An improved catalogue of putative synaptic genes defined exclusively by temporal transcription profiles through an ensemble machine learning approach

Flavio Pazos Obregón (fpazos@iibce.edu.uy)
Instituto de Investigaciones Biologicas Clemente Estable https://orcid.org/0000-0003-4980-1071
Pablo Soto
Instituto de Investigaciones Biologicas Clemente Estable
Martín Palazzo
Instituto de Investigaciones Biomédicas de Buenos Aires
Gustavo Guerberoff
Universidad de la Republica Facultad de Ingenieria
Patricio Yankilevich
Instituto de Investigaciones Biomédicas de Buenos Aires
Rafael Cantera
Instituto de Investigaciones Biologicas Clemente Estable

Research article

Keywords: Synaptic genes, Machine learning, Temporal transcription profiles, Gene function prediction, Drosophila melanogaster

Posted Date: August 30th, 2019

DOI: https://doi.org/10.21203/rs.2.13746/v1

Version of Record: A version of this preprint was published on December 23rd, 2019. See the published version at https://doi.org/10.1186/s12864-019-6380-z.

Abstract

Background . Assembly and function of neuronal synapses require the coordinated expression of a yet undetermined set of genes. Previously, we had trained an ensemble machine learning model to assign a probability of having synaptic function to every protein-coding gene in Drosophila melanogaster. This approach resulted in the publication of a catalogue of 893 genes which we postulated to be very enriched in genes with a still undocumented synaptic function. Since then, the scientific community has experimentally identified 79 new synaptic genes. Here we use these new empirical data to evaluate our original prediction. We also implement a series of changes to the training scheme of our model and using the new data we demonstrate that this improves its predictive power. Finally, we added the new synaptic genes to the training set and trained a new model, obtaining a new, enhanced catalogue of putative synaptic genes. Results . The retrospective analysis demonstrate that our original catalogue was significantly enriched in new synaptic genes. When the changes to the training scheme were implemented using the original training set we obtained even higher enrichment. Finally, applying the new training scheme with a training set including the 79 new synaptic genes, resulted in an enhanced catalogue of putative synaptic genes. Here we present this new catalogue and announce that a regularly updated version will be available online at: http://synapticgenes.bnd.edu.uy Conclusions . We show that training an ensemble of machine learning classifiers solely with the whole-body temporal transcription profiles of known synaptic genes resulted in a catalogue with a significant enrichment in undiscovered synaptic genes. Using new empirical data provided by the scientific community, we validated our original approach, improved our model an obtained an arguably more precise prediction. This approach reduces the number of genes to be tested through hypothesis-driven experimentation and will facilitate our understanding of neuronal function. Availability : http://synapticgenes.bnd.edu.uy

Background

The synapse, a specialized contact between neurons, is currently of fundamental importance for our understanding of learning, memory and other brain functions. Assembly and function of neuronal synapses require the coordinated expression of a yet undetermined set of genes, which for simplicity will be called here “synaptic genes”. There is a broad consensus that only a fraction of the total number of synaptic genes have been identified(1, 2). Due to the evolutionary conservation among synaptic genes, the knowledge obtained from studies in model organisms is very relevant for other species, including humans (2, 3).

Since the biological roles of the vast majority of known amino acid sequences remain partly or completely unknown (4), computational prediction of gene function is an open research problem of much relevance. In recent years diverse methodologies have been assayed, with a strong prevalence of machine learning approaches. The top-performing algorithms, architectures and training schemes are function-specific and context-dependent (5). In a previous study (6), we implemented an ensemble machine learning model that assigned a probability of being a “synaptic gene” to each protein-coding gene of *Drosophila melanogaster.* The features to infer the synaptic function were the whole-body transcription

levels of all protein-coding genes at 24 developmental stages, published by the modENCODE project (7). As far as we know, this is the only study that predicts gene function relying exclusively on temporal transcriptions profiles obtained through NGS technologies. After an exhaustive bibliographic review, a set of genes for which a function in synapse formation and/or maturation, and/or neurotransmission, and/or plasticity and/or maintenance had been experimentally demonstrated was selected as a positive example and included in the training set. Genes fulfilling any of two biological criteria defined *ad hoc* were selected as negatives examples (6). (Additional file 1). Our model intercepted the results of three learning algorithms: k-nearest neighbors (kNN) (8), Ranfom Forest (RF) (9), Support Vector Machines (SVM) (10). The classification threshold of the algorithms was set to meet the expected number of unknown synaptic genes (estimated *a priori)* at that time. We obtained a catalogue that we postulated to be highly enriched in genes for which a synaptic function was yet to be discovered.

Following the publication of that catalogue, scientists around the world have experimentally identified 79 new synaptic genes (NSG), giving us the opportunity to empirically evaluate the predictive power of the catalogue. Thereafter we tested a new training scheme and evaluated it by measuring the enrichment in NSG of the resulting catalogues. The tested training scheme, inspired by the principle of bagging (11), is meant to alleviate a probable bias of our model due to a relatively small training set. We found that the new training scheme resulted in a model producing a catalogue more enriched in NSG. Finally, we added the 79 NSG to the training set and trained a new model with the new training scheme. With this new model we obtained the new, enhanced catalogue of putative synaptic genes that we are publishing here. The whole procedure is schematized in *Figure 1.* A monthly updated version of this catalogue will be available online at: http://synapticgenes.bnd.edu.uy.

Results

Evaluation of the original catalogue

Since the publication of our original catalogue up to the preparation of this manuscript, we identified 79 *Drosophila* genes that had gathered enough experimental evidence to be considered a synaptic gene according to our previous criteria (6). Additional file 2lists these genes along with the references supporting their synaptic functions. Roughly a third of these NSG (28) were present in our original catalogue, giving an enrichment in NSG of 4.38 with a p-value < 10^{-10} *(Table 1).*

Table 1—Evaluation of our original prediction.

Improved training scheme

The changes to the training scheme of our model tested here are detailed in the Methods section and schematized in *Figure 2.*

To test if these changes improved the predictive performance of our approach, we trained a model implementing these changes with the original training set and then compared the results with those of the original model. As shown in *Table 1* and *Figure 3,* the changes resulted in a better predictive power measured as enrichment in NSG. This improvement is also observed when we considered each classification algorithm separately (Figure 3A-C). The performance of the intersection of the classifiers trained with the sub-samples of the training set was always better than that of the performance of the classifier trained with the full training set. By intersection of the classifiers for a given threshold we mean the set of genes that were assigned with a probability above the threshold by the 3 classifiers simultaneously.

Evaluation of the new classifiers

After demonstrating that the proposed changes to the training scheme and ensemble rules would have resulted in a series of catalogues more enriched in NSG, we incorporated the 79 NSG to the training set and repeated the whole procedure described above, obtaining 15 new classifiers. Each of these classifiers was evaluated with an independent test set, that was used to calculate the accuracy, the F1 score and the area under the ROC curve (*Fig.2* and *Table 2).* The obtained values were compared with those reported by other colleagues when training models to predict other biological functions (12–14).

Table 2. Evaluation of the 15 classifiers.

A new catalogue of putative synaptic genes

We trained a new model incorporating the 79 NSG to the training set and the changes to the training scheme. In our original work only those genes assigned with a probability of being synaptic of at least 0.9 by the three classifiers were included in the final catalogue. This high classification threshold was set to obtain a catalogue of a given size. Now the threshold was set at 0.95 because we aimed to obtain a smaller catalogue since there are fewer unknown synaptic genes. The resulting catalogue had 601 genes.

Enrichment of the new catalogue in synapse-related GO terms

To evaluate the quality of the new catalogue, we determined its enrichment in synapse-related GO terms. This could be done because we constructed our training set without taking into account Gene Ontology. We found that 83 of the 601 genes to which our 15 classifiers assigned a probability above 0.95 had some synapse-related GO annotation. To determine whether this is a significant enrichment, all the genes in our training set that have some synapse-related GO annotation must be removed from the background set. This analysis was performed with Gorilla (15) and the results are shown in *Table 3.* After excluding from the catalogue these 83 genes a final catalogue of 518 putative synaptic genes was obtained (Additional file 3).

Table 3. Enrichment of the new catalogue in synapse-related GO terms.

Regularly updated on-line catalogue

The model we are presenting here will be re-trained as new synaptic genes are identified. This will result in an updated catalogue that will be available here: http://synapticgenes.bnd.edu.uy. The updated list of synaptic genes used to train the model will be available at the same site.

Discussion

A catalogue obtained by training an ensemble machine-learning model that assigned each *Drosophila* protein-coding gene a probability of having synaptic function was published four years ago (6). Of note, the model was based exclusively on a whole-body temporal transcriptome. It was hypothesized that the catalogue was enriched in genes of relevance for neuronal synapses which were still not recognized as such. Since the publication of the catalogue, 79 NSG were experimentally identified by others with a variety of experimental methods. This offered a great opportunity to test both the predictive power of our machine learning approach and some changes to the training scheme that could improve the predictive power of our model.

The enrichment of the previous catalogue in genes for which a synaptic function was experimentally identified by other colleagues between 2015 and 2019 represents a good experimental validation of the predictive power of our machine learning approach. It is thus possible to predict gene function using machine learning based exclusively on temporal transcription data.

On the other hand, our original model assigned very low probabilities to some genes that were later proven to have synaptic functions. A possible explanation could be that our model cannot capture the entire diversity in expression profiles among the hundreds of genes required for assembly and function of neuronal synapses. The coordinated expression of hundreds of genes, during the two massive waves of synapse formation taking place during *Drosophila* development (6) probably includes genes that encode reppressors, which will result in very different transcription profiles. It is quite probable that any model exclusively trained with transcription profiles will fail to recognize genes with specular transcription profiles.

It is also worth noting that none of the 79 NGS belongs to the list of “non-synaptic genes” which had been used to train the algorithms. This provides unequivocal validation for the biological criteria used to select the negative examples of the training set.

Note that even when the original model and its improved version were based on the same algorithms and were trained with the same set of genes, the enrichment in NSG found in the catalogue obtained with the improved model was 38% higher. This is interpreted as a clear demonstration that the new training scheme really improved the predictive performance of our approach. A possible explanation is that the

tested training scheme alleviated a probable bias of our model due to a relatively small training set and increased its generalization capacity. Since there are hundreds of synaptic genes to be discovered this is an important feature (16).

Conclusions

We show here that a catalogue of *Drosophila* putative synaptic genes obtained by an ensemble machine learning model four years ago has a significant enrichment in genes whose synaptic function was discovered after the publication of the catalogue. We implemented a number of changes to the training scheme and thereafter trained this new model with the original training set. This generated a catalogue that was even more enriched in new synaptic genes than the original one, indicating that the improved model had better performance. Finally, we included all the new synaptic genes in the training set and trained a model that incorporates the proposed changes, obtaining a new, enhanced catalogue of putative synaptic genes.

We are making this catalogue available to the scientific community, firmly believing that this will facilitate the identification of genes important for the assembly and function of synapses, by means of gene silencing, mutant analysis, electrophysiology, neuroanatomy, behavioral assays and other traditional protocols, all of which will most likely lead to a better understanding of the function of the brain. The catalogue is available at: http://synapticgenes.bnd.edu.uy

Methods

Data

We used the developmental transcriptome of *Drosophila melanogaster* published by the MODENCODE Project(7). In this data, each sample consisted of total polyAAA-RNA isolated from 30 whole bodies obtained along the organism life cycle. The data set consist of the transcript levels of 15,398 genes expressed as fragments per kilo base of exon per million fragments mapped (FPKM). We excluded 1,756 genes that show transcript levels above zero only during adult life and normalized each gene's temporal series dividing it by its maximum value, thus obtaining for each gene a series of 24 values oscillating between 0 and 1. More details in (6).

Evaluation of the predictive power of our original model

Using the same *ad hoc* definition for "synaptic gene" that was adopted in our previous work, we performed a bibliographic revision and identified 79 new synaptic genes. Then we analyzed the enrichment of our original catalogue in these NSG using home-made scripts assuming an hypergeometric distribution.

A new training scheme

To obtain our original catalogue we had trained three learning algorithms (k-NN, RF and SVM) with an unbalanced training set, in which there were many more negative than positive examples. After a careful bibliographic revision, aimed to construct a list of genes whose importance for neuronal synapses had gathered strong experimental evidence, 92 positive examples were selected (*Additional file 1).*. As negative examples, 397 genes were selected based on two biological criteria: genes with no expression during developmental stages (when massive synaptogenesis is known to take place), or genes with no expression during adult life in one of the two sexes.

With the aim of improving the predictive power of our model, here we sub-sampled the original training set to construct five smaller, slightly different training sets (*See Fig. 1).*. To construct each of these training sets we picked out four fifths of the original positive examples and four fifths of the original negative examples. We repeated this procedure five times, leaving out a different fifth each time. Using the positive and negative examples left out when constructing each training set we defined a test set, used to independently calculate the accuracy, the AUC of the ROC and the F1 score of each classifier.

The hyper parameters of each classifier were set by grid search combined with 10-fold cross validation and its performance was evaluated by an independent test set. Each classifier assigned a different probability of being synaptic to each gene. To obtain our catalogues we considered, for each classification threshold, the intersection of the 15 results and then we took the mean probability assigned to each gene in the intersection.

All calculations were performed using Jupyter Notebooks and Sklearn (Pedregosa et al., 2011).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of Data and Materials

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests

Funding

FPO received financial assistance from PEDECIBA (Uruguay) and received funds from the Agencia Nacional de Investigación e Innovación (ANII, Uruguay). RC and GG received funds from the Sistema Nacional de Investigadores (Uruguay). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

FPO conceived the study, performed the experiments and data analysis and wrote the manuscript. PS discussed the experiments prepared the web site and corrected the manuscript. MP discussed the experiments and corrected the manuscript. GG supervised the project and corrected the manuscript. PY supervised the project and corrected the manuscript. RC conceived the study, supervised the project and corrected the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grant FSDA_1_2017_1_142427 from Agencia Nacional de Investigación e Innovación. FPO is the recipient of a PhD fellowship from Agencia Nacional de Investigación e Innovación and GG and RC were supported by Sistema Nacional de Investigadores (Uruguay).

List of Additional files

Additional file 1 - Original training set

Additional file 2 - New synaptic genes & references

Additional file 3—New catalogue of putative synaptic genes

References

1. Frank CA, Wang X, Collins CA, Rodal AA, Yuan Q, Verstreken P, et al. New approaches for studying synaptic development, function, and plasticity using Drosophila as a model system. J Neurosci Off J Soc Neurosci. 2013 Nov 6;33(45):17560–8.
2. Laßek M, Weingarten J, Volknandt W. The synaptic proteome. Cell Tissue Res. 2015 Jan 1;359(1):255–65.
3. Burkhardt P. The origin and evolution of synaptic proteins—choanoflagellates lead the way. J Exp Biol. 2015 Feb 15;218(4):506.

4. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018 Mar 16;46(5):2699.
5. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016 Sep 7;17:184.
6. Pazos Obregón F, Papalardo C, Castro S, Guerberoff G, Cantera R. Putative synaptic genes defined from a Drosophila whole body developmental transcriptome by a machine learning approach. BMC Genomics. 2015 Sep 15;16:694.
7. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, et al. The developmental transcriptome of Drosophila melanogaster. Nature. 2011 Mar 24;471(7339):473–9.
8. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am Stat. 1992;46(3):175–85.
9. Breiman L. Random Forests. Mach Learn. 2001 Oct 1;45(1):5–32.
10. Vapnik V. Statistical learning theory [Internet]. Wiley; 1998. Available from: http://www.amazon.ca/exec/obidos/redirect?tag = citeulike09–20&path = ASIN/0471030031
11. Dietterich TG. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. Springer Berlin Heidelberg; 2000. p. 1–15. (Lecture Notes in Computer Science).
12. Kacsoh BZ, Greene CS, Bosco G. Machine Learning Analysis Identifies Drosophila Grunge/Atrophin as an Important Learning and Memory Gene Required for Memory Retention and Social Learning. G3 GenesGenomesGenetics. 2017 Sep 9;7(11):3705–18.
13. Kerepesi C, Daróczy B, Sturm Á, Vellai T, Benczúr A. Prediction and characterization of human ageing-related proteins by using machine learning. Sci Rep. 2018 Mar 6;8(1):4094.
14. Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, et al. Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci. 2019 Feb 5;116(6):2344–53.
15. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10.
16. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning data mining, inference, and prediction. New York: Springer; 2009.

Tables

Due to technical limitations, the Table(s) are only available as a download in the supplemental files section.

Figures

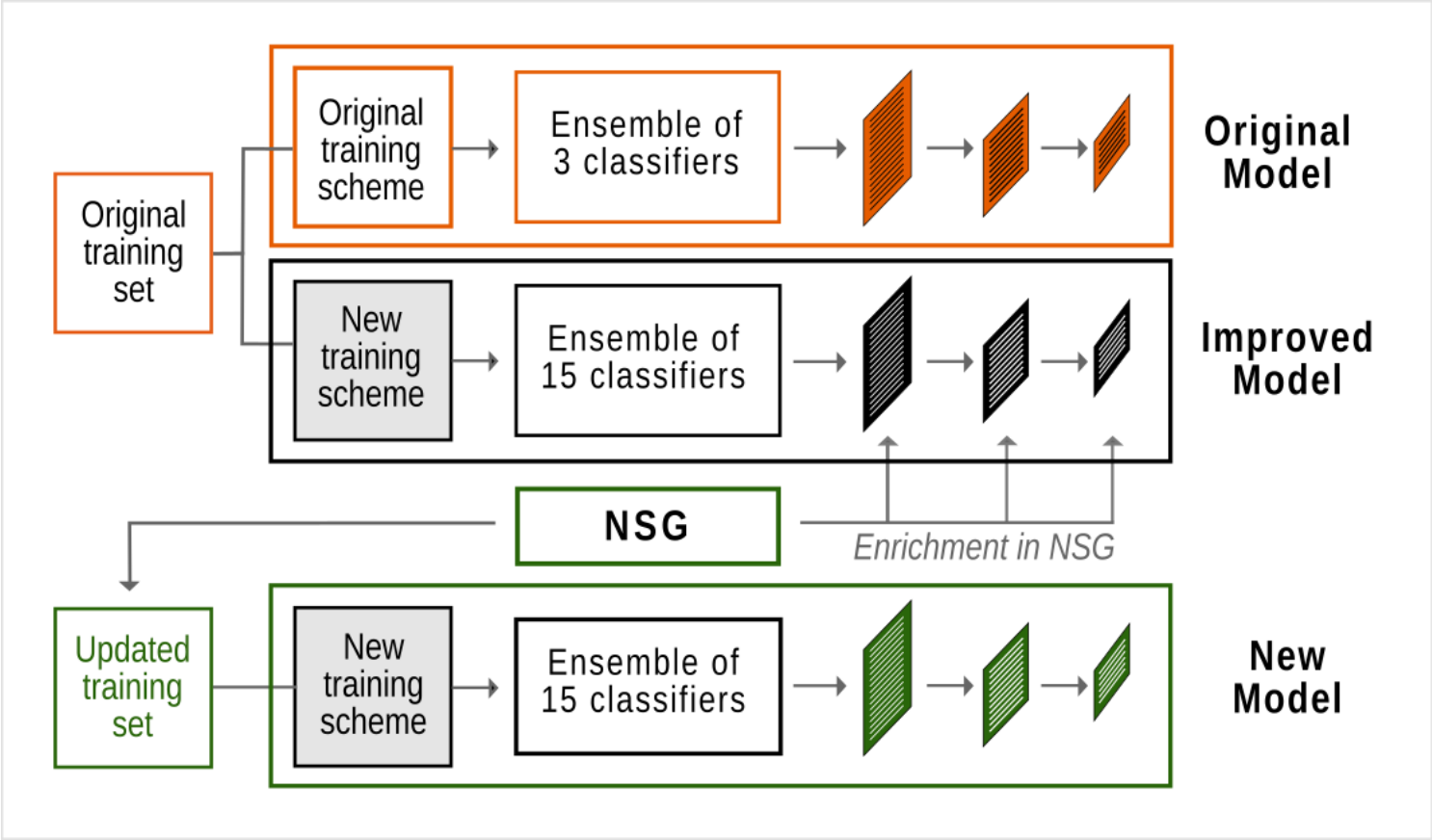

Figure 1

Scheme of the work-flow to obtain the new catalogue of putative synaptic genes. First we trained two models with the original training set, one with the original training scheme and one with the new training scheme (Figure 2). Then we compared the enrichment in new synaptic genes (NSG) of the catalogues resulting from each method. Once we tested that the new training scheme improves the prediction (Figure 3), we incorporated the 79 NSG to the training set, trained a new model with the new training scheme and obtained a new catalogue of putative synaptic genes.

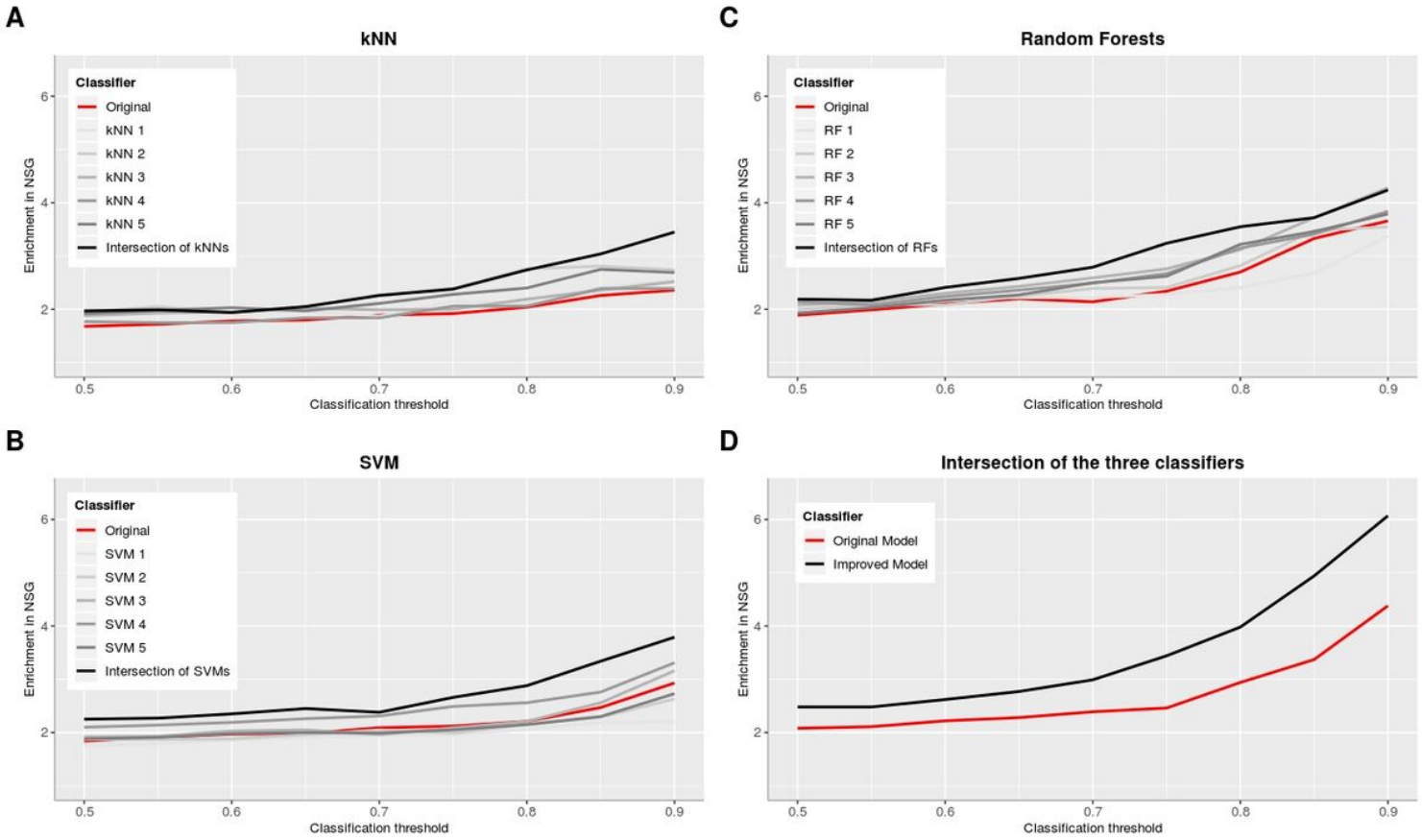

Figure 2

Comparison between the original and the improved models. Original model. With the whole training set we trained three algorithms: SVM, Random Forest and kNN. The hyper parameters of each classifier were set by exhaustive grid search combined with 10-fold cross validation over the training set. Finally, we increased the classification threshold of the classifiers and considered the intersection between the resulting catalogues. Improved model. First we sub-sampled five times the original training set, leaving out each time a different fifth of the positive and negative examples. By this procedure we obtained five smaller, slightly different training sets. Using the positive and negative examples left out in each iteration, we created five test sets, used to independently evaluate each classifier. With each training set we trained three algorithms: SVM, Random Forest and kNN, thus obtaining 15 different classifiers. The hyper parameters of each classifier were set by exhaustive grid search combined with 10-fold cross validation over the training set. After training, we evaluated each classifier (accuracy, ROC and F1) using a different test set. Finally, we increased the classification threshold of the classifiers and considered the intersection between the resulting catalogues.

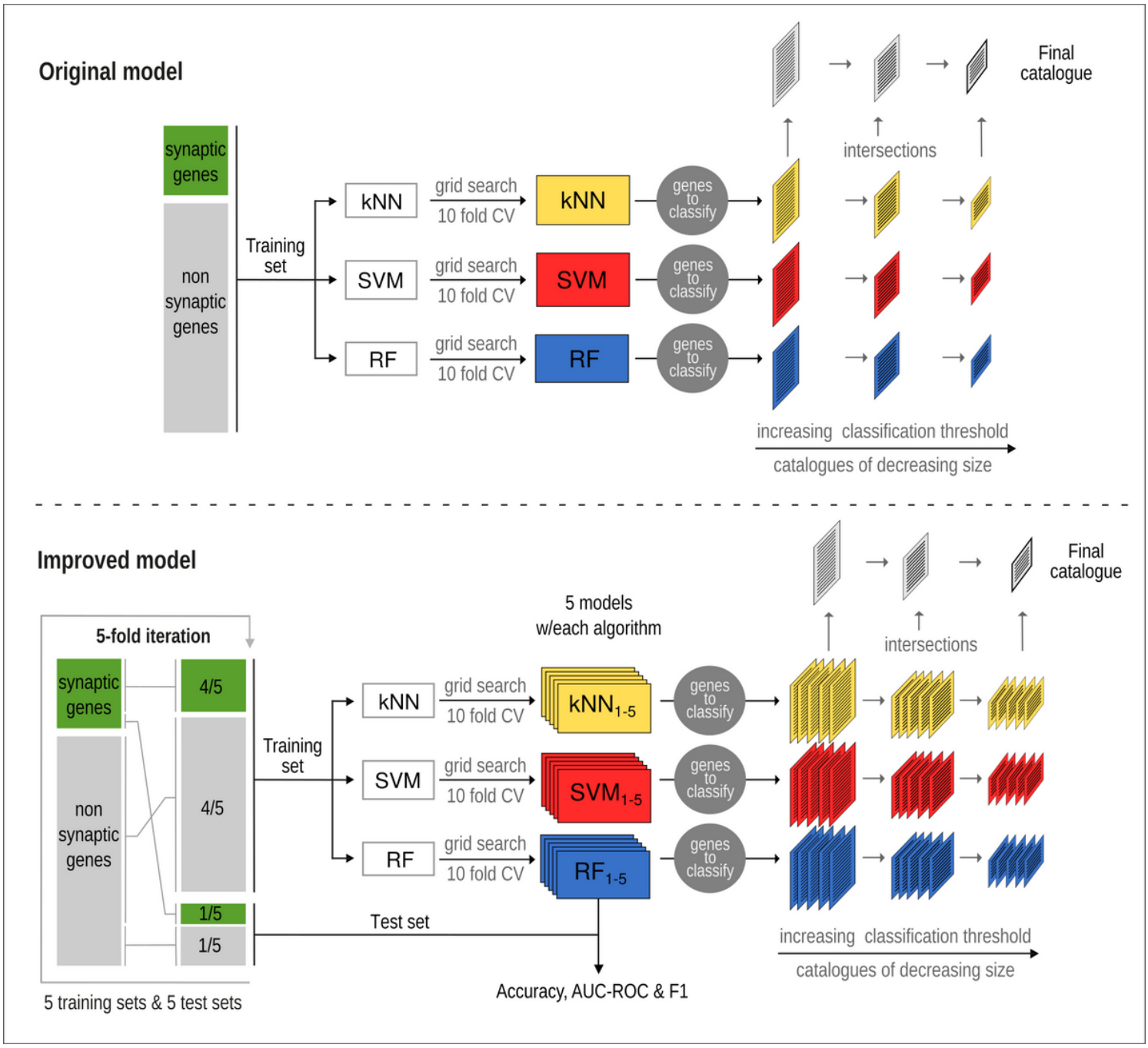

Figure 3

Comparison between the original and the improved model. Enrichment in NSG for an increasing classification threshold. A: kNN, B: SVM, C: Random Forest, D: Intersection of the three algorithms. In each panel, the red line represents the results of the original model, the gray lines represent the results of the models obtained after training the corresponding algorithm with each sub-sample of the original training set and the black line shows the result.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.png
- supplement1.png
- supplement2.png
- supplement4.xls
- supplement5.xls
- supplement6.xls