POSTER

# Biomarker Modelling in Omics Technologies Using Symbolic Regression

**DAVID EDUARDO ROJAS-VELÁZQUEZ**, Utrecht University, Utrecht, Netherlands

**ALBERTO PAOLO TONDA**, National Research Institute for Agriculture, Food and Environment, Paris, Ile-de-France, France

**ALEJANDRO LOPEZ-RINCON**, Utrecht University, Utrecht, Netherlands

**Open Access Support** provided by:

**National Research Institute for Agriculture, Food and Environment**

**Utrecht University**



**Published:** 14 July 2025

**Citation in BibTeX format**

GECCO '25 Companion: Genetic and Evolutionary Computation Conference Companion
*July 14 - 18, 2025*
*Malaga, Spain*

**Conference Sponsors:**
SIGEVO


https://doi.org/10.1145/3712255.3726746
ISBN: 9798400714641

# Biomarker Modelling in Omics Technologies Using Symbolic Regression

David Eduardo Rojas-Velazquez
e.d.rojasvelazquez@uu.nl
Pharmacology, Utrecht University
Utrecht, Netherlands

Alberto Tonda
alberto.tonda@inrae.fr
UMR 518 MIA-PS, INRAE, Université Paris-Saclay
Palaiseau, France

Alejandro Lopez-Rincon
a.lopezrincon@uu.nl
Pharmacology, Utrecht University
Utrecht, Netherlands

## Abstract

Omics data can contain predictive information of the onset of diseases and chronic conditions. Applying machine learning (ML) techniques to omics data is a promising venue of research, but domain data sets are typically high-dimensional and low-sample-size, presenting significant challenges to classic ML approaches. Another obstacle is the black-box nature of many ML algorithms, which prevents them from being deployed in medical practice. Symbolic regression (SR) is a possible solution to obtain human-interpretable models; but even equations cannot be easily understood, if they include hundreds or thousands of features. While feature selection can help reducing the number of features to be considered, most algorithms make unrealistic assumptions or bias the selection using a single classifier. In this work, we apply the Recursive Ensemble Feature Selection (REFS) algorithm, designed to avoid over-relying on a single ML model, with a modern SR algorithm, to obtain interpretable models predictive for different diseases, starting from real-world omics data. Experimental results for five different omics studies show that the completely open-source approach is competitive with the state-of-the-art in closed-source software. Comparing the same pipeline with REFS and more classic feature selection techniques shows that models created with REFS have a better performance.

## CCS Concepts

• **Applied computing** → **Bioinformatics**; • **Computing methodologies** → **Feature selection**; • **Software and its engineering** → **Genetic programming**.

## Keywords

Genetic Programming, Bioinformatics, Feature Selection

**ACM Reference Format:**
David Eduardo Rojas-Velazquez, Alberto Tonda, and Alejandro Lopez-Rincon. 2025. Biomarker Modelling in Omics Technologies Using Symbolic Regression. In *Genetic and Evolutionary Computation Conference (GECCO '25 Companion), July 14–18, 2025, Malaga, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3712255.3726746


*GECCO '25 Companion, July 14–18, 2025, Malaga, Spain*

ACM ISBN 979-8-4007-1464-1/2025/07
https://doi.org/10.1145/3712255.3726746

## 1 Introduction

Advancements in sequencing methods for *omics* data allow for a deeper and more precise sampling, facilitating the identification of specific biomarkers with different applications in the medical field [6]. A biomarker can be a criterion to diagnose or differentiate a disease, assessing its progression, determine the severity as well as predict treatment responses. Thus, in the scientific community there is a growing interest to find new biomarkers [8]. A major challenge in the field of biomarker discovery using machine and deep learning in medicine is the lack of interpretability. This issue arises from several factors, one of the most significant being the high-dimensional, low-sample-size (HDLSS) nature of the datasets used [12]. HDLSS datasets are defined by having a large number of features (high-dimensional) but a relatively small number of samples (low-sample-size). These datasets are common in *omics* studies, such as genomics, transcriptomics, and proteomics, where sequencing each sample is costly, and the number of variables can range from hundreds to thousands, depending on the technology used. The high-dimensional nature of the data is crucial for capturing complex biological processes. In this study, we compare the use of a feature selection algoritms (REFS) with PYSR in a single pipeline in comparison to using KBest algorithm with PYSR in 5 *omics* datasets raging from 785 to 23,514 features. Additionally, we compare our methodologies against Qlattice results in 2 of them, showing comparable results.

## 2 Background

PySR is an open-source tool designed to find simple and understandable mathematical formulas that best match a dataset. It is built in Python and uses a fast Julia backend. PySR employs genetic programming, an evolutionary method that evolves mathematical expressions to fit the data. This approach helps PySR create clear models that optimize specific goals [10]. Unlike complex "black-box" algorithms, PySR produces transparent mathematical models, which is especially important in fields like medicine, where understanding the underlying biological processes is crucial [10]. One weakness in PYSR is that it is not built for HDLSS data. Nevertheless, if we use REFS for the feature selection and then PYSR to build the models we can create an automatic pipeline for *omics* datasets.

## 3 Proposed approach

For each of the examples, we normalized the data with standard *z-score* normalization and use REFS for 10 runs in a 10-folds stratified cross-validation scheme. We made an exception in Acute Myeloid Leukemia, as we have only 6 samples in one class, we had to adjust the REFS algorithm to a 5-fold cross-validation. Then, we select the

best solution of features using Matthews Correlation Coefficient (MCC) to compensate for class imbalance. Although REFS is suitable for multi-label problems, we focused on binary label examples. We selected the following datasets as they belong different *omics* technologies and their medical importance (Table 1).

**Table 1: Different datasets with omics technologies used in the study.**

| Dataset | Measures | Feats. | Reference |
|---|---|---|---|
| Long COVID | Immune Exhaustion gene expression panel | 785 | GSE275334 |
| Dutch Hunger Winter | epigenomics | 2,165 | GSE275334 |
| Acute Myeloid Leukemia | transcriptomics | 23,514 | GSE229046 |
| Alzheimer's Disease | proteomics | 1,166 | [2] |
| Breast Cancer | multi-omics | 1,936 | [2] |

Then, following the procedure described in [2], we divide the reduced dataset into 5-folds, where we will use 4 folds for training and 1 fold for testing. Each time we run PYSR we start the models from scratch to avoid over-fitting and *data leakage* [2].

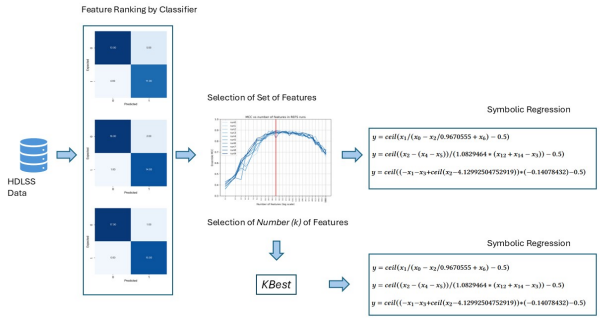


**Figure 1: Overall pipeline: we select the best set of features and the number of features($k$) based on MCC (REFS) and generate the models with PYSR. Using $k$ we generate another set of models based on the top features using f-score as metric.**

The parameters for PYSR are 500 iterations and binary_operators= ["+", "-", "*", "/"], unary_operators=["*exp*", *log*", "*round*", "*floor*", "*ceil*"]. This will give us 5 independent models, based on the set of features from REFS (Fig. 1). To select the most appropriate model on the Pareto front, we rely on PYSR's internal heuristic [4]. As a comparison, using $k$ obtained from REFS algorithm, we ran a second instance of PYSR with the *SelectKBest* algorithm from sci-kit learn to select the top set of features, based on the metric *F-score* [7].

## 3.1 Long COVID

*3.1.1 Dataset.* RNA was extracted from PBMCs of participants with long COVID (n=15), and healthy controls (n=18) with 785 genes with a Nanostring nCounter Immune Exhaustion Panel. Long COVID participants met the WHO's working case definition for Post COVID-19 Condition. We used the file *GSE275334_File_1_Normalised* available in the gene expression omnibus (GEO) with ascension number GSE275334 [5] as dataset.

## 3.2 Dutch Hunger Winter

*3.2.1 Dataset.* The study investigated genome-wide DNA methylation differences linked to early gestational famine exposure. The research compared individuals conceived during the Dutch Famine at the end of WWII (n=24) with their same-sex siblings (n=24) [9].

## 3.3 Acute Myeloid Leukemia

*3.3.1 Dataset.* We used as dataset the *Series Matrix File* with the accession number GSE229046 [11]. This dataset is a study that focuses on identifying predictive biomarkers for chemotherapy response and survival outcomes in pediatric B-cell Acute Lymphoid Leukemia (B-ALL). The research used mRNA sequencing on bone marrow samples from 28 newly diagnosed Hispanic children with B-ALL. 26 samples mRNA expression, where 20 samples were responders to treatment and 6 recurrent with 23,514 genes.

## 3.4 Alzheimer's disease

*3.4.1 Dataset.* Protein expression data from the cerebrospinal fluid of 137 subjects, encompassing 1166 proteins [1]. The data is divided into likelihood of a patient developing Alzheimer's disease (n=49) or not (n=88) [2].

## 3.5 Breast Cancer

*3.5.1 Dataset.* The dataset was obtained from Ciriello [3] and contains multi-omics data to identify regulatory interactions across different omics types (copy numbers, somatic mutations, gene expression, and protein expression) that could predict survival from 705 breast tumour samples of different patients with 1,936 features, with survival (n=611) and a fatal outcome(n=94) outcomes in breast cancer patients.

# 4 Experimental evaluation

## 4.1 Long COVID

Using the REFS algorithm, we reduced the number of genes from 785 to 7 for detecting Long COVID (Fig.2 ). The best model generated with REFS-PYSR is shown in Eq.2, while the best model with KBest-PYSR is shown in Eq.2. The ROC curves for the five models generated using REFS and KBest $k = 7$ (Fig. 3) are also provided.

$$y = ceil(x_1/(x_0 - x_2/0.9670 + x_6) - 0.5) \quad (1)$$

$$y = ceil((-x_1 - x_3 + ceil(x_2 - 4.13)) * (-0.1408) - 0.5) \quad (2)$$

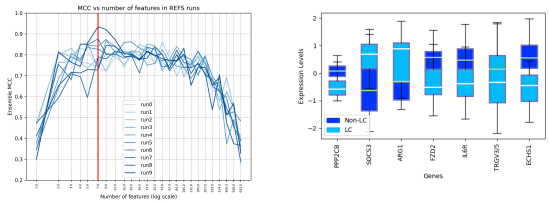


**Figure 2: Feature reduction from 785 to 7 genes to identify Long COVID and Boxplot of the selected 7 genes.**

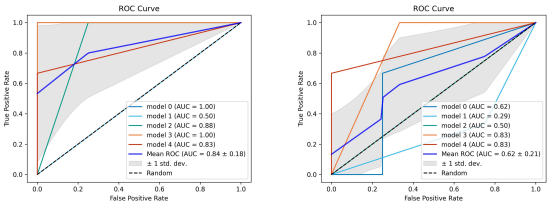


**Figure 3: ROC Curve of the 5 generated models, using REFS (Left) and using KBest (Right) to select the features.**

## 4.2 Dutch Hunger Winter

From the REFS algorithm we reduce the 2,165 CpG sites to 16 to differentiate between siblings that lived through the Dutch Hunger Winter (Fig. 4 ). The best model generated using REFS-PYSR is Eq. 4 and the best model with Kest-PYSR is Eq. 4. The ROC curves for the five models generated using REFS and KBest $k = 16$ (Fig. 5) are also provided.

$$y = ceil((x_2 - (x_4 - x_5))/(1.0829 * (x_{12} + x_{14} - x_3)) - 0.5) \quad (3)$$

$$y = ceil((x1 + x10 - x15 + x3 + x9) * 0.0169 + 0.0093) \quad (4)$$

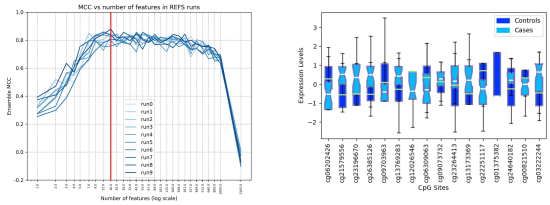


**Figure 4: Feature reduction from 2,165 to 16 CpG sites and Boxplot of the 16 CpG sites.**

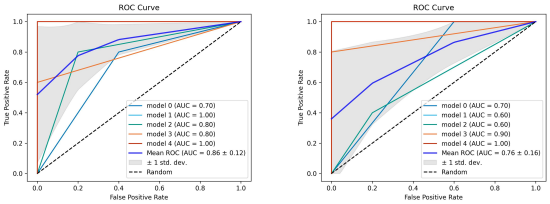


**Figure 5: ROC Curve of the 5 generated models, using REFS (Left) and using KBest (Right) to select the features.**

## 4.3 Acute Myeloid Leukemia Chemoresistance

From the REFS algorithm we reduce from 23,514 genes to 3 to find AML Chemoresistance biomarkers (Fig. 6 ). The best model generated using REFS-PYSR is Eq. 6 and the best model with KBest-PYSR is Eq. 6. The ROC curves for the five models generated using REFS and KBest $k = 3$ (Fig. 7) are also provided.

$$y = ceil(0.0082 * x_0 + 0.9492 * x_1 - 0.5) \quad (5)$$

$$y = ceil(0.2120 * (x_2 + exp(x0)) - 0.5) \quad (6)$$

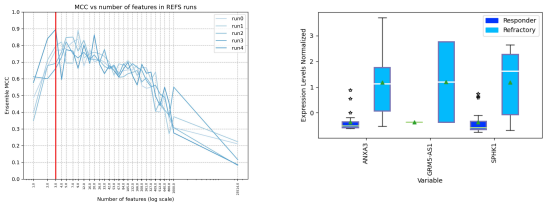


**Figure 6: Feature reduction from 23,514 genes to 3 meaningful genes and Boxplot of the 3 meaningful genes.**

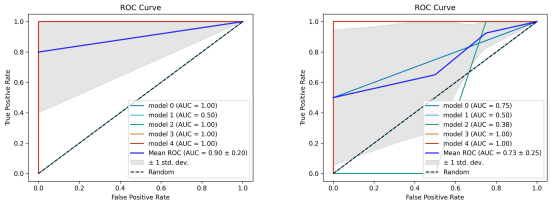


**Figure 7: ROC Curve of the 5 generated models, using REFS (Left) and using KBest (Right) to select the features.**

## 4.4 Alzheimer's Disease

From the REFS algorithm we reduced from 1,166 proteins to 20 to find AD biomarkers (Fig. 8). The best model generated using REFS-PYSR is Eq. 8 and the best model with KBest-PYSR is Eq. 9. The ROC curves for the five models generated using REFS and KBest $k = 20$ (Fig. 9) are also provided.

$$y = ceil(((-0.0377/x_7 + 0.0472/x_2) * ceil \quad (7)$$
$$(x_3 * 1.9030e - 5 - 0.5) + 2.0652e - 5) * x_0 * 0.74969 - 0.5)$$

$$y = ceil(x_{11} * 0.1617 + (x_0 + x_{13} * exp(x_4)) * 0.0470 - (- \quad (8)$$
$$(-0.0401) * x_3 * x_5 + ceil(x_{11} * 0.1617 - 0.5)) - 0.0171)$$

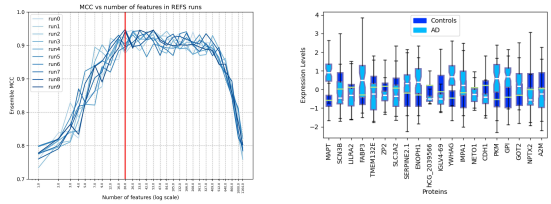


**Figure 8: Feature reduction from 1,166 to 20 proteins and Boxplot of the most 20 proteins.**

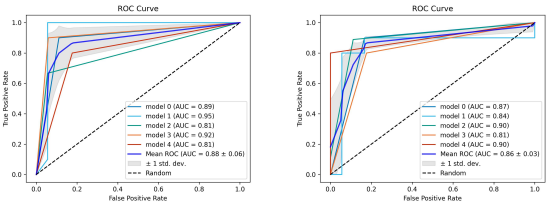


**Figure 9: ROC Curve of the 5 generated models, using REFS (Left) and using KBest (Right) to select the features.**

In comparison to Qlattice [2], our methodology is slightly better, as their reported best model is 0.92, whereas ours is 0.95. However, it is important to notice that 2 out 3 of the features selected by QLattice are the same as Eq. 8, as $x_0 = MAPT$ and $x_2 = LILRA2$, this serves as validation of the importance of these proteins.

## 4.5 Breast Cancer

From the REFS algorithm we reduced from 1,936 multi-omics features to 208 to find Breast Cancer biomarkers (Fig. 10). The best model generated using REFS-PYSR is Eq. 10 and the best model with KBest-PYSR is Eq. 11. The ROC curves for the five models generated using REFS and KBest $k = 208$ (Fig. 11) are also provided.

$$y = x_{31} * x_{12} * 0.00717484 \quad (9)$$

$$y = ceil(x_{166} * (-0.0476) * x_{148} * ceil(-x_{117} + x_{56} - 0.5) - 0.5) + 0.1028 \quad (10)$$

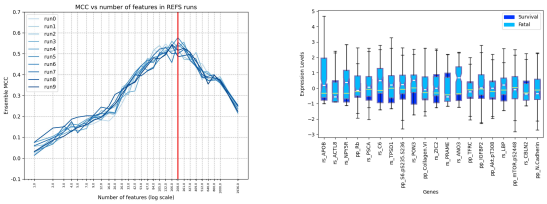


**Figure 10: Feature reduction 1,936 of top 208 meaningful proteins and Boxplot of the top 20 meaningful multi-omics features.**

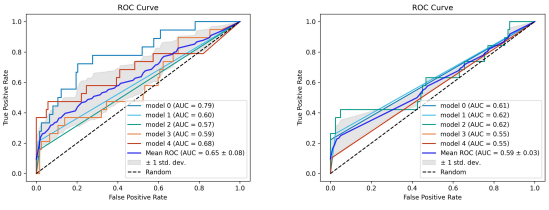


**Figure 11: ROC Curve of the 5 generated models, using REFS (Left) and using KBest (Right) to select the features.**

In comparison to Qlattice, our best model has 0.79 AUC in comparison to 0.66 AUC. Our mean AUC is 0.65 with a std of 0.08, in comparison to 0.635, with a std of 0.07 using REFS-PYSR. However, with the $k = 208$ best features we get an AUC mean of 0.70 with a std of 0.09 using MLP.

## 5 Conclusions and future work

The study demonstrates that the Recursive Ensemble Feature Selection (REFS) combined with PySR (symbolic regression) is effective in reducing high-dimensional omics data to a manageable number of features while maintaining or improving model performance. This approach outperforms traditional methods like KBest. The use of symbolic regression with PySR provides interpretable models, which is crucial for understanding the underlying biological mechanisms in medical applications. This interpretability is a significant advantage over black-box machine learning models. The methodology shows robustness across various omics datasets, including gene expression, epigenomics, transcriptomics, proteomics, and multi-omics data, indicating its potential for broad application in biomarker discovery.

## References

[1] Jakob M Bader, Philipp E Geyer, Johannes B Müller, Maximilian T Strauss, Manja Koch, Frank Leypoldt, Peter Koertvelyessy, Daniel Bittner, Carola G Schipke, Enise I Incesoy, et al. 2020. Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease. *Molecular systems biology* 16, 6 (2020), e9356.

[2] Niels Johan Christensen, Samuel Demharter, Meera Machado, Lykke Pedersen, Marco Salvatore, Valdemar Stentoft-Hansen, and Miquel Triana Iglesias. 2022. Identifying interactions in omics data for clinical biomarker discovery using symbolic regression. *Bioinformatics* 38, 15 (2022), 3749–3758.

[3] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, Matthew D Wilkerson, Suhn K Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, et al. 2015. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* 163, 2 (2015), 506–519.

[4] Miles Cranmer. 2023. Interpretable machine learning for science with PySR and SymbolicRegression. jl. *arXiv preprint arXiv:2305.01582* (2023).

[5] Natalie Eaton-Fitch, Penny Rudd, Teagan Er, Livia Hool, Lara Herrero, and Sonya Marshall-Gradisnik. 2024. Immune exhaustion in ME/CFS and long COVID. *JCI insight* 9, 20 (2024), e183810.

[6] Chao Li, Zhenbo Gao, Benzhe Su, Guowang Xu, and Xiaohui Lin. 2022. Data analysis methods for defining biomarkers from omics data. *Analytical and Bioanalytical Chemistry* 414, 1 (2022), 235–250.

[7] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[8] Adam S Ptolemy and Nader Rifai. 2010. What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. *Scandinavian Journal of Clinical and Laboratory Investigation* 70, sup242 (2010), 6–14.

[9] Elmar W Tobi, Jelle J Goeman, Ramin Monajemi, Hongcang Gu, Hein Putter, Yanju Zhang, Roderick C Slieker, Arthur P Stok, Peter E Thijssen, Fabian Müller, et al. 2014. DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nature communications* 5, 1 (2014), 5592.

[10] Alberto Tonda. 2025. Review of PySR: high-performance symbolic regression in Python and Julia.

[11] YX Torres-Llanos, J Zabaleta, N Cruz-Rodriguez, SM Quijano, PC Guzman, I de los Reyes, N Poveda, A Infante, L Lopez, and AL Combita. 2024. MIR4435-2HG as a novel predictive biomarker of chemotherapy response and death in pediatric B-cell ALL [RNA-seq]. *GEO Accession viewer* GSE229046 (2024). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE229046 Public on May 22, 2024.

[12] Kosuke Yoshida. 2018. Interpretable machine learning approaches to high-dimensional data and their applications to biomedical engineering problems. (2018).