BioPrediction: Democratizing Machine Learning in the Study of Molecular Interactions

Bruno Rafael Florentino[1], **Natan Henrique Sanches**[2], **Robson Parmezan Bonidia**[2*], **André C. P. L. F. de Carvalho**[2]

[1] Instituto de Física de São Carlos
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

[2] Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

brunorf1204@usp.br, natan.sanches@usp.br,

bonidia@usp.br, andre@icmc.usp.br

Abstract. *Given the increasing number of biological sequences stored in databases, there is a large source of information that can benefit several sectors such as agriculture and health. Machine Learning (ML) algorithms can extract useful and new information from these data, increasing social and economic benefits, in addition to productivity. However, the categorical and unstructured nature of biological sequences makes this process difficult, requiring ML expertise. In this paper, we propose and experimentally evaluate an end-to-end automated ML-based framework, named BioPrediction, able to identify implicit interactions between sequences, e.g., long non-coding RNA and protein pairs, without the need for end-to-end ML expertise. Our experimental results show that the proposed framework can induce ML models with high predictive accuracy, between 77% and 91%, which are competitive with state-of-the-art tools.*

Resumo. *Com o crescente número de sequências biológicas armazenadas em bancos de dados, existe uma grande fonte de informações que pode beneficiar diversos setores, como agricultura e saúde. Algoritmos de Aprendizado de Máquina (AM) podem extrair informações úteis e novas a partir delas, resultando em benefícios e produtividade. No entanto, a natureza categórica e não-estruturada dificulta esse processo, requerendo conhecimento especializado. Neste trabalho, é proposto um framework fim-a-fim baseado em AM automatizado, chamado BioPrediction, capaz de identificar interações implícitas entre sequências, por exemplo, pares de RNA longo não-codificante e proteínas, sem a necessidade de conhecimento especializado em AM de ponta a ponta. Como resultado, obteve-se um modelo robusto com acurácia balanceada entre 77% e 91% nos conjuntos de dados utilizados para validação, além de apresentar resultados competitivos com as ferramentas estado-da-arte.*

1. Informações Gerais

Com o advento das técnicas modernas de sequenciamento genético, houve um aumento considerável do número de sequências biológicas disponíveis em bancos de dados [Jiang et al. 2022, Hashemi et al. 2018]. Como resultado, todas as informações

*Autor Correspondente

extraídas das mais variadas espécies catalogadas estão concentradas nessas bases [Mingyue et al. 2019, P and M. 2021]. Levando isso em consideração, o desenvolvimento e a utilização de técnicas computacionais de alto desempenho para lidar com grandes volumes de dados é o caminho mais comum para a extração de informações [Chicco 2017].

Dentre os diversos subproblemas existentes que incluem sequências biológicas, podemos destacar aqueles que envolvem a interação de RNA longo não-codificante (*lncRNA*, do inglês "*long non-coding RNA*") e proteínas, os denominados LPIs (*do inglês, lncRNA–Protein Interaction*). É importante destacar que *lncRNA* é classe do material genético transcrito de tamanho maior que 200 nucleotídeos [Xu et al. 2022] e que não codifica proteínas [Statello et al. 2021]. Diferentemente do que era pensado até então, um *lncRNA* não pode ser categorizado como parte do DNA não fundamental aos seres vivos [Zhang et al. 2023], uma vez que é essencialmente complexos e desempenha diversas funções nos diferentes organismos [Kopp and Mendell 2018].

No que tange essas estruturas biológicas, os *lncRNA*s são responsáveis por regular estados da cromatina e a expressão de genes, tanto de regiões próximas a seu sítio de transcrição, quanto de regiões mais distantes. Além disso, também são responsáveis por interagir e regular o comportamento de proteínas e outros RNAs [Kopp and Mendell 2018]. Além disso, o nível de expressão de alguns *lncRNAs* está relacionado a vias da regulação inicial de alguns cânceres sólidos, atuando como biomarcadores [Cantile et al. 2021]. Observa-se, então, que o entendimento dessa classe de moléculas tem importância prognóstica.

Há, no entanto, uma série de dificuldades ao trabalhar com essas sequências utilizando técnicas computacionais. Uma das maiores complicações está na natureza dos dados: categóricos e não estruturados [Bonidia et al. 2022]. Considerando esse contexto, a fim de trabalhar com problemas de interação *lncRNA*-proteína utilizando aprendizado de máquina (AM), é necessário um tratamento prévio desses dados buscando extrair informações relevantes que possam ser utilizadas pelo modelo. Conhecido como engenharia de características, trata-se de um processo que normalmente é executado por especialistas e caracteriza-se como uma das etapas mais demoradas de AM [Waring et al. 2020].

Além disso, a engenharia de características envolve a seleção e extração cuidadosa das informações (ou características) mais relevantes presentes nos dados brutos. É nessa etapa que ocorre, por exemplo, a estruturação dos dados para a análise do problema [Bonidia et al. 2020][Muhammod et al. 2019], uma vez que os dados brutos precisam ser processados e transformados em características significativas. Essas características, por sua vez, são utilizadas pelos algoritmos de AM para realização de inferências (seja de classificação ou regressão) [Kreuzberger et al. 2023].

Alguns desafios adicionais em tarefas de AM são a seleção das características mais relevantes para inferência, e a identificação do modelo mais robusto para cada conjunto de dados [Bonidia et al. 2022]. Ambos desafios podem ser solucionados simultaneamente através da estratégia de Otimização Bayesiana (*Bayesian Optimization*) [Frazier 2018], uma abordagem muito utilizada na otimização de problemas que envolvem funções de alto custo computacional [Binois and Wycoff 2022]. Essa técnica pode ser observada na

prática em [Bonidia et al. 2022], sendo utilizada para encontrar a melhor combinação dos dois principais fatores citados: características mais relevantes e modelo mais robusto. Essa estratégia também permite lidar com dados de alta dimensionalidade, levando em consideração, no processo de seleção, a performance do modelo e a quantidade de tempo requerida para processamento dos dados.

A desproporção entre classes é outra questão importante que pode surgir para vários conjuntos de dados, potencialmente afetando o desempenho dos modelos de AM [Patel et al. 2020]. Geralmente, as técnicas utilizadas para lidar com esse tipo de problema envolvem o processo de sobreamostragem (do inglês *oversampling*) da classe minoritária, ou o de subamostragem (*undersampling*) da classe majoritária [Patel et al. 2020]. Normalmente, as duas técnicas ocorrem de forma aleatória [Wang et al. 2021], buscando dispensar vieses de seleção.

Por fim, um grande desafio na indução de modelos de AM a partir de dados biológicos, muitas vezes, é falta de interpretabilidade desses modelos por parte do profissional de Biologia que utiliza-o como ferramenta. A visão do modelo como uma ”caixa preta”é preocupante do ponto de vista médico/biológico, uma vez que esses resultados influenciarão na tomada de decisões que poderão afetar, por exemplo, outras pessoas ou entidades [Rudin 2019]. Na literatura, impulsionou-se o estudo de técnicas que ampliem a interpretabilidade de resultados gerados por algoritmos de AM, permitindo a existência de metodologias como o SHAP [Lundberg 2017] e o LIME [Ribeiro et al. 2016]. A aplicação dessas metodologias no modelo treinado permite que esses resultados sejam analisados com mais detalhes por parte do profissional que o utiliza, entendendo quais características foram as mais influentes no processo de classificação.

Considerando esses desafios, foi proposta uma extensão da ferramenta BioAutoML [Bonidia et al. 2022] capaz de extrair e selecionar o melhor conjunto de características, o melhor modelo de AM para os dados de entrada e ajustar esse modelo para tarefas de classificação de sequências biológicas, ou seja, um *pipeline* fim-a-fim, que abrange todas as etapas necessárias para executar esse tipo de tarefa. Essa extensão *open-source*, denominada *BioPrediction*, tem como objetivo ampliar a capacidade da ferramenta base ao lidar com interações entre sequências biológicas, sendo que nesse trabalho a validação foi limitada a interações entre *lncRNA*s e proteínas. Através da ampliação da etapa de engenharia de características da ferramenta base, *BioPrediction* é capaz de realizar a tarefa de classificação, distinguindo entre *lncRNA*-proteína interagentes e não interagentes. Esse artigo é guiado pela seguinte Questão de Pesquisa (QP):

QP: É possível criar um *framework* fim-a-fim, que trabalhe sem a intervenção de especialistas, para gerar um modelo de classificação e detecção de interações implícitas entre pares de sequências, e.g., *lncRNA*-proteína, com desempenho competitivo em relação as ferramentas manuais construídas por especialistas?

BioPrediction pode desempenhar um papel importante na democratização de AM para não especialistas, auxiliando no avanço de estudos relacionados ao metabolismo e compreensão mais profunda das vias envolvidas em doenças. Por fim, a ferramenta está disponível no GitHub por meio do link `https://github.com/Bonidia/BioPrediction`.

2. Trabalhos Relacionados

Nos últimos anos, a interação entre entidades biológicas tornou-se um assunto que recebeu muita atenção nos estudos de biólogos e acadêmicos de áreas correlatas. Vem em pauta, por exemplo, estudos relacionados à interação droga-droga [Li et al. 2020], gene-gene [Cole et al. 2017] e, finalmente, *lncRNA*-proteína [Ferrè et al. 2016]. Considerando o alto custo de tempo e recursos necessários em análises experimentais para esse contexto, impulsionou-se a aplicação de algoritmos de AM de alto desempenho para realização de predição de interações, com base em interações já descobertas [Yu et al. 2022].

No contexto de interações *lncRNA-proteína*, alguns trabalhos envolvem a utilização de modelos de aprendizado profundo. LPI-deepGBDT [Zhou et al. 2021], por exemplo, utiliza de árvores de decisão com *gradient boosting*, um algoritmo de AM que combina múltiplos modelos fracos para a criação de um modelo forte, robusto. Além disso, complementa o processo de predição com uma arquitetura profunda de mapeamento para identificar interações implícitas. Na mesma linha, EnANNDeep [Peng et al. 2022] propôs uma abordagem baseada em redes neurais e árvores de decisão profundas, complementando com aplicação de um classificador adaptativo baseado em K-vizinhos mais próximos.

Ainda abordando algoritmos de aprendizado profundo, métodos com uma abordagem direcionada ao processamento de redes complexas também surgiram, utilizando redes neurais em grafos para atingir seu objetivo. BiHo-GNN [Ma et al. 2023] é uma arquitetura profunda que integra propriedades de redes homogêneas e heterogêneas, com utilização da técnica de incorporação (*embedding*) para grafos bipartidos. Além dessa, ncRPI-LGAT [Han and Zhang 2023] utiliza de redes neurais para grafos com mecanismos de atenção para processamento de redes de interações *lncRNA*-proteína.

Por fim, existem trabalhos que buscam aplicar algoritmos clássicos de AM nesse contexto. SFPEL-LPI [Zhang et al. 2018] utiliza de algoritmos de agrupamento de dados e medidas de similaridade para atingir seus propósitos. O método extrai características de ambas moléculas isoladamente e, em seguida, identifica pares *lncRNA-lncRNA* e proteína-proteína similares com base nas características extraídas e em interações *lncRNA*-proteína conhecidas. Por fim, o resultado é processo com um *framework* baseado em aprendizado por agrupamento.

Mesmo com diversos estudos presentes na literatura, observa-se que todos os métodos propostos requerem a atuação de especialistas para seu uso. Na revisão, para o melhor do nosso conhecimento, não foi identificado nenhum método capaz de realizar o processamento fim-a-fim dos dados biológicos. Além disso, também não foram encontrados trabalhos que incluíssem um módulo dedicado à interpretabilidade dos resultados.

3. Metodologia

3.1. Conjuntos de Dados e Configuração Experimental

O conjunto de dados utilizado para as interações RNA-proteína (*LPI*s) foi obtido do estudo publicado em [Zhou et al. 2021]. Esse conjunto de dados foi dividido em cinco subproblemas, consistindo em três conjuntos de *LPI*s humanas e dois conjuntos de interações de plantas (Tabela 1). Cada subconjunto contém os dados em formato *FASTA*, separados

Tabela 1. Descrição dos *datasets* utilizados para a validação do *BioPrediction*

Dataset	Número de *lncRNA*s	Número de proteínas	Número de pares	Número de interações
1	935	59	55165	3479
2	885	84	74340	3265
3	990	27	26730	4158
4	109	35	3815	948
5	1704	42	71568	22133

em sequências primárias de RNAs e proteínas. Além disso, uma tabela complementar é disponibilizada, indicando quais moléculas interagem entre si.

Durante a análise dos subconjuntos de dados, observou-se que todos eles apresentam uma distribuição desbalanceada das classes, com percentual de interações positivas variando de 4% a 31%. Além disso, constatou-se que as sequências que representam os RNAs estão codificadas de duas maneiras: utilizando os nucleotídeos de DNA (A, T, C, G) ou diretamente na codificação de RNA (A, U, C, G). Isso indica a presença de diferentes formas de representação das sequências de RNA nos conjuntos de dados, um problema que precisou ser tratado antes da alimentação do modelo.

Em relação à divisão dos dados, cada subconjunto foi separado em 70% para treinamento e 30% para teste. Essa divisão é comumente utilizada para garantir uma avaliação adequada do desempenho do modelo treinado, permitindo que ele seja treinado em uma parte dos dados e validado em uma parte independente, ainda não vista. Além disso, cada conjunto de dados foi testado 20 vezes e, para fins de comparação, foi realizada uma análise estatística desses experimentos, como feito em [Zhou et al. 2021]. Dentre as medidas de avaliação, temos tanto medidas que não consideram dados desbalanceados, como acurácia, quanto medidas que os consideram, como a acurácia balanceada. Essa avaliação possibilitou obter uma estimativa realista do desempenho do modelo em dados não vistos anteriormente e, além disso, permite comparação direta com o estudo de [Zhou et al. 2021].

3.2. *BioPrediction* - Aprendizado Automatizado

Diante da entrada fornecida, *BioPrediction* realiza a execução automatizada de diferentes etapas no processo de aprendizado para a construção de um modelo robusto: engenharia de características, concatenação dos resultados, seleção de características, seleção e treinamento do modelo e, por fim, otimização dos hiperparâmetros. A construção do *framework* foi pensada visando minimizar a necessidade de interação humana e de conhecimento técnico do usuário. Entretanto, ele não utiliza abordagens como *deep learning* e *graph neural networks* (*GNNs*), que são técnicas atualmente utilizadas em diversos problemas de AM.

O módulo inicial de extração de características recebe como entrada um conjunto de sequências biológicas e realiza a extração dos padrões presentes em cada uma. Para cada tipo de sequência, é realizada uma verificação a fim de identificar eventuais erros no conjunto de bases, removendo possíveis sequências incorretas. Posteriormente, são extraídas as características – medidas descritivas de cada sequência utilizadas na posterior alimentação do modelo – para cada uma das sequências. Mais especificamente, essas características são previamente definidas e apresentadas na Tabela 2, retiradas de

[Bonidia et al. 2021], [Chen et al. 2018] e [Liu et al. 2014]. Para facilitar a usabilidade, todas as características extraídas são armazenadas em disco para permitir que o usuário tenha acesso a elas para aplicações futuras.

Tabela 2. Características Extraídas pelo *BioPrediction*

Descritor	Identificador	Aplicação
k-mer complementar reverso	rev_kmer	DNA
Composição de pseudodinucleotídeos	Pse_dnc	DNA
Correlação em série PseDNC	SCPseDNC	DNA
Correlação em série PseTNC	SCPseTNC	DNA
Autocovariância baseada em dinucleotídeos	DAC	DNA
Composição de ácido nucleico	NAC	DNA/RNA
Composição de dinucleotídeos	DNC	DNA/RNA
Composição de trinucleotídeos	TNC	DNA/RNA
Quadro de leitura aberta	ORF	DNA/RNA
Pontuação de Fickett	Fic	DNA/RNA
Entropia de Shannon	SE	DNA/RNA
Fourier Binário	BF	DNA/RNA
Fourier com números complexos	CF	DNA/RNA
EIIP + Fourier	F_EIIP	DNA/RNA
Composição de aminoácidos	AAC	Proteínas
Composição de dipeptídeos	DPC	Proteínas
Composição de tripeptídeos	TPC	Proteínas
Pares de aminoácidos com espaçamento K	Kgap	DNA/RNA/Proteínas
Entropia de Tsallis	TE	DNA/RNA/Proteínas

Após a extração das características, o próximo passo consiste na concatenação dessas para cada par RNA-proteína, ao mesmo tempo em que é atribuída sua respectiva classe. Dessa forma, é formado um único vetor representativo para cada par de moléculas, contendo as características concatenadas de cada molécula individual juntamente com a classe atribuída.

Posteriormente, inicia-se o processo de seleção de características e seleção de modelo, utilizando a técnica de Otimização Bayesiana [Frazier 2018] para identificar o conjunto de características descritivas mais adequado e o modelo mais robusto para os dados em questão. Essa seleção segue o mesmo procedimento descrito na ferramento modelo BioAutoML [Bonidia et al. 2022], uma vez que este é um problema de alta dimensionalidade e a escolha desse tipo de seleção é mantida devido a sua relevância.

Os modelos disponíveis no *BioPrediction* são todos baseados em árvores de decisão, selecionados devido à sua alta interpretabilidade, facilidade no treinamento e simplicidade estrutural. Dentre eles, podemos citar: *Random Forest* [Liaw and Wiener 2002], *CatBoost* [Prokhorenkova et al. 2018] e *LightGBM* [Ke et al. 2017].

O processo de seleção de modelo considera a opção de ativar a funcionalidade que lida com dados desbalanceados a partir da análise do conjunto de dados original, a fim de escolher a melhor técnica para abordar o problema em questão. Nesse caso, duas técnicas podem ser utilizadas: subamostragem aleatória [Hasib et al. 2020] ou *SMOTE*

[Bowyer et al. 2011]. Adicionalmente, caso essa opção esteja ativada, também acontece a estimação dos melhores hiperparâmetros para o classificador escolhido, visando otimizar o desempenho final da classificação. Esses parâmetros são obtidos a partir da documentação oficial de cada modelo, garantindo uma escolha fundamentada e embasada nos melhores valores disponíveis.

Após a seleção da melhor configuração do modelo, o próximo passo consiste em treiná-lo utilizando validação cruzada com 10 subpartes, de forma que o modelo consiga ser generalizável para outros conjuntos de dados. Esse processo gera, também, uma tabela contendo as medidas de desempenho do modelo durante o treinamento. Por fim, é realizada a avaliação final utilizando os dados de teste separados no início, com o modelo sendo utilizado para realização de classificações com base nesses dados. As medidas de desempenho também são calculadas para o conjunto de teste, com o intuito de avaliar a eficácia e a qualidade do modelo nesse conjunto de dados independente.

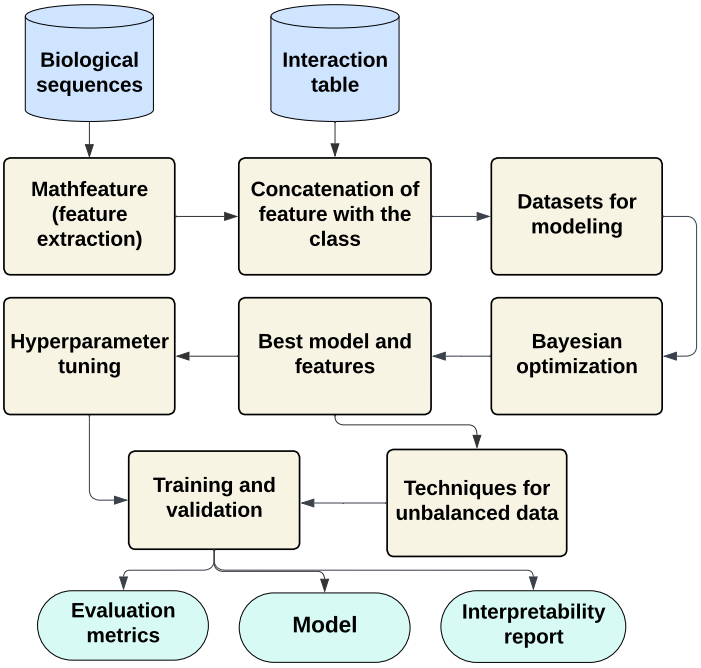

Figura 1. Fluxo de trabalho do *BioPrediction*: em azul, as entradas do modelo; em amarelo, os processos internos e, em verde, as saídas do *framework*.

No diagrama (Figura 1), é possível visualizar o fluxo de trabalho do *framework* desenvolvido nesta pesquisa. Em azul, estão representadas as entradas do sistema, que incluem as sequências biológicas dos *lncRNAs* e das proteínas, juntamente com a tabela de interações que consta quais sequências interagem entre si. Em seguida, em amarelo, são apresentados os processos internos da ferramenta. Essas etapas incluem a engenharia de características, a seleção do modelo, a seleção de características e a escolha da técnica mais apropriada para lidar com o desbalanço dos dados. Por fim, em verde, são mostradas as saídas geradas pelo *BioPrediction*. Essas saídas são compostas pelo modelo final desenvolvido, pelas medidas de validação utilizadas para avaliar o desempenho do sistema e pelo relatório de interpretabilidade, fornecendo novas percepções e explicações sobre as decisões tomadas pelo modelo na hora de realizar a classificação.

3.3. Interpretabilidade

BioPrediction também se encarrega da geração de um relatório explicativo ao usuário final, onde expõe brevemente quais foram as maiores considerações do modelo recém treinado no processo de classificação. Para isso, propõe uma análise de quais foram as características mais influentes em uma amostra de entradas de determinada classe. Cada classe é analisada individualmente.

Além disso, o módulo de interpretabilidade também propõe analisar como a magnitude dos valores de cada característica influencia na decisão do modelo (*i.e.* a baixa magnitude de uma determinada característica A está relacionada à classificação de pares com a classe C_1, enquanto a alta magnitude de A está relacionada à classificação com a classe C_2). Esses resultados são sumarizados com gráficos expositivos no relatório exportado junto ao modelo.

Para a realização dessa análise, utilizou-se o método SHAP [Lundberg 2017] como metodologia de interpretação, responsável por unificar muitos outros métodos já existentes na literatura (como o LIME [Ribeiro et al. 2016]). Esse método, ainda, possui um módulo exclusivo para algoritmos baseados em árvore [Lundberg et al. 2020], produzindo resultados consistentes em conjunto com os modelos treinados pelo *BioPrediction*.

Por SHAP, entende-se "Explicações Aditividas de Shapley"(do inglês, *Shapley Additive Explanations*), uma metodologia de interpretação de modelos de AM baseada na Teoria dos Jogos. A métrica de contribuição para classificação, atribuída a cada característica analisada, é definida pelo modelo de Shapley. Esse modelo permite a extração de valores Shapley, coeficientes numéricos utilizados para determinar a contribuição individual de um jogador quando dois ou mais outros jogadores colaboram entre si.

Dentro do contexto de AM, interpretam-se as características utilizadas para uma predição como os jogadores do modelo de Shapley. Os valores Shapley calculados descrevem, portanto, qual a contribuição individual de cada característica, considerando diferentes coalizões, para que o modelo tomasse aquela decisão de classificação.

4. Resultados

Nessa etapa, foram realizadas explorações experimentais dos dados referentes aos *LPIs*, seguindo a metodologia anteriormente estabelecida. Dentre as várias medidas de avaliação possíveis de serem utilizadas, daremos destaque à acurácia balanceada, uma vez que os dados apresentam desbalanceamento. Outras medidas, como a acurácia, podem levar a análises incorretas, pois não levam em consideração a proporção das classes.

4.1. Performance

Nessa seção, apresentamos uma análise estatística dos experimentos realizados nos diferentes conjuntos de dados. Avaliando a partir da acurácia balanceada, o modelo treinado pelo *framework* obteve resultados consistentes: 0.891 ± 0.003, 0.908 ± 0.002, 0.774 ± 0.009, 0.904 ± 0.007 e 0.856 ± 0.003 para os conjuntos de dados do primeiro ao quinto, respectivamente. A acurácia balanceada superior a 75% (0.75) em todos os conjuntos de dados utilizados nos experimentos indica que o modelo foi capaz de aprender informações relevantes desses dados. Também sugere que esse é capaz de distinguir de maneira satisfatória entre as classes presentes nos conjuntos de dados.

Para continuar nossa análise de resultados, é possível realizar uma comparação da performance dos modelos gerados com os dados disponíveis em [Zhou et al. 2021], cujo autor compilou os resultados de seis ferramentas em seus bancos de dados. Os resultados obtidos pela ferramenta *BioPrediction* serão comparados a fim de avaliar a eficácia da ferramenta desenvolvida. Os detalhes e os resultados dessa comparação podem ser encontrados na tabela complementar[1], além de constar nos gráficos a seguir.

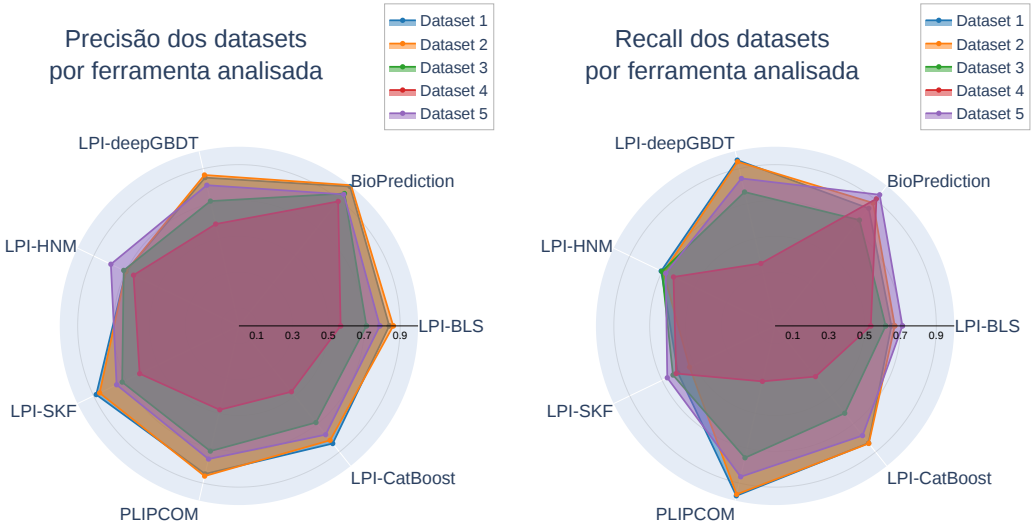

Figura 2. Comparação referente as medidas precisão e *recall*.

Destaca-se, na Figura 2, uma análise comparativa das medidas de avaliação relacionadas à precisão e sensibilidade (*recall*), respectivamente, entre as diferentes ferramentas testadas. Uma vez que alta precisão indica baixa taxa de falsos positivos e alta sensibilidade indica baixa taxa de falsos negativos, ambas medidas são importantes para avaliar o desempenho de um modelo. Nesse contexto, pode-se observar que o modelo treinado pelo *BioPrediction* possui maior precisão quando comparado aos seus semelhantes, o que demonstra melhor performance em minimizar falsos positivos. Por outro lado, sua sensibilidade é mais baixa do que os demais em três dos cinco conjuntos de dados avaliados, evidenciando um pequeno déficit em minimizar falsos negativos.

Na Figura 3, temos as medidas F1-*score* e acurácia. No estudo de AM, F1-*score* se caracteriza como uma função da precisão e da sensibilidade, utilizada para otimizar ambas medidas simultaneamente. Por outro lado, a acurácia indica qual a fração das classificações estão corretas, independente das classes. Nessa etapa, ao analisar a Figura 3 vemos que o modelo treinado pelo *framework* tem o melhor F1-*score* considerando todos os conjuntos de dados, mostrando um ótimo balanço precisão-sensibilidade. Também possui a melhor acurácia na maioria dos conjuntos de dados, indicando que a maior parte das predições realizadas nestes estão corretas.

Por fim, a Figura 4 exibe uma comparação das medidas *AUC* (área sob a curva *ROC*) e *AUPR* (área sobre a curva sensibilidade-precisão). Ambas as curvas mencionadas são construídas ao comparar duas grandezas em diferentes pontos de corte. No caso da curva ROC (*Receiver Operating Characteristic Curve*), trata-se da relação entre verda-

[1] https://github.com/Bonidia/BioPrediction/blob/main/ENIAC_table.pdf

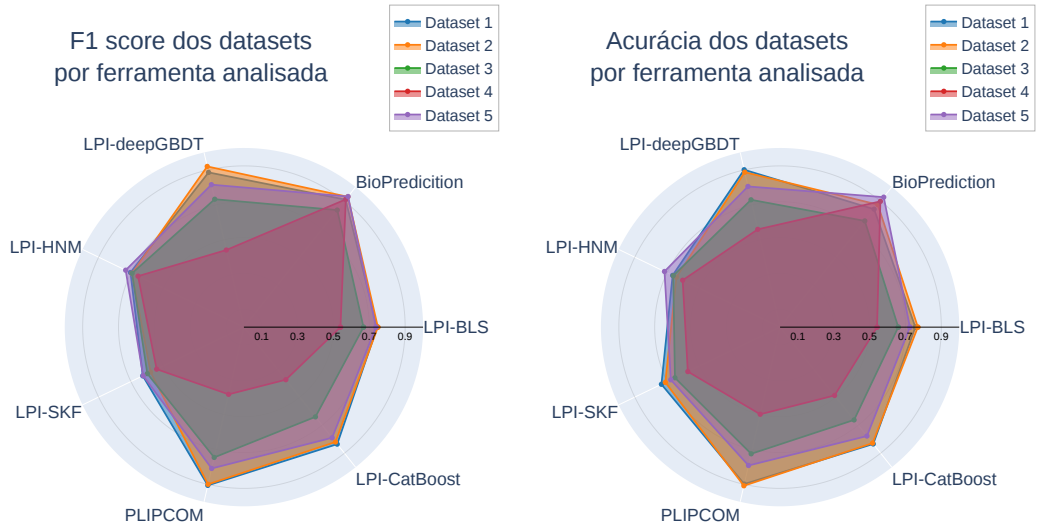

Figura 3. Comparação inter-ferramentas relativa as medidas *F1-score* e acurácia

deiros positivos e falsos positivos, de forma que área sobre essa curva é uma medida para mensurar a capacidade do modelo de distinguir entre as classes. Por outro lado, a curva sensibilidade-precisão é construída pelos vários cortes da sensibilidade e precisão, tendo propriedades similares à anterior, mas sendo indicada em problemas de dados desbalanceados. Como resultado, temos que o modelo desenvolvido pelo *framework* teve os maiores valores de *AUC* em todos os testes. Apesar disso, ficou com *AUPR* mais baixa que seus semelhantes em três dos cinco testes.

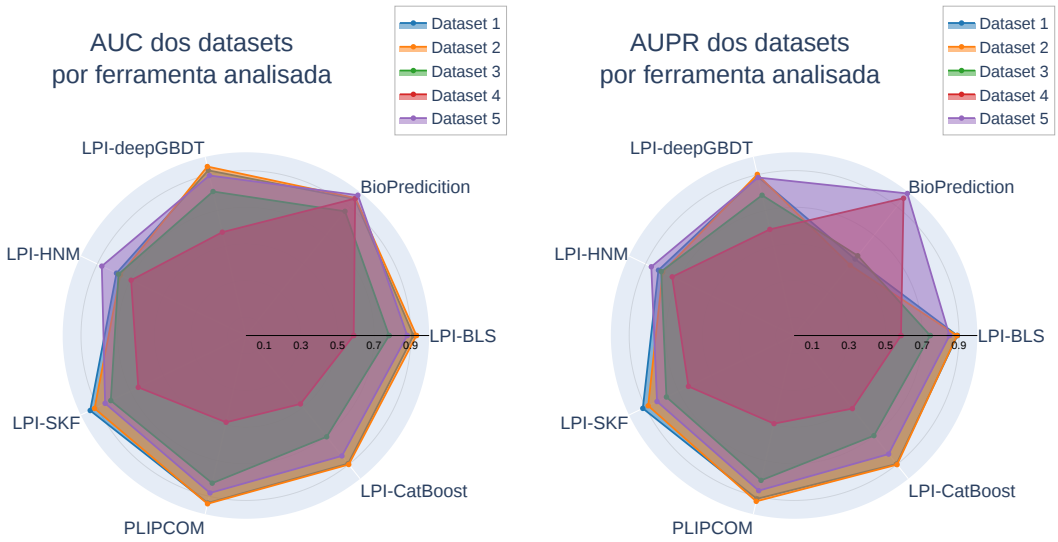

Figura 4. Comparação inter-ferramentas relativa as medidas *AUC* e *AUPR*

Dados os resultados acima, podemos realizar testes de hipóteses para avaliar a significância estatística dos resultados, aplicando, por exemplo, o teste de Friedman para as medidas F1-*score*, *AUC* e *AUPR*. A hipótese nula ($\mathcal{H}_0$) desse teste constata que a média de desempenho das ferramentas é idêntica, enquanto a hipótese alternativa ($\mathcal{H}_a$) considera que pelo menos uma população dentre as envolvidas tem média diferente das

demais. Utilizaremos um nível de significância $\alpha = 0.05$.

Considerando um nível de significância $\alpha = 0.05$, a hipótese nula ($\mathcal{H}_0$) é aceita para F1-*score* e para a *AUC* quando todas as ferramentas, exceto o *BioPrediction*, são consideradas. No entanto, quando o *BioPrediction* é incluído na análise, a hipótese nula ($\mathcal{H}_0$) é rejeitada. Com esse resultado, é possível afirmar com significância estatística que a média de desempenho do *BioPrediction*, considerando tais medidas, difere das demais ferramentas nesse quesito. Por outro lado, no caso da *AUPR* não há evidências para rejeitar a hipótese nula em nenhum dos casos, considerando ou não o *BioPrediction*.

Além disso, *BioPrediction* teve seu melhor desempenho nos conjuntos de dados 4 e 5, que possuem as maiores taxas de exemplos positivos. Nesses conjuntos de dados, cerca de 25% e 31% das amostras de pares são interagentes, enquanto os anteriores possuem 6%, 4% e 15% das amostras com essa condição, respectivamente. Isso fica evidente quando analisado o desempenho do *AUPR*, métrica capaz de avaliar dados desbalanceados, onde vemos que tem desempenho baixo para os três primeiros conjuntos de dados, portanto essa evidência sugere que o módulo de dados desbalanceados do *BioPrediction* pode ter dificuldade em lidar com desproporção entre classes muito acentuada, funcionando melhor para conjuntos de dados balanceados.

4.2. Interpretabilidade

Ao final do processo de treinamento e validação do modelo, é gerado um relatório de interpretabilidade que inclui gráficos para auxiliar o usuário final a interpretar as decisões tomadas pelo modelo treinado. Os gráficos buscam enfatizar, principalmente, quais foram as características que mais contribuíram para determinada classificação, e de que forma a distribuição de valores possíveis para cada característica influencia na classificação de cada classe.

Na Figura 5, observa-se um exemplo do gráfico responsável por descrever como cada característica influencia na classificação de uma determinada classe. Esse gráfico tem o objetivo de revelar padrões e analisar de que forma a magnitude de uma determinada característica está relacionada com a classe. No gráfico, cada ponto amostral é marcado com uma cor dentro de um espectro vermelho-azul. Pontos vermelhos representam alta magnitude da característica em questão, enquanto pontos azuis representam baixa magnitude. A distância no eixo horizontal do ponto em relação ao centro da distribuição (0.0) indica quão intensamente essa característica contribuiu, positivamente (valores SHAP positivos) ou negativamente (valores SHAP negativos), para a classificação final em inferências de determinada classe. Na lateral esquerda, estão as nove características mais influentes, acompanhada da análise particular de cada uma.

Além desse, a Figura 6 exibe um exemplo de outro tipo de gráfico gerado pelo módulo de interpretabilidade. Esse gráfico é responsável por demonstrar como cada característica contribuiu para uma classificação específica. É possível ver, no título do gráfico, a classe inferida e o número identificador que representa a amostra em questão. As características e seus valores são representados na lateral esquerda. No gráfico, as barras direcionadas indicam uma contribuição positiva (a favor, em vermelho) ou negativa (contra, em azul) de cada característica em prol da classe que a amostra foi classificada. Analogamente, as barras de maior comprimento são as que mais influenciaram na decisão do modelo.

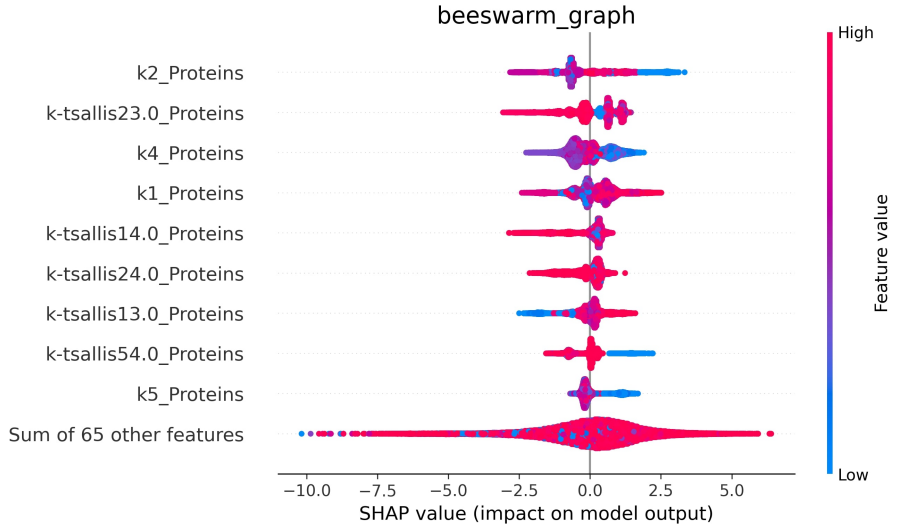

Figura 5. Exemplo de gráfico *beeswarm*, gerado pelo módulo de interpretabilidade do *BioPrediction*

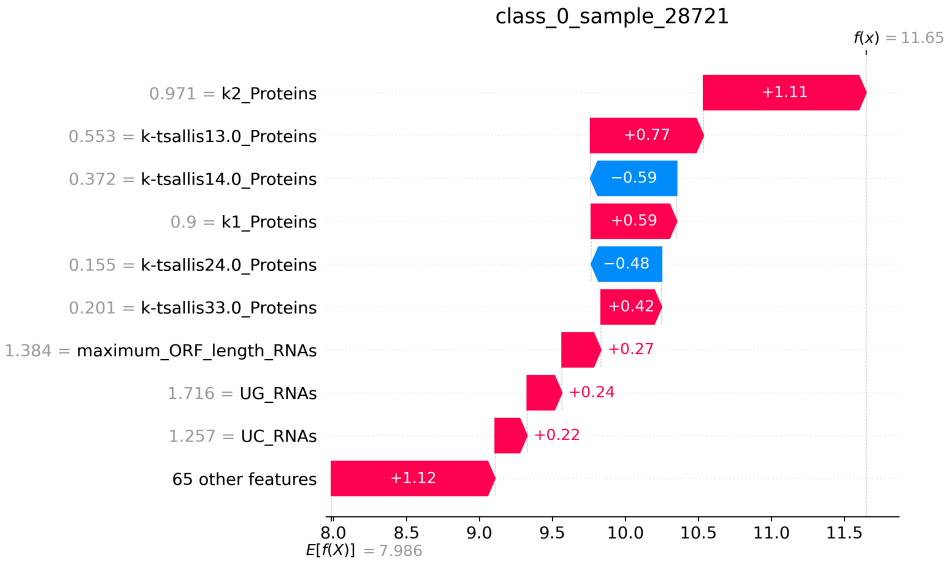

Figura 6. Exemplo de gráfico *cascade*, gerado pelo módulo de interpretabilidade do *BioPrediction*

5. Conclusão

Com base nos resultados obtidos, conclui-se que o *framework* baseado em AM fim-a-fim, *BioPrediction*, apresenta desempenho superior quando comparado a outras ferramentas existentes na literatura para a predição de interação *lncRNA*-proteínas. A análise estatística dos experimentos revelou que o modelo treinado alcançou acurácia balanceada consideravelmente alta dentre os conjuntos de dados, indicando sua capacidade de aprender informações relevantes e distinguir satisfatoriamente entre as classes presentes nos dados.

Os resultados obtidos nas outras medidas de avaliação (dentre elas precisão, F1-*score*, *AUC*, *AUPR* e sensibilidade) demonstraram que o modelo treinado pelo *framework*,

que reduz a necessidade de intervenção especializada, superou a performance das outras ferramentas consideradas, conforme evidenciado nas comparações com os modelos presentes no estado-da-arte. Esses resultados reforçam a eficácia do *framework* desenvolvido, destacando seu potencial para auxiliar na predição dessas interações.

Portanto, a criação de *BioPrediction* representa um avanço significativo no campo da predição de interação, e.g., *lncRNA*-proteínas, ao democratizar o acesso ao tratamento de dados biológicos e treinamento automatizado de modelos de AM, de forma que esses possuam performance competitiva quando comparados aos criados por especialistas. Por fim, sua implementação proporciona resultados consistentes e confiáveis, abrindo novas perspectivas para a investigação e compreensão dessas interações.

Referências

Binois, M. and Wycoff, N. (2022). A survey on high-dimensional gaussian process modeling with application to bayesian optimization. 2(2).

Bonidia, R. P., Domingues, D. S., Sanches, D. S., and de Carvalho, A. C. P. L. F. (2021). Mathfeature: feature extraction package for dna, rna and protein sequences based on mathematical descriptors. *Briefings in Bioinformatics*, page bbab434.

Bonidia, R. P., Sampaio, L. D. H., Domingues, D. S., Paschoal, A. R., Lopes, F. M., de Leon Ferreira de Carvalho, A. C. P., and Sanches, D. S. (2020). Feature extraction approaches for biological sequences: A comparative study of mathematical models. *bioRxiv*.

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).

Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

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).

Chen, Z., Zhao, P., Li, F., et al. (2018). ifeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 34(14):2499–2502.

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

Cole, B. S., Hall, M. A., Urbanowicz, R. J., Gilbert-Diamond, D., and Moore, J. H. (2017). Analysis of gene-gene interactions. *Curr. Protoc. Hum. Genet.*, 95(1):1.14.1–1.14.10.

Ferrè, F., Colantoni, A., and Helmer-Citterich, M. (2016). Revealing protein-lncRNA interaction. *Brief. Bioinform.*, 17(1):106–116.

Frazier, P. I. (2018). A tutorial on bayesian optimization.

Han, Y. and Zhang, S.-W. (2023). ncRPI-LGAT: Prediction of ncRNA-protein interactions with line graph attention network framework. *Comput. Struct. Biotechnol. J.*, 21:2286–2295.

Hashemi, F. S. G., Ismail, M. R., Yusop, M. R., Hashemi, M. S. G., Shahraki, M. H. N., Rastegari, H., Miah, G., and Aslani, F. (2018). Intelligent mining of large-scale bio-data: Bioinformatics applications. *Biotechnology & Biotechnological Equipment*, 32(1):10–29.

Hasib, K. M., Iqbal, M. S., Shah, F. M., Mahmud, J. A., Popel, M. H., Showrov, M. I. H., Ahmed, S., and Rahman, O. (2020). A survey of methods for managing the classification and solution of data imbalance problem. *CoRR*, abs/2012.11870.

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

Ke, G., Meng, Q., Finley, T., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30.

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

Kreuzberger, D., Kühl, N., and Hirschl, S. (2023). Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access*, 11:31866–31879.

Li, A., Li, M. K., Crowther, M., and Vazquez, S. R. (2020). Drug-drug interactions with direct oral anticoagulants associated with adverse events in the real world: A systematic review. *Thromb. Res.*, 194:240–245.

Liaw, A. and Wiener, M. (2002). Classification and regression by random forest. *R News*, 2(3):18–22.

Liu, B., Liu, F., Fang, L., et al. (2014). repdna: a python package to generate various modes of feature vectors for dna sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, 31(8):1307–1309.

Lundberg, Scott M e Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.

Ma, Y., Zhang, H., Jin, C., and Kang, C. (2023). Predicting lncRNA-protein interactions with bipartite graph embedding and deep graph neural networks. *Front. Genet.*, 14:1136672.

Mingyue, C., Le, C., and Kang, N. (2019). Microbiome big-data mining and applications using single-cell technologies and metagenomics approaches toward precision medicine. *Frontiers in Genetics*, 10.

Muhammod, R., Ahmed, S., Md Farid, D., Shatabda, S., Sharma, A., and Dehzangi, A. (2019). PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. *Bioinformatics*, 35(19):3831–3833.

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.

Patel, H., Rajput, D. S., Reddy, G. T., Iwendi, C., Bashir, A. K., and Jo, O. (2020). A review on classification of imbalanced data for wireless sensor networks. *International Journal of Distributed Sensor Networks*, 16(4):1550147720916404.

Peng, L., Tan, J., Tian, X., and Zhou, L. (2022). EnANNDeep: An ensemble-based lncRNA-protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models. *Interdiscip. Sci.*, 14(1):209–232.

Prokhorenkova, L., Gusev, G., Vorobev, A., et al. (2018). Catboost: unbiased boosting with categorical features. pages 6638–6648.

Ribeiro, M., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. pages 97–101.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding rnas and its biological functions. *Nature reviews Molecular cell biology*, 22(2):96–118.

Wang, L., Han, M., Li, X., Zhang, N., and Cheng, H. (2021). Review of classification methods on unbalanced data sets. *IEEE Access*, 9:64606–64628.

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.

Xu, J., Xu, J., Liu, X., and et al. (2022). The role of lncrna-mediated cerna regulatory networks in pancreatic cancer. *Cell Death Discovery*, 8:287.

Yu, H., Shen, Z.-A., Zhou, Y.-K., and Du, P.-F. (2022). Recent advances in predicting protein-lncRNA interactions using machine learning methods. *Curr. Gene Ther.*, 22(3):228–244.

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

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA protein interactions. *PLoS Comput. Biol.*, 14(12):e1006616.

Zhou, L., Wang, Z., Tian, X., et al. (2021). LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncrna–protein interaction identification. *BMC Bioinformatics*, 22:479.