# Scaling tree-based automated machine learning to biomedical big data with a dataset selector

*This manuscript (permalink) was automatically generated from trang1618/tpot-ds-ms@c77b3f7 on December 19, 2018.*

## Authors


- **Trang T. Le**[☯]
  0000-0003-3737-6565 · trang1618 · trang1618
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104
- **Weixuan Fu**[☯]
  0000-0002-6434-5468 · weixuanfu · weixuanfu
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104
- **Jason H. Moore**[†]
  0000-0002-5015-1099 · EpistasisLab · moorejh
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104 · Funded by National Institute of Health Grant Nos. LM010098, LM012601

[☯] — These authors contributed equally to this work.

[†] — Direct correspondence to jhmoore@upenn.edu.

# Abstract

Automated machine learning (AutoML) systems are helpful data science assistants designed to scan data for novel features, select appropriate supervised learning models and optimize their parameters. For this purpose, Tree-based Pipeline Optimization Tool (TPOT) was developed using strongly typed genetic programming to recommend an optimized analysis pipeline for the data scientist's prediction problem. However, TPOT may reach computational resource limits when working on big data such as whole-genome expression data. We introduce two new features implemented in TPOT that helps increase the system's scalability: Dataset selector and Template. Dataset selector (DS) provides the option to specify subsets of the features as separate datasets, assuming the signals come from one or more of these specific data subsets. Built in at the beginning of each pipeline structure, DS reduces the computational expense of TPOT to only evaluate on a smaller subset of data rather than the entire dataset. Consequently, DS increases TPOT's efficiency in application on big data by slicing the dataset into smaller sets of features and allowing genetic programming to select the best subset in the final pipeline. Template enforces type constraints with strongly typed genetic programming and enables the incorporation of DS at the beginning of each pipeline. We show that DS and Template help reduce TPOT computation time and potentially provide more interpretable results. Our simulations show TPOT-DS significantly outperforms a tuned XGBoost model and standard TPOT implementation. We apply TPOT-DS to real RNA-Seq data from a study of major depressive disorder. Independent of the previous study that identified significant association with depression severity of the enrichment scores of two modules, in an automated fashion, TPOT-DS corroborates that one of the modules is largely predictive of the clinical diagnosis of each individual.



# Author Summary

Big data have recently become prevalent in many fields including meteorology, complex physics simulations, large scale imaging, genomics, biomedical research, environmental research and more. However, big data present challenges for Automated Machine Learning (AutoML) tools that help data scientists find best analysis solution with the long runtime, high computational expense as well complex pipeline with low interpretability. TPOT, a Python AutoML tool that uses genetic programming to optimize machine learning pipelines for analyzing biomedical data, faces the same challenges in the early implementations. We developed two novel features for TPOT, Template and Dataset Selector, that leverage domain knowledge, greatly reduce the computational expense and flexibly extend TPOT's application to biomedical big data analysis.

# Introduction

For many bioinformatics problems of classifying individuals into clinical categories from high-dimensional biological data, performance of a machine learning (ML) model depend greatly on the problem it is applied to [1,2]. In addition, choosing a classifier is merely one step of the arduous process that leads to predictions. To detect patterns among features (*e.g.*, clinical variables) and their associations with the outcome (*e.g.*, clinical diagnosis), a data scientist typically has to design and test different complex machine learning (ML) frameworks that consist of data exploration, feature engineering, model selection and prediction. Automated machine learning (AutoML) systems were developed to automate this challenging and time-consuming process. These intelligent systems increase the accessibility and scalability of various machine learning applications by efficiently solving an optimization problem to discover pipelines that yield satisfactory outcomes, such as prediction accuracy. Consequently, AutoML allows data scientists to focus their effort in applying their expertise in other important research components such as developing meaningful hypotheses or communicating the results.

Various approaches have been employed to build AutoML systems for diverse applications. Auto-sklearn [3] and Auto-WEKA [4] use Bayesian optimization for model selection and hyperparameter optimization. Meanwhile, Recipe [6] optimizes the ML pipeline through grammar-based genetic programming and Autostacker [7] automates stacked ensembling. Both methods automate hyperparameter tuning and model selection using evolutionary algorithm. DEvol [8] designs deep neural network specifically via genetic programming. H2O.ai [9] automates data preprocessing, hyperparameter tuning, random grid search and stacked ensembles in a distributed ML platform in multiple languages. Finally, Xcessiv [10] provides web-based application for quick, scalable, and automated hyper-parameter tuning and stacked ensembling in Python.

Tree-based Pipeline Optimization Tool (TPOT) is a genetic programming-based AutoML system that uses genetic programming (GP) [11] to optimize a series of feature selectors, preprocessors and ML models with the objective of maximizing classification accuracy. While most AutoML systems primarily focus on model selection and hyperparameter optimization, TPOT also pays attention to feature selection and feature engineering by evaludating the complete pipelines based on their cross-validated score such as mean squared error or balanced accuracy. Given no a priori knowledge about the problem, TPOT has been showed to frequently outperform standard machine learning analyses [12,13]. Effort has been made to specialize TPOT for human genetics research, which results in a useful extended version of TPOT, TPOT-MDR, that features Multifactor Dimensionality Reduction and an Expert Knowledge Filter [14]. However, at the current stage, TPOT still requires great computational expense to analyze large datasets such as in genome-wide association studies (GWAS) or gene expression analyses. Consequently, application of TPOT on real-world datasets has been limited to small sets of features [15].

In this work, we introduce two new features implemented in TPOT that helps increase the system's scalability. First, the Dataset Selector (DS) allows the users to pass specific subsets of the features, reducing the computational expense of TPOT at the beginning of each pipeline to only evaluate on a smaller subset of data rather than the entire dataset. Consequently, DS increases TPOT's efficiency in application on large data sets by slicing the data into smaller sets of features (*e.g.* genes) and allowing genetic algorithm to select the best subset in the final pipeline. Second, Template enables the option for strongly typed GP, a method to enforce type constraints in genetic programming. By letting users specify a desired structure of the resulting machine learning pipeline, Template helps reduce TPOT computation time and potentially provide more interpretable results.

# Methods

We begin with description of the two novel additiona to TPOT, Dataset Selector and Template. Then, we provide detail of a real-world RNA-Seq expression dataset and describe a simulation approach to generate data comparable with the expression data. Finally, we discuss other methods and performance metrics for comparison. The `R` and `Python` scripts for simulation and analysis are publicly available on the GitHub repository https://github.com/lelaboratoire/tpot-ds.

## Tree-based Pipeline Optimization Tool

Tree-based Pipeline Optimization Tool (TPOT) automates the laborious process of designing a ML pipeline by representing pipelines as binary expression trees with ML operators as primitives. Pipeline elements include algorithms from the extensive library of scikit-learn [16] as well as other efficient implementations such as extreme gradient boosting. Applying GP with the NSGA-II Pareto optimization [17], TPOT optimizes the accuracy achieved by the pipeline while accounting for its complexity. Specifically, to automatically generate and optimize these machine learning pipelines, TPOT utilizes the Python package DEAP [18] to implement the GP algorithm. Implementation details can be found at TPOT's active Github repository https://github.com/EpistasisLab/tpot.

## Dataset Selector

TPOT's current operators include sets of feature pre-processors, feature transformers, feature selection techniques, and supervised classifiers and regressions. In this study, we introduce a new operator called Dataset Selector (DS) that enables biologically guided group-level feature selection. Specifically, taking place at the very first stage of the pipeline, DS passes only a specific subset of the features onwards, effectively slicing the large original dataset into smaller ones. Hence, with DS, users can specify subsets of features of interest to reduce the feature space's dimension at pipeline initialization. From predefined subsets of features, the DS operator allows TPOT to select the best subset that maximize average accuracy in k-fold cross validation (5-fold by default).

For example, in a gene expression analysis of major depressive disorder, a neuroscientist can specify collections of genes in pathways of interest and identify the important collection that helps predict the depression severity. Similarly, in a genome-wide association study of breast cancer, an analyst may assign variants in the data to different subsets of potentially related variants and detect the subset associated with the breast cancer diagnosis. In general, the DS operator allows for compartmentalization the feature space to smaller subsets based on *a priori* expert knowledge about the biomedical dataset. From here, TPOT selects the most relevant group of features, which can be utilized to motivate further analysis on that small group of features in biomedical research.

## Template

Parallel with the establishment of the Dataset Selector operator, we now offer TPOT users the option to define a Template that provides a way to specify a desired structure for the resulting machine learning pipeline, which will reduce TPOT computation time and potentially provide more interpretable results.

Current implementation of Template supports linear pipelines, or path graphs, which are trees with two nodes (operators) of vertex degree 1, and the other $n - 2$ nodes of vertex degree 2. Further, Template takes advantage of the strongly typed genetic programming framework that enforces data-type constraints [19] and imposes type-based restrictions on which element (*i.e.*, operator) type can be chosen at each node. In strongly typed genetic programming, while the fitness function and parameters remain the same, the initialization procedure and genetic operators (*e.g.*, mutation, crossover) must respect the enhanced legality constraints [19]. With a Template defined, each node in the tree pipeline is assigned one of the five major operator types: dataset selector, feature selection, feature transform, classifier or regressor. Moreover, besides the major operator types, each node can also be assigned more specifically as a method of an operator, such as decision trees for classifier. An example Template is Dataset selector → Feature transform → Decision trees.

## Datasets

We apply TPOT with the new DS operator on both simulated datasets and a real world RNA-Seq gene expression dataset. With both real-world and simulated data, we hope to acquire a comprehensive view of the strengths and limitations of TPOT in the next generation sequencing domain.

### Simulation methods

The simulated datasets were generated using the `R` package `privateEC`, which was designed to simulate realistic effects to be expected in gene expression or resting-state fMRI data. In the current study, to be consistent with the real expression dataset (described below), we simulate interaction effect data with $m$ = 200 individuals (100 cases and 100 controls) and $p$ = 5,000 real-

valued features with 4% functional (true positive association with outcome) for each training and testing set. Full details of the simulation approach can be found in Refs. [20,21]. Briefly, the privateEC simulation induces a differential co-expression network random normal expression levels and permute the values of targeted features within the cases to generate interactions. Further, by imposing a large number of background features (no association with outcome), we seek to assess TPOT-DS's performance in accommodating large numbers of non-predictive features.

To closely resemble the module size distribution in the RNA-Seq data, we first fit a $\Gamma$ distribution to the observed module sizes then sample from this distribution values for the simulated subset size, before the total number of features reaches 4,800 (number of background features). Then, the background features were randomly placed in each subset corresponding to its size. Also, for each subset $S_i, i = 1, \ldots, n$, a functional feature $s_j$ belongs to the subset with the probability

$$P(s_j \in S_i) \sim 1.618^{-i} \tag{1}$$

where 1.618 is an approximation of the golden ratio and yields a reasonable distribution of the functional features: they are more likely to be included in the earlier subsets (subset 1 and 2) than the later ones.

### Real-world RNA-Seq expression data

We employed TPOT-DS on an RNA-Seq expression dataset of 78 individuals with major depressive disorder (MDD) and 79 healthy controls (HC) from Ref. [20]. Gene expression levels were quantified from reads of 19,968 annotated protein-coding genes and underwent a series of preprocessing steps including low read-count and outlier removal, technical and batch effect adjustment, and coefficient of variation filtering. Consequently, whole blood RNA-Seq measurements of 5,912 genes were obtained and are now used in the current study to test for association with MDD status. We use the 23 subsets of interconnected genes called depression gene modules (DGMs) identified from the RNA-Seq gene network module analysis [20] as input for the DS operator. We remark that these modules were constructed by an unsupervised machine learning method with dynamic tree cutting from a co-expression network. As a result, this prior knowledge of the gene structure does not depend on the diagnostic phenotype and thus yields no bias in the downstream analysis of TPOT-DS.

## Performance assessment

For each simulated and real-world dataset, after randomly splitting the entire data in two balanced smaller sets (75% training and 25% holdout), we trained TPOT-DS with the Template `Dataset Selector-Transformer-Classifier` on training data to predict class (*e.g.*, diagnostic phenotype in real-world data) in the holdout set. We assess the performance of TPOT-DS by quantifying its ability to correctly select the most important subset (containing most functional features) in 100 replicates of TPOT runs on simulated data with known underlying truth. To prevent

potential overfitting, we select the pipeline that is closest to the 90th percentile of the cross-validation accuracy to be optimal. We compare the out-of-sample (holdout) accuracy of TPOT-DS's optimal pipeline on the holdout set with that of standard TPOT (with `Transformer-Classifier` Template, no DS operator) and eXtreme Gradient Boosting [22], or XGBoost, which is a fast and an efficient implementation of the gradient tree boosting method that has shown much utility in many winning Kaggle solutions [23] and been successfully incorporated in several neural network architectures [24,25]. In the family of gradient boosted decision trees, XGBoost accounts for complex non-linear interaction structure among features and leverages gradient descents and boosting (sequential ensemble of weak classifiers) to effectively produce a strong prediction model. To obtain the optimal performance for this baseline model, we tune XGBoost hyperparameters using TPOT Template with only one classifier `XGBClassifier`, which is imported from the `xgboost` python package. Because of stochasticity in the optimal pipeline from TPOT-DS, standard TPOT and the tuned XGBoost model, we fit these models on the training data 100 times and compare 100 holdout accuracy values from each method. We choose accuracy to be the metric for comparison because phenotype is balanced in both simulated data and real-world data. Detailed code needed to reproduce the results has been made available on the GitHub repository https://github.com/trang1618/tpot-ds.

### Manuscript drafting

This manuscript is collaboratively written using Manubot [26], a software that supports open paper writing via GitHub using the Markdown language. Manubot uses continuous integration to monitor changes and automatically update the manuscript. Consequently, the latest version of this manuscript is always available at https://trang1618.github.io/tpot-ds-ms/.

## Results

Our main goal is to test the performance of methods to identify features that discriminate between groups and optimize the classification accuracy.

### Evaluation of TPOT's subset selection ability

#### Simulated data

We compare the accuracy of each method for a simulated dataset with moderate interaction effect. We assign values of the effect size in the simulations to generate adequately challenging datasets so that the methods' accuracies stay moderate and do not cluster around 0.5 or 1. The data set is split into 75% training and 25% holdout. The three models, TPOT-DS, standard TPOT and XGBoost, are built from the training dataset, then the trained model is applied to the independent holdout data to obtain the generalization accuracy. Our simulation design produces a reasonable distribution of the functional features in all subsets, of which proportions are shown in Table [S1].

According to Eq. 1, the earlier the subset, the more functional features it has. Therefore, our first aim is to determine how well TPOT-DS can identify the first subset ($S_1$) that contains the largest number of informative features. The general workflow of TPOT-DS is shown in Figure 1 along with the optimal pipeline found with the specified template `Dataset Selector-Transformer-Classifier` in simulated data (top) and real-world expression data (bottom).

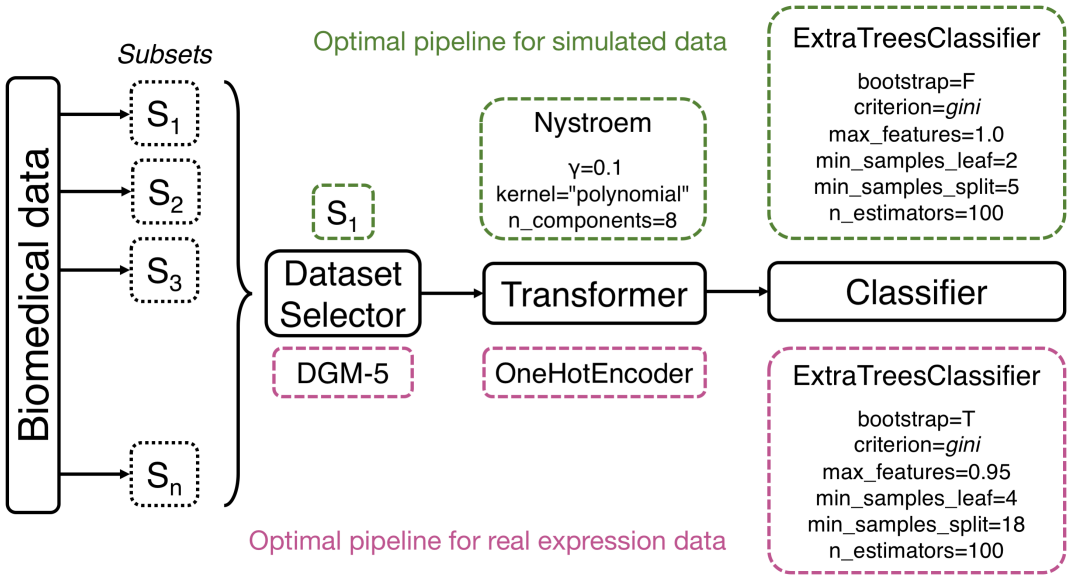


Figure 1: TPOT-DS's workflow and example pipelines. Optimal pipeline with optimized parameters are shown for simulated data (top) and real-world data (bottom).

For simulated dataset, the optimal pipeline selects subset $S_1$ then constructs an approximate feature map for a linear kernel with Nystroem, which uses a subset of the data as basis for the approximation. The final prediction is made with an extra-trees classifier that fits a number of randomized decision trees on various sub-samples of the dataset with the presented optimized parameters (Fig. 1).

In 100 replications, TPOT-DS correctly selects subset $S_1$ in 75 resulting pipelines (Fig. 2), with the highest average holdout accuracy (0.69 across all 75 pipelines).

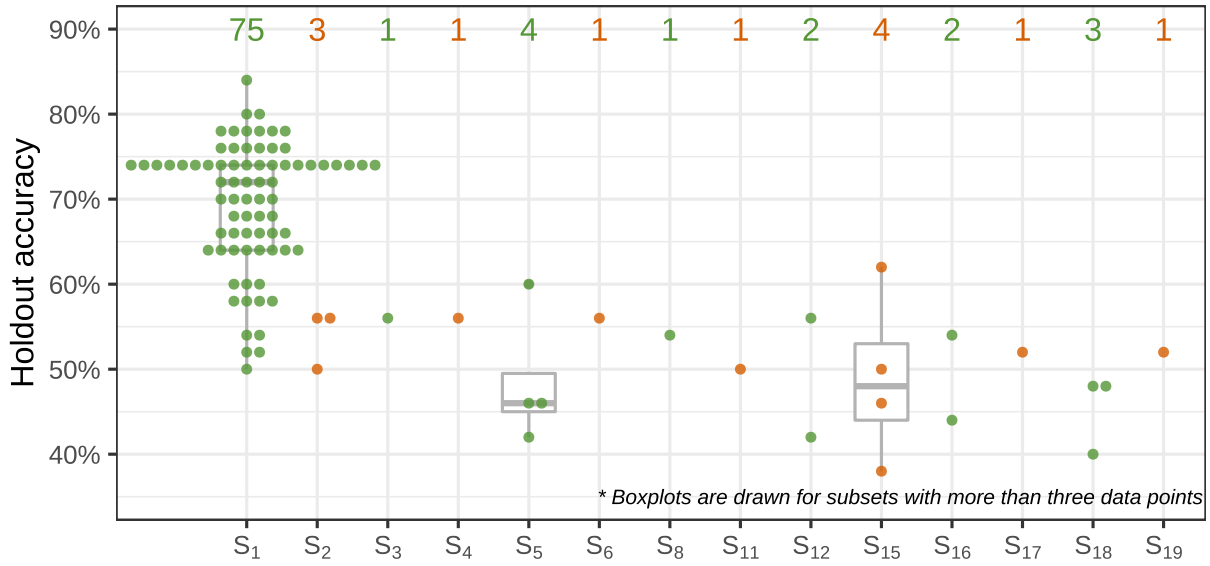


Figure 2: TPOT-DS's holdout accuracy in simulated data with selected subset. Number of pipeline inclusions of each subset in 100 replications is displayed above the boxplots. Subset *s1* is the most frequent to be included in the final pipeline and yields the best prediction accuracy in the holdout set.

### RNA-Seq expression data

We apply standard TPOT, TPOT-DS and XGBoost to the RNA-Seq study of 78 major depressive disorder (MDD) subjects and 79 healthy controls (HC) described in [20]. The dataset contains 5,912 genes after preprocessing and filtering (see Methods for more detail). We excluded 277 genes that did not belong to 23 subsets of interconnected genes (DGMs) so that the dataset remains the same across the three methods. As with simulated data, all models are built from the training dataset (61 HC, 56 MDD), then the trained model is applied to the independent holdout data (18 HC, 22 MDD) to obtain the generalization accuracy.

The most optimal pipeline selects subset DGM-5 then scales each expression feature by its maximum absolute value (Fig. 1). Similar to the best pipeline for simulated data, the final prediction is made with an extra-trees classifier with a different set of optimized parameters (Fig. 1).

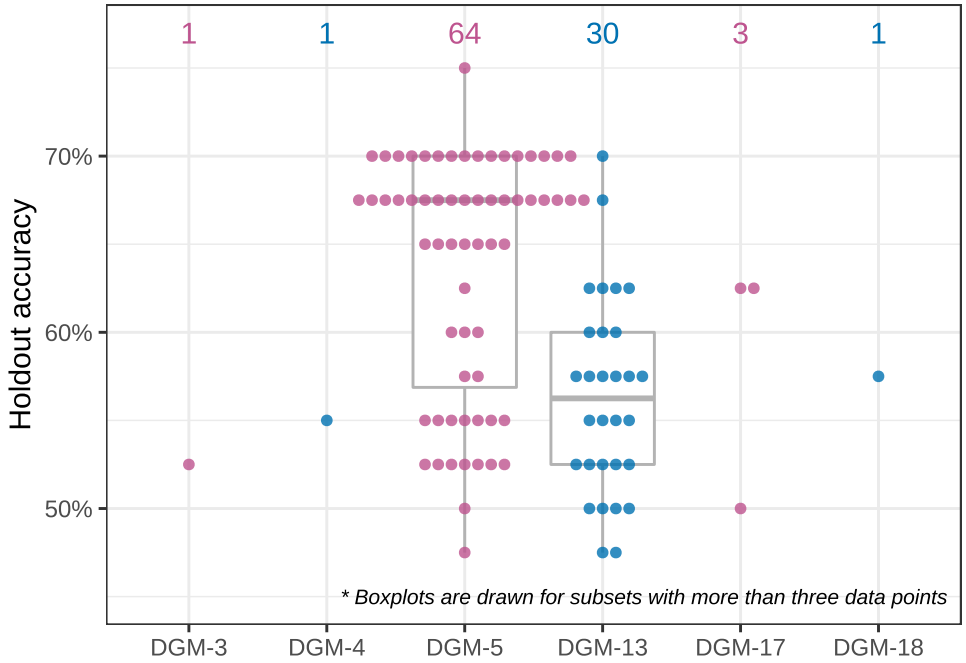


Figure 3: TPOT-DS's holdout accuracy in RNA-Seq expression data with selected subset. Number of pipeline inclusions of each subset in 100 replications is displayed above the boxplots. Subsets DGM-5 and DGM-13 are the most frequent to be included in the final pipeline. Pipelines that include DGM-5 on average produces higher MDD predition accuracy in the holdout set.

In 100 replications, TPOT-DS selects DGM-5 (291 genes) 64 times to be the subset most predictive of the diagnosis status (Fig. 3), with the highest average holdout accuracy of 0.636 across 64 pipelines. In the previous study with a modular network approach, we showed that DGM-5 has statistically significant associations with depression severity measured by the Montgomery-Åsberg Depression Scale (MADRS). Although there is no direct link between the top genes of the module (Fig. 4a) and MDD in the literature, many of these genes interact with other MDD-related genes. For example, NR2C2 and TCF7L1 interact with FKBP5 gene whose association with MDD has been strongly suggested [27,28,29]. Many of DGM-5's top genes were also shown to have statistically significant association with diagnosis phenotypes from a univariate analysis after multiple hypothesis testing correction [20]. Further, with 82% overlap of DGM-5's genes in a separate dataset from the RNA-Seq study by Mostafavi et al. [30], this gene collection's enrichment score was also shown to be significantly associated with the diagnosis status in this independent dataset.

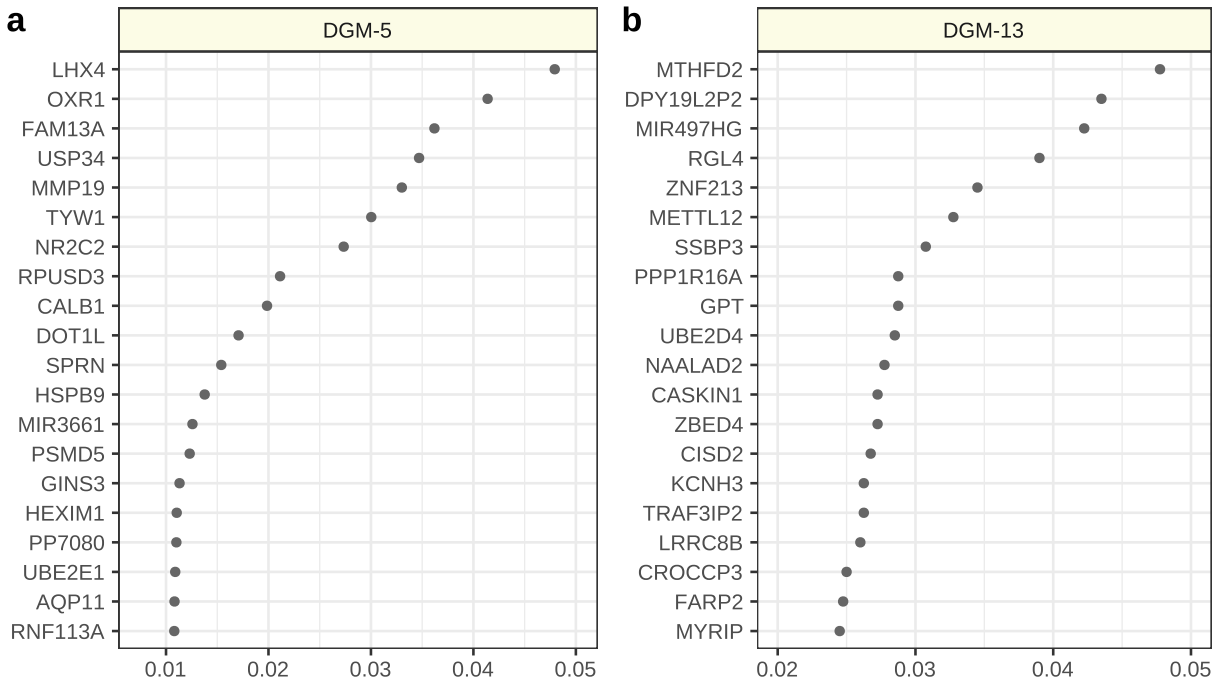


Figure 4: Importance scores of the top twenty expression features in the best pipeline that selects DGM-5 and one that selects DGM-13. Comprehensive importance scores of the all expression features computed from the final classifiers of the best pipelines are provided in Table S2.

After DGM-5, DGM-13 (134 genes) was selected by TPOT-DS 30 times (Fig. 3), with an average holdout accuracy of 0.563 across 30 pipelines. Previous network approach did not find statistically significant association between this module's enrichment score and the MADRS. Gene set enrichment analysis reported DGM-13's involvement in axon guidance and developmental biology pathways with Reactome-FDR $q$-value $< 0.05$ [20].

## Accuracy assessment

For the simulated data, across all 100 model fits, the optimal TPOT-DS pipeline yields an average holdout prediction accuracy of 0.65, while the standard TPOT without DS and tuned XGBoost models respectively report an average holdout accuracy of 0.48 and 0.49 (Fig. 5). This overfitting in the performance of these other two models is likely due to the models' high flexibility that *over-learns* the training data, especially with the presence of many noisy background variables.

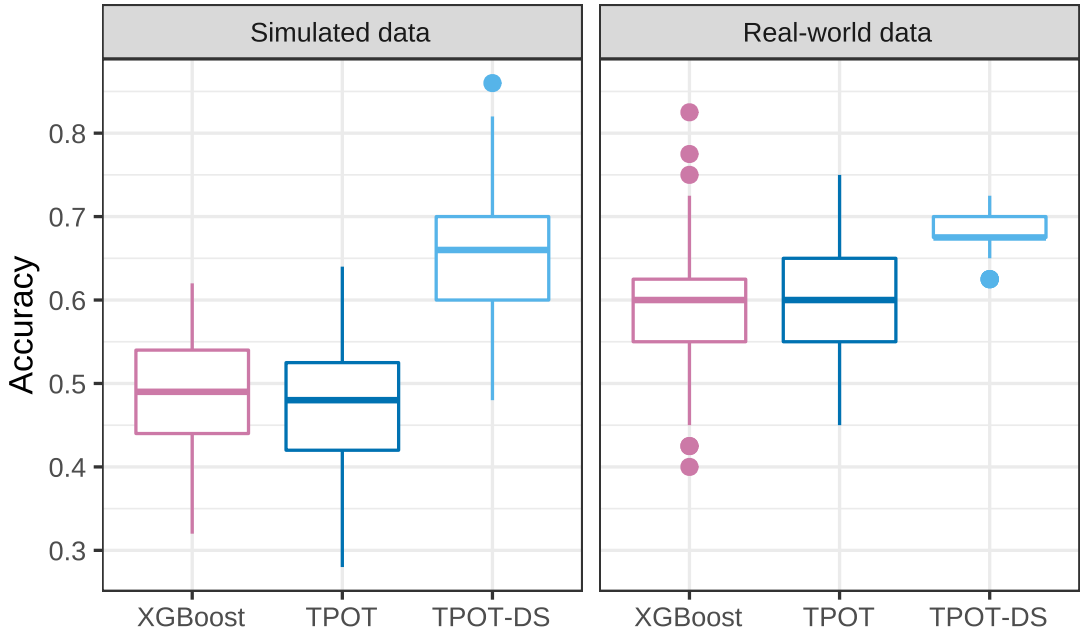


Figure 5: Performance comparison of three models: tuned XGBoost, optimal pipeline from standard TPOT and optimal pipeline from TPOT-DS.

Meanwhile, for the real-world expression data, the optimal TPOT-DS pipeline yields an average holdout prediction accuracy of 0.68, while the standard TPOT without DS and tuned XGBoost models respectively produces an average holdout accuracy of 0.60 and 0.59 across all 100 model fits (Fig. 5). In summary, the optimal models from standard TPOT and XGBoost perform better in real-world data compared to simulated data but still worse than that of TPOT-DS. In both datasets, separate Welch two-sample one-sided *t*-tests show TPOT-DS optimal pipelines significantly outperform those of XGBoost and standard TPOT (all *p* values $< 10^{-15}$).

### Computational expense

For a dataset of the size simulated in our study (*m*=200 samples and *p* = 5000 attributes), TPOT-DS has a 65-minute runtime on a low performance computing machine with an Intel Xeon E5-2690 2.60GHz CPU, 28 cores and 256GB of RAM, whereas standard TPOT has a 18.5-hour runtime, approximately 17 times slower. On the same low performance computing machine (Intel Xeon E5-2690 2.60GHz CPU, 28 cores and 256GB RAM), each replication of TPOT-DS on the expression data takes on average 40 minutes, whereas standard TPOT takes 13.3 hours, approximately 20 times slower.

## Discussion

To our knowledge, TPOT-DS is the first AutoML tool to offer the option of feature selection at the group level. Previously, it was computationally expensive for any AutoML program to process biomedical big data. TPOT-DS is able to identify the most meaningful group of features to include in the prediction pipeline. We assessed TPOT-DS's out-of-sample prediction accuracy compared to

standard TPOT and XGBoost, another state-of-the-art machine learning method. We applied TPOT-DS to real-world expression data to demonstrate the identification of biologically relevant groups of genes.

Implemented with a strongly typed GP, Template provides more flexibility by allowing users to pre-specify a particular pipeline structure based on their knowledge, which speeds up AutoML process and provides potentially more interpretable results. For example, in high-dimensional data, dimensionality reduction or feature selection algorithms are preferably included at the beginning of the pipelines via Template to identify important features and, meanwhile, reduce computation time. For datasets with categorical features, preprocessing operators for encoding those features, like one-hot encoder, should be specified in the pipeline structure to improve pipelines' performance. Template was utilized in this study to specify the DS as the first step of the pipeline, which enables the comparison between the two TPOT implementations, with and without DS.

We simulated data of the similar scale and chalenging enough for the models to have similar predictive power as in the real-world RNA-Seq data. TPOT-DS correctly selects the first subset (containing the most important features) 75% of the time with high holdout accuracy (0.69). When another subset is chosen in the final pipeline, this method still produces holdout accuracy comparable to that of standard TPOT and XGBoost (0.565 - 0.575). For the RNASeq gene expression data, the best TPOT-DS pipeline selects DGM-5 and reports competitive holdout accuracy with standard TPOT (0.636 vs. 0.642) but with a smaller feature space (291 vs. 5,635 genes) and 20 times more computationally efficient. The best pipeline from TPOT-DS also produces comparable holdout accuracy with XGBoost (0.75 vs. 0.725).

Interestingly enough, TPOT-DS repeatedly selects DGM-5 to include in the final pipeline. In a previous study, we showed DGM-5 and DGM-17 enrichment scores were significantly associated with depression severity [20]. We also remarked that DGM-5 contains many genes that are biologically relevant or previously associated with mood disorders [20] and its enriched pathways such as apoptosis indicates a genetic signature of MDD pertaining shrinkage of brain region-specific volume due to cell loss [31,32].

TPOT-DS also selects DGM-13 as a potentially predictive group of features with smaller average holdout accuracy compared to DGM-5 ($0.563 < 0.636$). While many of the top genes do not have direct disease association in MalaCards, several have been linked to depression in animal studies such as PPP1R16A [33] and MXRA8 [34]. Further, the RGL4 gene, a Ral guanine nucleotide dissociation stimulator, was found to have a rare protein disruptive variant in at least one suicide patient among 60 other mutations [35]. The lack of previously found association of these genes with the phenotype is likely because MDD is a complex disorder of heterogeneous etiology [36]. Hence, the clinical diagnosis is the accumulative result of coordinated variation of many genes in the module, especially ones with high importance scores. Future studies to refine and characterize genes in DGM-13 as well as DGM-5 may deploy expression quantitative trait loci (e-QTL) or interaction QTL analysis to discover disease-associated variants [37].

Complexity-interpretability trade-off is an important topic to discuss in the context of AutoML. While arbitrarily-shaped pipelines may yield predictions competitive to human-level performance, these pipelines are often too complex to be interpretable. Vice versa, a simpler pipeline with defined steps of operators may be easier to interpret but yield suboptimal prediction accuracy. Finding the balance between pipeline complexity, model interpretation and generalization remains a challenging task for AutoML application in biomedical big data. With DS, each pipeline individual of a TPOT generation during optimization holds lower complexity due to the selected subset's lower dimension compared to that of the entire dataset. We hope that, with the complexity reduction from imposing a strongly-type GP template and DS, a small loss in dataset-specific predictive accuracy can be compensated by considerable increase in interpretability and generalizability. In this study, the resulting TPOT-DS pipelines are more interpretable with only two simple optimized operators after the DS: a transformer and a classifier. In the case of the expression analysis, these pipelines also highlight two small sets of interconnected genes that contain candidates for MDD and related disorders. Additionally, complexity reduction results in more efficient computation, which is strongly desirable in biomedical big data analysis.

A limitation of the DS analysis is the required pre-definition of subsets prior to executing TPOT-DS. While this characteristic of an intelligent system is desirable when *a prior* knowledge on the biomedical data is available, it might pose as a challenge when this knowledge is inadequate, such as when analyzing data of a brand-new disease. Nevertheless, one can perform a clustering method such as *k*-means to group features prior to performing TPOT-DS on the data. Another limitation of the current implementation of TPOT-DS is its restricted ability to select only one subset. A future design to support tree structures for Template will enable TPOT-DS to identify more than one subset that have high predictive power of the outcome. A new operator that combines the data subsets will prove useful in this design. Extensions of TPOT-DS will also involve overlapping subsets, which will require pipeline complexity reformulation beyond the total number of operators included in a pipeline. Specifically, in the case of overlapping subsets, the number of features in the selected subset(s) is expected to be an element of the complexity calculation. Extension of TPOT-DS to GWAS is straightforward. However, because of the low predictive power of variants in current GWAS, alternative metrics beside accuracy, balanced accuracy or area under the receiving operator characteristic curve will need to be designed and included in the fitness function of TPOT's evolutionary algorithm.

In this study, we developed two new operators for TPOT, Dataset Selector and Template, to enhance its performance on high-dimensional data by simplifying the pipeline structure and reducing the computational expense. Dataset Selector helps users leverage domain knowledge to narrow down important features for further interpretation, and Template largely increases flexibility of TPOT via customizing pipeline structure. Future extension and integration of these two operators have the potential to enrich the application of AutoML on different real world biomedical problems.

# References


1. **PMLB: a large benchmark suite for machine learning evaluation and comparison**
Randal S. Olson, William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, Jason H. Moore
*BioData Mining* (2017-12) https://doi.org/gfrbw5
DOI: 10.1186/s13040-017-0154-4 · PMID: 29238404 · PMCID: PMC5725843

2. **Data-driven advice for applying machine learning to bioinformatics problems.**
Randal S Olson, William La Cava, Zairah Mustahsan, Akshay Varik, Jason H Moore
*Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2018) https://www.ncbi.nlm.nih.gov/pubmed/29218881
PMID: 29218881 · PMCID: PMC5890912

3. https://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning

4. **Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms**
Chris Thornton, Frank Hutter, Holger H. Hoos, Kevin Leyton-Brown
*arXiv* (2012-08-18) https://arxiv.org/abs/1208.3719v2

5. http://www.jmlr.org/papers/v18/16-261.html

6. **RECIPE: A Grammar-Based Framework for Automatically Evolving Classification Pipelines**
Alex G. C. de Sá, Walter José G. S. Pinto, Luiz Otavio V. B. Oliveira, Gisele L. Pappa
*Lecture Notes in Computer Science* (2017) https://doi.org/gfjzg2
DOI: 10.1007/978-3-319-55696-3_16

7. **Autostacker: A Compositional Evolutionary Learning System**
Boyuan Chen, Harvey Wu, Warren Mo, Ishanu Chattopadhyay, Hod Lipson
*arXiv* (2018-03-02) https://arxiv.org/abs/1803.00684v1

8. https://github.com/joeddav/devol

9. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

10. https://github.com/reiinakano/xcessiv

11. **Genetic Programming: An Introduction: on the Automatic Evolution of Computer Programs and Its Applications**
Wolfgang Banzhaf, Frank D. Francone, Robert E. Keller, Nordin Peter
*ACM Digital Library* (1998) https://dl.acm.org/citation.cfm?id=280485

12. **Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science**
Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, Jason H. Moore
*Proceedings of the 2016 on Genetic and Evolutionary Computation Conference - GECCO '16* (2016) https://doi.org/gfgqv2
DOI: 10.1145/2908812.2908918

13. **Identifying and Harnessing the Building Blocks of Machine Learning Pipelines for Sensible Initialization of a Data Science Automation Tool**
Randal S. Olson, Jason H. Moore
*arXiv* (2016-07-29) https://arxiv.org/abs/1607.08878v1

14. **Toward the automated analysis of complex diseases in genome-wide association studies using genetic programming**
Andrew Sohn, Randal S. Olson, Jason H. Moore
*Proceedings of the Genetic and Evolutionary Computation Conference on - GECCO '17* (2017) https://doi.org/gfgqv3
DOI: 10.1145/3071178.3071212

15. **Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients**
Trang T. Le, Nigel O. Blackwood, Jaclyn N. Taroni, Weixuan Fu, Matthew K. Breitenstein
*arXiv* (2018-03-08) https://arxiv.org/abs/1803.04487v1

16. https://hal.archives-ouvertes.fr/hal-00650905/

17. **A fast and elitist multiobjective genetic algorithm: NSGA-II**
K. Deb, A. Pratap, S. Agarwal, T. Meyarivan
*IEEE Transactions on Evolutionary Computation* (2002-04) https://doi.org/bnw2vv
DOI: 10.1109/4235.996017

18. http://www.jmlr.org/papers/v13/fortin12a.html

19. **Strongly Typed Genetic Programming**
David J. Montana
*Evolutionary Computation* (1995-06) https://doi.org/ct3mnb
DOI: 10.1162/evco.1995.3.2.199

20. **Identification and replication of RNA-Seq gene network modules associated with depression severity**
Trang T. Le, Jonathan Savitz, Hideo Suzuki, Masaya Misaki, T. Kent Teague, Bill C. White, Julie H. Marino, Graham Wiley, Patrick M. Gaffney, Wayne C. Drevets, … Jerzy Bodurka

*Translational Psychiatry* (2018-09-05) https://doi.org/gd7jx7
DOI: 10.1038/s41398-018-0234-3 · PMID: 30185774 · PMCID: PMC6125582

21. **Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure**
Caleb A Lareau, Bill C White, Ann L Oberg, Brett A McKinney
*BioData Mining* (2015-02-03) https://doi.org/gb5fpr
DOI: 10.1186/s13040-015-0040-x · PMID: 25685197 · PMCID: PMC4326454

22. **XGBoost**
Tianqi Chen, Carlos Guestrin
*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016) https://doi.org/gdp84q
DOI: 10.1145/2939672.2939785

23. https://www.kaggle.com/

24. **Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation**
Huiting Zheng, Jiabin Yuan, Long Chen
*Energies* (2017-08-08) https://doi.org/gbtqbr
DOI: 10.3390/en10081168

25. **A Novel Image Classification Method with CNN-XGBoost Model**
Xudie Ren, Haonan Guo, Shenghong Li, Shilin Wang, Jianhua Li
*Digital Forensics and Watermarking* (2017) https://doi.org/gfgvf3
DOI: 10.1007/978-3-319-64185-0_28

26. https://greenelab.github.io/meta-review/

27. **Variations in FKBP5 and BDNF genes are suggestively associated with depression in a Swedish population-based cohort**
Catharina Lavebratt, Elin Åberg, Louise K. Sjöholm, Yvonne Forsell
*Journal of Affective Disorders* (2010-09) https://doi.org/b726hn
DOI: 10.1016/j.jad.2010.02.113 · PMID: 20226536

28. **Modulation of glucocorticoid receptor nuclear translocation in neurons by immunophilins FKBP51 and FKBP52: Implications for major depressive disorder**
Erick T. Tatro, Ian P. Everall, Marcus Kaul, Cristian L. Achim
*Brain Research* (2009-08) https://doi.org/b9cn9m
DOI: 10.1016/j.brainres.2009.06.036 · PMID: 19545546 · PMCID: PMC2724600

29. **Polymorphisms in FKBP5 are associated with increased recurrence of depressive episodes and rapid response to antidepressant treatment**

Elisabeth B Binder, Daria Salyakina, Peter Lichtner, Gabriele M Wochnik, Marcus Ising, Benno Pütz, Sergi Papiol, Shaun Seaman, Susanne Lucae, Martin A Kohli, … Bertram Muller-Myhsok
*Nature Genetics* (2004-11-21) https://doi.org/bh28xx
DOI: 10.1038/ng1479 · PMID: 15565110

30. **Type I interferon signaling genes in recurrent major depression: increased expression detected by whole-blood RNA sequencing**
S Mostafavi, A Battle, X Zhu, JB Potash, MM Weissman, J Shi, K Beckman, C Haudenschild, C McCormick, R Mei, … DF Levinson
*Molecular Psychiatry* (2013-12-03) https://doi.org/f6qdpr
DOI: 10.1038/mp.2013.161 · PMID: 24296977 · PMCID: PMC5404932

31. **A meta-analysis examining clinical predictors of hippocampal volume in patients with major depressive disorder.**
Margaret C McKinnon, Kaan Yucel, Anthony Nazarov, Glenda M MacQueen
*Journal of psychiatry & neuroscience : JPN* (2009-01) https://www.ncbi.nlm.nih.gov/pubmed/19125212
PMID: 19125212 · PMCID: PMC2612082

32. **Increased apoptosis in patients with major depression: A preliminary study.**
E Eilat, S Mendlovic, A Doron, V Zakuth, Z Spirer
*Journal of immunology (Baltimore, Md. : 1950)* (1999-07-01) https://www.ncbi.nlm.nih.gov/pubmed/10384158
PMID: 10384158

33. **A Molecular Signature of Depression in the Amygdala**
Etienne Sibille Ph.D., Yingjie Wang M.S., Jennifer Joeyen-Waldorf B.S., Chris Gaiteri B.S., Alexandre Surget Ph.D., Sunghee Oh B.S., Catherine Belzung Ph.D., George C. Tseng Ph.D., David A. Lewis M.D.
*American Journal of Psychiatry* (2009-09) https://doi.org/fmbrrk
DOI: 10.1176/appi.ajp.2009.08121760 · PMID: 19605536 · PMCID: PMC2882057

34. **Effect of Acute Stressor and Serotonin Transporter Genotype on Amygdala First Wave Transcriptome in Mice**
Christa Hohoff, Ali Gorji, Sylvia Kaiser, Edith Willscher, Eberhard Korsching, Oliver Ambrée, Volker Arolt, Klaus-Peter Lesch, Norbert Sachser, Jürgen Deckert, Lars Lewejohann
*PLoS ONE* (2013-03-11) https://doi.org/gfkt9h
DOI: 10.1371/journal.pone.0058880 · PMID: 23536833 · PMCID: PMC3594195

35. **High-Coverage Whole-Exome Sequencing Identifies Candidate Genes for Suicide in Victims with Major Depressive Disorder**
Dóra Tombácz, Zoltán Maróti, Tibor Kalmár, Zsolt Csabai, Zsolt Balázs, Shinichi Takahashi, Miklós Palkovits, Michael Snyder, Zsolt Boldogkői

*Scientific Reports* (2017-08-02) https://doi.org/gbrpc3
DOI: 10.1038/s41598-017-06522-3 · PMID: 28769055 · PMCID: PMC5541090

36. **Genetic Studies of Major Depressive Disorder: Why Are There No Genome-wide Association Study Findings and What Can We Do About It?**
Douglas F. Levinson, Sara Mostafavi, Yuri Milaneschi, Margarita Rivera, Stephan Ripke, Naomi R. Wray, Patrick F. Sullivan
*Biological Psychiatry* (2014-10) https://doi.org/gd8sv4
DOI: 10.1016/j.biopsych.2014.07.029 · PMID: 25201436 · PMCID: PMC4740915

37. **An interaction quantitative trait loci tool implicates epistatic functional variants in an apoptosis pathway in smallpox vaccine eQTL data**
CA Lareau, BC White, AL Oberg, RB Kennedy, GA Poland, BA McKinney
*Genes & Immunity* (2016-04-07) https://doi.org/f9mvjk
DOI: 10.1038/gene.2016.15 · PMID: 27052692