# AmyloGram 2.0: MBO in the prediction of amyloid proteins

Dominik Rafacz* [1] Stefan Rödiger [2] Małgorzata Kotulska [3] Michał Burdukiewicz [1]

[1]Warsaw University of Technology [2]Brandenburg Technical University Cottbus-Senftenberg [3]Wrocław University of Science and Technology

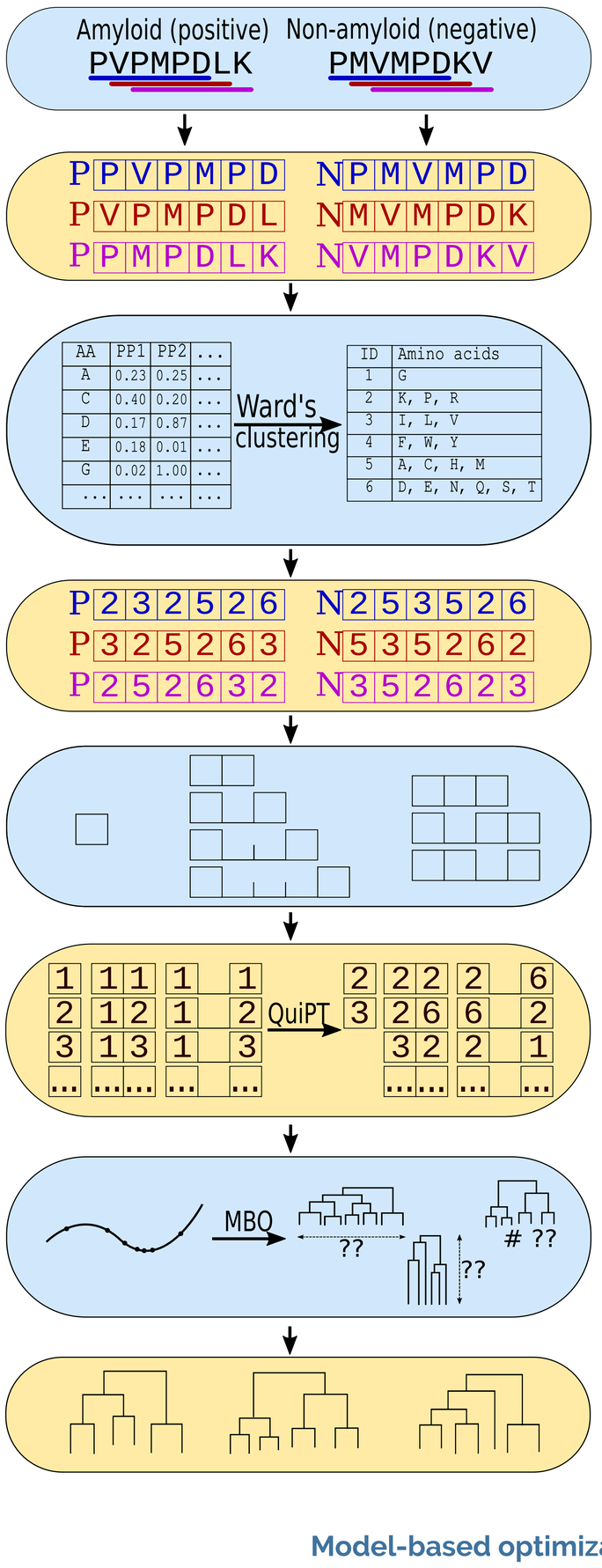
*dominikrafacz@gmail.com ♢ DominikRafacz@github.com

## Introduction

Amyloids are self-aggregating proteins associated with neurodegenerative disorders. The *in silico* identification of amyloid proteins is challenging because their amino acid composition can be extremely variable. Recently, we were able to identify motifs occurring in amyloid sequences and create a machine learning tool, AmyloGram [1], which has outperformed other predictors of amyloids. AmyloGram focuses on identifying amino acid motifs responsible for aggregation, thus providing researches with insights about structural sources of amyloidogenesis.

## AmyloGram workflow

AmyloGram 2.0 is an improved version of the AmyloGram. The improvement in the 2.0 version is the usege of Bayesian hyperparameter tuning.




Source data: peptides with known amyloidicity status. Overlapping hexamers are marked by horizontal lines.

Extraction of overlapping hexamers with ascribed the amyloid status taken from their source peptide (P-positive, N - negative).

Clusterization of amino acids into an encoding using a combination of various physicochemical properties (PP).

Reduction of the amino acid alphabet in hexamers.

Extraction of n-grams. From each hexamer, we extracted continuous and discontinous n-grams with the length n = 1, 2 or 3.

Selection of informative n-grams with Quick Permutation Test (QuiPT).

Seeking for optimal hyperparameters for a classifier using model based optimization (MBO).

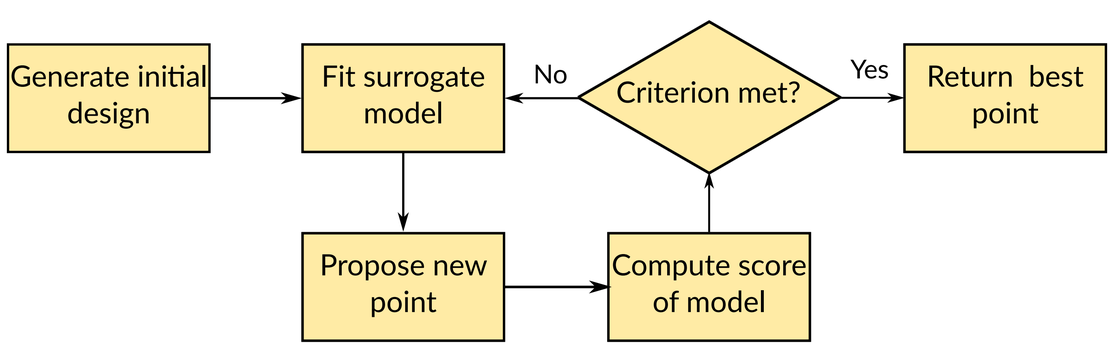
Training of a random forest classifier using the parameters determined in the previous step.

## Model-based optimization

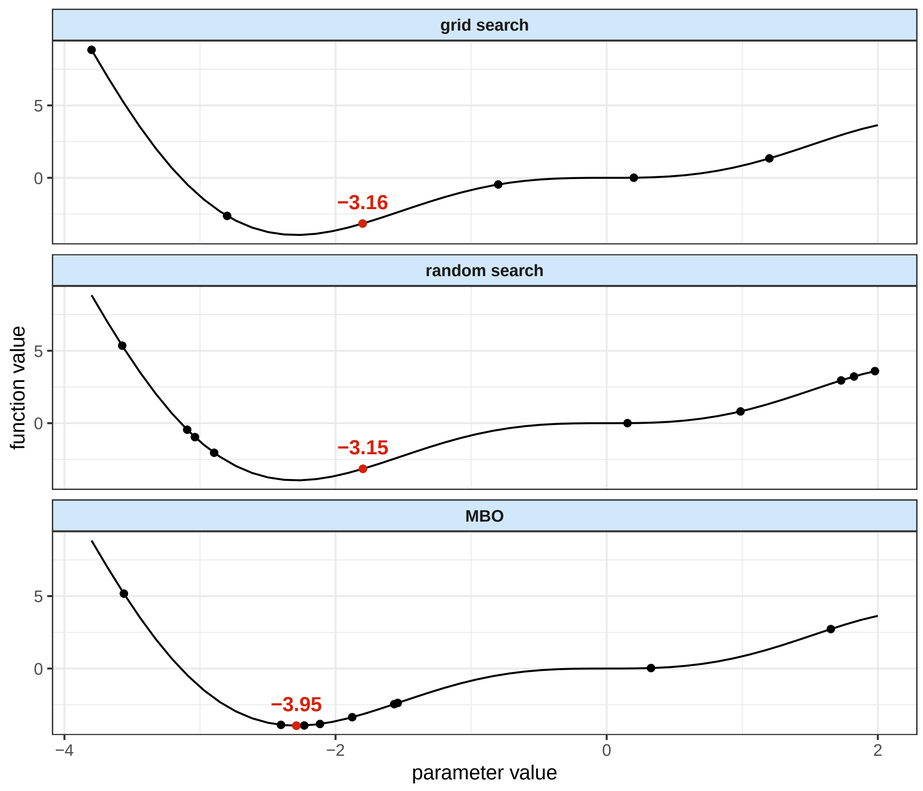
The model-based optimization method is one of the possible techniques of discovering optimal hyperparameters for a model in machine learning. Here we treat the model as a function from space of hyperparameters into space of possible model performance values. After initializing surrogate model $\hat{f}$, the following steps are repeated until some of stopping criterions are met:

1. A set of points $\mathbf{x}^{(\mathbf{i})}$ in the hyperparameters space is proosed basing on values of $\hat{f}$.
2. Value of performance of destination model $f$ is calculated in proposed points $\mathbf{x}^{(\mathbf{i})}$ (model is fit with given set of hyperparameters).
3. Points $\mathbf{x}^{(\mathbf{i})}$ with corresponding values $\mathbf{y}^{(\mathbf{i})}$ of model $f$ are used to fit surrogate model $\hat{f}$.

After reaching certain stoping criterion, the best point $\mathbf{x}^{(\mathbf{i})}$ is returned.
MBO framework used by us is `mlrMBO` [2].

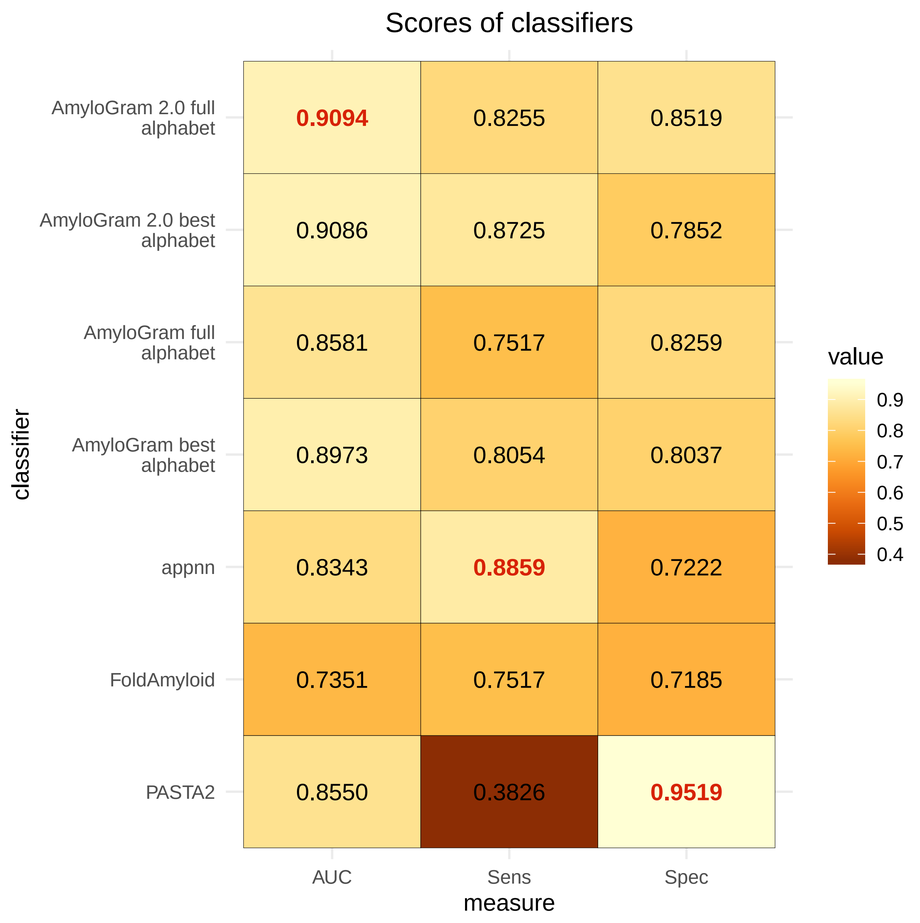



## Comparison of methods of hyperparameters tuning




Grid search method seeks for the optimal value of hyperparameter only in equally-spaced points that are defined by the user. Random search method relies on randomly chosen points in parameter space and is empirically better than the previous one. The model-based optimization seeks for improvement of prediction performance using results of its previous iterations.

## AmyloGram 2.0 performance




## Results and availability

On the pep424 dataset AmyloGram 2.0 reached AUC 0.91. Moreover, the new version of AmyloGram can detect aggregation-prone regions in proteins and explain which amino acid motifs are contributing to the amyloidogenicity. We cross-validate motifs and protein regions detected by our tool with experimental data. MBO method allowed us to reach better results in the prediction of amyloids, but feature engineering (development of new alphabets and n-grams) requires further adjustment of the tuning method.

AmyloGram is available primarily as the web server (under the following link: http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/) but can be also accessed as standalone software and the R package.

## Funding

- Students' Union of Warsaw University of Technology,
- Innovative Regional Growth Core PRÆMED.BIO (Bundesministerium für Bildung und Forschung)

## Bibliography

[1] Michał Burdukiewicz, Piotr Sobczyk, Stefan Rödiger, Anna Duda-Madej, Paweł Mackiewicz, and Małgorzata Kotulska. Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports*, 7(1):12961, October 2017.

[2] Bernd Bischl, Jakob Richter, Jakob Bossek, Daniel Horn, Janek Thomas, and Michel Lang. mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. March 2017.