# BioAutoML: Democratizing Machine Learning in Life Sciences

**Robson Parmezan Bonidia**[1,2],
**André Carlos Ponce de Leon Ferreira de Carvalho**[1]

[1]Institute of Mathematics and Computer Sciences, University of São Paulo,
São Carlos, 13566-590, São Paulo, Brazil
[2]Department of Computer Science, Federal University of Technology-Paraná (UTFPR)
Cornélio Procópio 86300-000, Brazil

rpbonidia@gmail.com, andre@icmc.usp.br

***Abstract.*** *Recent technological advances have allowed an exponential expansion of biological sequence data, and the extraction of meaningful information through Machine Learning (ML) algorithms. This knowledge improved the understanding of the mechanisms related to several fatal diseases, e.g., Cancer and COVID-19, helping to develop innovative solutions, such as CRISPR-based gene editing, coronavirus vaccine, and precision medicine. These advances benefit our society and economy, directly impacting people's lives in various areas, such as health care, drug discovery, forensic analysis, and food analysis. Nevertheless, ML approaches to biological data require representative, quantitative, and informative features. Necessarily, as many ML algorithms can handle only numerical data, sequences need to be translated into a feature vector. This process, known as feature extraction, is a fundamental step for elaborating high-quality ML-based models in bioinformatics, by allowing the feature engineering stage, with the design and selection of suitable features. Feature engineering, ML algorithm selection, and hyperparameter tuning are often time-consuming processes that require extensive domain knowledge and are performed by a human expert. To deal with this problem, we developed a new package, BioAutoML, which automatically runs an end-to-end ML pipeline. BioAutoML extracts numerical and informative features from biological sequence databases, automating feature selection, recommendation of ML algorithm(s), and tuning of hyperparameters, using Automated ML (AutoML). Our experimental results demonstrate the robustness of our proposal across various domains, such as SARS-CoV-2, anticancer peptides, HIV sequences, and non-coding RNAs. BioAutoML has a high potential to significantly reduce the expertise required to use ML pipelines, aiding researchers in combating diseases, particularly in low- and middle-income countries. This initiative can provide biologists, physicians, epidemiologists, and other stakeholders with an opportunity for widespread use of these techniques to enhance the health and well-being of their communities.*

## 1. Introdução, Desafios Centrais e Motivação

Devido à expansão e complexidade inerente dos dados biológicos, métodos de Inteligência Artificial (IA), especificamente Aprendizado de Máquina (AM), têm

demonstrado ampla aplicabilidade nas ciências biológicas [Volkamer et al. 2023]. Os algoritmos de AM podem extrair conhecimento útil e significativo de dados biológicos [Chen et al. 2021], acelerando descobertas, reduzindo despesas de pesquisa e aumentando a eficiência científica [Sharma et al. 2021]. Esses avanços beneficiam diretamente a sociedade, a economia e a vida das pessoas. Além disso, algoritmos de AM foram utilizados com sucesso para mitigar o impacto de problemas relacionados com a saúde, como a Pandemia de COVID-19 [Kamalov et al. 2023], diagnóstico de câncer [Painuli et al. 2022] e tecnologia de edição genética baseada em CRISPR/Cas9 [Mitrofanov et al. 2020].

Apesar de sua ampla aplicação, projetar soluções de AM robustas e confiáveis geralmente requer conhecimentos que não são comumente encontrados em pesquisadores da área de biologia e saúde, causando graves desigualdades [Rubeis et al. 2022]. Por exemplo, desigualdade de acessibilidade (isso cria uma disparidade em quem pode usar ferramentas poderosas, muitas vezes desfavorecendo aqueles que trabalham em instituições menores ou em regiões com poucos recursos); e desigualdade de conhecimento (a complexidade dos algoritmos de AM e as habilidades necessárias constituem uma barreira, limitando o potencial para pesquisa inovadora). Neste contexto, democratizar a IA implica conceder acessibilidade ao AM para indivíduos que não são especialistas, por exemplo, indivíduos sem formação em ciência de dados, matemática ou informática. Esta democratização deve capacitar diversas partes interessadas com contribuições únicas baseadas na sua aptidão, disponibilidade, dedicação e rapidez, proporcionando oportunidades iguais ao nível global.

Além desses desafios, um dos principais obstáculos à aplicação de algoritmos de AM para sequências biológicas é a natureza não estruturada de muitos desses dados, uma vez que a maioria dos algoritmos, inclusive aqueles que produzem modelos interpretáveis, só funcionam com dados estruturados. Este problema pode ser resolvido com técnicas de extração de características para representar os dados originalmente não estruturados em um formato estruturado. No entanto, as características devem capturar as informações relevantes presentes na sequência biológica, uma vez que o desempenho preditivo do modelo induzido por um algoritmo de AM depende fortemente da representatividade do vetor de características de entrada. Esses processos geralmente exigem amplo conhecimento especializado, executado manualmente por um especialista humano, sendo uma das etapas mais demoradas.

Para mitigar essas limitações, métodos de AM automatizado (AutoML, do inglês *Automated Machine learning*) estão sendo usados para democratizar o acesso e o uso eficaz de algoritmos de AM por não especialistas. O AutoML tem sido aplicado a dados de sequências biológicas, proporcionando soluções robustas, como autoBioSeqpy [Jing et al. 2020], AutoGenome [Liu et al. 2021], iLearn [Chen et al. 2019] e iLearnPlus [Chen et al. 2021]. No entanto, a maioria delas aplicam propostas de AutoML que não automatizam todo o processo, conhecido como AM ponta a ponta, nem consideram as especificidades dos dados de sequência. As duas primeiras dessas ferramentas cobrem apenas a etapa de modelagem de dados. As duas últimas, iLearn e iLearnPlus, incluem mais etapas, mas não automatizam a extração de características de dados não estruturados. No entanto, conforme a *International Data Corporation*

*(IDC)*[1], até 2025, cerca de 80% dos dados gerados serão não estruturados.

Essas limitações motivaram o desenvolvimento de um novo pacote de software de código aberto, chamado BioAutoML[2][3], projetado para facilitar a extração e seleção automática de características de dados sequenciais, considerando uma variedade de aspectos. Além disso, oferece recomendações de algoritmos e ajuste de hiperparâmetros específicos para a classificação multi-classe e binária em contextos biológicos, simplificando significativamente o processo de análise de dados complexos nesse campo. BioAutoML é uma ferramenta completa de AM ponta a ponta para experimentos usando sequências biológicas. Sendo assim, esta tese tem a seguinte hipótese:

- **Hipótese:** BioAutoML pode recomendar pipelines eficientes e robustos para representar sequências biológicas, automatizar a seleção de características, recomendação de algoritmos e ajuste de hiperparâmetros. Isso reduz o demorado estágio de pré-processamento, ao mesmo tempo que mantém ou melhora o desempenho dos modelos preditivos, diminuindo, consequentemente, o conhecimento necessário para usar pipelines de AM para análise de sequências biológicas.

Finalmente, para apoiar nossa proposta, conduzimos uma revisão sistemática da literatura, durante a qual identificamos 29 estudos que desenvolveram soluções para análise de sequências biológicas, incluindo pacotes, servidores web e kits de ferramentas. Apesar da existência de estudos, os usuários ainda precisam ter um entendimento e conhecimento técnico do campo para sua execução. BioAutoML destaca-se como uma solução completamente automatizada, cobrindo todo o espectro de análise de sequências biológicas com AM.

## 2. Resumo dos Resultados, Discussões e Principais Contribuições

BioAutoML não apenas automatiza tarefas complexas, mas também permite que pesquisadores sem conhecimento de domínio apliquem algoritmos de AM para análise de dados de sequência biológica. A capacidade de gerar um pipeline de AM automatizado de ponta a ponta reduz a carga trabalhosa do pré-processamento manual de dados. As contribuições desta pesquisa são multifacetadas, estendendo-se desde avanços teóricos até aplicações práticas. Elas são resumidas da seguinte forma:

- Uma revisão sistemática da literatura para apresentar, resumir e estudar ferramentas (ou pacotes e servidores web) baseadas em AM que têm como proposta fornecer diversos descritores de características para classificação de sequências biológicas;
- Um novo pipeline de extração de características usando recursos matemáticos;
- Uma nova técnica de extração de características baseada na entropia de Tsallis;
- Um novo pacote, chamado MathFeature, que introduz abordagens matemáticas para a extração de características de sequências biológicas, permitindo análises mais profundas, uma capacidade não disponível em pacotes existentes;

[1]https://www.idc.com/
[2]https://github.com/Bonidia/BioAutoML
[3]https://bonidia.github.io/BioAutoML-WP/

- BioAutoML representa, segundo o conhecimento dos autores, a automação mais extensa de pipelines até o momento para sequências biológicas, abrangendo engenharia de características, seleção de algoritmos de AM e ajuste de hiperparâmetros.

Adicionalmente, a proposta central desta tese, BioAutoML, demonstrou alcançar resultados robustos em diversos domínios de problemas, evidenciando casos de sucesso em áreas como SARS-CoV-2, peptídeos anticancerígenos, peptídeos pró-inflamatórios, sequências de HIV-1, proteínas secretadas não clássicas, promotores sigma70, pontos de recombinação, pequenos RNAs não codificantes, longos RNAs não codificantes, RNAs circulares e outros. BioAutoML apresenta potencial de diminuir substancialmente a experiência necessária para operar pipelines de AM. Este apoio ajuda os investigadores a abordar diversas questões, incluindo doenças que afetam profundamente a vida humana, dando aos biólogos e outras partes interessadas uma oportunidade para a utilização generalizada destas técnicas.

Além disso, ao longo de um período de estágio na Alemanha, no *Helmholtz Centre for Environmental Research — UFZ*, Leipzig, o BioAutoML foi submetido a testes com problemas reais, a fim de aprimorar suas funcionalidades, adaptando-as para enfrentar os desafios práticos encontrados por biólogos, microbiologistas, e virologistas em seu trabalho diário. Por fim, os resultados desta tese também geraram prêmios, bolsas e publicações em revistas científicas de alto impacto. Os artigos e ferramentas associados à tese, até a escrita do artigo, conquistaram 119 estrelas no GitHub e aproximadamente 123 citações. Os principais artigos derivados desta tese têm um Fator de Impacto (IF, do inglês *Impact Factor*) acumulado de 63,064. Um resumo dos artigos e prêmios é apresentado a seguir:

- **Google Latin America Research Awards (LARA), 2021**: BioAutoML foi eleito pelo LARA-Google entre as 24 ideias mais promissoras da América Latina (24 projetos premiados, de uma base de 700 inscrições);
- **AutoAI-Pandemics (Democratizando AM para não especialistas, 2023)**, selecionado como um dos projetos mais promissores dentre um total de 221 propostas de 47 países em uma competição global realizada pela *Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network* – AI4PEP, conquistando um financiamento de 362.500 dólares canadenses.
- **Prêmio Helmholtz para Pesquisador Visitante** (*Helmholtz Information & Data Science Academy (HIDA)*);
- **Prêmio de Pesquisa e Treinamento FEMS** (*Federation of European Microbiological Societies*);
- **BioPrediction (uma versão aprimorada do BioAutoML para trabalhar com interações entre sequências biológicas)** foi selecionado para participar do *Prototypes for Humanity 2023*, durante a COP28-Dubai, escolhido entre 3.000 inscritos, de mais de 100 países, destacando-se entre os 100 melhores do mundo.
- **Finalista (TOP 15 de 82)** no *Falling Walls Lab Brazil 2022*;
- **Artigos**: (1) Briefings in Bioinformatics (IF 2020: 11.622) [Bonidia et al. 2021]; (2) Entropy (IF 2021: 2.738) [Bonidia et al. 2022a];

(3) IEEE Access (IF 2019: 3.745) [Bonidia et al. 2020]; (4) Nucleic Acids Research (IF 2020: 16.971) [Alkhnbashi et al. 2021]; (5) Briefings in Bioinformatics (IF 2021: 13.994) [Bonidia et al. 2022b, Bonidia et al. 2022c].

Esses resultados relatam que BioAutoML pode inovar como a análise de dados de sequência biológica é conduzida, democratizando o acesso a ferramentas avançadas de AM para pesquisadores e comunidades ao redor do mundo, desbloqueando o potencial para inovações em vários campos da biologia e medicina. Com o BioAutoML, pesquisadores em laboratórios menores, sem acesso extensivo a especialistas em dados, podem realizar análises complexas de metabolismo e expressão gênica. Grupos comunitários e pesquisadores em regiões menos desenvolvidas podem utilizar o BioAutoML para impulsionar pesquisas na função e a evolução dos genes e das proteínas.

O BioAutoML também serve como uma ponte, permitindo que investigadores sem formação avançada em computação apliquem técnicas sofisticadas de AM. Ao oferecer uma maneira simplificada de realizar análises complexas de dados biológicos, BioAutoML promove uma inclusão mais ampla de pesquisadores de diferentes origens e recursos, fortalecendo o empenho global em ciência e saúde. Finalmente, não é mais aceitável que as aplicações de AM nas Ciências da Vida, ou em qualquer outro campo, permaneçam confinadas ao conhecimento especializado. Isto significa uma mudança da exclusividade para a acessibilidade, tornando o AM um recurso partilhado para a melhoria coletiva da ciência e da sociedade.

## Agradecimentos

Este projeto recebe apoio financeiro de uma variedade de fontes, incluindo o Centro Internacional de Pesquisa para o Desenvolvimento do Canadá (IDRC) sob o número de concessão 109981, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Universidade de São Paulo. Além disso, o projeto é beneficiário do *Google Latin America Research Awards* (LARA), do Prêmio Pesquisador Visitante da Helmholtz e do Prêmio de Pesquisa e Treinamento da FEMS.

## Referências

Alkhnbashi, O. S., Mitrofanov, A., Bonidia, R., et al. (2021). CRISPRloci: comprehensive and accurate annotation of CRISPR–Cas systems. *Nucleic Acids Research*, 49(W1):W125–W130.

Bonidia, R. P., Avila Santos, A. P., de Almeida, B. L., Stadler, P. F., Nunes da Rocha, U., Sanches, D. S., and De Carvalho, A. C. (2022a). Information theory for biological sequence classification: A novel feature extraction technique based on tsallis entropy. *Entropy*, 24(10):1398.

Bonidia, R. P., Domingues, D. S., Sanches, D. S., and de Carvalho, A. C. (2022b). Mathfeature: feature extraction package for dna, rna and protein sequences based on mathematical descriptors. *Briefings in Bioinformatics*, 23(1):bbab434.

Bonidia, R. P., Machida, J. S., Negri, T. C., Alves, W. A. L., Kashiwabara, A. Y., Domingues, D. S., De Carvalho, A., Paschoal, A. R., and Sanches, D. S. (2020). A novel decomposing model with evolutionary algorithms for feature selection in long non-coding rnas. *IEEE Access*, 8:181683–181697.

Bonidia, R. P., Sampaio, L. D. H., Domingues, D. S., Paschoal, A. R., Lopes, F. M., de Carvalho, A. C. P. L. F., and Sanches, D. S. (2021). Feature extraction approaches for biological sequences: a comparative study of mathematical features. *Briefings in Bioinformatics*. bbab011.

Bonidia, R. P., Santos, A. P. A., de Almeida, B. L. S., Stadler, P. F., da Rocha, U. N., Sanches, D. S., and de Carvalho, A. C. P. L. F. (2022c). BioAutoML: automated feature engineering and metalearning to predict noncoding RNAs in bacteria. *Briefings in Bioinformatics*, 23(4):bbac218.

Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., Akutsu, T., Daly, R., Webb, G., Zhao, Q., Kurgan, L., and Song, J. (2021). iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Research*. gkab122.

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., Zhu, Y., Powell, D. R., Akutsu, T., Webb, G. I., Chou, K.-C., Smith, A. I., Daly, R. J., Li, J., and Song, J. (2019). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings in Bioinformatics*, 21(3):1047–1057.

Jing, R., Li, Y., Xue, L., Liu, F., Li, M., and Luo, J. (2020). autobioseqpy: a deep learning tool for the classification of biological sequences. *Journal of Chemical Information and Modeling*, 60(8):3755–3764.

Kamalov, F., Cherukuri, A. K., Sulieman, H., Thabtah, F., and Hossain, A. (2023). Machine learning applications for covid-19: a state-of-the-art review. *Data Science for Genomics*, pages 277–289.

Liu, D., Xu, C., He, W., Xu, Z., Fu, W., Zhang, L., Yang, J., Wang, Z., Liu, B., Peng, G., et al. (2021). Autogenome: an automl tool for genomic research. *Artificial Intelligence in the Life Sciences*, 1:100017.

Mitrofanov, A., Alkhnbashi, O. S., Shmakov, S. A., Makarova, K., Koonin, E., and Backofen, R. (2020). CRISPRidentify: identification of CRISPR arrays using machine learning approach. *Nucleic Acids Research*, 49(4):e20–e20.

Painuli, D., Bhardwaj, S., et al. (2022). Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review. *Computers in Biology and Medicine*, 146:105580.

Rubeis, G., Dubbala, K., and Metzler, I. (2022). "democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term. *Frontiers in Genetics*, 13:902542.

Sharma, M. et al. (2021). Emerging trends of bioinformatics in health informatics. In *Computational Intelligence in Healthcare*, pages 343–367. Springer.

Volkamer, A., Riniker, S., Nittinger, E., Lanini, J., Grisoni, F., Evertsson, E., Rodríguez-Pérez, R., and Schneider, N. (2023). Machine learning for small molecule drug discovery in academia and industry. *Artificial Intelligence in the Life Sciences*, 3:100056.