Data integration through heterogeneous ensembles for protein function prediction

Linhua Wang[1#], Jeffrey Law[2], T. M. Murali[3] and Gaurav Pandey[1*]

1 Department of Genetics and Genomic Sciences and Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

2 Genetics, Bioinformatics, and Computational Biology Ph.D. Program, Virginia Tech, Blacksburg, VA, USA

3 Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

* Corresponding author: gaurav.pandey@mssm.edu

Current address: Baylor College of Medicine, Houston, TX, USA

Introduction

An exciting aspect of the data revolution in biomedical sciences has been the collection of multiple data modalities reflecting diverse aspects of entities of interest. For example, multiple aspects of proteins, such as amino acid sequences, three-dimensional structures, and protein-protein interactions, are used to study their function[1-3] and other properties. Several data integration strategies have been developed to leverage the inherent diversity and complementarity of these modalities to build comprehensive predictive model(s) for properties of biomedical entities[4-6]. However, most of these strategies focus on building *uniform* integrated representations[7], such as networks[2, 8, 9], kernels[10] and tensors[11], an approach sometimes referred to as *early or intermediate integration*[5, 12]. While such uniform integration is expected to reinforce the consensus among the modalities, this approach may not be ideal when different classes of prediction methods are effective for individual data types, such as deep learning[13] for medical images[14] and clinical text[15], XGBoost[16] for structured data[17, 18] and label propagation for networks[19]. We refer to dataset- or data type-specific models inferred using these types of methods as *local models.*

To better utilize the properties of *local models*, we propose a potentially more effective approach to data integration and predictive modeling by assimilating these models into *heterogenous ensembles*[20]. These ensembles can incorporate a large number and variety of base predictors, including local models, and can benefit from both the consensus and diversity among these predictors. Due to these properties, heterogeneous ensembles have been effective at improving predictive performance for protein function prediction (PFP)[20-22], DREAM Challenges[23], and other applications[24-26].

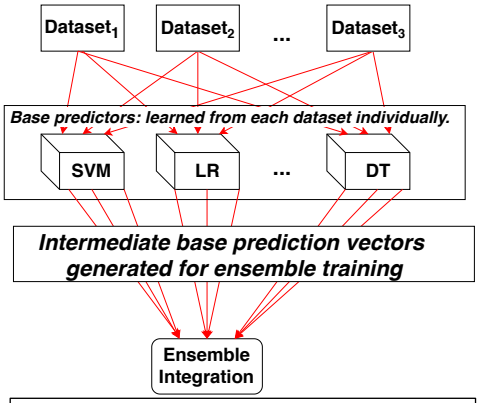

Figure 1: Overview of the Ensemble Integration (EI) approach. One *base predictor*/*local model* each is trained on the individual data sets using ten standard classification algorithms, such as Support Vector Machine (SVM), Logistic Regression (LR) and Decision Tree (DT). The predictions generated by these *base predictors* are then *integrated* using heterogeneous ensemble methods, such as Ensemble Selection developed by Caruana et al (CES) and Stacking using eight classifiers, as in our previous work[22].

Here, we present a novel approach named Ensemble-based data Integration (EI) for predictive modeling. EI leverages established heterogeneous ensemble methods, specifically stacking[27] and ensemble selection (ES)[28], which are generally applied to base predictors derived from the same data set[20]. In the data integration scenario, EI uses these methods to aggregate local models inferred from different data sets to develop an ensemble predictor for the target problem (**Figure 1**). Note that this approach has been referred to as *late integration* in the literature, although we deploy it more systematically and at a larger scale than in previous work[4-6]. We tested EI's effectiveness for the challenging problem of protein function prediction (PFP)[1-3] using diverse networks from the STRING database[29], and compared its performance with data integration algorithms developed for networks[30, 31].

Materials and Methods

Ensemble Integration. Figure 1 shows an overview of the ensemble integration (EI) approach. First, we trained one local model each using ten Weka[32]-based standard binary classification algorithms, such as Support Vector Machine (SVM), Logistic Regression (LR) and Decision Tree (DT), from base predictor training subsets of each of the individual datasets being integrated. We used random under-sampling[33] to balance the positive and negative classes before training these base predictors. Next, to construct the integrated ensemble, we executed a second set of algorithms, namely ES developed by Caruana *et al*[34, 35] and Stacking[27], on the predictions generated by these local models on separate ensemble training subsets of the individual datasets. We used eight standard classification algorithms for Stacking, as in previous work[22]. This whole procedure was conducted within a five-fold nested cross-validation setup typically used for heterogeneous ensemble learning (described in detail in ref. 20), which reduces adverse effects of overfitting while ensuring fair evaluation of the ensemble methods being tested. The source code implementing this EI framework is publicly available[36].

Baseline data/predictor integration methods. We compared EI's performance with those of the Mashup[30] and deepNF[31] early network integration algorithms. Specifically, we tested the same ten types of classifiers on the Mashup- and deepNF-integrated versions of the individual networks. This baseline's design enabled a fair comparison with EI, since each approach represent a single level of integration, i.e., early versus late, respectively.

As an alternate baseline, we also tested the integration of base predictors inferred from individual datasets into dataset-specific heterogeneous ensembles[22]. This baseline enabled us to compare the above data integration strategies with the individual datasets being integrated.

In both the above baselines, as in EI, random under-sampling[33] was applied to balance the positive and negative classes before training the base/individual predictors.

Experimental data. We tested all the above integration approaches for PFP, specifically GO term annotation prediction[1-3] from six human STRING datasets/networks[29]. These included protein-protein interactions (*PPI*), interactions in *curated databases*, *co-expression* networks, and genomic *neighborhood*, *co-occurrence* and *fusion* interactions. For consistency, these datasets were exactly the same as those used to evaluate Mashup[30] and deepNF[31] in their respective publications. For all the datasets, both individual and integrated, we used each protein's adjacency vector as its feature values vector for training and evaluating predictive models. This feature encoding has recently been shown to be effective for network-based gene classification[37].

Our PFP experiments focused on 112 GO Molecular Function and Biology Process terms that were assigned to at least a thousand human genes each in September 2017 with a non-IEA evidence code. For each GO term, we defined proteins annotated to it as positive examples, and any that were neither annotated to the term, nor its ancestors or descendants, as negative examples[38]. For each term, only proteins that were either positive or negative examples were used in all the experiments. If any of these proteins were not covered by an individual data set, the protein was assigned a corresponding feature vector consisting of all zeros.

We also calculated specific properties of the GO terms considered using the GOATOOLS package (version 0.8.4)[39]. The depth of term t was defined as the length of the shortest path from the root of the corresponding ontology to t, while its information content (IC) is defined as $-\log_{10}(p(t))$, where $p(t)$ is the probability of a human protein being annotated with t.

Evaluation methodology. In all our assessments, we used the recommended F_{max} evaluation measure[40], i.e., the maximum value of the F-measure across all prediction score thresholds. For each GO term, we compared the representative predictors in the following three categories:

- The heterogeneous ensemble algorithm yielding the best-performing EI predictor.
- The best-performing base predictor on the Mashup- and deepNF-integrated datasets.
- The best-performing heterogeneous ensemble algorithm on each individual dataset.

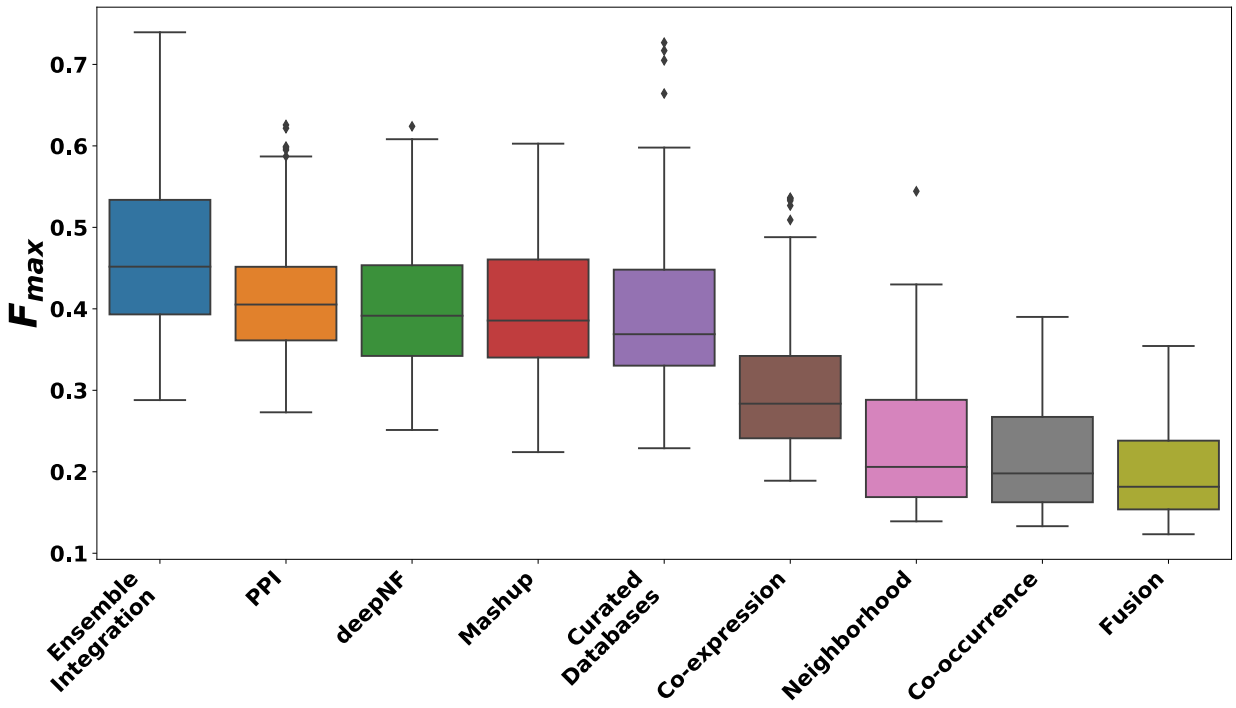

Figure 2: Comparison of the distributions of prediction performances, measured in terms of F_{max} values, of the various data integration approaches, as well as six individual STRING datasets/networks, when predicting annotations to 112 GO Molecular Function and Biology Process terms. Shown here are the best-performing Ensemble Integration (EI) algorithms tested, the best-performing individual predictors (the same ones used in the first layer of EI (Figure 1)) after early network integration using the deepNF and Mashup algorithms, and the best-performing heterogeneous ensembles (the same ones used in the second layer of EI (Figure 1)) on the individual datasets. The best-performing implementations of each approach were identified for each GO term separately.

This methodology enables a fair comparison among the representative predictors of the EI and baseline approaches. Finally, we statistically compared the performances of these approaches across all the GO terms tested using the recommended Friedman and Nemenyi tests[41].

Results

Figure 2 shows the distributions of the F_{max} scores obtained by the best implementations of the different integration approaches for the each of GO terms tested. EI achieved the highest median F_{max} score overall. It significantly outperformed the best individual predictors on the integrated Mashup ($p=1.55x10^{-15}$) and deepNF ($p=2.12x10^{-12}$) networks, as well the best-performing heterogeneous ensembles on every individual dataset ($p \leq 9.14x10^{-8}$). This improvement is likely due to the ability of EI to encapsulate local information in the individual datasets into the most effective base predictors before their assimilation into heterogeneous ensembles. It is also notable that the heterogeneous ensembles on the individual PPI datasets performed slightly better than predictors from the Mashup and deepNF networks (p=0.008 and 0.092, respectively), reaffirming the utility of these ensembles for predictive modeling.

It is well-known that the performance of PFP algorithms can vary substantially across GO terms depending on their information content (IC) (inversely related to its size, i.e., the number of proteins annotated to a term) and depth in the respective GO hierarchy (specificity of the term)[40, 42]. Thus, we also compared EI, deepNF and Mashup, the main data integration strategies evaluated in this study, with regard to how their performance varied with these properties of GO terms. In terms of both these properties, EI consistently performed better than both deepNF and Mashup at increasing depths (**Figure 3(A)**) and levels of IC (**Figure 3(B)**). These trends indicate

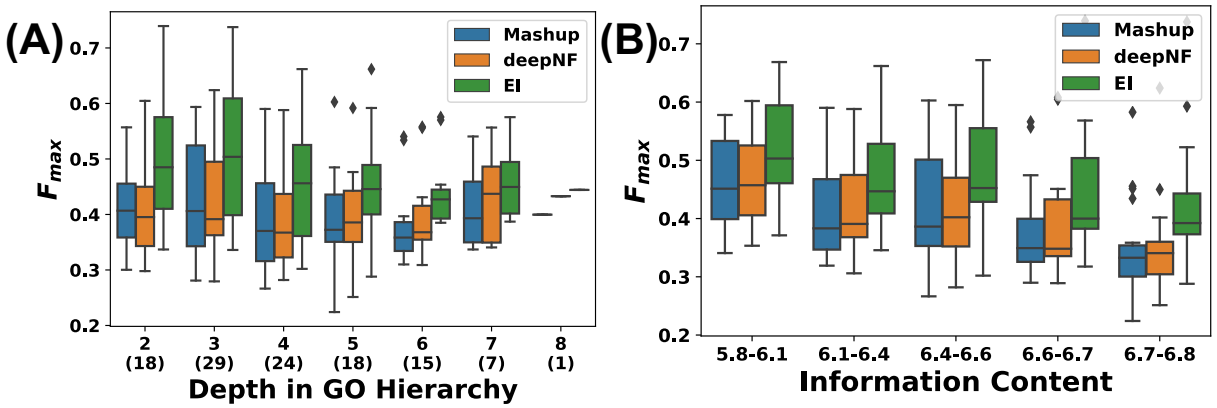

Figure 3: Variation of performance of our EI approach, as well as the deepNF and Mashup algorithms, in terms of **(A)** the depth and **(B)** information content (discretized into equally sized bins covering 22 or 23 terms each) of the 112 GO terms tested in this work. The number of GO terms falling under each depth is shown below the corresponding value in (A). Both depth and information content were calculated using the GOATOOLS package (version 0.8.4). In terms of both these properties, EI consistently performs better than deepNF and Mashup.

that EI can perform well even for GO terms that are deeper or have fewer annotated genes than terms near the root of the hierarchy. This is again likely due to its ability to capture and aggregate information local to individual datasets, which is not enabled sufficiently by early integration.

Discussion

Here, we proposed a novel approach to late data integration for predictive modeling, namely Ensemble Integration (EI), where *local models* derived from individual datasets are integrated into heterogeneous ensembles to develop predictive models. Through the problem of protein function prediction, we demonstrated that EI performs better than early or no data integration approaches.

This proof of principle study has limitations, which offer several avenues for future work. Foremost, our assessments only included a relatively small number of GO terms with at least a thousand human genes annotated to each of them. Developing EI methods for more specific GO terms and other sparse labels, which have many fewer positive examples, is an important challenge. Moreover, even though the individual STRING datasets considered in this study were natively structured as networks, we used their adjacency matrices as sources of feature vectors for traditional classification algorithms such as SVM, LR and DT. Although this feature-based representation has been recently shown to be more effective for (gene) classification[37], this assessment may need to be re-evaluated in the context of network integration. Additionally, in the current implementation of EI, we assigned all-zero feature vectors to proteins disconnected from an individual network. In future work, it will be important to investigate other strategies for handling such missing values. It is also important to repeat our assessments for different subsets of individual STRING networks and GO evidence codes to eliminate non-obvious circularities between the two data sources. Finally, it will also be useful to study the mathematical properties of EI that led to its improved performance over other data integration algorithms in our experiments, as well as the methodologies to interpret the ensembles EI yields.

Acknowledgements

This work was supported by NIH grant R01GM114434, an IBM faculty award, and IARPA, under Cooperative Agreement Number [W911NF-17-2-0105]. It was also enabled by Scientific Computing resources at Mount Sinai. We thank colleagues in the FunGCAT/IGACAT team and at Mount Sinai for discussions and suggestions. The views and conclusions herein do not represent those of the U.S. Government, which is authorized to reproduce and distribute reprints notwithstanding any copyright annotation thereon.

References

1. Pandey G, Kumar V, Steinbach M. Computational Approaches for Protein Function Prediction: A Survey. 2006.
2. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Molecular systems biology. 2007;3(1).
3. Rentzsch R, Orengo CA. Protein function prediction--the power of multiplicity. Trends Biotechnol. 2009;27(4):210-9. Epub 2009/03/03. doi: 10.1016/j.tibtech.2009.01.002. PubMed PMID: 19251332.
4. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet. 2017;8:84. Epub 2017/07/04. doi: 10.3389/fgene.2017.00084. PubMed PMID: 28670325; PMCID: PMC5472696.
5. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion. 2019;50:71-91.
6. Gligorijevic V, Przulj N. Methods for biological data integration: perspectives and challenges. J R Soc Interface. 2015;12(112). Epub 2015/10/23. doi: 10.1098/rsif.2015.0571. PubMed PMID: 26490630; PMCID: PMC4685837.
7. Tiwari P, Viswanath S, Lee G, Madabhushi A, editors. Multi-modal data fusion schemes for integrated classification of imaging and non-imaging biomedical data. 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2011 30 March-2 April 2011.
8. Ray B, Ghedin E, Chunara R. Network inference from multimodal data: A review of approaches from infectious disease transmission. J Biomed Inform. 2016;64:44-54. Epub 2016/09/11. doi: 10.1016/j.jbi.2016.09.004. PubMed PMID: 27612975.
9. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333-7. Epub 2014/01/28. doi: 10.1038/nmeth.2810. PubMed PMID: 24464287.
10. Yu S, Tranchevent L-C, De Moor B, Moreau Y. Kernel-based data fusion for machine learning: Springer; 2013.
11. Fang J. Tightly integrated genomic and epigenomic data mining using tensor decomposition. Bioinformatics. 2019;35(1):112-8. Epub 2018/06/26. doi: 10.1093/bioinformatics/bty513. PubMed PMID: 29939222; PMCID: PMC6298052.
12. Pavlidis P, Weston J, Cai J, Noble WS. Learning gene functional classifications from multiple data types. J Comput Biol. 2002;9(2):401-11. Epub 2002/05/23. doi: 10.1089/10665270252935539. PubMed PMID: 12015889.
13. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM, Xie W, Rosen GL, Lengerich BJ, Israeli J, Lanchantin J, Woloszynek S, Carpenter AE, Shrikumar A, Xu J, Cofer EM, Lavender CA, Turaga SC, Alexandari AM, Lu Z, Harris DJ, DeCaprio D, Qi Y, Kundaje A, Peng Y, Wiley LK, Segler MHS, Boca SM, Swamidass SJ, Huang A, Gitter A, Greene CS. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface. 2018;15(141):20170387. doi: doi:10.1098/rsif.2017.0387.
14. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017;19:221-48. doi: 10.1146/annurev-bioeng-071516-044442. PubMed PMID: 28301734; PMCID: PMC5479722.
15. Liu F, Chen J, Jagannatha A, Yu H. Learning for biomedical information extraction: Methodological review of recent advances. arXiv preprint arXiv:160607993. 2016.
16. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA. 2939785: ACM; 2016. p. 785-94.

17. Reinstein I. XGBoost, a Top Machine Learning Method on Kaggle, Explained 2017 [cited 2019 June 26]. Available from: https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html.
18. Morde V, Setty VA. XGBoost Algorithm: Long May She Reign! 2019 [cited 2019 June 26]. Available from: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.
19. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet. 2017;18(9):551-62. Epub 2017/06/14. doi: 10.1038/nrg.2017.38. PubMed PMID: 28607512.
20. Whalen S, Pandey OP, Pandey G. Predicting protein function and other biomedical characteristics with heterogeneous ensembles. Methods. 2016;93:92-102.
21. Stanescu A, Pandey G. Learning Parsimonious Ensembles for Unbalanced Computational Genomics Problems. Pac Symp Biocomput. 2017;22:288-99. doi: 10.1142/9789813207813_0028. PubMed PMID: 27896983; PMCID: PMC5147733.
22. Wang L, Law J, Kale SD, Murali TM, Pandey G. Large-scale protein function prediction using heterogeneous ensembles. F1000Res. 2018;7:ISCB Comm J-1577. doi: 10.12688/f1000research.16415.1. PubMed PMID: 30450194.
23. Sieberts SK, Zhu F, Garcia-Garcia J, Stahl E, Pratap A, Pandey G, Pappas D, Aguilar D, Anton B, Bonet J, Eksi R, Fornes O, Guney E, Li H, Marin MA, Panwar B, Planas-Iglesias J, Poglayen D, Cui J, Falcao AO, Suver C, Hoff B, Balagurusamy VS, Dillenberger D, Neto EC, Norman T, Aittokallio T, Ammad-Ud-Din M, Azencott CA, Bellon V, Boeva V, Bunte K, Chheda H, Cheng L, Corander J, Dumontier M, Goldenberg A, Gopalacharyulu P, Hajiloo M, Hidru D, Jaiswal A, Kaski S, Khalfaoui B, Khan SA, Kramer ER, Marttinen P, Mezlini AM, Molparia B, Pirinen M, Saarela J, Samwald M, Stoven V, Tang H, Tang J, Torkamani A, Vert JP, Wang B, Wang T, Wennerberg K, Wineinger NE, Xiao G, Xie Y, Yeung R, Zhan X, Zhao C, Members of the Rheumatoid Arthritis Challenge C, Greenberg J, Kremer J, Michaud K, Barton A, Coenen M, Mariette X, Miceli C, Shadick N, Weinblatt M, de Vries N, Tak PP, Gerlag D, Huizinga TW, Kurreeman F, Allaart CF, Louis Bridges S, Jr., Criswell L, Moreland L, Klareskog L, Saevarsdottir S, Padyukov L, Gregersen PK, Friend S, Plenge R, Stolovitzky G, Oliva B, Guan Y, Mangravite LM, Bridges SL, Criswell L, Moreland L, Klareskog L, Saevarsdottir S, Padyukov L, Gregersen PK, Friend S, Plenge R, Stolovitzky G, Oliva B, Guan Y, Mangravite LM. Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. Nat Commun. 2016;7:12460. doi: 10.1038/ncomms12460. PubMed PMID: 27549343; PMCID: PMC4996969.
24. Altmann A, Rosen-Zvi M, Prosperi M, Aharoni E, Neuvirth H, Schulter E, Buch J, Struck D, Peres Y, Incardona F, Sonnerborg A, Kaiser R, Zazzi M, Lengauer T. Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy. PLoS One. 2008;3(10):e3470. doi: 10.1371/journal.pone.0003470. PubMed PMID: 18941628; PMCID: PMC2565127.
25. Niculescu-Mizil A, Perlich C, Swirszcz G, Sindhwani V, Liu Y, Melville P, Wang D, Xiao J, Hu J, Singh M, editors. Winning the KDD cup orange challenge with ensemble selection. Proceedings of the 2009 International Conference on KDD-Cup 2009-Volume 7; 2009: JMLR. org.
26. Pandey G, Zhang B, Chang AN, Myers CL, Zhu J, Kumar V, Schadt EE. An integrative multi-network and multi-classifier approach to predict genetic interactions. PLoS computational biology. 2010;6(9). Epub 2010/09/15. doi: 10.1371/journal.pcbi.1000928. PubMed PMID: 20838583; PMCID: 2936518.
27. Sesmero MP, Ledezma AI, Sanchis A. Generating ensembles of heterogeneous classifiers using Stacked Generalization. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2015;5(1):21-34. doi: 10.1002/widm.1143.

28. Tsoumakas G, Partalas I, Vlahavas I, editors. A taxonomy and short review of ensemble selection. Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications; 2008.
29. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):D447-52. Epub 2014/10/30. doi: 10.1093/nar/gku1003. PubMed PMID: 25352553; PMCID: PMC4383874.
30. Cho H, Berger B, Peng J. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. Cell Systems. 2016;3(6):540-8.e5. doi: 10.1016/j.cels.2016.10.017.
31. Gligorijević V, Barot M, Bonneau R. deepNF: deep network fusion for protein function prediction. Bioinformatics. 2018;34(22):3873-81. doi: 10.1093/bioinformatics/bty440.
32. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009;11(1):10-8. doi: 10.1145/1656274.1656278.
33. Chawla NV. Data Mining for Imbalanced Datasets: An Overview. In: Maimon O, Rokach L, editors. Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer US; 2005. p. 853-67.
34. Caruana R, Munson A, Niculescu-Mizil A. Getting the Most Out of Ensemble Selection. Proceedings of the Sixth International Conference on Data Mining. 1193228: IEEE Computer Society; 2006. p. 828-33.
35. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A. Ensemble selection from libraries of models. Proceedings of the twenty-first international conference on Machine learning; Banff, Alberta, Canada. 1015432: ACM; 2004. p. 18.
36. Wang L. Source code for Ensemble Integration 2019 [cited 2019 September 25]. Available from: https://github.com/GauravPandeyLab/ensemble_integration.
37. Liu R, Mancuso CA, Yannakopoulos A, Johnson KA, Krishnan A. Supervised-learning is an accurate method for network-based gene classification. bioRxiv. 2019, 10.1101/721423:721423. doi: 10.1101/721423.
38. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 2008;9 Suppl 1:S4. Epub 2008/07/22. doi: 10.1186/gb-2008-9-s1-s4. PubMed PMID: 18613948; PMCID: PMC2447538.
39. Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, Dampier W, Dessimoz C, Flick P, Tang H. GOATOOLS: A Python library for Gene Ontology analyses. Scientific Reports. 2018;8(1):10872. doi: 10.1038/s41598-018-28948-z.
40. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Toronen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kassner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Honigschmid P, Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Bjorne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJ, Skunca N, Supek F, Bosnjak M, Panov P, Dzeroski S, Smuc T, Kourmpetis YA, van Dijk AD, ter Braak CJ, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD, Friedberg I. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10(3):221-7. doi: 10.1038/nmeth.2340. PubMed PMID: 23353650; PMCID: PMC3584181.

41. Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets. J Mach Learn Res. 2006;7:1-30.
42. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo da CE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, Salakoski T, Ginter F, Fang H, Smithers B, Oates M, Gough J, Toronen P, Koskinen P, Holm L, Chen CT, Hsu WL, Bryson K, Cozzetto D, Minneci F, Jones DT, Chapman S, Bkc D, Khan IK, Kihara D, Ofer D, Rappoport N, Stern A, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC, Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P, ElShal S, Tranchevent LC, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero AE, Sasidharan R, Yang H, Paccanaro A, Gillis J, Sedeno-Cortes AE, Pavlidis P, Feng S, Cejuela JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T, Marcet-Houben M, Supek F, Gong Q, Ning W, Zhou Y, Tian W, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giollo M, Piovesan D, Tosatto SC, Del Pozo A, Fernandez JM, Maietta P, Valencia A, Tress ML, Benso A, Di Carlo S, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW, van Dijk AD, Gemovic B, Glisic S, Perovic V, Veljkovic V, Veljkovic N, Almeida ESDC, Vencio RZ, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z, Sternberg MJ, Wass MN, Huntley RP, Martin MJ, O'Donovan C, Robinson PN, Moreau Y, Tramontano A, Babbitt PC, Brenner SE, Linial M, Orengo CA, Rost B, Greene CS, Mooney SD, Friedberg I, Radivojac P. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016;17(1):184. doi: 10.1186/s13059-016-1037-6. PubMed PMID: 27604469; PMCID: PMC5015320.