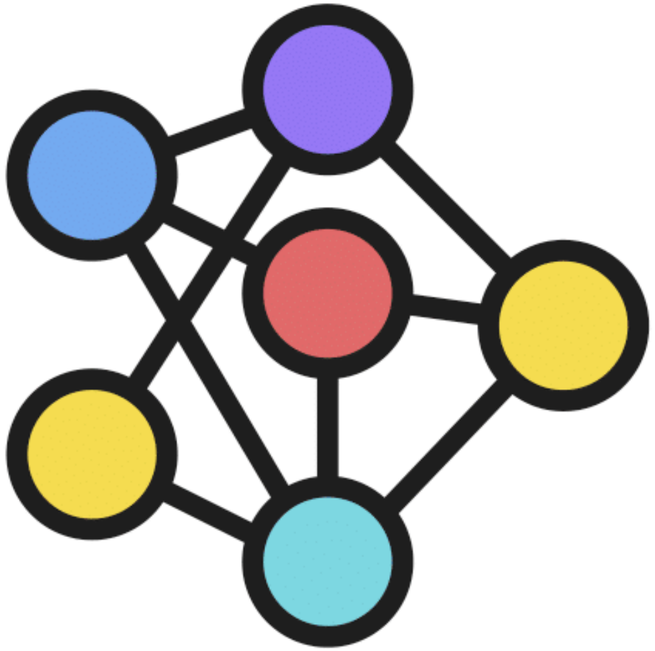


BioAutoML

Democratizing Machine Learning in Life Sciences

Robson Bonidia and André de Carvalho

Universidade de São Paulo - ICMC

Universidade Tecnológica Federal do Paraná

Introduction

Due to biological data's expansion and inherent complexity, **Artificial Intelligence (AI) methods, specifically Machine Learning (ML),** have shown broad applicability in the biological sciences.

ML algorithms can extract useful and meaningful knowledge from biological data, **accelerating discoveries, reducing research expenses, and increasing scientific efficiency.**

These advancements directly benefit **society, the economy, and people’s lives.**

Challenge

Despite its broad application, **designing robust and reliable ML solutions often requires expertise not commonly found among researchers in biology and health**, leading to serious inequalities.

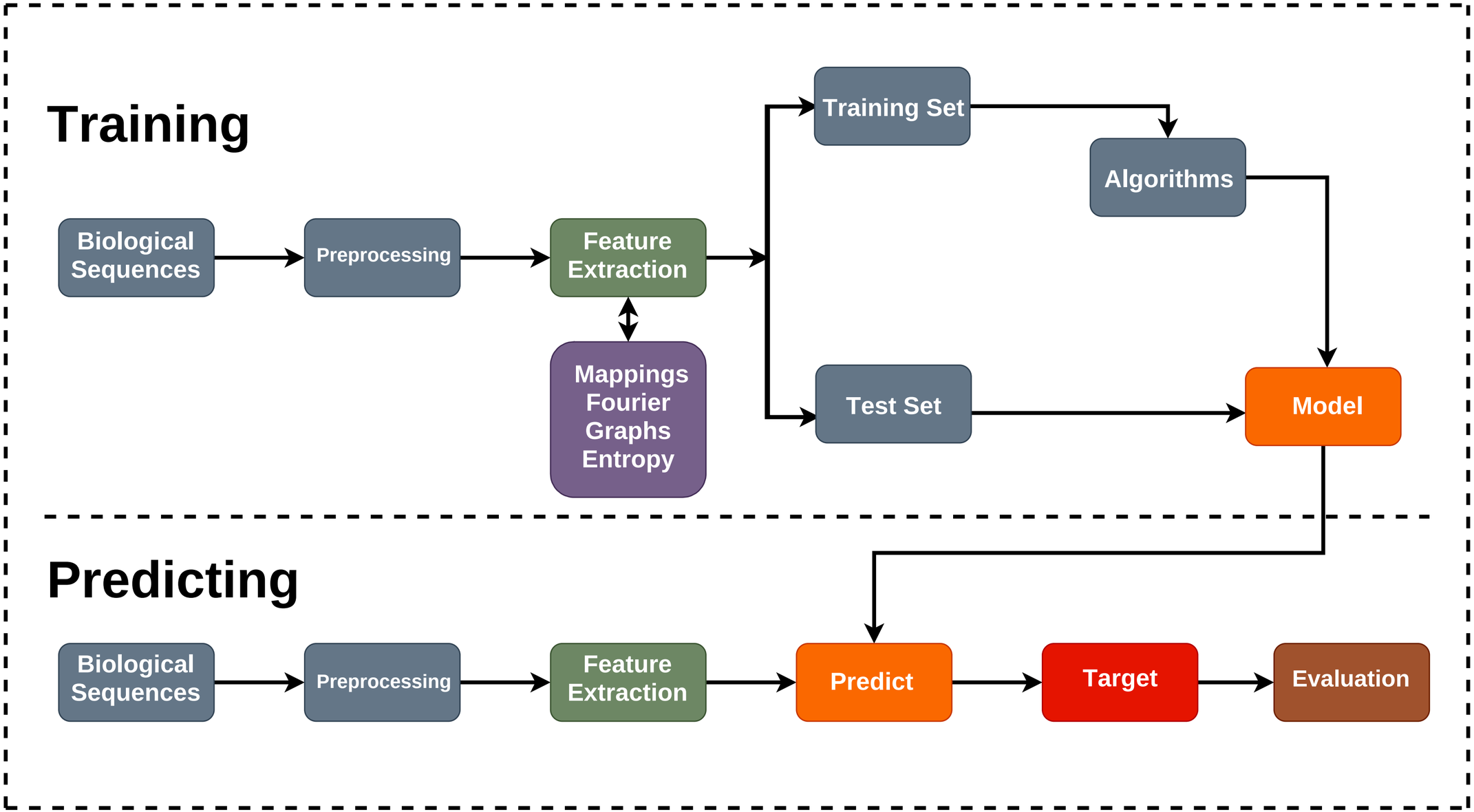
For example, **accessibility inequality** (this creates a disparity in who can use powerful tools, often disadvantaging those working in smaller institutions or regions with few resources)

Knowledge inequality (the complexity of ML algorithms and the skills required constitute a barrier, limiting the potential for innovative research).

Democratization

In this context, **democratizing AI** implies granting ML **accessibility to individuals who are not experts**, for example, individuals without training in data science, mathematics, or computer science.

Democratization



Related Works

Study	Feature Engineering	ML algorithm	Tuning
PseAAC	-	-	
propy	-	-	-
PseKNC-General	-	-	-
SPiCE	-	-	-
Pse-in-One	-	-	-
repDNA	-	-	-
Rcpi	-	-	-
BioSeq-Analysis	-	-	
PyFeat	-	-	-
iLearn	-	**V**	**V**
iLearnPlus	-	**V**	**V**
BioSeq-BLM	-	-	-
autoBioSeqpy	-	**V**	**V**
AutoGenome	-	**V**	**V**
BioAutoML	**V**	**V**	**V**

Our Proposal

BioAutoML, is designed to facilitate the **automatic extraction and selection** of features from sequential data, considering a variety of aspects, **recommendations for ML algorithms**, and specific **hyperparameter tuning** for multi-class and binary classification in biological contexts.

Our Proposal

Feature Engineering

Feature Extraction

Biological Sequences

FASTA

MathFeature

Mathematical descriptors

Conventional descriptors

Feature Vector

Selection and Recommendation

BioAutoML

AUTOMATED FEATURE ENGINEERING

Feature Vector

Selected Features

Algorithms Recommendation

Ensemble of Algorithms

ML - Algorithm Tuning

BioAutoML

AUTOMATED FEATURE ENGINEERING

OR

Selected Features

Algorithms

Hyperparameters Tuning of Algorithms

Combination of the Models induced by the tuned algorithms

Confusion Matrix

Descriptors Importance

Predictive Performace

Trained Model

Metalearning

Our Proposal

Package	Mathematical Descriptors	Conventional Descriptors	Number of Descriptors Calculated
MathFeature	20	17	37
PROFEAT	0	2	2
PseAAC	0	2	2
propy	0	5	5
PseKNC-General	0	5	5
SPiCE	0	4	4
ProtrWeb	0	5	5
ProFET	2	3	5
Pse-in-One	0	5	5
repDNA	0	5	5
Rcpi	0	3	3
repRNA	0	5	5
BioSeq-Analysis	0	9	9
iFeature	1	4	5
PyBioMed	0	7	7
Seq2Feature	0	0	0
PyFeat	1	8	9
iLearn	2	13	15

Main Results

BioAutoML has demonstrated robust results across various problem domains, presenting success cases in areas such as:

- SARS-CoV-2;
- Anticancer peptides;
- Pro-inflammatory peptides;
- HIV-1 sequences;
- Non-classical secreted proteins;
- Sigma70 promoters;
- Recombination sites;
- Small non-coding RNAs, long non-coding RNAs, circular RNAs, and others.

Main Contributions

Democratizing Access to Machine Learning:

- By making BioAutoML **accessible to non-experts,** we address the significant accessibility and knowledge inequalities in the use of ML in biological research.

- Automating complex tasks such as **feature selection, algorithm recommendation, and hyperparameter tuning** reduces the time and expertise required to analyze biological sequences.

- This **increases scientific efficiency, accelerates discoveries, and can lead to significant advancements** in understanding and addressing critical biological and health issues.

Main Contributions

Promoting Inclusivity and Innovation:

- BioAutoML can promote **broader inclusion of researchers from diverse backgrounds and resources**, strengthening the global effort in science and health.

- This can lead to breakthroughs that directly **benefit society, the economy, and people's lives.**

- This signifies a shift from **exclusivity to accessibility,** making ML a shared resource for the collective improvement of science and society.

Our Impact

+20.000
Accesses to our solutions and articles

+7
Awards, Recognitions, and Grants

+136
citations in our articles

+123
Stars on GitHub

+63
Accumulated Impact Factor (IF)

+7
Articles published in high-impact journals

Main Articles

[Link][IF 2021: 13.994] BONIDIA, ROBSON P et al. BioAutoML: automated feature engineering and metalearning to predict noncoding RNAs in bacteria. Briefings in Bioinformatics, v. 1, p. 1-13, 2022.

[Link][IF 2021: 13.994] BONIDIA, ROBSON P et al. MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors. Briefings in Bioinformatics, v. 1, p. 1-10, 2022.

[IF 2021: 2.738] BONIDIA, ROBSON et al. Information Theory for Biological Sequence Classification: A Novel Feature Extraction Technique Based on Tsallis Entropy. Entropy, v. 24, p. 1398, 2022.

Main Recognitions

Google Latin America Research Awards (LARA): BioAutoML was selected by LARA-Google as one of the 24 most promising ideas in Latin America (24 awarded projects, from a pool of 700 entries);

AutoAI-Pandemics (Democratizing ML for Non-Experts, 2023), was selected as one of the most promising projects among a total of 221 proposals from 47 countries in a global competition organized by the AI4PEP, securing funding of 362,500 Canadian dollars;

Helmholtz Visiting Researcher Award and FEMS Research & Training Grant/Award;

Finalists (Top 15 of 82) — Falling Walls Lab Brazil 2022, DWIH São Paulo, Falling Walls Foundation, DAAD The German Center for Science and Innovation.

Acknowledgments

HiDA HELMHOLTZ Information & Data Science Academy

CAPES

Federation of European Microbiological Societies

ICMC USP SÃO CARLOS

LARA

AutoAI-Pandemics
Democratizing Machine Learning

GLOBAL SOUTH AI4PEP NETWORK · SUD GLOBAL RESEAU AI4PEP

IDRC · CRDI
International Development Research Centre
Centre de recherches pour le développement international

Canada

Thank You