Breaking Barriers: Democratizing Machine Learning for RNA-Protein Interaction Prediction in Life Sciences*

Bruno R. Florentino[1], **Robson P. Bonidia** [1,2], **André C. P. L. F. de Carvalho** [1]

[1]University of São Paulo, São Carlos, 13566-590, São Paulo, Brazil
[2]Department of Computer Science, Federal University of Technology-Paraná Cornélio Procópio 86300-000, Brazil

Abstract. *As biological sequence storage grows, extracting information becomes crucial for advances in health. The complexity of these sequences requires sophisticated techniques such as Machine Learning (ML). Nevertheless, developing strong ML solutions demands specialized knowledge that is often beyond the reach of many life sciences researchers, further widening disparities. Considering this, we present BioPrediction, an end-to-end ML framework that creates models to identify interactions between sequences, such as non-coding RNA (ncRNA) and protein pairs, without human intervention. The results highlight its superior performance over expert-crafted models across multiple datasets. This automation opens novel avenues for unraveling complex interactions and exploring disease mechanisms.*

Resumo. *À medida que o armazenamento de sequências biológicas aumenta, extrair informações torna-se crucial para avanços na saúde. A complexidade dessas sequências exige técnicas sofisticadas, como Aprendizado de Máquina (AM). No entanto, desenvolver soluções fortes de AM demanda conhecimento especializado, muitas vezes fora do alcance de muitos pesquisadores das ciências da vida, ampliando ainda mais as disparidades. Considerando isso, apresentamos o BioPrediction, um framework de AM ponta a ponta que cria modelos para identificar interações entre sequências, como pares de RNA não codificante e proteínas, sem intervenção humana. Os resultados destacam seu desempenho superior sobre modelos criados por especialistas em múltiplos conjuntos de dados. Essa automação abre novos caminhos para desvendar interações complexas e explorar mecanismos de doenças.*

1. Introduction

The advancement of technologies for exploring the cellular environment offers researchers new methods to study disease pathways, such as in oncology [Shaath et al. 2022] and infectious diseases [Zhong et al. 2021]. A notable innovation is Next Generation Sequencing (NGS) for DNA sequencing, which significantly amplifies the volume of biological data in databases [Jiang et al. 2022]. Consequently, information about various organisms is now available across numerous datasets [P and M. 2021]. To

*Note: The main author led the creation of BioPrediction with guidance from the other authors. BioPrediction evolved from BioAutoML [Bonidia et al. 2022], a tool for classifying RNAs into different categories. It's also a key part of the main author's ongoing PhD research. Some parts of this article are currently being reviewed by the Computational and Structural Biotechnology Journal (Impact Factor: 6).

effectively extract and utilize this wealth of data, is essential the development of computational tools capable of analyzing these biological sequences for various applications [Chicco 2017], e.g., ML techniques.

A sub-problem involving the analysis of biological sequences with ML is the interaction between non-coding RNAs (ncRNA) and proteins, commonly referred to as RPIs. ncRNAs are a class of genetic material that cannot simply be categorized as part of the non-essential DNA in the genome [Zhang et al. 2023], as they play a complex role with numerous functions in the organism [Kopp and Mendell 2018]. Different structures are present in ncRNAs. Among them, are Long Non-Coding RNAs (lncRNAs), which play a crucial role in regulating genetic expressions and chromatin [Kopp and Mendell 2018]. Furthermore, the expression levels of some lncRNAs are directly related to the initial regulation pathways of solid cancers, conferring them a significant role as biomarkers [Cantile et al. 2021]. Exploring these interactions can generate significant benefits in several fields, such as (1) cancer research and treatment [Shaath et al. 2022], (2) genetic disorders [Ferre et al. 2016, Qin et al. 2021], (3) viral infections [Wang et al. 2020], and (4) human disease [Armaos et al. 2021].

Despite the existence of numerous ML-based approaches for predicting RPIs, there is substantial scope for enhancing their robustness and generalization capabilities. Furthermore, working with biological sequences presents additional challenges, including dealing with categorical and non-structured data, which complicates the analysis [Bonidia et al. 2022]. Specifically, in the context of RNA-protein interactions, utilizing ML approaches requires the extraction of relevant features from both molecules to develop an effective predictive model. This development step is called feature engineering, typically performed by experts, and the most time-consuming step in ML [Waring et al. 2020]. Although libraries and environments are widely available, users often face challenges when beginning their research or developing projects with ML due to a lack of expertise in the field [Dwivedi et al. 2023].

This challenge has led to an increasing debate in the literature on the democratization of Artificial Intelligence (AI), focusing on various facets, one of which is the democratization of the development of ML models [Seger et al. 2023]. In this context, developing open-source tools and integrating automated pipelines are crucial steps toward this democratization [Seger et al. 2023, Thirunavukarasu et al. 2023, Vanschoren 2023], aiming to enable non-experts to leverage the benefits of AI. Considering this, we introduce BioPrediction, an end-to-end framework built to conduct feature engineering and ML model training autonomously, based on user input, to accurately predict RPIs. This framework works without the need for direct user intervention.

BioPrediction encompasses the entire ML model development process, including feature extraction and selection, optimal model identification, hyperparameter tuning, and interpretability reports. This framework provides that researchers or other interested parties who are not AI specialists can develop a model suitable for their data and predict interactions between biological molecules. Our research is guided by the following Research Question (RQ):

RQ: Is it possible to develop an autonomous, comprehensive ML framework that operates independently of expert input, aiming to generate classification and detection models for interactions between sequence pairs, like ncRNA-protein, that perform on par with those designed by specialists?

To the best of our knowledge, this is the first study to propose an end-to-end framework to classify interactions between biological sequences, competitive with models developed by experts. Finally, BioPrediction could be a step in the democratization of ML for studying interactions between molecular sequences, facilitating progress in metabolism research, and offering insights into disease-associated pathways. Our framework is available on GitHub[1].

2. Workflow: BioPrediction

BioPrediction has an automated workflow for building an end-to-end ML pipeline to predict interactions, along with a report designed to explore some characteristics of the model. To initiate the ML model construction, it is essential to input the path to the data, which consists of three main files: the list of known interactions and dictionaries containing the sequences for all proteins and RNAs. Afterward, the feature extraction module starts, where features are obtained to characterize each biological sequence. More specifically, there are two main types of features: structural and topological.

Structural features refer to those extracted directly from the primary sequence of each molecule, including examples such as amino acid frequencies, Shannon entropy, and physicochemical properties of each amino acid, such as hydrophobicity (H1). On the other hand, topological features are derived from the interaction network present exclusively in the training set. These features include the number of interactions and other graph measures, such as centrality and betweenness. Thus, each RNA and protein has a set of numerical columns that characterize their various properties.

Next, datasets are constructed for the modeling stage by concatenating the features of proteins and RNAs with the interaction table, creating a table where each row contains the features of the sequences and the label associated with that pair. In total, 5 subsets of features were created, four exclusively with structural features and one exclusive for topological features. The subsequent step involves training partial models for each feature set, aiming to reduce the dimensionality of the problem and, consequently, improve efficiency in the final execution. After constructing these partial models, the probability of belonging to the interaction class is used as the new compressed feature.

This procedure is repeated for all feature sets, resulting in the creation of a final dataset with all compressed features, which will be trained again to combine partial decisions into a final one. Finally, using the training sets, the model is constructed to make the definitive decision on which class each interaction pair will be classified into. Both the partial model and the final model are based on decision trees, such as Random Forest, Catboost, and XGBoost. Once the model is ready, an interpretability report based on the SHAP Values library is generated to elucidate the decision-making process, and a usability report is created to clarify the metrics and properties of the model to the user.

[1] https://github.com/0nurB/BioPredictionRPI-1.0

2.1. Validation

To assess our framework's efficacy, we benchmarked it against other tools designed for RPI prediction across five distinct datasets (RPI369, RPI488, RPI1087, RPI2241, and NPInter), all referenced in the RPITER article [Peng et al. 2019]. Our goal was to compare the performance of BioPrediction with that of the RPITER model and additional tools cited in the original publication. The dimensions of each dataset are detailed in Table 1. In all experiments, the mean and standard deviation will result from 20 executions of BioPrediction. It's important to clarify that our goal is not to surpass all studies in the literature. Instead, we aim to create a framework that does not require specialized knowledge for execution, offering performance similar to models developed by experts.

Table 1. Summary of datasets in the experiment.

Dataset	Interaction pairs	Non-interaction pairs	RNAs	Proteins
RPI369	369	369	332	338
RPI488	243	245	25	247
RPI1807	1807	1436	1078	3131
RPI2241	2241	2241	841	2042
NPInter	10412	10412	4636	449

3. Summary of Results, Discussions and Main Contributions

In this section, we provide a synthesis of experimental results, highlighting how BioPrediction compares with studies from the existing literature. This assessment includes RPI369, RPI488, RPI1807, RPI2241, and NPInter, as detailed in Table 2. Overall, BioPrediction demonstrated competitive performance across these datasets when measured against the studies cited, as verified by the Mann-Whitney U one-sided test with a significance level (alpha) of 0.05.

In essence, although certain metrics may show higher values in isolation, there is no statistically significant difference in the overall effectiveness of the models. Furthermore, in our comprehensive analysis across four studies and five datasets, we meticulously examined a total of 120 metrics. Impressively, only 29% of these metrics showed performance exceeding that of BioPrediction by more than 1%. This finding serves as preliminary evidence of BioPrediction's capability to match, and in some cases, rival the performance of models developed by domain experts.

Our comprehensive validation across various datasets underscores BioPrediction's robustness, establishing it as a flexible framework that enables non-experts to construct robust predictive models. This level of accessibility diminishes the need for advanced ML expertise and promotes a cooperative atmosphere. Despite areas for improvement, comprehensive evidence suggests that BioPrediction not only competes but also challenges the dominance of expert-developed models, affirming its position as a viable alternative for biological interaction prediction tasks.

Finally, BioPrediction was selected to participate in Prototypes for Humanity 2023[2], during COP28-Dubai, chosen from among 3000 entries, from more than 100 countries, standing out among the 100 best in the world. This study is also part of a comprehensive set of solutions that prioritize positive outcomes with a significant impact on

[2]https://www.prototypesforhumanity.com/project/bioprediction-framework/

Table 2. Performance measuring accuracy (ACC), precision (Pre), recall (Rec), specificity (Spec), Matthews correlation coefficient (MCC), and Area Under the Curve (AUC).

Dataset	Study	ACC	Pre	Rec	Spec	MCC	AUC
RPI369	RPITER	72.8	70.1	79.7	65.9	46.1	82.1
	IPMiner	70.0	**84.0**	78.4	56.0	42.8	70.0
	RPISeq-RF	71.3	72.4	71.6	70.2	42.6	71.3
	lncPro	50.2	51.2	23.7	**77.1**	00.9	46.8
	BioPrediction	**79.1 ± 2.0**	75.8 ± 3.0	**88.7 ± 3.1**	69.2 ± 6.0	**60.4 ± 4.1**	**89.5 ± 1.9**
RPI488	RPITER	**89.3**	94.3	83.9	**94.7**	**79.3**	91.1
	IPMiner	**89.3**	**95.1**	**94.6**	83.5	**79.3**	89.3
	RPISeq-RF	88.3	93.5	92.8	83.1	77.1	88.3
	lncPro	85.6	94.0	77.0	**94.7**	72.5	**92.9**
	BioPrediction	88.7 ± 1.4	92.2 ± 2.3	84.8 ± 0.8	92.5 ± 2.7	78.0 ± 2.6	90.1 ± 1.5
RPI1807	RPITER	96.8	95.9	**98.6**	94.6	93.6	**99.0**
	IPMiner	96.8	95.5	96.5	**97.8**	93.5	96.6
	RPISeq-RF	**97.0**	96.2	97.0	97.6	**93.9**	96.9
	lncPro	47.2	53.2	44.5	50.6	-4.9	50.6
	BioPrediction	95.3 ± 0.2	**96.3 ± 0.5**	95.3 ± 0.4	95.3 ± 0.7	90.5 ± 0.4	98.3 ± 0.3
RPI2241	RPITER	**89.0**	87.1	**91.7**	86.3	**78.1**	**95.7**
	IPMiner	86.1	**88.2**	87.7	84.1	72.4	86.1
	RPISeq-RF	85.1	86.3	86.1	83.8	70.2	85.1
	lncPro	60.6	63.2	51.8	69.5	21.6	64.4
	BioPrediction	84.8 ± 0.3	86.3 ± 1.0	82.9 ± 0.8	**86.7 ± 1.3**	69.8 ± 0.7	92.4 ± 0.2
NPInter	RPITER	95.5	93.9	**97.3**	93.7	91.0	98.5
	IPMiner	**95.7**	**95.6**	95.6	**95.8**	**91.4**	95.7
	RPISeq-RF	94.3	93.6	93.7	94.9	88.5	94.3
	lncPro	50.8	50.5	73.9	27.6	1.7	51.7
	BioPrediction	95.3 ± 0.1	94.8 ± 0.1	95.8 ± 0.1	94.7 ± 0.1	90.5 ± 0.1	**98.5 ± 0.1**

society, linked to AutoAI-Pandemics[3], which was selected as one of the most promising proposals (out of 221 entries) in a global competition, held by the Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network (AI4PEP)[4].

Acknowledgments

This research is funded by Canada's International Development Research Centre (IDRC) (Grant No. 109981).

References

Armaos, A., Zacco, E., Sanchez de Groot, N., and Tartaglia, G. G. (2021). Rna-protein interactions: Central players in coordination of regulatory networks. *BioEssays*, 43(2):2000118.

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. (2022). BioAutoML: automated feature engineering and metalearning to predict noncoding RNAs in bacteria. *Briefings in Bioinformatics*, 23(4).

Cantile, M., Di Bonito, M., Tracey De Bellis, M., and Botti, G. (2021). Functional interaction among lncrna hotair and micrornas in cancer and other human diseases. *Cancers*, 13(3).

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(35).

[3]http://autoaipandemics.icmc.usp.br/

[4]https://ai4pep.org/brazil/

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., et al. (2023). "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

Ferre, F., Colantoni, A., and Helmer-Citterich, M. (2016). Revealing protein–lncrna interaction. *Briefings in bioinformatics*, 17(1):106–116.

Jiang, P., Sinha, S., Aldape, K., et al. (2022). Big data in basic and translational cancer research. *Nature Reviews Cancer*, 22:625–639.

Kopp, F. and Mendell, J. T. (2018). Functional classification and experimental dissection of long noncoding rnas. *Cell*, 172(3):393–407.

P, B. and M., G. (2021). Worldwide protein data bank (wwpdb): A virtual treasure for research in biotechnology. *Eur J Microbiol Immunol (Bp)*, 11(4):77–86.

Peng, C., Han, S., Zhang, H., and Li, Y. (2019). Rpiter: A hierarchical deep learning framework for ncrnaprotein interaction prediction. *Int J Mol Sci*, 20(5):1070.

Qin, W., Cho, K. F., Cavanagh, P. E., and Ting, A. Y. (2021). Deciphering molecular interactions by proximity labeling. *Nature methods*, 18(2):133–143.

Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., and Dafoe, A. (2023). Democratising ai: Multiple meanings, goals, and methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 715–722, New York, NY, USA. Association for Computing Machinery.

Shaath, H., Vishnubalaji, R., Elango, R., Kardousha, A., Islam, Z., Qureshi, R., Alam, T., Kolatkar, P. R., and Alajez, N. M. (2022). Long non-coding rna and rna-binding protein interactions in cancer: Experimental and machine learning approaches. In *Seminars in Cancer Biology*, volume 86, pages 325–345. Elsevier.

Thirunavukarasu, A., Elangovan, K., Gutierrez, L., Li, Y., Tan, I., Keane, P., Korot, E., and Ting, D. (2023). Democratizing artificial intelligence imaging analysis with automated machine learning: Tutorial. *J Med Internet Res*, 25:e49949.

Vanschoren, J. (2023). Democratising artificial intelligence to accelerate scientific discovery. In *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*. OECD Publishing, Paris.

Wang, Y., Wang, Y., Luo, W., Song, X., Huang, L., Xiao, J., Jin, F., Ren, Z., and Wang, Y. (2020). Roles of long non-coding rnas and emerging rna-binding proteins in innate antiviral responses. *Theranostics*, 10(20):9407.

Waring, J., Lindvall, C., and Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, 104:101822.

Zhang, W., Wang, J., Li, B., Sun, B., Yu, S., Wang, X., and Zan, L. (2023). Long noncoding rna bnip3 inhibited the proliferation of bovine intramuscular preadipocytes via cell cycle. *International Journal of Molecular Sciences*, 24(4).

Zhong, Y., Xu, F., Wu, J., Schubert, J., and Li, M. M. (2021). Application of next generation sequencing in laboratory medicine. *Ann Lab Med*, 41(1):25–43.