação

Para avaliar os métodos propostos e compará-los com a literatura, utilizamos as métricas quantitativas $F_{\max}$, $F^*_{\max}$, $wF_{\max}$, $S_{\min}$ e AuPRC, amplamente utilizadas na literatura. As definições e equações de cada uma dessas medidas podem ser encontradas no texto da tese. Além dessas medidas, apresentamos também análises estatísticas no texto da tese que não serão detalhadas neste resumo.

[1]https://biofunctionprediction.org/cafa

Em complemento das métricas usadas na literatura, propusemos a medida IAuPRC, que representa a área sob a curva precisão e revocação interpolada (do inglês, *Interpolated Area under Precision-Recall Curve*). O objetivo dessa medida é não gerar punições para métodos capazes de realizar classificações que resultam em altos valores de revocação, pois altos valores de revocação indicam baixa presença de falsos negativos, como é feito pela métrica AuPRC.

3.5. Métodos Comparados

Nesta subseção, apresentamos brevemente os métodos comparados com a abordagem proposta. As comparações foram divididas em duas categorias: aprendizado de máquina e *ensemble* de aprendizado de máquina com alinhamento local. Todos os métodos comparados foram reexecutados e avaliados sob o mesmo protocolo experimental, garantindo uma comparação justa e controlada.

3.6. Aprendizado de Máquina

Para a avaliação dos modelos SUPERMAGO e SUPERMAGOv2, comparamos os métodos desenvolvidos com as seguintes abordagens estado da arte: PFmulDL [Xia et al. 2022], DeepGO [Kulmanov et al. 2018], DeepGOCNN [Kulmanov and Hoehndorf 2019], TALE [Cao and Shen 2021], ATGO [Zhu et al. 2022], PU-GO [Zhapa-Camacho et al. 2024], PROTGOAT [Chua et al. 2024], InterLabelGO [Liu et al. 2024], TEMPROT [Oliveira et al. 2023] e MAGO [Oliveira et al. 2024].

3.7. *Ensemble* de Aprendizado de Máquina com Alinhamento Local

Para a avaliação dos modelos SUPERMAGO+ e SUPERMAGOv2+, comparamos os métodos desenvolvidos com as seguintes abordagens estado da arte: DeepGOPlus [Kulmanov and Hoehndorf 2019], TALE+ [Cao and Shen 2021], ATGO+ [Zhu et al. 2022], PU-GO+Diamond [Zhapa-Camacho et al. 2024], InterLabelGO+ [Liu et al. 2024], TEMPROT+ [Oliveira et al. 2023] e MAGO+ [Oliveira et al. 2024].

4. Resultados Experimentais

Neste resumo, apresentamos os resultados considerando a métrica $F_{\max}$, que é a principal medida utilizada na literatura. A comparação com as demais métricas e análises relacionadas a avaliação por similaridade de sequência, domínio, nível da ontologia, frequência dos termos, tipos de termos, análises estatísticas e estudo de ablação estão disponíveis no texto da tese.

As Tabelas 1 e 2 apresentam os resultados dos métodos de aprendizado de máquina e *ensemble* de aprendizado de máquina com alinhamento local para as três ontologias, respectivamente. Os resultados indicam que o SUPERMAGOv2 e o SUPERMAGOv2+ apresentaram os melhores resultados entre os métodos avaliados, com destaque para CC considerando os modelos *ensemble*, com um aumento de 1,2 pontos percentuais. Os testes estatísticos indicam que as diferenças observadas nas Tabelas 1 e 2 são estatisticamente significativas na maioria dos cenários avaliados, com a análise completa disponível no texto da tese.

Tabela 1. Resultados em $F_{\max}$ dos métodos de aprendizado de máquina. O melhor resultado por coluna está em negrito e o segundo melhor está sublinhado.

Método	CC	FM	PB
PFmulDL	0,707	0,693	0,460
DeepGO	0,655	0,465	0,372
DeepGOCNN	0,678	0,688	0,497
TALE	0,693	0,685	0,427
ATGO	0,724	0,749	0,505
PU-GO	0,692	0,703	0,463
PROTGOAT	0,763	0,794	0,572
InterLabelGO	<u>0,778</u>	<u>0,807</u>	<u>0,638</u>
TEMPROT	0,731	0,762	0,514
MAGO	0,755	0,793	0,555
SUPERMAGO	0,769	0,802	0,582
SUPERMAGOv2	**0,786**	**0,814**	**0,644**

Tabela 2. Resultados em $F_{\max}$ dos métodos de *ensemble* de aprendizado de máquina com alinhamento local. O melhor resultado por coluna está em negrito e o segundo melhor está sublinhado.

Método	CC	FM	PB
DIAMOND	0,699	0,742	0,588
BLASTp	0,712	0,763	0,576
DeepGOPlus	0,738	0,784	0,601
TALE+	0,753	0,770	0,580
ATGO+	0,743	0,787	0,569
PU-GO+Diamond	0,751	0,793	0,620
InterLabelGO+	0,775	0,806	<u>0,647</u>
TEMPROT+	0,752	0,794	0,588
MAGO+	0,774	0,801	0,615
SUPERMAGO+	<u>0,777</u>	<u>0,809</u>	0,619
SUPERMAGOv2+	**0,789**	**0,815**	**0,652**

Por fim, também foram desenvolvidas versões eficientes em memória dos métodos propostos, denominadas SUPERMAGO+Web e SUPERMAGOv2+Web. Essas versões utilizam modelos de linguagem mais compactos, substituindo o ESM2 T36 pelo ESM2 T12, além do uso de quantização e técnicas de adaptação de baixo custo, reduzindo significativamente os requisitos computacionais, ao mesmo tempo em que mantêm desempenho próximo às versões originais.

Avaliações experimentais demonstraram que as versões otimizadas apresentam resultados comparáveis aos métodos completos e não possuem diferenças estatisticamente significativas em relação a eles. Com base nessas versões, foi desenvolvido um servidor *Web*[2] público que permite a predição de funções proteicas a partir de sequências no formato FASTA, possibilitando o uso prático dos métodos pela comunidade científica com

[2]https://supermago.ic.unicamp.br

baixo custo computacional.

5. Publicações

A pesquisa realizada nesta tese resultou diretamente na publicação de um relatório técnico, quatro artigos apresentados em conferências e três artigos publicados em revistas internacionais. Além do escopo da tese, nossa contribuição na área de aprendizado de máquina resultou em cinco artigos apresentados em conferências e um artigo publicado em revista internacional adicionais durante o período do doutorado. A Tabela 3 apresenta um panorama geral dos resultados obtidos.

Tabela 3. Sumarização das publicações realizadas durante o período do doutorado, apresentando a quantidade de artigos publicados em cada veículo. A coluna "Tese" se refere a quantidade de publicações ligadas diretamente com a tese, enquanto a coluna "Extra" indica as publicações adicionais na área de aprendizado de máquina.

Periódico	**Qualis**	**Tese**	**Extra**
Applied Soft Computing	A1	–	1
BMC Bioinformatics	A1	1	–
IEEE Access	A1	1	–
PROTEINS: Structure, Function, and Bioinformatics	A2	1	–
Conferência	**Qualis**	**Tese**	**Extra**
Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)	A1	1	–
Brazilian Symposium on Bioinformatics (BSB)	B1	1	1
Brazilian Conference on Intelligent Systems (BRACIS)	A4	1	1
International Conference on Artificial Intelligence and Soft Computing (ICAISC)	A4	1	–
International Conference on Computer Vision Theory and Applications (VISAPP)	A3	–	1
International Conference on Agents and Artificial Intelligence (ICAART)	A4	–	2
Relatório Técnico		**Tese**	**Extra**
Relatório técnico de Projeto Final de Graduação		1	–
Total		**Tese**	**Extra**
Número de publicações		8	6

Todos os métodos desenvolvidos nesta tese, bem como modelos e recursos associados, estão disponíveis publicamente para garantir reprodutibilidade e facilitar o uso pela comunidade científica. A Tabela 4 apresenta os principais recursos disponibilizados.

Tabela 4. Recursos públicos disponibilizados durante a tese.

Recurso	**Link**
Servidor *Web*	https://supermago.ic.unicamp.br
SUPERMAGOv2	https://github.com/gabrielbianchin/SUPERMAGOv2
SUPERMAGO	https://github.com/gabrielbianchin/SUPERMAGO
MAGO	https://github.com/gabrielbianchin/MAGO
TEMPROT	https://github.com/gabrielbianchin/TEMPROT
TEMPO / DS-TEMPO	https://github.com/gabrielbianchin/TEMPO-and-DSTEMPO-MF
ICAISC Model	https://github.com/gabrielbianchin/ProteinFunctionTransformers
ESM2 Models	https://huggingface.co/collections/gabrielbianchin/esm2-models
Conjunto de dados	https://zenodo.org/records/10982903

6. Conclusões

Nesta tese, investigamos a combinação de modelos Transformers pré-treinados com alinhamento local para a predição de funções de proteínas segundo a Ontologia Genética. Foram propostos os métodos SUPERMAGO e SUPERMAGOv2, bem como suas variações em *ensemble* e versões eficientes em memória.

Os resultados experimentais demonstraram que SUPERMAGOv2 e SUPERMAGOv2+ superam consistentemente as abordagens mais avançadas da literatura, estabelecendo novo estado da arte tanto no cenário de aprendizado de máquina quanto no de ensembles com alinhamento local.

Além do desempenho superior, demonstramos que é possível reduzir significativamente os requisitos computacionais por meio de técnicas de compressão e adaptação de modelos, mantendo desempenho estatisticamente equivalente. Como perspectivas futuras, destacam-se a integração de dados estruturais adicionais e o avanço em interpretabilidade dos modelos.

Referências

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, 25(17):3389–3402.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive Protein Alignments at Tree-of-Life Scale using DIAMOND. *Nature Methods*, 18(4):366–368.

Cao, Y. and Shen, Y. (2021). TALE: Transformer-based Protein Function Annotation with Joint Sequence–Label Embedding. *Bioinformatics*, 37(18):2825–2833.

Chua, Z. M., Rajesh, A., Sinha, S., and Adams, P. D. (2024). PROTGOAT: Improved Automated Protein Function Predictions Using Protein Language Models. *bioRxiv*, pages 1–15.

Consortium, G. O. (2004). The Gene Ontology (GO) Database and Informatics Resource. *Nucleic Acids Research*, 32(suppl_1):D258–D261.

Dobson, C. M. (1999). Protein Misfolding, Evolution and Disease. *Trends in Biochemical Sciences*, 24(9):329–332.

Elnaggar, A., Essam, H., Salah-Eldin, W., Moustafa, W., Elkerdawy, M., Rochereau, C., and Rost, B. (2023). Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. *arXiv:2301.06568*, pages 1–29.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2021). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE.

Kulmanov, M. and Hoehndorf, R. (2019). DeepGOPlus: Improved Protein Function Prediction from Sequence. *Bioinformatics*, 36(2):422–429.

Kulmanov, M., Khan, M. A., and Hoehndorf, R. (2018). DeepGO: Predicting Protein Functions from Sequence and Interactions using a Deep Ontology-Aware Classifier. *Bioinformatics*, 34(4):660–668.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Costa, A. d. S., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2023). Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science*, 379(6637):1123–1130.

Liu, Q., Zhang, C., and Freddolino, L. (2024). InterLabelGO+: unraveling label correlations in protein function prediction. *Bioinformatics*, 40(11):btae655.

Oliveira, G. B., Pedrini, H., and Dias, Z. (2023). TEMPROT: Protein Function Annotation using Transformers Embeddings and Homology Search. *BMC Bioinformatics*, 24(1):1–16.

Oliveira, G. B., Pedrini, H., and Dias, Z. (2024). 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.

Radivojac, P. (2013). A (Not So) Quick Introduction to Protein Function Prediction. *Indiana University, USA*.

Ranjan, A., Fernández-Baca, D., Tripathi, S., and Deepak, A. (2021). An Ensemble Tf-Idf Based Approach to Protein Function Prediction via Sequence Segmentation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(5):2685–2696.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All You Need. In *30th Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Xia, W., Zheng, L., Fang, J., Li, F., Zhou, Y., Zeng, Z., Zhang, B., Li, Z., Li, H., and Zhu, F. (2022). 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.

Zhapa-Camacho, F., Tang, Z., Kulmanov, M., and Hoehndorf, R. (2024). Predicting Protein Functions using Positive-Unlabeled Ranking with Ontology-Based Priors. *bioRxiv*, pages 1–9.

Zhu, Y.-H., Zhang, C., Yu, D.-J., and Zhang, Y. (2022). Integrating Unsupervised Language Model with Triplet Neural Networks for Protein Gene Ontology Prediction. *PLoS Computational Biology*, 18(12):e1010793.