SYSTEMS MODELING: APPROACHES AND APPLICATIONS

EDITED BY: Alberto Jesus Martin, Ernesto Perez-Rueda and Daniel Garrido
PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-413-9
DOI 10.3389/978-2-88966-413-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

SYSTEMS MODELING: APPROACHES AND APPLICATIONS

Topic Editors:
Alberto Jesus Martin, Universidad Mayor, Chile
Ernesto Perez-Rueda, Universidad Nacional Autónoma de México Merida, Mexico
Daniel Garrido, Pontificia Universidad Católica de chile, Chile

Citation: Martin, A. J., Perez-Rueda, E., Garrido, D., eds. (2021). Systems Modeling: Approaches and Applications. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-413-9

Table of Contents

04 ***Editorial: Systems Modeling: Approaches and Applications***
Alberto J. Martin, Ernesto Perez-Rueda and Daniel Garrido

06 ***Identification of Modules With Similar Gene Regulation and Metabolic Functions Based on Co-expression Data***
Edgardo Galán-Vásquez and Ernesto Perez-Rueda

18 ***Integration of Biological Networks for* Acidithiobacillus thiooxidans *Describes a Modular Gene Regulatory Organization of Bioleaching Pathways***
María Paz Cortés, Vicente Acuña, Dante Travisany, Anne Siegel, Alejandro Maass and Mauricio Latorre

29 ***Challenges of Integrative Disease Modeling in Alzheimer's Disease***
Sepehr Golriz Khatami, Christine Robinson, Colin Birkenbihl, Daniel Domingo-Fernández, Charles Tapley Hoyt and Martin Hofmann-Apitius

42 ***Development of Supervised Learning Predictive Models for Highly Non-linear Biological, Biomedical, and General Datasets***
David Medina-Ortiz, Sebastián Contreras, Cristofer Quiroz and Álvaro Olivera-Nappa

58 ***A Middle-Out Modeling Strategy to Extend a Colon Cancer Logical Model Improves Drug Synergy Predictions in Epithelial-Derived Cancer Cell Lines***
Eirini Tsirvouli, Vasundra Touré, Barbara Niederdorfer, Miguel Vázquez, Åsmund Flobak and Martin Kuiper

frontiers
in Molecular Biosciences

EDITORIAL
published: 09 December 2020
doi: 10.3389/fmolb.2020.620464

Editorial: Systems Modeling: Approaches and Applications

Alberto J. Martin[1*†], *Ernesto Perez-Rueda*[2*†] and *Daniel Garrido*[3*†]

[1] Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile, [2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Unidad Académica de Yucatán, Mérida, Mexico, [3] Departamento de Ingeniería Química y Bioprocesos, Pontificia Universidad Católica, Santiago, Chile

Keywords: systems modeling, biological networks, computational methods, network inference, network analysis

Editorial on the Research Topic

Systems Modeling: Approaches and Applications

Edited and reviewed by:
Giulia Palermo,
University of California, Riverside,
United States

****Correspondence:***
Alberto J. Martin
alberto.martin@umayor.cl
Ernesto Perez-Rueda
ernesto.perez@iimas.unam.mx
Daniel Garrido
dgarridoc@ing.puc.cl

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *22 October 2020*
Accepted: *18 November 2020*
Published: *09 December 2020*

Citation:
Martin AJ, Perez-Rueda E and Garrido D (2020) Editorial: Systems Modeling: Approaches and Applications.
Front. Mol. Biosci. 7:620464.
doi: 10.3389/fmolb.2020.620464

INTRODUCTION

Systems Biology, a relatively recent discipline relies on computational modeling as one of its main tools. Ever appearing computational approaches allow us to raise new hypotheses that were unfeasible to test a few years ago. Given the broad range of applications of systems biology, we considered necessary to increase the coverage of tools and their applications in several areas, such as medicine, biotechnology and engineering.

The main goal of the Research Topic (Systems Modeling: Approaches and Applications) was to provide an overview covering both research articles and reviews. In this regard, the collection highlights the impact of computational tools and the usefulness of modeling to decipher the inner workings of biological systems.

Galán-Vásquez and Perez-Rueda evaluated co-expression networks for 17 bacterial organisms via weighted gene co-expression network analysis and clustered into modules of genes with similar expression patterns for each species, to determine relevant modules through a hypergeometric approach based on a set of transcription factors and enzymes for each genome.

Next, Cortés et al., constructed the regulatory and metabolic networks of the bacterium Acidithiobacillus thiooxidans, using an *in silico* semi-automatic genome scale approach. The authors provide an elegant identification of confident connections between both networks (V-shapes), identifying a sub-network of transcriptional factors (34 regulators) regulating genes (61 operons) encoding for proteins involved in biomining-related pathways. In contrast, pathways involved in iron homeostasis and oxidative stress damage are mainly regulated by unique primary regulators, conferring Licanantay an efficient, and specific metal resistance response.

In the third article, Khatami et al. make an excellent review describing the models to characterize Alzheimer's Disease. In this context, integrative models can be sorted in hypothetical models and data-driven models. The latter group split into two subgroups: (i) Models that use traditional statistical methods such as linear models, (ii) Models that take advantage of more advanced artificial intelligence approaches such as machine learning. The review highlights advancements of integrative modeling in the field of AD research.

Medina-Ortiz et al. explored an approach of unsupervised learning algorithms, and a new methodology designed to find optimum partitions within highly non-linear datasets that allow deconvoluting variables and improve performance metrics in supervised learning classification or regression models. These algorithms provide an excellent approach to generate predictive models for highly non-linear datasets; with not significant human input, which guarantees a higher usability in the biological, biomedical, and protein engineering community with no specific knowledge in the machine learning area.

Finally, Tsirvouli et al. show how a relatively large manually curated logical model can be efficiently enhanced further by including components highlighted by a multi-omics data analysis of data from Consensus Molecular Subtypes covering colorectal cancer; finding that the approach can benefit *in silico* experiments on cancer cell lines.

We believe as Editors of this topic, that the original aims have been fulfilled. We consider that the five articles (four original and one review), cover diverse descriptions and proposals to evaluate the modeling to understand the complexity of the biological systems. We must appreciate the works and authors for their excellent contributions that allow for inspiration for other professors in the field.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

The Editors thank all the contributors and the grants IN-209620 (Dirección General de Asuntos del Personal Académico-Universidad Nacional Autónoma de México) to EP-R, Fondecyt projects 1190074 and 1181089 awarded respectively to DG and AM, Proyecto Estrategico CYTED P918PTE0261 and 919PTE0588, and Seed Fund Ingenieria UC (DG).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 13 December 2019
doi: 10.3389/fmolb.2019.00139

Identification of Modules With Similar Gene Regulation and Metabolic Functions Based on Co-expression Data

Edgardo Galán-Vásquez[1]* *and Ernesto Perez-Rueda*[2,3]*

[1] *Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Ciudad Universitaria, Universidad Nacional Autónoma de México, Ciudad de México, Mexico,* [2] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Unidad Académica Yucatán, Mérida, Mexico,* [3] *Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile*

Biological systems respond to environmental perturbations and to a large diversity of compounds through gene interactions, and these genetic factors comprise complex networks. In particular, a wide variety of gene co-expression networks have been constructed in recent years thanks to the dramatic increase of experimental information obtained with techniques, such as microarrays and RNA sequencing. These networks allow the identification of groups of co-expressed genes that can function in the same process and, in turn, these networks may be related to biological functions of industrial, medical and academic interest. In this study, gene co-expression networks for 17 bacterial organisms from the COLOMBOS database were analyzed via weighted gene co-expression network analysis and clustered into modules of genes with similar expression patterns for each species. These networks were analyzed to determine relevant modules through a hypergeometric approach based on a set of transcription factors and enzymes for each genome. The richest modules were characterized using PFAM families and KEGG metabolic maps. Additionally, we conducted a Gene Ontology analysis for enrichment of biological functions. Finally, we identified modules that shared similarity through all the studied organisms by using comparative genomics.

Keywords: transcription factors, gene expression, metabolism, gene co-expression networks, WGCNA

OPEN ACCESS

Edited by:
Elena Papaleo,
Danish Cancer Society Research Center (DCRC), Denmark

Reviewed by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China
Zhen Su,
China Agricultural University (CAU), China

Correspondence:
Edgardo Galán-Vásquez
edgardo.galan@iimas.unam.mx
Ernesto Perez-Rueda
ernesto.perez@iimas.unam.mx

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *28 August 2019*
Accepted: *18 November 2019*
Published: *13 December 2019*

Citation:
Galán-Vásquez E and Perez-Rueda E (2019) Identification of Modules With Similar Gene Regulation and Metabolic Functions Based on Co-expression Data. Front. Mol. Biosci. 6:139. doi: 10.3389/fmolb.2019.00139

INTRODUCTION

Organisms are dynamic systems that respond to intracellular and extracellular signals through the regulated expression of their genes. In recent years, a large number of experiments utilizing high-throughput technologies, including microarrays and RNA sequencing (RNA-seq), have been performed to analyze this differential expression, allowing the identification of genes co-expressed in a particular condition. Recent approaches have shown that there are underlying properties that can only be explained by studying organisms as complex systems (Kitano, 2002; Trewavas, 2006). In this context, a systematic analysis to understand the gene expression in a particular genome is through Gene Co-expression Networks (GCNs), where the network $G = (V, E)$ is composed of a set of nodes (V) that represent the genes and a set of edges (E) that indicate significant co-expression relationships (Stuart et al., 2003; Junker and Schreiber, 2008). These types of networks maintain the structural properties of real networks, such as scale-free topology, which means that there are some highly, connected nodes, namely hubs, and a large number of nodes with a small number of connections (Van Noort et al., 2004; Tsaparas et al., 2006).

In this regard, different algorithms have been developed to reconstruct GCNs; in particular, Weighted Gene Co-expression Network Analysis (WGCNA) allows the construction of networks by considering not only the co-expression patterns between two genes but also the overlapping of neighbor genes (Zhang and Horvath, 2005). Thus, highly correlated genes are clustered into large modules based on similarities in their expression profiles. These modules are often enriched for genes that share similar biological functions (Mueller et al., 2017; van Dam et al., 2018). WGCNA also compares different GCNs to identify conserved modules between species or cell types (Yang et al., 2014; Bakhtiarizadeh et al., 2018; Hosseinkhan et al., 2018). GCNs have been used to identify genes with similar expression patterns in a set of samples, allowing the prediction of gene functions at the genome level, the functional discovery of unknown genes and their associations with diseases (Carlson et al., 2006; Emilsson et al., 2008; Amar et al., 2013).

To date, two highly conserved processes between the organisms have been identified: metabolism and gene regulation (McAdams et al., 2004; Peregrín-Alvarez et al., 2009). Both processes are mediated by specific proteins; on one hand, for metabolism, enzymes catalyze the transformation of one compound to another. On the other hand, gene expression at the transcriptional level is regulated by proteins called transcription factors (TFs). In recent works, a compendium of TF families for different organisms has been identified; and other studies have revealed promiscuity of different enzymes related to metabolism. Therefore, due to the relevance of these two types of protein-encoding genes, it is important to evaluate how the gene expression patterns are distributed in functional modules.

In this study, a gene co-expression network for 17 bacterial organisms from the COLOMBOS database using WGCNA was identified. To do this, the genes were clustered into modules with similar expression patterns. These modules were exhaustively analyzed considering the repertoire of enzymes and TFs, suggesting that these proteins are involved in similar functional processes. Additionally, to determine what functional classes are overrepresented in the respective modules, an enrichment analysis was conducted. This study provides insights into how regulatory proteins and metabolic maps are expressed in different organisms.

MATERIALS AND METHODS

Datasets

The gene expression dataset was obtained from the COLlections of Microarrays for Bacterial OrganismS (COLOMBOS) dataset and included gene expression data for 17 different bacterial organisms with 31,982 genes and 11,224 contrasts (http://colombos.net/). In brief, COLOMBOS is a compendium of data obtained from microarray and RNA-seq experiments performed under different experimental conditions. These data are further curated and normalized, considering the following principles: (1) raw intensities are preferred as data source, (2) no local background or mismatch probe correction procedures are performed, (3) quantile normalization for high-density oligonucleotide experiments are performed, and (4) logratios are created for single-channel data according to the condition contrast definitions and combined with the dual channel measurements (Moretto et al., 2016).

Thus, we analyzed with principal components analysis (PCA) the microarray compendia of each species to identify outlier samples, i.e., those samples with a substantial difference in expression value compared with other samples. In a posterior step, the dataset results were inspected via the goodSamplesgenes function of the WGCNA R package to inspect data for missing value, and for genes with zero variance, the genes and samples identified as good genes and good samples were conserved (Largfelder and Holvarth, 2008). Finally, the total number of genes and samples considered for each organism were: Ban: 5,027 genes and 53 samples; Bce: 5,200 genes and 159 samples; Bsu: 4,176 genes and 762 samples; Bth: 4,763 genes and 217 samples; Cac: 3,777 genes and 218 samples; Cje: 1,572 samples and 103 samples; Eco: 4,321 samples, and 2,415 samples; Hpy: 1,600 genes and 83 samples; Lrh: 2,731 genes and 49 samples; Mtu: 4,068 genes and 709 samples; Pae: 5,564 genes and 375 samples; Stm: 4,466 genes and 74 samples; Sfl: 3,786 genes and 23 samples; Sme: 6,218 genes and 270 samples; Spd: 1,884 genes and 40 samples; Ttj: 2,173 genes and 303 samples; and Ype: 3,730 genes and 22 samples (**Table 1**). The gene expression dataset for each organism is provided as **Supplementary S1**.

Construction of Co-expression Networks

The gene co-expression networks were constructed with the WGCNA program, which allow network construction, module detection, gene selection, calculations of topological properties, and data simulation, among others (Largfelder and Holvarth, 2008). First, the scale-free topology properties of biological networks were added by calculating the power (β) using the pickSoftThereshold function, see **Table 1** for the β value per organism. Then, we constructed an adjacency matrix for each bacterium, using signed correlation networks, where nodes with negative correlation are considered unconnected; as well as, the pairwise biweight midcorrelation coefficients between all genes. This correlation method was considered because it is more powerful than the Spearman and Pearson correlation methods (Song et al., 2012; Bakhtiarizadeh et al., 2018). Then, the adjacency matrix was transformed into a Topological Overlap Matrix (TOM), where a higher TOM value allowed identification of gene modules for each pair of genes with strong interconnectivity. Therefore, it was used signed correlation networks, pairwise biweight midcorrelation coefficients and β value.

Finally, the genes were clustered into modules with similar expression patterns by using the average linkage hierarchical clustering algorithm (flashClust function) and the cutreeDynamic function was used to cut the branches of the resulting dendrogram that results in the generation of gene modules. To do this, it was used 1-TOM as a distance matrix with a minimum module size equal to 20. Therefore, the modules with highly correlated eigengenes were merged, based on a minimum height of 0.25 (mergeCloseModules function). Each module was identified with a color, where the gray color is reserved for uncorrelated genes (Horvath, 2011) and discarded; whereas the rest of modules were renamed with a number (**Table S1**).

TABLE 1 | Overview of dataset and co-expression modules in this study.

Organism (KEGG ID)	No. of samples*	No. of modules	Avg Size/SD **	No. of ORFs/% of coverage	No. of TFs in modules	No. of enzymes in modules	Power β***
B. anthracis strain Ames (Ban)	53	6	837.83/849.14	5,508/91.27 (5,027)	333	802	12
B. cereus ATCC 14579 (Bce)	159	26	200/230.77	5,366/97.9 (5,200)	339	811	12
B. subtilis 168 (Bsu)	762	38	109.89/67.52	4,220/98.96 (4,176)	285	759	12
B. thetaiotaomicron VPI-5482 (Bth)	217	12	396.9/356.56	4,816/98.9 (4,763)	223	660	10
C. acetobutylicum ATCC 824 (Cac)	218	7	539.57/529.80	3,778/99.99 (3,777)	254	611	14
C. jejuni NCTC 11168 (Cje)	103	20	78.6/54.1	1,654/95.0 (1,572)	35	413	10
E. coli K-12 MG1655 (Eco)	2,415	58	74.5/60.49	4,600/93.9 (4,321)	335	892	14
H. pylori 26695 (Hpy)	83	8	200/157.18	1,600/100 (1,600)	19	350	9
L. rhamnosus GG (Lrh)	49	11	248.27/210.82	2,944/92.96 (2,731)	188	507	12
M. tuberculosis H37Rv (Mtu)	709	29	140.27/173.83	4,096/99.3 (4,068)	245	751	10
P. aeruginosa PAO1 (Pae)	375	20	278.2/347.78	5,570/99.9 (5,564)	468	1,002	12
S. enterica LT2 (Stm)	74	20	223.3/251.72	4,548/98.2 (4,466)	328	896	9
S. flexneri 301 (Sfl)	23	5	757.2/505.02	4,313/88.0 (3,786)	271	776	12
S. meliloti 1021 (Sme)	270	15	414.53/649.46	6,218/100 (6,218)	372	797	12
S.pneumoniae D39 (Spd)	40	9	209.33/134.51	1,911/98.59 (1,884)	98	414	8
T. thermophilus HB8 (Ttj)	303	11	197.54/166.66	2,173/100 (2,173)	92	523	12
Y. pestis C092 (Ype)	22	11	339.09/160.73	3,979/94.39 (3,756)	238	739	14

For each species, we show the final number of experiments analyzed after PCA, the total number of modules identified, the average size of the modules**, the coverage of genes included in the modules in relation to the total number of ORFs, the total of TFs and enzymes, and the lowest possible power term where topology approximates fits a scale-free network***.*

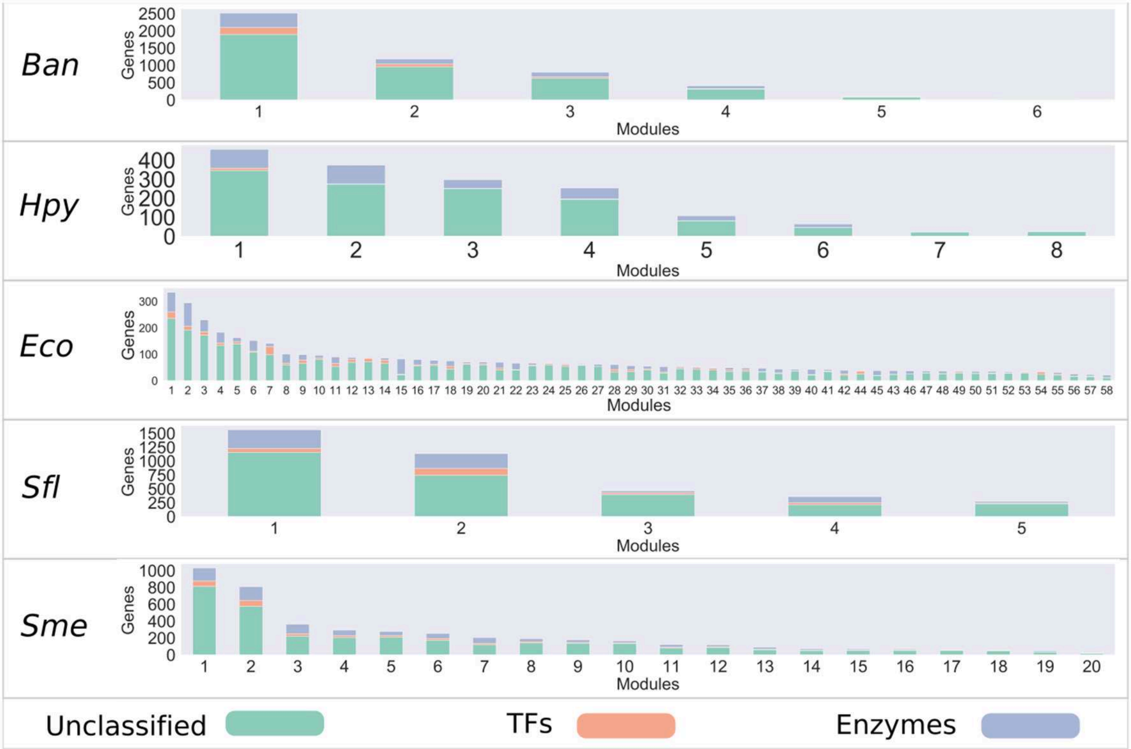

FIGURE 1 | Bacteria co-expression modules. On the x-axis are shown the modules identified with the WGCNA package, identified with a number. The distribution of modules is represented in decreasing order, where the y-axis represents the number of genes per module. Each module is made up of a set of genes associated with TFs (orange), metabolic enzymes (blue), and unclassified genes (green).

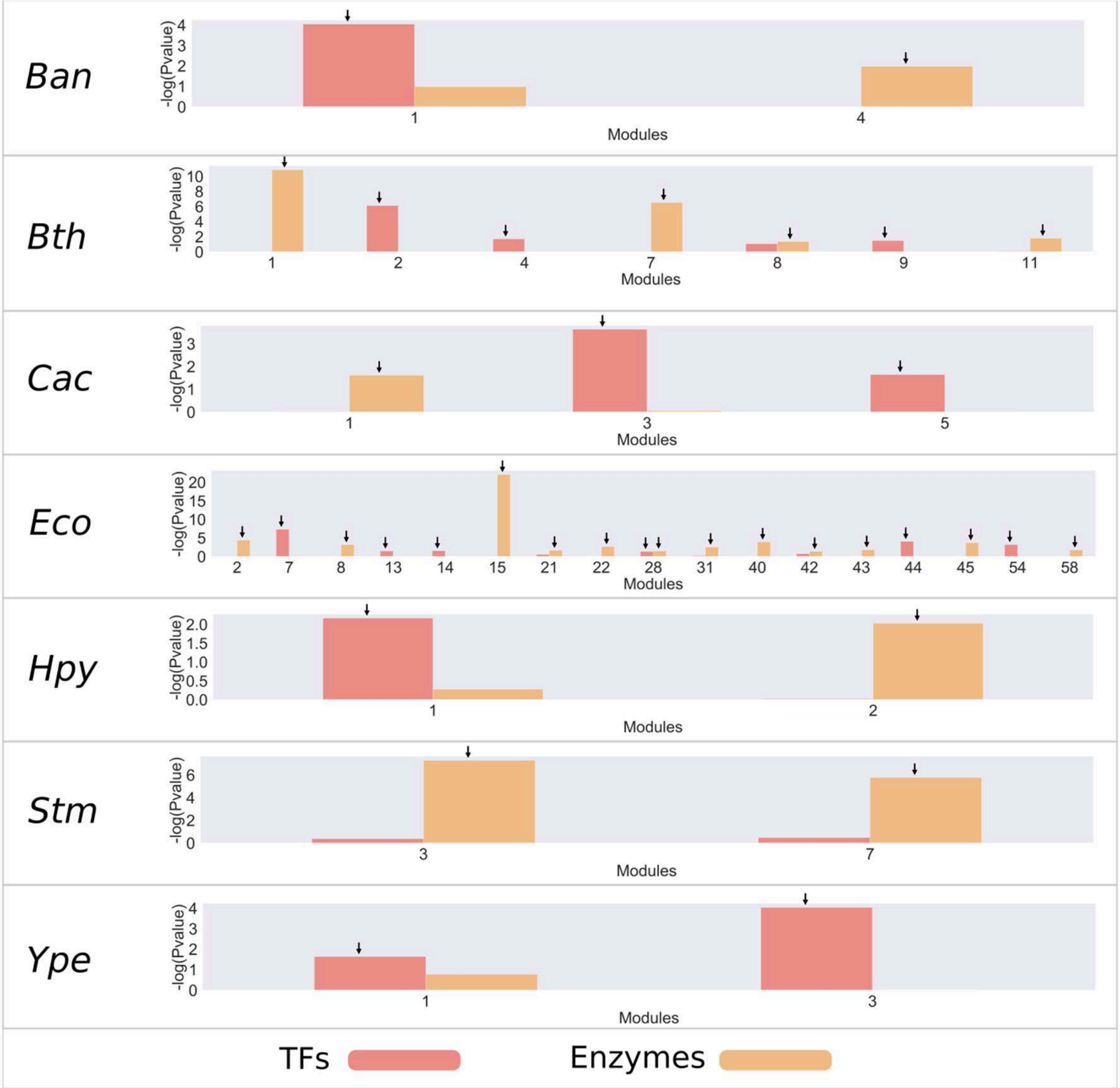

FIGURE 2 | Enrichment of TFs and metabolic enzymes. Modules with a –log10 (*P*-value) >1.5 (corresponding to a *P*-value <0.05) were selected as enriched and are indicated by an arrow on the bar. The red bars represent modules enriched with TF families, and the orange bars represent modules enriched with enzymes.

To perform an analysis of hubs on the modules of interest, these were exported using the exportNetworkToCytoscape function and we selected the 100 most highly correlated genes for each module. The hubs were defined as the most highly connected nodes within the module, so we calculated the degree of connectivity for each node (K), which is defined as the number of edges adjacent to each node (Junker and Schreiber, 2008) (**Figure S1**). A general version of all scripts were included in **Supplementary S2**.

Distribution of TFs and Enzymes

For each genome, we associated the Enzyme Commission number (E.C. number) using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000). Then, each enzyme with an E.C. number was associated with its respective metabolic map. In a similar manner, for TFs we used the compendium of TFs predicted by Rivera-Gómez et al. (2017); assigned from the hidden Markov model (HMM) profiles. To determine the abundance and distribution of each dataset, an incidence rate of the genome and a heatmap for each genome were determined.

Enrichment Analysis

To evaluate the functional association between the modules and TFs and enzymes, an enrichment analysis using a hypergeometric test was conducted. The resulting distribution thus describes the probability of finding x domains associated with a particular category in a list of interest k, from a set of N domains containing m domains that are associated with the same category. We set

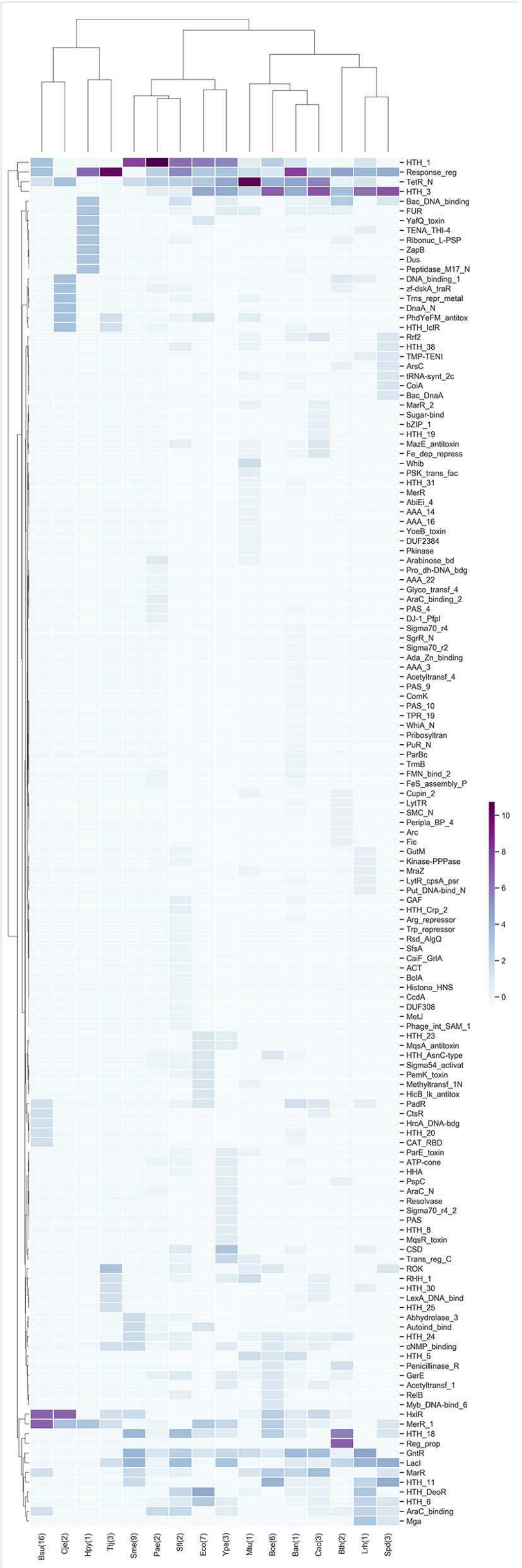

FIGURE 3 | TF families identified as frequent in the enriched modules. *Z*-score hierarchical clustering based on Euclidean distance measure and Ward's method for linkage analysis. Each row represents the PFAM and each column represents the most enriched module for that bacterial species.

statistical significance at a *P*-value of <0.05. All analyses were performed in Python (https://www.python.org/).

Similarity Analysis

To determine the similarity degree between the different enriched modules, orthologous proteins between each pair of genomes were identified. Orthologs were accepted if they had an *e*-value <1e-6, sequence identity >30%, and alignment length >60% of the individual proteins. Then, the Jaccard index was calculated for each pair of modules, which is defined as the size of the intersection that represents the orthologs between each pair of modules of two organisms, divided by the union size of the sample sets.

Functional Annotation Analysis

To identify the biological process in each module, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov/), which is a gene functional classification system that integrates a set of functional annotation tools (Huang et al., 2009).

RESULTS AND DISCUSSION

Construction of Gene Co-expression Networks

In order to determine which genes share similar co-expression patterns in bacteria, a set of co-expression networks was inferred for 17 different bacteria with WGCNA R package (Largfelder and Holvarth, 2008), based on the information deposited in the COLOMBOS database (Moretto et al., 2016). We considered signed networks, because this method takes into account the sign of the underlying correlation coefficient and it has been shown that these networks can identify modules with more significant enrichment of functional groups (Medina and Lubovac-Pilav, 2016; Liu et al., 2018). Based on this approach, the reconstructed co-expression networks had a coverage of around 90% of the predicted open reading frames (ORFs) for each of the bacteria analyzed. In addition, modules inferred showing different sizes, for instance, *Escherichia coli* (Eco) contains the highest number of modules with 58, while for *Shigella flexneri* (Sfl) only 5 modules were identified (see **Figure 1** and **Table S1** and **Figure S2**).

It has been described that, i.e., more samples usually lead to more robust and refined results (Horvath, 2011). However, in the case of the dataset used in our study, the number of samples did not reflect the number of Gene Expression Omnibus (GEO) series used for each bacterium, and this would have influenced the number of modules identified for each organism, as in the case of *Bacillus anthracis* strain Ames (Ban), for which the samples belonged to 4 GEO series, or *Helicobacter pylori* 26695 (Hpy), for which the samples belonged to 8 GEO series, while *Salmonella enterica* LT2 (Stm) samples came from 16 GEO series.

Highly Enriched Modules in TFs and Metabolism Terms

Two processes highly conserved between all the organisms are metabolism and gene regulation, which are mediated by enzymes that catalyze metabolic reactions and by DNA-binding TFs, respectively (Browning and Busby, 2004; Peregrín-Alvarez et al., 2009). In order to identify if metabolism and regulation-related genes share similar co-expression patterns, their distributions into the modules were mapped. Therefore, a collection of TFs, which were identified by homology from a dataset compendium of TFs previously characterized together with family-specific HMM profiles, as well as a compendium of metabolic enzymes of the KEGG repertoire for each one of the 17 bacteria, was used to integrate the information for the inferred modules.

We found that both enzymes and TFs are distributed in almost every co-expression module. This finding is consistent with previous works on modules of co-expression of *E. coli*, where TFs are distributed in all the modules, which allows them to be regulated (Sastry et al., 2019). However, there are modules that have a greater proportion of TFs or enzymes, and this leads us to think that some modules may be more relevant than others in the context of gene regulation or metabolism (**Figure 1**).

To identify relevant modules that consider those regulatory mechanisms and metabolism, an analysis of enrichment was carried out by using a hypergeometric test with the set of TFs and the enzymes associated with metabolism for each of the modules (**Figure 2** and **Figure S3**). From this analysis, we found that most bacteria have an average 2 modules enriched with TFs, with the exception of *E. coli* K-12 MG1655 (Eco), which has 11 modules enriched, and *S. enterica* LT2 (Stm), which does not contain modules enriched with TFs. On the other hand, bacteria contain an average of 4 modules enriched for metabolic enzymes; where *E. coli* is the only species with more modules, with 17. In contrast, *Yersinia pestis* (Ype) does not contain modules enriched with metabolic enzymes.

The most enriched modules with TFs contain on average 27% of the predicted genes with this function. Meanwhile, the modules enriched with metabolic enzymes contain on average 19% genes predicted to be related to metabolism in each organism. Specifically, *B. anthracis* strain Ames (Ban), *H. pylori* 26695 (Hpy), and *S. flexneri* 301 (Sfl) contain around 50% of all predicted TFs. In the same way, *Bacteroides thetaiotaomicron* VPI-5482 (Bth), *Clostridium acetobutylicum* ATCC824 (Cac), *Lactobacillus rhamnosus* GG (Lrh), and *Sinorhizobium meliloti* (Sme) modules contain around 30% of the genes associated with metabolic enzymes.

Based on the modules identified, diverse and interesting findings emerged, such as the fact that there is at least one module with a high percentage of TFs and enzymes, and this led us to evaluate if the richer modules also have a preference for a particular TF family or metabolic maps.

TFs and Metabolism Terms More Abundant

The TFs of each of the highest enrichment modules were classified using the families described in the PFAM database, and the z-scores of the frequency of the families were clustered hierarchically based on Euclidean distance measure and Ward's method for linkage analysis. We determined that the families most frequently present in these modules belong to Response_reg, LysR (HTH_1), Cro-C1 (HTH_3), TetR_N, and GntR (**Figure 3**), and these findings are in agreement with previous results for families more abundant in bacteria (Perez-Rueda et al., 2018).

In this regard, the Response_reg family is related to the two-component systems of bacteria, in which a signal is received from a sensor protein (i.e., the two components). This family of regulators allows the organism to adapt to a wide range of environments, stressors, and growth conditions (Skerker et al., 2005). Another family identified in the modules corresponds to TetR_N, which was one of the most abundant within our study; it is involved in regulating antibiotic resistance, catabolic pathways, biosynthesis of antibiotics, osmotic stress response and pathogenicity. These regulators typically function as repressors (Ramos et al., 2005; Cuthbertson and Nodwell, 2013).

Other families of regulators identified as abundant in the modules were LysR (HTH_1), a family of TFs involved in the regulation of a wide variety of processes that includes the regulation of amino acid biosynthesis and catabolism, stress responses and cell detoxification (Maddocks and Oyston, 2008); and Cro-C1 (HTH_3), which is part of the binary switch that regulates lytic/lysogenic growth of phages by differential binding to the operator sites (Steinmetzer et al., 2002).

In *Bacillus subtilis* 168 (Bsu) and *Campylobacter jejuni* NCTC 11168 (Cje), the abundant families are HxlR, which includes activators involved in the detoxification of formaldehyde, and MerR_1, which responds to environmental stimuli, such as heavy metals, oxidative stress or antibiotics and a subgroup of transcription activators that respond to metal ions (Brown et al., 2003). Meanwhile, in *B. thetaiotaomicron* VPI-5482 (Bth) the most abundant families are HTH_18, which is related to the arabinose operon regulatory protein AraC (Gallegos et al., 1993), and Reg_prop, which is part of a hybrid two-component system and are a key part of this species' ability to sense and degrade complex carbohydrates in the gut (Lowe et al., 2012).

In the same context, the metabolic enzymes were classified according to the KEGG maps, and the z-scores of the frequency of each metabolic map were clustered, similar to our groupings for TF families. In general, we identified that the central metabolism pathways that includes glycolysis/gluconeogenesis, the citrate cycle (TCA cycle) and pyruvate metabolism are expressed independently of the experimental conditions analyzed, similar to the case for nucleotide metabolism. Another conserved cluster is related to carbohydrate metabolism and includes amino sugar and nucleotide sugar metabolism, starch and sucrose metabolism, galactose metabolism, fructose and mannose metabolism and pentose and glucuronate interconversions (**Figure 4**).

In **Figure 4**, there are well-defined clusters, such as the one in *B. anthracis* str. Ames (Ban) that contains maps belonging to xenobiotic biodegradation and metabolism of xenobiotics by cytochrome P450 and to drug metabolism by cytochrome P450, which is mediated by a class II P450 system in this organism (De Mot and Parre, 2002). In addition, in *Mycobacterium tuberculosis* H37Rv (Mtu) we identified maps related to glycerolipid metabolism, which is used to generate glycerols from the host's fatty acids, the vitamin B6 metabolic pathway, which is essential for survival and virulence (Dick et al., 2010), and a nitrogen metabolic pathway

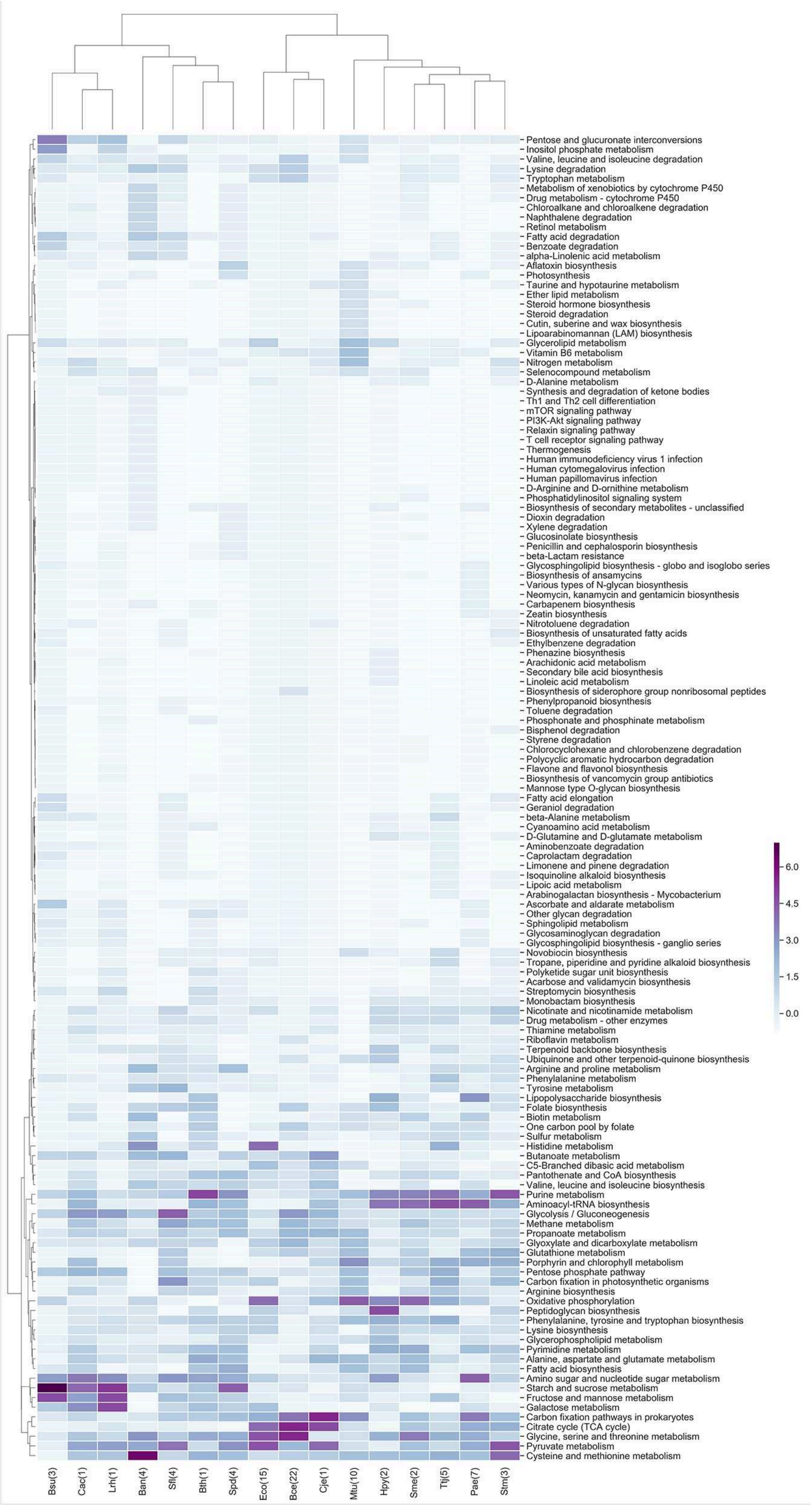

FIGURE 4 | Metabolic maps more frequent in the enriched modules. *Z*-score hierarchical clustering based on Euclidean distance measure and Ward's method for linkage analysis. Each row represents a metabolic map (KEGG), and each column represents the most enriched module, with E.C. numbers for each species.

that is essential for growth and virulence of this bacterium (Gouzy et al., 2014).

In summary, we identified diverse families of TFs and metabolic maps common to all modules in the organisms analyzed, suggesting that common regulatory processes governing a large diversity of metabolic genes expressed under different conditions, and by consequence the global response could be similar even when the organisms employ a diverse repertoire of genes, i.e., not homologous genes. This led us to evaluate the similarity between these modules.

Metabolism and Similar Regulation

To determine the organisms with similar regulation, we calculated the Jaccard index between each pair of modules enriched with TFs, using the number of orthologs shared between each pair of organisms, additionally each module was analyzed by means of Gene Ontology using DAVID (Huang et al., 2009). The Jaccard index matrix was used to build a circos plot (**Figure 5A**), showing similar modules between *S. flexneri* 301 (Sfl), *B. anthracis* (Ban), and *Y. pestis* C092 (Ype), which are characterized as having genes related to biosynthetic process, regulation of cellular process and regulation of primary metabolic processes.

The second group contains *Pseudomonas aeruginosa* PAO1 (Pae), *B. thetaiotaomicron* VPI-5482 (Bth), *M. tuberculosis* H37Rv (Mtu), *Thermus thermophilus* HB8 (Ttj), *C. acetobutylicum* ATCC824 (Cac), *E. coli* K-12 MG1655 (Eco), which include gene related to regulation of cellular and metabolic process, single-organism localization and cellular process and regulation of metabolic process. Finally, the third group consists of *Bacillus cereus* ATCC14579 (Bce), *H. pylori* 26695 (Hpy), *C. jejuni* NCTC 11168 (Cje), *B. subtilis* 168 (Bsu), *L. rhamnosus* GG (Lrh), *Streptococcus pneumoniae* D39 (Spd), *S. meliloti* 1021 (Sme), which have gene related to regulation of cellular process, single-organism metabolic process and nitrogen compound metabolic process.

On the other hand, in the modules related to metabolism, we used the Jaccard index between each pair of modules enrichment with enzymes to identify the similar modules (**Figure 5B**). Based on this approach, we identified that *S. meliloti* 1021 (Sme) is a module that contains a high proportion of orthologs with the other modules, where genes related to cellular metabolic process, primary metabolic process, nitrogen compound metabolic process and organism substance metabolic process were identified. This result could be associated to the prevalence of genetic redundancy in this bacterium, an in particular to those genes involved in a variety of metabolic pathways, including central carbon metabolism, transport, and amino acid biosynthesis (diCenzo and Finan, 2015); and the number of genes with some regulatory mechanisms identified in one of the three replicons, and the function of regulated genes was found to be in accordance with the overall replicon functional signature: housekeeping functions for the chromosome, metabolism for the chromid, and symbiosis for the megaplasmid (Galardini et al., 2015).

This group include *C. jejuni* NCTC 11168 (Cje), *B. thetaiotaomicron* VPI-5482 (Bth), *S. enterica LT2* (Stm), *P. aeruginosa PA01* (Pae), *C. acetobutylicum* ATCC824 (Cac), *H. pylori* 26695 (Hpy), *S. flexneri* 301 (Sfl), which are characterized by genes related to cellular metabolic process, single-organism cellular process, biosynthetic process and organic substance metabolic process. Finally, this group includes *E. coli* K-12 MG1655 (Eco), *B. anthracis* strain Ames (Ban), *T. thermophilus* HB8 (Ttj), *B. subtilis* 168 (Bsu), *M. tuberculosis* H37Rv (Mtu), *B. cereus* ATCC 14579 (Bce), *L. rhamnosus* GG (Lrh), *S. pneumoniae*

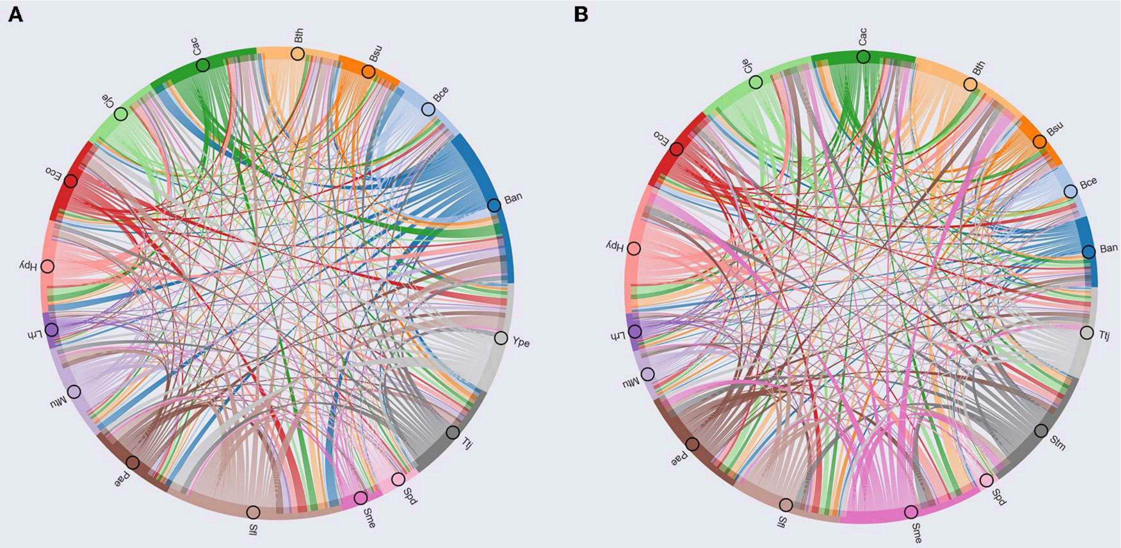

FIGURE 5 | Circos based in Jaccard index. **(A)** Circos based on TFs; **(B)** Circos based on metabolic maps.

D39 (Spd); these species have gene related to catabolic process, single-organism metabolic process, and establishment of localization.

In addition, enriched modules were analyzed to determine those genes with greater connectivity. To this end, we used the first 100 nodes that most correlate in each module where the identified genes had the highest connectivity or highest node degree, which describes the number of interactions or edges adjacent to the node (**Table S2**). Many of the most highly connected nodes are related to nitrogen compound metabolic process, biosynthetic process, cellular metabolic process, primary metabolic process, and single-organism metabolic process, although in some cases the most important hub genes encode for hypothetical proteins, which would allow future analysis to determine their functional role.

From this analysis, in the case of the module 2 enriched with TFs of *S. flexneri* 301 (Sfl), the most highly connected genes were SF2819, an activator of the L-fucose operon from the DeoR family, and SF2545, a polyphosphate kinase [E.C. 2.7.4.1] involved in the nitrogen compound metabolic process and biosynthetic process, respectively; in addition, two hypothetical proteins, SF1784 and SF3500 were also identified as highly

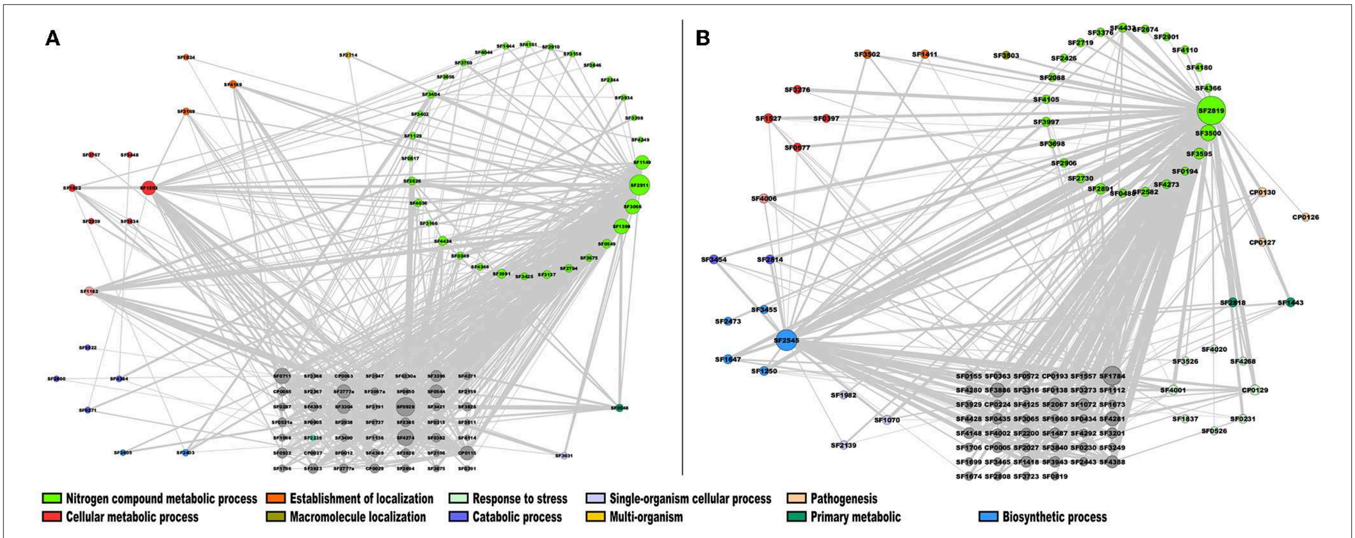

FIGURE 6 | Co-expression network of *S. flexneri*. The most highly correlated genes were plotted in Cytoscape (Smoot et al., 2010). The size of the modules corresponds to their degree of connectivity, while the widths of the edges represent the weights of the correlations, gray nodes do not have an assigned function. **(A)** TFs; **(B)** metabolism modules.

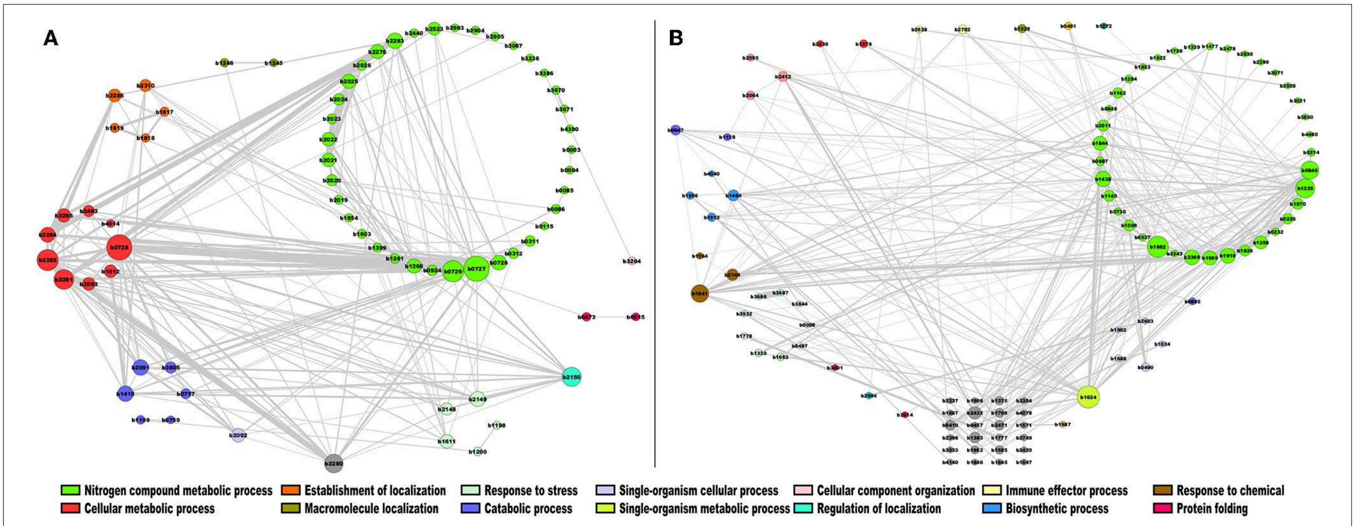

FIGURE 7 | Co-expression network of *E. coli*. The most highly correlated genes were plotted in Cytoscape (Smoot et al., 2010). The sizes of the modules correspond to their degrees of connectivity, while the widths of the edges represent the weights of the correlations, gray nodes do not have an assigned function. **(A)** TFs; **(B)** metabolism modules.

connected genes (**Figure 6A**). In module 4, that was enriched with enzymes, the genes with the highest connectivity were SF2911, which encodes a phosphoglycerate kinase [E.C. 2.7.2.3] involved in nitrogen compound metabolic process; SF0929, which encodes an aminopeptidase N [E.C. 3.4.11.2] involved in the Glutathione metabolism; and SF4274, a NAD(P)H dehydrogenase (quinone) [EC:1.6.5.2] involved in Metabolic pathways (**Figure 6B**). This result correlates with the fact that glutathione and quinone metabolism play a major role in the defense against redox cycling-derived oxidative stress (Kelly et al., 2019), reinforcing the notion that common expression patterns identified in this work correlates with similar protein roles in the cell.

In the case of module 7 enriched with TFs in *E. coli*, we identified the following genes with the highest connectivity: *ydgJ* (b1624), a probable D-galactose 1-dehydrogenase, involved in single-organism metabolic process (Reed et al., 2003); *ribC* (b1662) (for riboflavin synthase), which catalyzes the final step in riboflavin biosynthesis (Eberhardt et al., 1996); *ogt* (b1335), which encodes a methyltransferase enzyme for the repair of alkylated DNA (Taira et al., 2013); and *deoR* (b0840), which is involved in the negative expression of genes related to transport and catabolism of deoxyribonucleoside nucleotides (Garces et al., 2008). These highly correlated genes are mainly involved in biosynthetic processes and nitrogen compound metabolic processes, as shown in **Figure 7A**. In this regard, DeoR and regulated genes have been involved in DNA damage response by drugs, modifying the nucleotide level modulation (Sangurdekar et al., 2011), suggesting that b1335 and b0840 are functionally closer. Therefore, the other genes identified in this module could also participate in a similar response, however further evidence is necessary. On the other hand, in module 15, which is enriched with enzymes, the genes with the highest connectivity were *sucB* (b0727), *sucC* (b0728), and *sucD* (b0729), which are associated with the citrate cycle, an important aerobic pathway for the final steps of the oxidation of carbohydrates and fatty acids (Buck et al., 1986); *nuoH* (b2282), *nuoI* (b2281), *nuoJ* (b2280), and *nuoG* (b2283), involved in the oxidative phosphorylation pathway (Bongaerts et al., 1995) (**Figure 7B**).

CONCLUSIONS

In this work, we identified and analyzed modules considered relevant from a metabolic and regulatory point of view in a set of bacteria, using a weighted gene co-expression analysis method. Based on this analysis, we identified some modules enriched with TFs and metabolic enzymes. In the case of regulation, we identified TFs from the families Response_reg, TetR_N, LysR, and HTH_3, which are mainly related to biological processes, such as biosynthetic processes, cellular metabolic processes, nitrogen compound metabolic processes and primary metabolic processes. On the other hand, the modules enriched with enzymes are associated mainly with primary metabolic, organic substance metabolic, cellular metabolic and nitrogen compound metabolic processes. Our approach also identified genes with similar expression patterns and involved in similar metabolic or regulatory roles, such as DeoR and Ogt. In summary, this analysis allowed us to determine that, despite the diversity of experimental information available for each organism, these mechanisms are similar in all of the organisms, and this will allow us to address new experimental results, such as the use of gene expression data in metagenomic studies.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

EG-V performed the experiments, analyzed the data, and wrote the paper. EP-R analyzed the data and wrote the paper.

FUNDING

This work was supported by the DGAPA of the Universidad Nacional Autonoma de Mexico (IN-201117) and CYTED (P918PTE0261).

ACKNOWLEDGMENTS

We thank Katya Rodriguez, A. C. Poot-Hernandez, and Elsa Beatriz Gongora for carefully reading the manuscript; and Israel Sanchez, Joaquin Morales, and Sandra Sauza for their technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00139/full#supplementary-material

Supplementary Material S1 | Dataset processed by PCA and goodSamplesGenes.

Supplementary Material S2 | General collection of scripts to process, analyze, and visualize the dataset of previously processed.

Table S1 | Collection of co-expression modules for each of the organisms.

Table S2 | Collection of hub genes for each organism.

Figure S1 | Workflow of the analysis procedure.

Figure S2 | Co-expression modules for all organisms. On the x-axis are shown the modules identified with the WGCNA package, identified with a number. The distribution of modules is represented in decreasing order, where the y-axis represents the number of genes per module.

Figure S3 | Enrichment of TFs and metabolic enzymes for all organisms. Modules with a –log10 (*P*-value) >1.5 (corresponding to a *P*-value <0.05) were selected as enriched and are indicated by an arrow on the bar.

REFERENCES

Amar, D., Safer, H., and Shamir, R. (2013). Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput. Biol.* 9:e1002955. doi: 10.1371/journal.pcbi.1002955

Bakhtiarizadeh, M. R., Hosseinpour, B., Shahhoseini, M., Korte, A., and Gifani, P. (2018). Weighted gene co-expression network analysis of endometriosis and identification of functional modules associated with its main hallmarks. *Front. Genet.* 9:453. doi: 10.3389/fgene.2018.00453

Bongaerts, J., Zoske, S., Weidner, U., and Linden, G. (1995). Transcriptional regulation of the proton translocating NADH dehydrogenase (nuoA-N) of Escherichia coli by electron acceptors, electron donors and gene regulators. *Mol. Microbiol.* 16, 521–534. doi: 10.1111/j.1365-2958.1995.tb02416.x

Brown, N. L., Stoyanov, J. V., Kidd, S. P., and Hobman, J. L. (2003). The MerR family of transcriptional regulators. *FEMS Microbiol. Rev.* 27, 145–163. doi: 10.1016/S0168-6445(03)00051-2

Browning, D. F., and Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* 2:57. doi: 10.1038/nrmicro787

Buck, D., Spencer, M. E., and Guest, J. R. (1986). Cloning and expression of the succinyl-CoA synthetase genes of *Escherichia coli* K12. *Microbiology* 132, 1753–1762. doi: 10.1099/00221287-132-6-1753

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., and Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* 7:40. doi: 10.1186/1471-2164-7-40

Cuthbertson, L., and Nodwell, J. R. (2013). The TetR family of regulators. *Microbiol. Mol. Biol. Rev.* 77, 440–475. doi: 10.1128/MMBR.00018-13

De Mot, R., and Parre, A. H. (2002). A novel class of self-sufficient cytochrome P450 monooxygenases in prokaryotes. *Trends Microbiol.* 10, 502–508. doi: 10.1016/S0966-842X(02)02458-7

diCenzo, G. C., and Finan, T. M. (2015). Genetic redundancy is prevalent within the 6.7 Mb *Sinorhizobium meliloti* genome. *Mol. Genet. Genomics* 290, 1345–1356. doi: 10.1007/s00438-015-0998-6

Dick, T., Manjunatha, U., Kappes, B., and Gengenbacher, M. (2010). Vitamin B6 biosynthesis is essential for survival and virulence of *Mycobacterium tuberculosis*. *Mol. Microbiol.* 78, 980–988. doi: 10.1111/j.1365-2958.2010.07381.x

Eberhardt, S., Richter, G., Gimbel, W., Werner, T., and Bacher, A. (1996). Cloning, sequencing, mapping and hyperexpression of the ribC gene coding for riboflavin synthase of *Escherichia coli*. *Eur. J. Biochem.* 242, 712–719. doi: 10.1111/j.1432-1033.1996.0712r.x

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., et al. (2008). Genetics of gene expression and its effect on disease. *Nature* 452:423. doi: 10.1038/nature06758

Galardini, M., Brilli, M., Spini, G., Rossi, M., Roncaglia, B., Bani, A., et al. (2015). Evolution of intra-specific regulatory networks in a multipartite bacterial genome. *PLoS Comput. Biol.* 11:e1004478. doi: 10.1371/journal.pcbi.1004478

Gallegos, M. T., Michan, C., and Ramos, J. L. (1993). The XylS/AraC family of regulators. *Nucleic Acids Res.* 21, 807–810. doi: 10.1093/nar/21.4.807

Garces, F., Fernández, F. J., Gómez, A. M., Pérez-Luque, R., Campos, E., Prohens, R., et al. (2008). Quaternary structural transitions in the DeoR-type repressor UlaR control transcriptional readout from the L-ascorbate utilization regulon in *Escherichia coli*. *Biochemistry* 47, 11424–11433. doi: 10.1021/bi800748x

Gouzy, A., Poquet, Y., and Neyrolles, O. (2014). Nitrogen metabolism in *Mycobacterium tuberculosis* physiology and virulence. *Nat. Rev. Microbiol.* 12:729. doi: 10.1038/nrmicro3349

Horvath, S. (2011). *Weighted Network Analysis: Applications in Genomics and Systems Biology*. New York, NY: Springer Science & Business Media.

Hosseinkhan, N., Mousavian, Z., and Masoudi-Nejad, A. (2018). Comparison of gene co-expression networks in *Pseudomonas aeruginosa* and *Staphylococcus aureus* reveals conservation in some aspects of virulence. *Gene* 639, 1–10. doi: 10.1016/j.gene.2017.10.005

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Junker, B. H., and Schreiber, F. (Eds.). (2008). *Analysis of Biological Networks*. Vol. 2. Hoboken, NJ: Wiley-Interscience, 31–59. doi: 10.1002/9780470253489

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kelly, R. A., Leedale, J., Calleja, D., Enoch, S. J., Harrell, A., Chadwick, A. E., et al. (2019). Modelling changes in glutathione homeostasis as a function of quinone redox metabolism. *Sci. Rep.* 19:6333. doi: 10.1038/s41598-019-42799-2

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492

Largfelder, P., and Holvarth, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Liu, W., Li, L., Long, X., You, W., Zhong, Y., Wang, M., et al. (2018). Construction and analysis of gene co-expression networks in escherichia coli. *Cells* 7:19. doi: 10.3390/cells7030019

Lowe, E. C., Baslé, A., Czjzek, M., Firbank, S. J., and Bolam, D. N. (2012). A scissor blade-like closing mechanism implicated in transmembrane signaling in a bacteroides hybrid two-component system. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7298–7303. doi: 10.1073/pnas.1200479109

Maddocks, S. E., and Oyston, P. C. (2008). Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* 154, 3609–3623. doi: 10.1099/mic.0.2008/022772-0

McAdams, H. H., Srinivasan, B., and Arkin, A. P. (2004). The evolution of genetic regulatory systems in bacteria. *Nat. Rev. Genet.* 5, 169–178. doi: 10.1038/nrg1292

Medina, I. R., and Lubovac-Pilav, Z. (2016). Gene co-expression network analysis for identifying modules and functionally enriched pathways in type 1 diabetes. *PLoS ONE* 11:e0156006. doi: 10.1371/journal.pone.0156006

Moretto, M., Sonego, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeida, D., et al. (2016). COLOMBOS v3. 0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res.* 44, D620–D623. doi: 10.1093/nar/gkv1251

Mueller, A. J., Canty-Laird, E. G., Clegg, P. D., and Tew, S. R. (2017). Cross-species gene modules emerge from a systems biology approach to osteoarthritis. *NPJ Syst. Biol. Appl.* 3:13. doi: 10.1038/s41540-017-0014-3

Peregrín-Alvarez, J. M., Sanford, C., and Parkinson, J. (2009). The conservation and evolutionary modularity of metabolism. *Genome Biol.* 10:R63. doi: 10.1186/gb-2009-10-6-r63

Perez-Rueda, E., Hernandez-Guerrero, R., Martinez-Nuñez, M. A., Armenta-Medina, D., Sanchez, I., and Ibarra, J. A. (2018). Abundance, diversity and domain architecture variability in prokaryotic DNA-binding transcription factors. *PLoS ONE* 13:e0195332. doi: 10.1371/journal.pone.0195332

Ramos, J. L., Martínez-Bueno, M., Molina-Henares, A. J., Terán, W., Watanabe, K., Zhang, X., et al. (2005). The TetR family of transcriptional repressors. *Microbiol. Mol. Biol. Rev.* 69, 326–356. doi: 10.1128/MMBR.69.2.326-356.2005

Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O. (2003). An expanded genome-scale model of *Escherichia coli* K-12 (i JR904 GSM/GPR). *Genome Biol.* 4:R54. doi: 10.1186/gb-2003-4-9-r54

Rivera-Gómez, N., Martínez-Núñez, M. A., Pastor, N., Rodriguez-Vazquez, K., and Perez-Rueda, E. (2017). Dissecting the protein architecture of DNA-binding transcription factors in bacteria and archaea. *Microbiology* 163, 1167–1178. doi: 10.1099/mic.0.000504

Sangurdekar, D. P., Zhang, Z., and Khodursky, A. B. (2011). The association of DNA damage response and nucleotide level modulation with the antibacterial mechanism of the anti-folate drug trimethoprim. *BMC Genomics* 12:583. doi: 10.1186/1471-2164-12-583

Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., et al. (2019). The *Escherichia coli* transcriptome consists of independently regulated modules. *bioRxiv* 620799. doi: 10.1101/620799

Skerker, J. M., Prasol, M. S., Perchuk, B. S., Biondi, E. G., and Laub, M. T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. *PLoS Biol.* 3:e334. doi: 10.1371/journal.pbio.0030334

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2010). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* 13:328. doi: 10.1186/1471-2105-13-328

Steinmetzer, K., Behlke, J., Brantl, S., and Lorenz, M. (2002). CopR binds and bends its target DNA: a footprinting and fluorescence resonance energy transfer study. *Nucleic Acids Res.* 30, 2052–2060. doi: 10.1093/nar/30.9.2052

Stuart, J., Segal, E., Koller, D., and Stuart, K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255. doi: 10.1126/science.1087447

Taira, K., Kaneto, S., Nakano, K., Watanabe, S., Takahashi, E., Arimoto, S., et al. (2013). Distinct pathways for repairing mutagenic lesions induced by methylating and ethylating agents. *Mutagenesis* 28, 341–350. doi: 10.1093/mutage/get010

Trewavas, A. (2006). A brief history of systems biology. *Plant Cell* 18, 2420–2430. doi: 10.1105/tpc.106.042267

Tsaparas, P., Marino-Ramirez, L., Bodenreider, O., Koonin, E. V., and Jordan, K. (2006). Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evol. Biol.* 6:70. doi: 10.1186/1471-2148-6-70

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinformatics* 19, 575–592. doi: 10.1093/bib/bbw139

Van Noort, V., Snel, B., and Huynen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* 5, 280–284. doi: 10.1038/sj.embor.7400090

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* 5:3231. doi: 10.1038/ncomms4231

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:Article17. doi: 10.2202/1544-6115.1128

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 10 January 2020
doi: 10.3389/fmolb.2019.00155

Integration of Biological Networks for *Acidithiobacillus thiooxidans* Describes a Modular Gene Regulatory Organization of Bioleaching Pathways

María Paz Cortés[1,2†], *Vicente Acuña*[1†], *Dante Travisany*[1,2], *Anne Siegel*[3,4], *Alejandro Maass*[1,2,5*] and *Mauricio Latorre*[1,2,6,7*]

[1] Center for Mathematical Modeling, Universidad de Chile and UMI CNRS 2807, Santiago, Chile, [2] Center for Genome Regulation, Universidad de Chile, Santiago, Chile, [3] IRISA, UMR 6074, CNRS, Rennes, France, [4] INRIA, Dyliss Team, Centre Rennes-Bretagne-Atlantique, Rennes, France, [5] Department of Mathematical Engineering, Universidad de Chile, Santiago, Chile, [6] Laboratorio de Bioinformática y Expresión Génica, INTA, Universidad de Chile, Santiago, Chile, [7] Instituto de Ciencias de la Ingeniería, Universidad de O'Higgins, Rancagua, Chile

OPEN ACCESS

Edited by:
Alberto Jesus Martin,
Centro de Genómica y Bioinformática,
Universidad Mayor, Chile

Reviewed by:
Mark Dopson,
Linnaeus University, Sweden
Marc Delarue,
Institut Pasteur, France

***Correspondence:**
Alejandro Maass
amaass@dim.uchile.cl
Mauricio Latorre
mauricio.latorre@uoh.cl

†These authors have contributed equally to this work

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: 01 September 2019
Accepted: 13 December 2019
Published: 10 January 2020

Citation:
Cortés MP, Acuña V, Travisany D,
Siegel A, Maass A and Latorre M
(2020) Integration of Biological
Networks for Acidithiobacillus
thiooxidans Describes a Modular
Gene Regulatory Organization of
Bioleaching Pathways.
Front. Mol. Biosci. 6:155.
doi: 10.3389/fmolb.2019.00155

Acidithiobacillus thiooxidans is one of the most studied biomining species, highlighting its ability to oxidize reduced inorganic sulfur compounds, coupled with its elevated capacity to live under an elevated concentration of heavy metals. In this work, using an *in silico* semi-automatic genome scale approach, two biological networks for *A. thiooxidans* Licanantay were generated: (i) An affinity transcriptional regulatory network composed of 42 regulatory family genes and 1,501 operons (57% genome coverage) linked through 2,646 putative DNA binding sites (arcs), (ii) A metabolic network reconstruction made of 523 genes and 1,203 reactions (22 pathways related to biomining processes). Through the identification of confident connections between both networks (V-shapes), it was possible to identify a sub-network of transcriptional factor (34 regulators) regulating genes (61 operons) encoding for proteins involved in biomining-related pathways. Network analysis suggested that transcriptional regulation of biomining genes is organized into different modules. The topological parameters showed a high hierarchical organization by levels inside this network (14 layers), highlighting transcription factors CysB, LysR, and IHF as complex modules with high degree and number of controlled pathways. In addition, it was possible to identify transcription factor modules named primary regulators (not controlled by other regulators in the sub-network). Inside this group, CysB was the main module involved in gene regulation of several bioleaching processes. In particular, metabolic processes related to energy metabolism (such as sulfur metabolism) showed a complex integrated regulation, where different primary regulators controlled several genes. In contrast, pathways involved in iron homeostasis and oxidative stress damage are mainly regulated by unique primary regulators, conferring Licanantay an efficient, and specific metal resistance response. This work shows new evidence in terms of transcriptional regulation at a systems level and broadens the study of bioleaching in *A. thiooxidans* species.

Keywords: *Acidithiobacillus thiooxidans*, biological networks, co-regulation, bioleaching, biotechnology

INTRODUCTION

Acidithiobacillus thiooxidans belongs to the *Acidithiobacillia* class of proteobacteria (Williams and Kelly, 2013). It is an autotrophic Gram-negative bacterium that obtains energy from the oxidation of reduced inorganic sulfur compounds (RISC). *Acidithiobacillus thiooxidans* capacity to produce sulfuric acid, especially during the control of biochemical steps related to elemental sulfur oxidation pathways and the acidification of the media (Mohapatra et al., 2008) have positioned this bacterium as one of the most studied organism in the field of bioleaching processes (Chen et al., 2015; Yan et al., 2015; Quatrini et al., 2017; Zhou et al., 2017).

Recently, *A. thiooxidans* Licanantay was presented as one of the most relevant participants of a consortium of five natural copper-bioleaching acidophilic bacteria (Latorre et al., 2016). This bacterium was isolated directly from a copper mine in the north of Chile. Its genome sequence revealed an elevated number of genes associated with RISC oxidation: several HDR complex genes, two gene copies for the sulfur oxidizing complex (Sox) and one archaeal type sulfur oxygenase reductase gene (*sor*) (Travisany et al., 2014), attributes directly correlated with its efficiency in copper recovery. In addition, Licanantay has an elevated capacity to survive under elevated concentrations of copper, arsenic, and chloride in relation to other biomining species and produces high quantities of glutathione (Martínez et al., 2013), a crucial metabolite directly or indirectly related to iron and RISC oxidation in bioleaching species.

A complete genome comparative analysis between nine draft genomes of *A. thiooxidans* postulates that the genetic diversity of this species might be correlated with geographic location and geochemical conditions (Zhang et al., 2016). In this study, the comparison between Licanantay and the reference strain AT19377 reaffirms the fact that the Chilean bacterium has a higher number of unique genes, which may confer an adaptive advantage to extreme environmental conditions for Licanantay compared to other *A. thiooxidans* strains.

In addition, a set of environmental resistance elements and metabolic pathways presumed relevant to its performance in bioleaching processes have been assigned to this bacterium, most of them related to the oxidation of RISC, metal resistance, biofilm formation, and energy production (Latorre et al., 2016). These results position *A. thiooxidans* Licanantay as an excellent model to study genomic and metabolic features in terms of gene regulation and metabolic pathways related to the adaptation of this bacterium to the environment of a copper mine.

Using bioinformatics tools in combination with a manual curation of regulatory patterns, a great amount of information can be extracted from the genome sequence and further summarized in an affinity transcriptional regulatory network (Balleza et al., 2008). These models depict the total set of statistically significant affinity relations between annotated transcription factors and their binding sites in promoter regions of operons. It is important to remark that this affinity relation does not necessarily imply that the regulatory relation is effectively used for a given set of conditions. Indeed, the regulatory process also depends on other factors that vary depending on the conditions imposed on the cell, and only expression experiments can confirm such relation (Potash, 2007). However, the strategy of generating affinity networks has been widely used in bacterial organisms as a starting point to identify a global regulatory organization. Affinity networks provide relevant information about the topological configuration of gene regulation at a system level and allows the importance of specific regulatory elements and its putative gene/operons targets to be identified (Balázsi et al., 2008; Latorre et al., 2014; Yus et al., 2019).

On the other hand, the study of a metabolic network is key to gaining insight regarding phenotypic features of an organism. The reconstruction of metabolic networks at the genome scale, i.e., incorporating all available information, allows us to have a global, and comprehensive picture of metabolism. These genome-scale reconstructions are considered specific knowledge repositories of studied organisms where information regarding their metabolism is organized and new data can be later integrated (Feist et al., 2009). This can be particularly useful to guide and contribute to the systematic study of less-studied organisms, as is the case of biomining organisms in general.

In this work, using a systems biology approach, genome-scale metabolic, and regulatory networks were integrated. The main objective of this article was to generate information on the transcriptional mechanism able to control the expression of elements involved in metabolic pathways related to bioleaching in *A. thiooxidans* Licanantay. To this end, the minimal configuration able to maintain bacterial-relevant functions was described and we showed that this gene regulatory organization strongly depended on different types of modules.

RESULTS AND DISCUSSION

A. thiooxidans Licanantay Affinity Transcriptional Regulatory Network

In order to understand the global transcriptional regulatory organization in *A. thiooxidans* Licanantay, a genome-scale affinity transcriptional regulatory network was generated. The complete model had a genome coverage of 57% and was composed of 1,543 nodes (42 corresponding to transcriptional factor nodes) and 2,646 arcs (putative binding sites) (**Figure 1**). The degree distributions (in-degree, out-degree, and total degree) showed a typical shape in which most nodes have a low degree and only a few nodes are highly connected (Albert, 2005). This characteristic is typical in power low distributions observed in other bacterial transcriptional regulatory networks. In terms of the interconnectivity between transcriptional factors, the network model contains at least three types of regulators (Schröder and Tauch, 2010). First, a global set of regulators, like LysR (global metabolism), and IHF (DNA structural organization), which are highly interconnected in the network. As shown in **Figure 1**, these two regulators present a multi-level regulation cascade structure, representative of a classical chain transcriptional regulatory process. Second, a set of master regulators (moderately connected), such as AtoC (acetoacetate metabolism) and MerR (metal resistance). Acetoacetate was

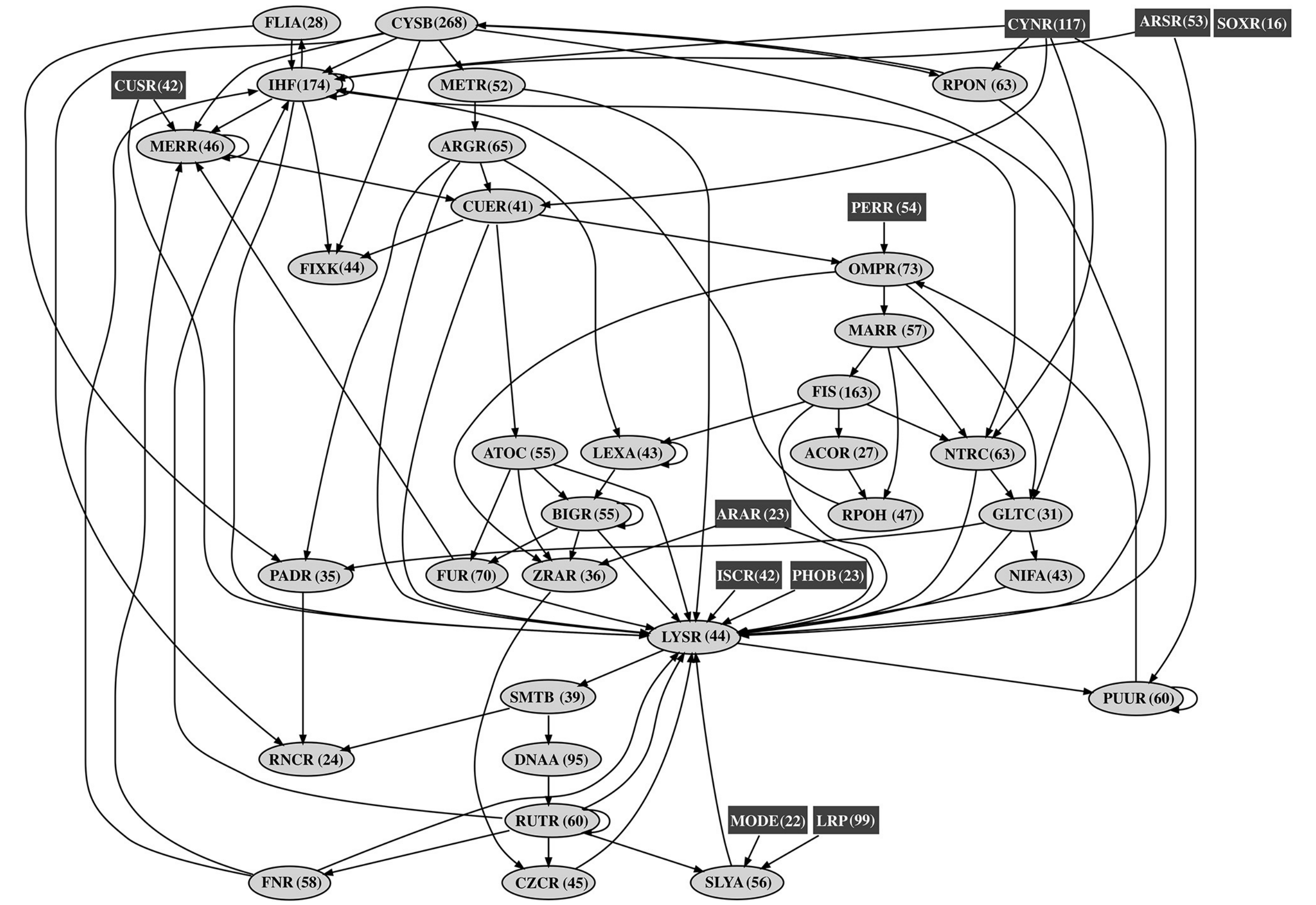

FIGURE 1 | *Acidithiobacillus thiooxidans* Licanantay affinity transcriptional regulatory network. The figure shows the interconnectivity (black arrows) between transcriptional factor nodes. Rectangular nodes (dark gray) correspond to transcriptional factors not regulated by others (origons). Oval nodes (light gray) represent transcriptional factors member of chain regulatory cascades. The number in parenthesis next to each transcriptional factor name is the number of operon targets for that transcriptional factor in the affinity network.

identified as a biofilm inhibitor (Horne et al., 2018), an important bacterial process during ore bioleaching (Bellenberg et al., 2014). The MerR family is highly conserved in other biomining organisms, including strains of *At. ferrooxidans* (Hödar et al., 2012). Finally, the third class corresponded to local regulators, highlighting the proteins Fur and CueR (also a MerR family member), controllers of metal homeostasis and oxidative stress damage, two main cellular processes considering the mining environment where Licanantay was isolated (Latorre et al., 2016), during oxidative dissolution, autotrophic organisms are able to use ferrous iron and reduced sulfur compounds as electron donors. In addition, the model contained a total of 10 regulators not controlled by other transcription factors (isolated). This type of element is called Origons (Balázsi et al., 2005) and represents topological units of environmental signal processing, able to directly transduce stimulus into gene expression control (direct and fast response). Inside this group, Licanantay had the CusR transcription factor, one of the main regulators of copper homeostasis in Gram-negative bacteria (Rensing and Franke, 2007). Considering the elevated concentration of copper in the mine, the presence of the CusR origon gives Licanantay an efficient and fast control over copper homeostasis, in particular over the expression of CopA ATPase involved in this metal efflux (Solioz and Stoyanov, 2003). Finally, the node with the highest out-degree (275) was CysB, making it one of the main Hubs inside the network. This regulator is known as a master regulator of genes encoding for proteins involved in sulfur metabolism, particularly, its assimilation (van der Ploeg et al., 2001) and also iron starvation (Imperi et al., 2010). For *A. thiooxidans* species, sulfur metabolism plays a crucial role in the acquisition of electrons for their autotrophic growth (Wang et al., 2018).

A. thiooxidans Licanantay Metabolic Network

Efforts have been made to reconstruct the metabolic networks for a few bioleaching bacteria (Hold et al., 2009; Merino et al., 2010; Bobadilla Fazzini et al., 2013; Merino Santis et al., 2015). These reconstructions were developed with the objective of generating metabolic models that allow the prediction of growth rates in different scenarios through metabolic flux analysis. With the exception of *At. ferrooxidans* ATCC23270, for which a genome-scale reconstruction was built (Campodonico et al., 2016), bioleaching bacterial metabolic reconstructions corresponded to reduced and simplified representations of their networks in all cases. This was also the case for *A. thiooxidans* Licanantay, for which we previously built a small stoichiometric model used to predict its growth rate in different media containing different reduced sulfur compounds for oxidation (Bobadilla Fazzini et al., 2013). This model incorporates a total of 181 metabolic reactions associated with RISC oxidation, central metabolism, amino acids, and nucleotides biosynthesis pathways.

For the work presented here, we revisited the analysis of *A. thiooxidans* Licanantay metabolic network, this time aiming at a global genome-scale reconstruction in order to later link metabolic genes through the regulatory network of the bacteria. To do this, we followed a semi-automatic approach starting by a full genome re-annotation in order to make the most of the available data. This new annotation resulted in 564 unique Enzyme Commission (EC) numbers, 20% of which were absent from our previous annotation and an improved annotation of 81 genes previously identified as hypothetical protein coding genes.

Our current genome-scale metabolic network reconstruction was made of 1,203 reactions, associated with 523 genes coding for enzymes and transport proteins. This reconstruction included all enzymatic reactions incorporated in the previous stoichiometric model as well as additional relevant reactions and pathways, e.g., the biosynthesis of spermidine, a metabolite that has been linked to sulfur-oxidation in a previous metabolomic study on this bacterium (Martínez et al., 2013) as well as to pH homeostasis and oxidative stress management (Samartzidou et al., 2003; Ferrer et al., 2016).

Interestingly, sulfur metabolism and siderophore biosynthesis were both highly connected in the metabolic network. Sulfur metabolism is directly connected to the capacity to produce cysteine in bacterial species. This amino acid can be used to synthetize Fe-S clusters, the principal co-factor of the HDR complex. Competition for iron occurs in acidic environments, where the capacity to produce and recognized different siderophores could be an adaption to respond to different iron concentrations (Bonnefoy and Holmes, 2012). In addition, biofilm processes in the model are related to routes involved in lysine degradation. This amino acid inhibits coaggregation and synergy in biofilm formation (Sharma et al., 2005; Okuda et al., 2012). The capacity of biomining organisms to produce biofilms is one of the critical and most studied areas of bioleaching. The active presence of lysine degradation pathways in Licanantay, supports the high capacity of this bacterium to recover copper during the process.

In previous work, a number of metabolic processes were linked to the bioleaching capacity of a bacterial consortium that has *A. thiooxidans* Licanantay as one of its members (Latorre et al., 2016). These processes included known key bioleaching steps, such as iron and RISC oxidation as well as related metabolic features such as sulfur assimilation, biosynthesis of essential components and precursors, electron transfer and energy generation, and biofilm formation.

The next step in the current study, was to consider a subset of these metabolic categories to focus our analysis of *A. thiooxidans* regulation on particularly relevant processes related to bioleaching that could be subject to co-regulation. This subset was composed of six sub-categories, selected because they corresponded to well-described bioleaching metabolic pathways part of the *A. thiooxidans* metabolic network. **Figure 2** shows these pathways in the context of the *A. thiooxidans* global metabolic network. They corresponded to RISC oxidation, sulfur assimilation, heme, NAD, and spermidine biosynthesis processes.

RISC oxidation (orange pathways in **Figure 2**) by sulfur-oxidizing bacteria such as *A. thiooxidans* is key for bioleaching operations. It results in the release of sulfuric acid which helps maintain the acidic condition required for bioleaching to occur. RISC is the only electron donors utilized by *A. thiooxidans.* Thus, sulfur oxidation was strongly linked to general metabolic pathways related to energy generation (depicted in red in **Figure 2**) that involve steps to harness energy through proton gradient and reducing power generation. Given that NAD(H) is a main reducing power carrier, its biosynthesis pathway was also considered in this analysis (blue pathway in **Figure 2**).

For sulfur oxidizers, as has been previously pointed out for *At. ferrooxidans* (Valdés et al., 2003), a balance should take place in the use of sulfur as an energy source and in the assimilation processes. Moreover, RISC oxidation is a complex process whose associated metabolic pathways have not being fully elucidated to date. Different pathways have been proposed for the *Acidithiobacillus* species (Wang et al., 2018) including a pathway that involves the assimilation enzymes APS kinase and PAPS reductase (Yin et al., 2014). Based on these considerations sulfur assimilation metabolic pathways were also considered in this analysis (pink pathways in **Figure 2**). Spermidine biosynthesis depicted in light blue in **Figure 2**, was also included given the previously mentioned link of this metabolite to sulfur oxidation.

Finally, the biosynthesis of heme was also included in this pathway selection (green pathways in **Figure 2**). Heme is an essential component of several proteins involved in electron transport chains which are key for *A. thiooxidans* energy generation. Heme is also a cofactor of enzymes involved in oxidative damage protection (Frankenberg et al., 2003). Minerals, which are abundant in bioleaching environments, are known to promote the formation of ROS species (Schoonen et al., 2006; Cárdenas et al., 2012), making protection mechanisms against them essential for *A. thiooxidans* survival. Additionally, as an iron-containing cofactor heme plays a role in iron homeostasis.

Co-regulatory Integrative Network Analysis

As stated above, the affinity transcriptional network represents the set of all transcriptional regulatory relations between all transcription factors annotated in the genome and their putative target operons. Each relation was represented in the network by a directed arc from the operon coding for the regulator to the target operon (which can also code for another regulator). Thus, indirect regulation of the bioleaching sub-category via regulatory cascades can be defined as paths in the network.

It is important to declare that the set of arcs in the affinity transcriptional network is considered as an overrepresentation of

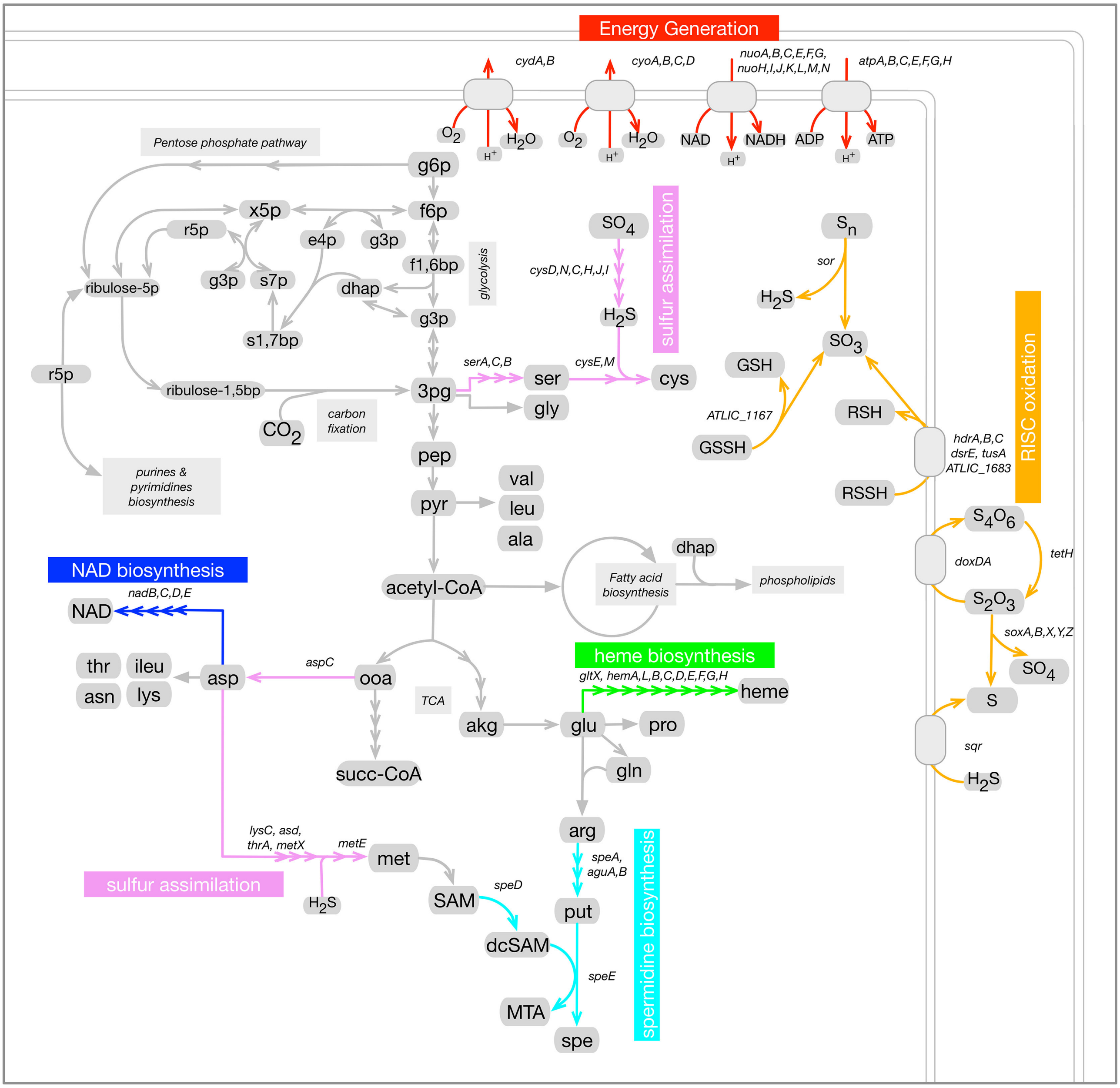

FIGURE 2 | Selected metabolic pathways related to bioleaching in the context of *A. thiooxidans* Licanantay metabolic network (sub-categories). Six metabolic processes were selected for this study: RISC oxidation (orange); Sulfur assimilation (violet); Energy generation (red); heme biosynthesis (green); spermidine biosynthesis (cyan); and NAD biosynthesis (blue). Genes associated to each of these pathways are listed next to the corresponding reactions.

the true transcriptional regulations occurring in the bacterium (Acuña et al., 2016). There are two main reasons for this overrepresentation. The first is purely methodological: some of the relations are simply false positives of the method that identifies binding sites. The second one is biological: even in the case that the transcription factors could effectively bind in a promoter region of a specific operon, bacteria only activate or repress this regulatory mechanism as required according to environmental conditions.

Considering these two statements, in order to give a new layer of likelihood to the regulatory relations effectively occurring in Licanantay for a given set of conditions, a method that selects feasible paths (i.e., regulatory cascades) was applied (Acuña et al., 2016). Under a parsimony principle, the method considers an arc between a transcription factor and its operon target as confident when it is part of a topological substructure (called V-shape) which is useful to coordinate the co-expression of operons in the same pathway.

Using this method, sets of confident regulatory relations to coordinate the co-expression of operons in each one of the six selected metabolic sub-categories related to bioleaching processes were identified (see Materials and Methods for details). The union of these subnetworks can be considered a (transcriptional) "co-regulatory network" for bioleaching processes. This co-regulatory network is presented in **Figure 3**. The network was composed of 95 nodes, 34 of which were transcription factor families and 61 were metabolic operons from the bioleaching sub-categories, and 148 arcs, in which 57 corresponded to regulations between transcription factors and 91 to regulations of metabolic operons.

A first topological analysis of this network showed that it was almost hierarchical, presenting only a very small number of cycles (three in total), a classical organization for biological networks, which confer a directional regulatory collaboration between the transcription factors to the system. In fact, removing only two arcs broke every directed cycle in the network (i.e., the minimum feedback arc set had a size of two). This fact showed that the network can be organized into 14 levels with only two feedback arcs, as showed in **Figure 3**. The order of levels depended on which arcs were considered feedback arcs, and the figure represents only one possible configuration of these levels.

Even if the co-regulatory network was hierarchical (i.e., with almost no directed cycles inside), its organization was far from being like a tree graph. Indeed, if we consider only the 57 arcs between transcription factors, the number of them that need to be removed to obtain a tree was 24 or 42.1%. This means that the metabolic consequences of the regulation were not segregated by network level, having many arcs that cross from one branch to another or that jump directly to a distant level. This observation is also consistent with the fact that the regulation of metabolic operons of each sub-category was not separated by the hierarchy and, on the contrary, were spread along the entire network. This implies that the regulation process cannot be viewed as a sum of independent parts controlled by a central mechanism, but rather as a complex regulation of the different connected modules inside the network.

Regulatory Bioleaching Modules

As shown in **Figure 3**, there were eight transcription factors (dark-gray boxes) that were not controlled by any other regulator inside the co-regulatory network. We referred to them as *primary regulators* in our model, representing the first modular structure inside the network. For each primary regulator, its potential to regulate metabolic operons of each bioleaching-related pathway, either by direct binding or by chain regulatory cascades, was computed (**Table 1**). We found no clear exclusive distribution between primary regulators and metabolic bioleaching sub-categories. Most of these transcription factors regulate at least one operon of each sub-category, and, with the exception of primary regulators CysB and CynR and the energy generation sub-category, most primary regulators could not regulate an entire sub-category. These results reinforce the idea that regulation of bioleaching metabolism has a complex modular structure, where an important part of operons related to bioleaching processes are transcriptionally controlled by different primary regulators.

In order to examine this point in depth, we analyze whether single primary regulators have some specific property to control an exclusive sub-category. Thus, for each sub-category, we computed the number of operons that were controlled exclusively by only one primary regulator (**Table 2**). The percentage of operons exclusively regulated by one primary regulator was slightly greater for the biosynthesis sub-categories (NAD, heme, and spermidine). In contrast, sulfur assimilation and RISC oxidation processes were mostly composed by operons regulated by two or more primary regulators. This observation suggests that metabolic processes related to bioleaching in Licanantay present a bias in the transcriptional regulation according to specific sub-categories. It was possible to identify at least two regulatory modules. The first one was composed of specific primary regulators controlling an important number of operons related to biosynthesis of molecules involved mainly in redox reactions (NAD) (Gazzaniga et al., 2009) and iron homeostasis (heme and spermidine) (Bergeron et al., 2001; Quatrini et al., 2005; Richard et al., 2019). Considering the elevated capacity to tolerate high amounts of metals and oxidative stress damage by Licanantay (Latorre et al., 2016), the presence of unique primary regulators controlling gene expression of resistance mechanisms, provides the system with a fast and efficient transcriptional activation. The second module was comprised of several primary regulators controlling the expression of different energy related processes. This configuration suggests an important regulatory redundancy inside this module. The involvement of different primary regulators grants the system alternatives to produce or consume energy, ensuring the correct functioning of Licanantay. This contrasts with the previous module of biosynthesis of resistance-related molecules, which was highly specific in terms of transcriptional and metabolic response.

In most cases, primary regulators are not able to control an entire bioleaching sub-category. Thus, an analysis of the regulatory capacity of small subsets of regulators was performed in order to determine the degree of regulation specificity of each sub-category. This was done by computing sets of minimum transcription factors able to control an entire sub-category (**Supplementary Table 1**). Results showed that, except for spermidine biosynthesis, there was a unique minimum set of primary regulators for each sub-category. This analysis also highlighted the transcriptional factor CysB as the most represented regulator connecting genes involved in the selected bioleaching metabolic pathways, appearing in almost all the six sub-categories inside the group of minimal regulators.

In addition to the analysis of primary regulators inside the network, it was also possible to classify transcription factors according to number of connections. Three regulators stand-out due to their elevated number of connections: CysB with out-degree 14, LysR with in-degree 9, and IHF with a total degree of 16 (in-degree 6 and out-degree 10). These transcription factors had an affinity with operons in different metabolic bioleaching sub-categories. As mentioned, CysB directly controlled 4 of them, IHF 5 and LYSR 3. Thus, CysB, IHF, and LysR were considered complex regulator modules. In addition, *ihf* and *lysR* genes were controlled by other 6 and 9 transcription factors, respectively. On the other hand, IHF regulated three other regulators (being one of them LysR) while LysR could also regulated PuuR. These

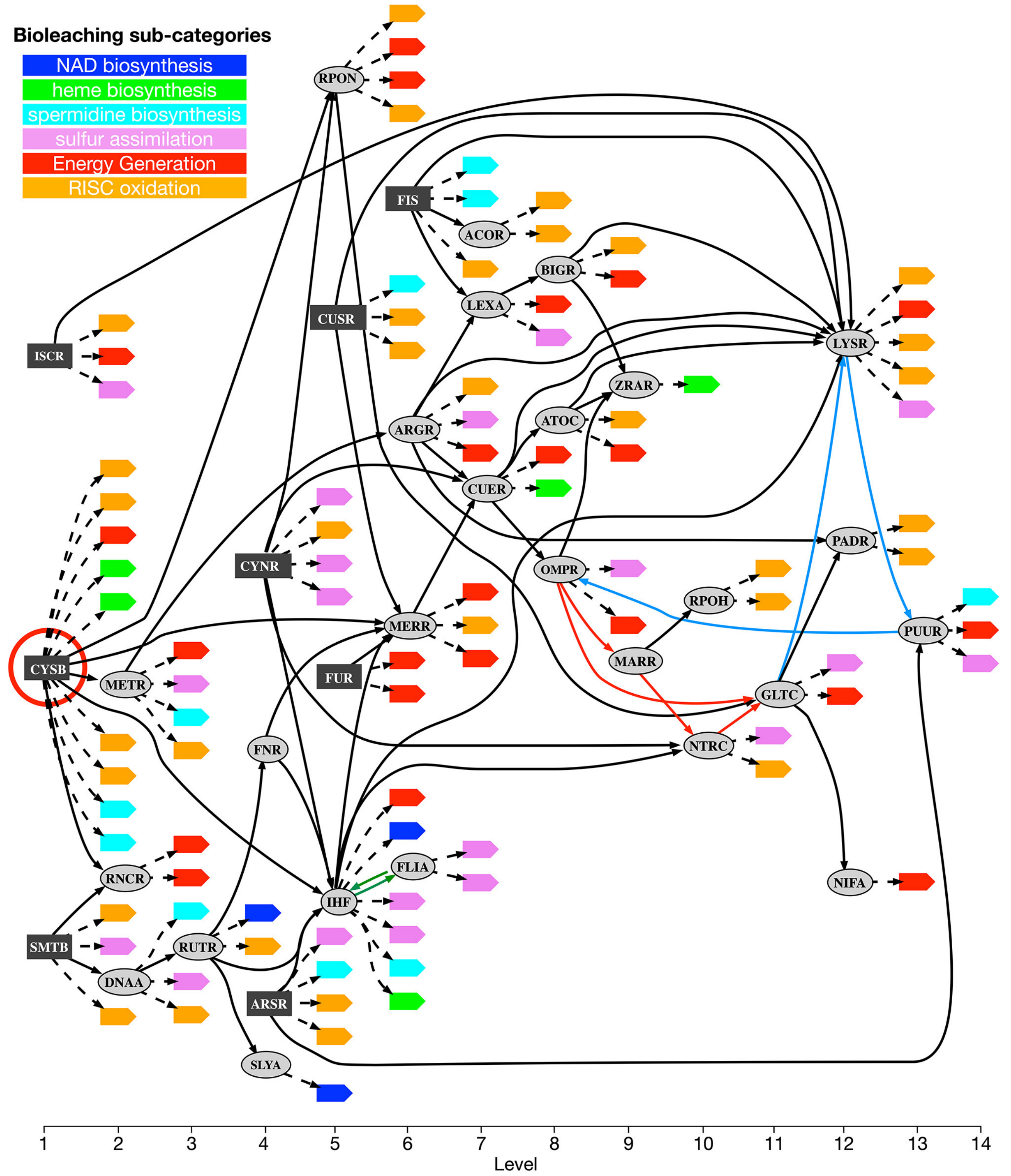

FIGURE 3 | *Acidithiobacillus thiooxidans* Licanantay bioleaching co-regulatory network. Transcriptional factors in the co-regulatory network are depicted as rectangular (dark gray) and oval nodes (light gray). Rectangular nodes correspond to primary regulators while oval nodes are transcriptional factors member of chain regulatory cascades. Leaf nodes are target operons colored according to their metabolic bioleaching sub-categories. Solid arcs represent regulation between transcriptional factors and dotted arcs represent regulation of metabolic operons. Hierarchical levels are listed at the bottom of the figure. Red circle highlights CysB transcription factor. Colored arcs (red, green, and light blue) correspond to connections forming directed cycles in the network. There are three directed cycles: a small one between FLIA and IHF (green arcs) and two larger ones that share the path made by light-blue arcs. Thus, removing any green and light blue pair of arcs breaks all directed cycles (minimum feedback arc sets).

TABLE 1 | Transcriptional regulatory representation of each primary regulator over the metabolic bioleaching sub-category.

	Operons	CysB	SMTB	CynR	ArsR	CusR	Fur	Fis	IscR
NAD biosynthesis	3	1	3	1	1	0	0	0	0
Heme biosynthesis	5	5	3	3	3	2	2	1	1
Spermidine biosynthesis	7	5	3	2	3	2	1	2	1
Sulfur assimilation	11	9	8	7	7	4	4	5	5
Energy generation	13	13	10	10	9	9	9	7	5
RISC oxidation	22	17	13	13	12	12	10	11	8
TOTAL	61	50	40	36	35	29	26	26	20

Column Operons corresponds to the total number of operons for each metabolic bioleaching sub-category. The rest of the columns show the number of operons regulated by each listed primary regulator (transcription factor).

TABLE 2 | Total number of operons for each metabolic bioleaching sub-category regulated by only one primary regulator.

	Operons	Total	CysB	SMTB	CynR	ArsR	CusR	Fur	Fis	IscR
NAD biosynthesis	3	2 (67%)	0	2	0	0	0	0	0	0
Heme biosynthesis	5	2 (40%)	2	0	0	0	0	0	0	0
Spermidine biosynthesis	7	3 (43%)	2	0	0	0	1	0	0	0
Sulfur assimilation	11	3 (27%)	1	1	0	0	0	0	0	1
Energy generation	13	1 (8%)	1	0	0	0	0	0	0	0
RISC oxidation	22	7 (32%)	4	1	0	0	1	0	1	0

Total percentage was calculated as the proportion between the operon regulated by only one primary regulator and the total number of operons belonging to each sub-category.

two regulators have been widely studied in other bacteria species, both controlling central metabolism and general processes in the cell (Schell, 1993; Lynch et al., 2003). While IHF and LysR had a high connectivity in the network, neither were central to maintaining the connectivity of the network (not belonging to minimal sets of regulators). On the contrary, the transcriptional regulator CysB in all the topological studies made, and under all the different analyses was a fundamental module in the transcriptional regulation of Licanantay. As showed in **Figure 3**, CysB also belonged to the first hierarchical level of organization inside the co-regulatory network, positioning this regulator as the main module involved in bioleaching processes.

CONCLUSION

The availability of genome sequences has opened an interesting field in systems biology to study global gene regulatory organization in bacteria of biotechnological interest. Through the identification of sets of confident regulatory relations, the integration of information from two biological network was achieved: (i) affinity transcriptional regulatory network and (ii) metabolic network. As a result, the first co-regulatory network model describing the global transcriptional regulation of different bioleaching metabolic pathways in the bacterium *A. thiooxidans* Licanantay was generated. The topological analysis of the network indicates that the global transcriptional regulation is a result of the combination of different specific modules. The first type corresponds to primary regulators (transcription factors not controlled by another regulator). Inside this group, CysB appeared as the most relevant module inside the network, also classified as the most represented primary regulator controlling a huge part of the network. Another two types of modules were identified in terms of bioleaching pathway regulation distribution. Metabolic processes involved in energy production demonstrated a complex integrated regulation, where different primary regulators controlled the expression of several genes (complex modules). In contrast, bioleaching pathways related to metal homeostasis and oxidative stress damage were mainly regulated by unique primary regulators (individual modules). The presence of both modules showed that at least two types of regulation were present in the bioleaching bacteria. Complex modules provide a wide set of alternatives related to energy requirements of the network. Individual modules on the other hand, highlight an efficient and specific metal resistance capacity to survive under the extreme environmental condition present in mines. These results bring us closer to having an complete view of *A. thiooxidans* metabolism and regulation. Moving forward and applying systems biology methodologies to the study of additional key bioleaching bacteria can inform and aid the rational design of effective biomining consortia for bioleaching processes. Finally, this integrative systems biology strategy should not be restricted to biomining related bacteria, but can also be applied to other sequenced bacterial genomes to construct new co-regulatory networks.

MATERIALS AND METHODS

Genome Annotation and Metabolic Network Reconstruction

Coding sequences (CDSs) previously identified in *A. thiooxidans* Licanantay draft genome (Travisany et al., 2014) were re-annotated. This was done through Blast searches against the nr, KEGG, UNIPROT, and COG databases. A gbk file with this new annotation was generated and used as the input to generate an automatic metabolic network reconstruction using Pathway-Tools v 21.0 (Karp et al., 2002). The cutoff score used for pathway prediction was 0.4. Additionally, a set of metabolic pathways previously suggested as relevant in bioleaching processes (Latorre et al., 2016) as well as central metabolism pathways were manually curated (**Supplementary Table 2**).

Affinity Transcriptional Regulatory Network

The generation of the Affinity transcriptional regulatory network was based on previously reported protocols (Latorre et al., 2014; DebRoy et al., 2016). Briefly, candidate transcription factors present in the genome of *A. thiooxidans* Licanantay were identified using the following protocol: First, using the results of genome annotation, each candidate must have at least a Helix-Turn-Helix domain, previously identified using HMMER software with Pfam database. Then, when information from UniProt-KB was available, amino acids in specific locations were manually searched (**Supplementary Table 3**). A position-specific scoring matrix (PSSM) was associated with each transcription factor candidate that fulfilled the previous requirements (**Supplementary Table 4**). Specifically, a Regprecise (Novichkov et al., 2013) or Prodoric (Münch et al., 2003) PSSM was downloaded or generated using MEME (Bailey et al., 2009) with promoter consensus sequences.

Operons and intergenic regions were retrieved from previous research (Travisany et al., 2014) in the following manner: an operon was defined as a cluster of co-regulated consecutive genes that share the same direction such that the maximum intergenic region between two consecutive genes contained <50 bp. Any region larger than 50 bp was considered a putative promoter intergenic region.

In order to identify putative binding sites for the transcription factor candidates, affinity relations between candidates and promoter intergenic sequences were obtained. To that end, individual occurrences of the associated PSSM motifs in the promoter intergenic sequences were computed using FIMO (Grant et al., 2011). Thus, an affinity relation between a transcription factor candidate and an operon was defined when at least one match ($p \leq 1e^{-5}$) of the PSSM associated with the transcription factor was obtained in the correspondent promoter region of the operon.

The affinity network is a directed graph that encompasses the set of all affinity relations. In this network there are two types of nodes: (a) transcription factor nodes that correspond to families of transcription factors and (b) operon nodes which are operons being regulated by transcription factors. Note that when there are several genes coding for transcription factors of the same family, there is only a single node representing the family in the affinity network. There are also two types of arcs: (a) arcs from a transcription factors node to an operon node, which indicate that there is an affinity relation between the transcription factors and the operon; and (b) arcs between two transcription factors nodes A and B, which indicate that there is an affinity relation between transcription factor A and an operon containing a gene which codes for a transcription factor B.

Co-regulation Network

A set of transcriptional regulations was defined as a regulatory mechanism that coordinates the expression of operons related to bioleaching in *A. thiooxidans* Licanantay.

To obtain this set of regulations, metabolic operons from six pathways previously associated with bioleaching were selected (Latorre et al., 2016). These pathways are NAD Biosynthesis, Heme Biosynthesis, Spermidine Biosynthesis, Sulfur assimilation, Energy generation, and RISC oxidation. Operons from these pathways that are regulated by at least one transcription factor were identified in the affinity network.

Under the assumption that the co-expression of operons belonging to the same metabolic pathway must be coordinated by a common factor (by directly regulating their expression or by regulatory cascades of transcription factors) and using a methodology that maximizes parsimony, arcs that are likely part of this co-regulation were selected from the affinity network (Acuña et al., 2016). Following this methodology, arcs in the affinity network were classified in four groups of the same size according to the *p*-value computed for the corresponding binding (between the transcription factor and the operon target). Then, weights of 1, 2, 4, and 8 were associated with arcs in each one of the four categories (a weight of 1 was given to arcs with lower *p*-values and a weight of eight to arcs with higher *p*-values).

To find common regulators of bioleaching related operons, the concept of a V-shape was used, which has been previously defined (Acuña et al., 2016). A V-shape is a subgraph that connects two given nodes (A and B) in a graph. It is composed of the union of two directed paths ending, respectively at A and B and starting at some node C with no other node in common. If a V-shape exists that connects A and B, then its starting point is a common regulator candidate. If more than one V-shape exists, then a parsimonious solution should be a combination of selecting a V-shape that uses less arcs and a V-shape having arcs with the smallest *p*-values. A way to consider both criteria is to consider V-shapes of minimum total weight (considering the total weight of a V-shape as the sum of the weight of its arcs).

According to the method explained, the set of all minimum weight V-shapes connecting two metabolic operons of the same metabolic category was computed involving 498 affinity relations, all of them having an original $p < 0.00010$. A histogram of the *p*-value obtained for the 498 affinity relations showed a decreasing tendency in the interval [0–0.00008] and a high peak in the interval [0.00008–0.00010]. In order to assure a confident set of relations, we applied a correction by removing arcs with an original $p > 0.00009$ from the network. As a result, a set of 457 arcs was selected from the affinity network, involving 63 operons coding for transcription factors, and 91 operons coding for metabolic genes. Operons coding for transcription factors were annotated according to the family of transcription factors they belong to, resulting in a total of 34 families. Finally, the co-regulatory network contained 95 nodes: 34 transcription factor family nodes and 61 metabolic operon nodes. Arcs in this network were defined according to selection by V-shapes. That is, if an arc from a transcription factor A to a target operon B was selected from the affinity network, then the co-regulatory network included an arc from the family of A to B (or to the family of B, if B itself was also a transcription factor). Thus, a total of 148 arcs were computed between the 95 nodes previously defined.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

MC, VA, and ML designed the research and analyzed the data. MC, DT, and VA generated the bacterial network models. AS and AM supported the bioinformatics metabolic and regulatory networks reconstruction, respectively. MC, DT, VA, and ML wrote the paper. AM and ML take responsibility for the manuscript. All authors read and approved final content.

FUNDING

This work was supported by the CONICYT-PIA Program AFB170001 of the Center for Mathematical Modeling; FONDECYT N° 1190742; Center for Genome Regulation FONDAP 15090007; Concurso Apoyo a la Formación de Redes Internacionales para Investigadores(as) en Etapa Inicial N° REDI170193. Mesa Minería - Consorcio de Universidades del Estado de Chile - CUECH. Fondo de Innovación para la Competitividad - FIC2018 - 6ta región. Gobierno Regional Chile.

ACKNOWLEDGMENTS

We also acknowledge the support of the National Laboratory of High Performance Computing NLHPC (PIA ECM-02.-CONICYT).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00155/full#supplementary-material

REFERENCES

Acuña, V., Aravena, A., Guziolowski, C., Eveillard, D., Siegel, A., and Maass, A. (2016). Deciphering transcriptional regulations coordinating the response to environmental changes. *BMC Bioinform.* 17:35. doi: 10.1186/s12859-016-0885-0

Albert, R. (2005). Scale-free networks in cell biology. *J. Cell Sci.* 118, 4947–4957. doi: 10.1242/jcs.02714

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Balázsi, G., Barabási, A.-L., and Oltvai, Z. N. (2005). Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proc. Natl. Acad. Sci U.S.A.* 102, 7841–7846. doi: 10.1073/pnas.0500365102

Balázsi, G., Heath, A. P., Shi, L., and Gennaro, M. L. (2008). The temporal response of the *Mycobacterium tuberculosis* gene regulatory network during growth arrest. *Mol. Syst. Biol.* 4:225. doi: 10.1038/msb.2008.63

Balleza, E., Lopez-Bojorquez, L. N., Martínez-Antonio, A., Resendis-Antonio, O., Lozada-Chávez, I., Balderas-Martínez, Y. I., et al. (2008). Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiol. Rev.* 33, 133–151. doi: 10.1111/j.1574-6976.2008.00145.x

Bellenberg, S., Díaz, M., Noël, N., Sand, W., Poetsch, A., Guiliani, N., et al. (2014). Biofilm formation, communication and interactions of leaching bacteria during colonization of pyrite and sulfur surfaces. *Res. Microbiol.* 165, 773–781. doi: 10.1016/j.resmic.2014.08.006

Bergeron, R. J., Xin, M. G., Weimar, W. R., Smith, R. E., and Wiegand, J. (2001). Significance of asymmetric sites in choosing siderophores as deferration agents. *J. Med. Chem.* 44, 2469–2478. doi: 10.1021/jm010019s

Bobadilla Fazzini, R. A., Cortés, M. P., Padilla, L., Maturana, D., Budinich, M., Maass, A., et al. (2013). Stoichiometric modeling of oxidation of reduced inorganic sulfur compounds (Riscs) in *Acidithiobacillus thiooxidans*. *Biotechnol. Bioeng.* 110, 2242–2251. doi: 10.1002/bit.24875

Bonnefoy, V., and Holmes, D. S. (2012). Genomic insights into microbial iron oxidation and iron uptake strategies in extremely acidic environments. *Environ. Microbiol.* 14, 1597–1611. doi: 10.1111/j.1462-2920.2011.02626.x

Campodonico, M. A., Vaisman, D., Castro, J. F., Razmilic, V., Mercado, F., Andrews, B. A., et al. (2016). *Acidithiobacillus ferrooxidans's* comprehensive model driven analysis of the electron transfer metabolism and synthetic strain design for biomining applications. *Metab. Eng. Commun.* 3, 84–96. doi: 10.1016/j.meteno.2016.03.003

Cárdenas, J. P., Moya, F., Covarrubias, P., Shmaryahu, A., Levicán, G., Holmes, D. S., et al. (2012). Comparative genomics of the oxidative stress response in bioleaching microorganisms. *Hydrometallurgy* 127, 162–167. doi: 10.1016/j.hydromet.2012.07.014

Chen, P., Yan, L., Wu, Z., Xu, R., Li, S., Wang, N., et al. (2015). Draft genome sequence of extremely acidophilic bacterium *Acidithiobacillus ferrooxidans* DLC-5 isolated from acid mine drainage in Northeast China. *Genom. Data* 6, 267–268. doi: 10.1016/j.gdata.2015.10.018

DebRoy, S., Saldaña, M., Travisany, D., Montano, A., Galloway-Peña, J., Horstmann, N., et al. (2016). A multi-serotype approach clarifies the catabolite control protein a regulon in the major human pathogen group A *Streptococcus*. *Sci. Rep.* 6:32442. doi: 10.1038/srep32442

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129–143. doi: 10.1038/nrmicro1949

Ferrer, A., Rivera, J., Zapata, C., Norambuena, J., Sandoval, Á., Chávez, R., et al. (2016). Cobalamin protection against oxidative stress in the acidophilic iron-oxidizing bacterium *Leptospirillum* group II CF-1. *Front. Microbiol.* 7:748. doi: 10.3389/fmicb.2016.00748

Frankenberg, N., Moser, J., and Jahn, D. (2003). Bacterial heme biosynthesis and its biotechnological application. *Appl. Microbiol. Biotechnol.* 63, 115–127. doi: 10.1007/s00253-003-1432-2

Gazzaniga, F., Stebbins, R., Chang, S. Z., McPeek, M. A., and Brenner, C. (2009). Microbial NAD metabolism: lessons from comparative genomics. *Microbiol. Mol. Biol. Rev.* 73, 529–541. doi: 10.1128/MMBR.00042-08

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064

Hödar, C., Moreno, P., Di Genova, A., Latorre, M., Reyes-Jara, A., Maass, A., et al. (2012). Genome wide identification of *Acidithiobacillus ferrooxidans* (ATCC 23270) transcription factors and comparative analysis of ArsR and MerR metal regulators. *Biometals* 25, 75–93. doi: 10.1007/s10534-011-9484-8

Hold, C., Andrews, B. A., and Asenjo, J. A. (2009). A stoichiometric model of *Acidithiobacillus ferrooxidans* ATCC 23270 for metabolic flux analysis. *Biotechnol. Bioeng.* 102, 1448–1459. doi: 10.1002/bit.22183

Horne, S. M., Schroeder, M., Murphy, J., and Prüβ, B. M. (2018). Acetoacetate and ethyl acetoacetate as novel inhibitors of bacterial biofilm. *Lett. Appl. Microbiol.* 66, 329–339. doi: 10.1111/lam.12852

Imperi, F., Tiburzi, F., Fimia, G. M., and Visca, P. (2010). Transcriptional control of the *pvdS* iron starvation sigma factor gene by the master regulator of sulfur metabolism CysB in *Pseudomonas aeruginosa*. *Environ. Microbiol.* 12, 1630–1642. doi: 10.1111/j.1462-2920.2010.02210.x

Karp, P. D., Paley, S., and Romero, P. (2002). The Pathway Tools software. *Bioinformatics* 18, S225–S232. doi: 10.1093/bioinformatics/18.suppl_1.S225

Latorre, M., Cortés, M. P., Travisany, D., Di Genova, A., Budinich, M., Reyes-Jara, A., et al. (2016). The bioleaching potential of a bacterial consortium. *Bioresour. Technol.* 218, 659–666. doi: 10.1016/j.biortech.2016.07.012

Latorre, M., Galloway-Peña, J., Roh, J. H., Budinich, M., Reyes-Jara, A., Murray, B. E., et al. (2014). *Enterococcus faecalis* reconfigures its transcriptional regulatory network activation at different copper levels. *Metallomics* 6, 572–581. doi: 10.1039/c3mt00288h

Lynch, T. W., Read, E. K., Mattis, A. N., Gardner, J. F., and Rice, P. A. (2003). Integration host factor: putting a twist on protein–DNA recognition. *J. Mol. Biol.* 330, 493–502. doi: 10.1016/S0022-2836(03)00529-1

Martínez, P., Gálvez, S., Ohtsuka, N., Budinich, M., Cortés, M. P., Serpell, C., et al. (2013). Metabolomic study of Chilean biomining bacteria *Acidithiobacillus ferrooxidans* strain Wenelen and *Acidithiobacillus thiooxidans* strain Licanantay. *Metabolomics* 9, 247–257. doi: 10.1007/s11306-012-0443-3

Merino Santis, M. P., Andrews Farrow, B., and de Leuze, J. (2015). Stoichiometric model and flux balance analysis for a mixed culture of *Leptospirillum ferriphilum* and *Ferroplasma acidiphilum. Biotechnol. Prog.* 31, 307–315. doi: 10.1002/btpr.2028

Merino, M. P., Andrews, B. A., and Asenjo, J. A. (2010). Stoichiometric model and metabolic flux analysis for *Leptospirillum ferrooxidans*. *Biotechnol. Bioeng.* 107, 696–706. doi: 10.1002/bit.22851

Mohapatra, B. R., Gould, W. D., Dinardo, O., and Koren, D. W. (2008). An overview of the biochemical and molecular aspects of microbial oxidation of inorganic sulfur compounds. *CLEAN Soil Air Water* 36, 823–829. doi: 10.1002/clen.200700213

Münch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., et al. (2003). PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.* 31, 266–269. doi: 10.1093/nar/gkg037

Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., et al. (2013). RegPrecise 3.0–a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genome.* 14:745. doi: 10.1186/1471-2164-14-745

Okuda, T., Kokubu, E., Kawana, T., Saito, A., Okuda, K., and Ishihara, K. (2012). Synergy in biofilm formation between *Fusobacterium nucleatum* and *Prevotella* species. *Anaerobe* 18, 110–116. doi: 10.1016/j.anaerobe.2011.09.003

Potash, J. (2007). Unraveling transcriptional networks. *Nat. Methods* 4:198. doi: 10.1038/nmeth0307-198

Quatrini, R., Escudero, L. V., Moya-Beltrán, A., Galleguillos, P. A., Issotta, F., Acosta, M., et al. (2017). Draft genome sequence of *Acidithiobacillus thiooxidans* CLST isolated from the acidic hypersaline Gorbea salt flat in northern Chile. *Stand. Genomic Sci.* 12:84. doi: 10.1186/s40793-017-0305-8

Quatrini, R., Jedlicki, E., and Holmes, D. S. (2005). Genomic insights into the iron uptake mechanisms of the biomining microorganism *Acidithiobacillus ferrooxidans*. *J. Ind. Microbiol. Biotechnol.* 32, 606–614. doi: 10.1007/s10295-005-0233-2

Rensing, C., and Franke, S. (2007). Copper homeostasis in *Escherichia coli* and other *Enterobacteriaceae*. *EcoSal Plus* 2, 1–16. doi: 10.1128/ecosalplus.5.4.4.1

Richard, K. L., Kelley, B. R., and Johnson, J. G. (2019). Heme uptake and utilization by gram-negative bacterial pathogens. *Front. Cell. Infect. Microbiol.* 9:81. doi: 10.3389/fcimb.2019.00081

Samartzidou, H., Mehrazin, M., Xu, Z., Benedik, M. J., and Delcour, A. H. (2003). Cadaverine inhibition of porin plays a role in cell survival at acidic pH. *J. Bacteriol.* 185, 13–19. doi: 10.1128/JB.185.1.13-19.2003

Schell, M. A. (1993). Molecular biology of the LysR family of transcriptional regulators. *Annu. Rev. Microbiol.* 47, 597–626. doi: 10.1146/annurev.mi.47.100193.003121

Schoonen, M. A. A., Cohn, C. A., Roemer, E., Laffers, R., Simon, S. R., and O'Riordan, T. (2006). Mineral-induced formation of reactive oxygen species. *Rev. Mineral. Geochem.* 64, 179–221. doi: 10.2138/rmg.2006.64.7

Schröder, J., and Tauch, A. (2010). Transcriptional regulation of gene expression in *Corynebacterium glutamicum*: the role of global, master and local regulators in the modular and hierarchical gene regulatory network. *FEMS Microbiol. Rev.* 34, 685–737. doi: 10.1111/j.1574-6976.2010.00228.x

Sharma, A., Inagaki, S., Sigurdson, W., and Kuramitsu, H. K. (2005). Synergy between *Tannerella forsythia* and *Fusobacterium nucleatum* in biofilm formation. *Oral Microbiol. Immunol.* 20, 39–42. doi: 10.1111/j.1399-302X.2004.00175.x

Solioz, M., and Stoyanov, J. V. (2003). Copper homeostasis in *Enterococcus hirae*. *FEMS Microbiol. Rev.* 27, 183–195. doi: 10.1016/S0168-6445(03)00053-6

Travisany, D., Cortés, M. P., Latorre, M., Di Genova, A., Budinich, M., Bobadilla-Fazzini, R. A., et al. (2014). A new genome of *Acidithiobacillus thiooxidans* provides insights into adaptation to a bioleaching environment. *Res. Microbiol.* 165, 743–752. doi: 10.1016/j.resmic.2014.08.004

Valdés, J., Veloso, F., Jedlicki, E., and Holmes, D. (2003). Metabolic reconstruction of sulfur assimilation in the extremophile *Acidithiobacillus ferrooxidans* based on genome analysis. *BMC Genome.* 4:51. doi: 10.1186/1471-2164-4-51

van der Ploeg, J. R., Eichhorn, E., and Leisinger, T. (2001). Sulfonate-sulfur metabolism and its regulation in *Escherichia coli*. *Arch. Microbiol.* 176, 1–8. doi: 10.1007/s002030100298

Wang, R., Lin, J.-Q., Liu, X.-M., Pang, X., Zhang, C.-J., Gao, X.-Y., et al. (2018). Sulfur oxidation in the acidophilic autotrophic *Acidithiobacillus* spp. *Front. Microbiol.* 9, 3290. doi: 10.3389/fmicb.2018.03290

Williams, K. P., and Kelly, D. P. (2013). Proposal for a new class within the phylum *Proteobacteria*, *Acidithiobacillia* classis nov., with the type order *Acidithiobacillales*, and emended description of the class *Gammaproteobacteria*. *Int. J. Syst. Evol. Microbiol.* 63, 2901–2906. doi: 10.1099/ijs.0.049270-0

Yan, L., Zhang, S., Wang, W., Hu, H., Wang, Y., Yu, G., et al. (2015). Draft genome sequence of *Acidithiobacillus ferrooxidans* YQH-1. *Genom. Data* 6, 269–270. doi: 10.1016/j.gdata.2015.10.009

Yin, H., Zhang, X., Li, X., He, Z., Liang, Y., Guo, X., et al. (2014). Whole-genome sequencing reveals novel insights into sulfur oxidation in the extremophile *Acidithiobacillus thiooxidans*. *BMC Microbiol.* 14:179. doi: 10.1186/1471-2180-14-179

Yus, E., Lloréns-Rico, V., Martínez, S., Gallo, C., Eilers, H., Blötz, C., et al. (2019). Determination of the gene regulatory network of a genome-reduced bacterium highlights alternative regulation independent of transcription factors. *Cell Syst.* 9, 143–158. doi: 10.1016/j.cels.2019.07.001

Zhang, X., Feng, X., Tao, J., Ma, L., Xiao, Y., Liang, Y., et al. (2016). Comparative genomics of the extreme acidophile *Acidithiobacillus thiooxidans* reveals intraspecific divergence and niche adaptation. *Int. J. Mol. Sci.* 17:E1355. doi: 10.3390/ijms17081355

Zhou, Q., Gao, J., Li, Y., Zhu, S., He, L., Nie, W., et al. (2017). Bioleaching in batch tests for improving sludge dewaterability and metal removal using *Acidithiobacillus ferrooxidans* and *Acidithiobacillus thiooxidans* after cold acclimation. *Water Sci. Technol.* 76, 1347–1359. doi: 10.2166/wst.2017.244

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Molecular Biosciences

REVIEW
published: 14 January 2020
doi: 10.3389/fmolb.2019.00158

Challenges of Integrative Disease Modeling in Alzheimer's Disease

Sepehr Golriz Khatami[1,2], Christine Robinson[1,2], Colin Birkenbihl[1,2], Daniel Domingo-Fernández[1,2], Charles Tapley Hoyt[1,2] and Martin Hofmann-Apitius[1,2]*

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany,
[2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

Dementia-related diseases like Alzheimer's Disease (AD) have a tremendous social and economic cost. A deeper understanding of its underlying pathophysiologies may provide an opportunity for earlier detection and therapeutic intervention. Previous approaches for characterizing AD were targeted at single aspects of the disease. Yet, due to the complex nature of AD, the success of these approaches was limited. However, in recent years, advancements in integrative disease modeling, built on a wide range of AD biomarkers, have taken a global view on the disease, facilitating more comprehensive analysis and interpretation. Integrative AD models can be sorted in two primary types, namely hypothetical models and data-driven models. The latter group split into two subgroups: (i) Models that use traditional statistical methods such as linear models, (ii) Models that take advantage of more advanced artificial intelligence approaches such as machine learning. While many integrative AD models have been published over the last decade, their impact on clinical practice is limited. There exist major challenges in the course of integrative AD modeling, namely data missingness and censoring, imprecise human-involved priori knowledge, model reproducibility, dataset interoperability, dataset integration, and model interpretability. In this review, we highlight recent advancements and future possibilities of integrative modeling in the field of AD research, showcase and discuss the limitations and challenges involved, and finally, propose avenues to address several of these challenges.

Keywords: Alzheimer's disease, challenges, integrative disease modeling, hypothetical, data-driven

OPEN ACCESS

Edited by:
Daniel Garrido,
Pontifical Catholic University of Chile, Chile

Reviewed by:
Lin Tan,
Qingdao Municipal Hospital, China
Owen T. Carmichael,
Pennington Biomedical Research Center, United States

****Correspondence:***
Sepehr Golriz Khatami
sepehr.golriz.khatami@scai-extern.fraunhofer.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *02 September 2019*
Accepted: *18 December 2019*
Published: *14 January 2020*

Citation:
Golriz Khatami S, Robinson C, Birkenbihl C, Domingo-Fernández D, Hoyt CT and Hofmann-Apitius M (2020) Challenges of Integrative Disease Modeling in Alzheimer's Disease. Front. Mol. Biosci. 6:158. doi: 10.3389/fmolb.2019.00158

INTRODUCTION

Alzheimer's Disease (AD) manifests in a collection of symptoms including the deterioration of cognition, memory, and behavior which often leads to interference with activities of daily living. In 2017, AD ranked among the top five causes of death worldwide, with 2.44 million (4.5%) deaths from AD[1,2]. Worldwide, there are currently around 50 million people living with AD, and every 3 s a person develops this condition. It is estimated that only a quarter of those living with AD are diagnosed, and more than 17 million healthcare workers annually invest 18 billion hours of care, at a cost of more than one trillion US dollars to tackle AD-associated problems[3,4]. Extrapolating these statistics to the coming decades suggests the immense socioeconomic impact of AD on all involved

[1] https://ourworldindata.org/causes-of-death
[2] https://www.thestreet.com/world/leading-causes-of-death-world-14869811
[3] https://www.alz.co.uk/research/statistics
[4] https://ourworldindata.org/causes-of-death

parties: patients, caregivers, healthcare systems, and indirectly, the economy. Thus, strategies to reduce the global emotional and financial burden of AD are of great importance. To develop such strategies, a deeper understanding of the pathophysiology underlying AD is necessary and may lead to opportunities for earlier detection and therapeutic interventions.

In general, AD progression is categorized into three clinical disease stages: (i) During the pre-symptomatic phase, individuals may have already developed pathological changes that underlie AD, but remain cognitively normal, (ii) in the prodromal phase, often referred to as mild cognitive impairment (MCI), the first cognitive symptoms, commonly episodic memory deficits, appear. These symptoms can be acute, but they do not yet meet the criteria for dementia, (iii) in the dementia stage, impairments are severe enough to interfere with daily life (Jack et al., 2010).

Understanding of the etiology of AD is complicated due to the existence of dysregulations at different biological scales, ranging from genetic mutations to structural and functional alterations of the brain (Aisen et al., 2017). For this reason, significant efforts have been made in recent years to discover candidate markers for disease-related pathological changes throughout all modalities, including neuro-imaging, cerebrospinal fluid (CSF) samples and a broad variety of -omics data. Studies have successfully identified multiple biomarkers for neurodegeneration and AD (Blennow and Zetterberg, 2018). However, effectively translating extensive biomarker screenings into clinical application remains a challenging task, because individual biomarkers can only provide a highly incomplete view on such a multifactorial disease (Younesi and Hofmann-Apitius, 2013). For instance, while multiple associations between genetic variants and AD have been established (Jansen et al., 2019; Kunkle et al., 2019), none of these associations fully describe disease pathogenesis. As a result, one of the major challenges in AD research is translating diverse biomarker signals available into multimodal, multiscale models of disease pathogenesis.

In recent years, a new translational research paradigm called "integrative disease modeling" has emerged, to address this challenge (Younesi and Hofmann-Apitius, 2013). It aims at modeling heterogeneous measurements across different biological scales, in order to provide a holistic picture of biomarker intercorrelations in the disease of study. To this end, advanced high-throughput technologies and neuroimaging procedures are being used to collect data from multiple modalities. These diverse data need to be integrated, that is, combined in a way that preserves the structure and meaning in the data, using computational algorithms. Only then can they provide a solid basis for further analysis such as reasoning, simulation, and visualization. In order to contribute to understanding of the complex pathophysiology of the disease, the results should be actionable and thus must be interpretable. Integrative disease modeling, by collecting, integrating, analyzing, and ultimately interpreting the measurements, facilitates the understanding of the pathophysiology of complex diseases like AD (Hampel et al., 2017).

Existing integrative models in the context of AD can be placed in two primary categories, namely hypothetical models and data-driven models (**Table 1**). Hypothetical models are non-numerical and rely on reasoning over findings of previously published studies (Jack et al., 2010), rather than large amounts of data. By including this prior knowledge, these models try to detail the temporal changes of AD biomarkers relative to each other as well as to clinical disease stages and trial endpoints.

TABLE 1 | Organization of and references for data-driven integrative AD models.

Data-driven integrative AD models			References
Traditional			Caroli and Frisoni, 2010; Jack et al., 2011, 2012
Machine learning	Generative		Fonteijn et al., 2012; Chen et al., 2016; Khanna et al., 2018; Oxtoby et al., 2018; Basu et al., 2019; De Jong et al., 2019; Gootjes-Dreesbach et al., 2019; Martinez-Murcia et al., 2019
	Discriminative	Supervised	Hinrichs et al., 2010; Magnin et al., 2010; Rao et al., 2011; Zhang et al., 2011; Da et al., 2013; Li et al., 2013
		Unsupervised	Nettiksimmons et al., 2014; Gamberger et al., 2017; Toschi et al., 2019

We subdivide data-driven integrative AD models which into two subgroups. While the first group uses simple statistical approaches (e.g., simple linear models), the second group uses more advanced techniques (e.g., machine learning). The advanced machine learning models include generative and discriminative models, the latter of which can be classified as either supervised or unsupervised models.

By contrast, data-driven integrative models take advantage of developments in computational approaches and big data. For the sake of this review, we will distinguish between two subcategories of data-driven models. The first covers traditional statistical methods of generally lower complexity, such as linear models. Often, these models are used to estimate biomarker trajectories by regressing measured data against a prespecified dependent variable, such as a clinical readout or the disease stage (Bateman et al., 2012). The second subtype exploits more advanced artificial intelligence approaches such as machine learning. Within this subtype, models can be characterized as discriminative or generative. Discriminative models are designed to discriminate between groups (e.g., cases and controls) and can be further described as supervised or unsupervised, depending on whether they rely on labeled (Hinrichs et al., 2011; Da et al., 2013) or unlabeled (Toschi et al., 2019) data. Generative models contribute to disease understanding by automatically learning the inherent distribution of a dataset and its feature interdependencies (Oxtoby et al., 2018). An exemplary application is the extraction of disease progression signatures as demonstrated by the ensemble of Bayesian networks developed by Khanna et al. (2018).

Integrative AD modeling faces many challenges. Hypothetical models, by their nature, are time-intensive to construct and require specialist knowledge. Their primary role in AD research is to provide ideas for future experiments. Likewise in data-driven modeling, several challenges at each step of the process (i.e., collection, integration, analysis, and interpretation) must be addressed. Data missingness and data censoring are significant bottlenecks in data collection as well as analysis and interpretation. Meanwhile, the heterogeneity and complexity of biological data are major impediments to data integration, which forms the basis for all data-driven approaches. Furthermore, data mapping, data labels, and biased data are additional barriers to robust data analysis and interpretation. Finally, insufficient numbers of subjects restrict the statistical power of data-driven integrative AD models. These fundamental challenges explain why, at this point in time, although many integrative AD models have been published over the last decade, their impact on clinical practice is limited.

In this review, we highlight recent advancements and future possibilities of integrative modeling, discuss the limitations and challenges involved, and finally, propose avenues to address several of these challenges, in the context of AD research.

INTEGRATIVE AD MODELS

As already mentioned, integrative AD models can be characterized as either hypothetical or data-driven, each of which has strengths and weaknesses. In the following, we compare different models of each type and discuss their benefits and limitations. Finally, we elaborate on how associated limitations and challenges could be handled.

Hypothetical Models

In hypothetical modeling, a model is generated about an object of study, direct knowledge of which is difficult to obtain. These models provide hypotheses about the object (Gladun, 1997). In integrative AD modeling, researchers develop so-called cascade models, in which the measurements of a set of biomarkers are normalized and their trajectories are plotted on a common time scale, aligned to disease stages (Jack et al., 2010, 2013). These models are typically developed by reviewing the available knowledge and reasoning over observations from previously published studies. They are not directly informed by measured data.

One of the first hypothetical integrative AD models was developed by Jack et al. (2013) [revised from a previous model (Jack et al., 2010)]. This model hypothesized the temporal changes of the five most studied biomarkers of AD pathology in relation to estimated years from expected symptom onset and in relation to other biomarkers. These biomarkers are CSF amyloid-beta protein (CSF Aβ_{1-42}) and tau protein (CSF tau) levels, amyloid-beta PET imaging (PET Aβ), Fluorodeoxyglucose-PET imaging, and structural MRI readouts. In this cascade model, the authors presumed that biomarker trajectories should exhibit a sigmoid-shaped curve. This imposition is a direct result of the limited sensitivity of measurements at time extremes, which the authors addressed by taking the floor of the measurements at early timepoints, and the ceiling of the measurements at late timepoints. The authors hypothesized that the two amyloid-beta (Aβ) biomarkers (i.e., CSF Aβ_{1-42} and PET Aβ imaging) gradually approach an abnormal state while the subject remains in a cognitively normal state. After a lag period, the length of which varies from patient to patient, and in later disease stages, CSF tau, Fluorodeoxyglucose-PET, and structural MRI biomarkers follow the same pattern and begin the transition to an abnormal state. Similarly, Frisoni et al. (2010) established a theoretical progression of cognitive and biological markers (primarily imaging features) based not only on the clinical disease stages, but also patient age at AD diagnosis and time since diagnosis. Although both models captured earliest detectable changes in amyloid markers, Frisoni et al. (2010) additionally theorized that these changes plateau by the MCI stage, when the individuals are no longer cognitively normal. Furthermore, they suggested that F-fluorodeoxyglucose PET is abnormal by the MCI stage and continues to change well into the dementia stage. Structural changes appear later, following a temporal pattern mirroring tau pathology deposition, which slightly differs from the Jack et al. models (Jack et al., 2010, 2013).

While hypothetical models cannot be directly applied, they can be used to suggest directions for future experiments that themselves would address diagnosis, prediction, or decision making tasks (Gladun, 1997). However, there are a number of challenges relating to the construction of hypothetical models. In the following, we discuss these challenges and propose ways to address some of them.

Challenges of Hypothetical Models

The exclusive reliance of hypothetical models on literature presents several challenges. First, relevant literature must be identified. Second, the scientific knowledge contained in the literature must be extracted in a meaningful form. Finally, the knowledge has to be modeled.

In order to build a hypothetical model, a researcher must identify a set of relevant publications, called a literature corpus, which accurately reflects AD knowledge. This corpus should be representative of the relevant aspects of AD, contain the most up-to-date publications, and not be biased toward subfields or trends. However, the number of new AD publications has increased each year since 2005, and there were nearly 15,000 such publications in 2017 alone (Dong et al., 2019). With such publication rates, it is challenging for researchers to manually create high quality corpora (Rodriguez-Esteban, 2015), Moreover, manual generation of these corpora is susceptible to bias, because researchers may tend to draw more heavily from authors or subfields with which they are more familiar (Atkins et al., 1992). The size of a corpus will also be limited by the time and resources available to the researchers. However, text mining has been used effectively to automatically classify relevant literature, based on titles and abstracts (e.g., see Simon et al., 2018), and to prioritize texts (Singh et al., 2015). Publications identified by this classification can be directly taken as the corpus or used as a more manageable set of publications from which the domain experts can appropriately select. Hypothetical models are susceptible to biases present in the literature (Boutron and Ravaud, 2018), but a well-designed, computationally selected corpus can mitigate the effects of those biases.

Once the corpus has been identified, the challenge of knowledge extraction remains. The goal here is to recover the knowledge contained in the publications in a meaningful way. Conducting this task manually is a time-consuming process that requires a high degree of domain knowledge. Here, text mining poses the opportunity to extract knowledge in a computable form (Gyori et al., 2017; Lamurias and Couto, 2019). Moreover, it significantly reduces the amount of time required to read publications, which enables significantly larger corpora to be used in the building of hypothetical models.

Finally, in order to build hypothetical models, the information gleaned from the literature corpus must be organized in a coherent way. The entities and the relationships between them should all be represented. Mind maps provide a non-automated way of generating a knowledge model, driven by domain-expert knowledge (Kudelic et al., 2011). However, if automated information extraction strategies were used on the literature corpus, then knowledge graphs are well-suited for storing the extracted knowledge (Gyori et al., 2017). A major advantage of this strategy is that the knowledge graph is computable, meaning downstream machine learning tasks can be carried out for knowledge discovery. Furthermore, knowledge graphs support hypothesis generation by enabling researchers to assess whether their hypotheses are compatible with existing knowledge (Humayun et al., 2019).

Automated methods of corpus identification, knowledge extraction, and knowledge modeling provide a means of mitigating the challenges of hypothetical modeling. They reduce the time burden, mitigate the risk of bias in manual methods, and generate computable knowledge representations. This can yield more reliable hypothetical AD models.

Hypothetical models are non-numerical and rely exclusively on qualitative information, gleaned from a review of previous findings. This limits their usability solely to eliciting hypotheses for future experiments. They are neither predictive nor can they be used for analysis of any kind of data. They are meant to represent a kind of "typical" AD progression, without reflecting individual deviations from that. Given the broad biological heterogeneity observed among AD subjects, and the increasing relevance of personalized medicine (Reitz, 2016), there is a need for models that are capable of achieving this.

Data-driven models built on data collected in longitudinal cohort studies can serve to support or challenge hypotheses generated by hypothetical models (Petrella et al., 2019). Data-driven models are appropriate for a wide range of tasks that lie beyond the scope of what hypothetical models are designed for. For example, using data models can capture individual subject particularities that hypothetical models cannot (see e.g., Young et al., 2015). In the following, we discuss data-driven models and their challenges in depth.

Data-Driven Models

In contrast to hypothetical models, data-driven integrative models are directly derived from datasets comprising readouts of multiple biomarkers. Such models can be applied to a broad variety of tasks ranging from predictive modeling e.g., predicting patient diagnosis (Ding et al., 2018) or age at disease onset (Chuang et al., 2016; Peng et al., 2016) to discovering patterns in the data that shed light on biomarker interdependencies and disease underlying mechanisms. Since these models use extensive data, they are not limited by preconceived notions in the way that hypothetical integrative models are.

Data-driven AD models can be classified into two primary subtypes based on the statistical approaches and algorithms applied (**Table 1**). The first subtype use traditional statistical methods such as linear modeling, and the second employs artificial intelligence and more specifically machine learning approaches.

Traditional Statistical Models

In AD modeling, traditional statistical approaches, such as linear mixed-effects models, are often used to estimate biomarker trajectories (Caroli and Frisoni, 2010; Jack et al., 2011, 2012). In these models, measured data, are regressed against a prespecified variable, such as disease stage, to detail the temporal changes of AD biomarkers during the course of disease. Essentially, these models provide empirical testing of hypothetical multiple biomarker trajectory plots.

Jack et al. (2012) used linear mixed-effects models to investigate the shape of five important AD biomarker trajectories (i.e., Aβ_{42}, tau, amyloid, fluorodeoxyglucose PET, and structural MRI) as a function of a cognitive test score, the Mini-Mental State Exam (MMSE). This model parameterization enabled them to assess within-subject rates of biomarker changes with respect to changes of the MMSE score. They found that lower baseline MMSE scores are correlated with worse baseline biomarker values and that higher rates of biomarker change were associated with worsening MMSE score. This model constructed the biomarker trajectories without making any assumptions about the shapes of the trajectories. This contrasts with the authors' earlier hypothetical biomarker cascade model, which imposed a sigmoid trajectory curve.

While the shapes of the trajectories in this data-driven model agree with the assumptions made in the hypothetical exemplar, the model has several limitations, pertaining to model design choices and deficiencies in the data. The authors chose to use the MMSE score as the independent variable. This choice was made because the MMSE score provides a linear measure of disease progression that was available across all datasets. However, this introduces challenges in the estimation of trajectories in early disease stages, because MMSE scores in cognitively normal patients are relatively stable over time (Tombaugh, 2005), yielding only a narrow range of values. Moreover, especially when studying early disease stages, the model additionally suffers from possible absence of information on future disease developments of a subject. This absence of data on future disease outcome is related to data censoring, which will be addressed in more detail later.

In their data-driven model (Jack et al., 2011), Jack et al. aimed to unravel the temporal order of biomarker trajectories becoming abnormal, rather than only describing the shape of their trajectories. They used the prevalence of biomarker abnormalities at different disease stages to empirically assess the temporal ordering of their trajectories. They employed generalized estimating equations, a generalized linear model for longitudinal data that can deal with correlated observations, to evaluate and compare the proportion of abnormal observations per biomarker. The proper choice of a cut-off defining when biomarker measures are considered to be abnormal is a point of debate and making this choice requires critical judgement. To differentiate between normal and abnormal biomarkers, Jack et al. (2011) determined a cut-off by looking at an independent post-mortem cohort. However, since, by construction, results were highly sensitive to the selected cut-off for each biomarker, the temporal resolution of the model is limited.

While the proportion of patients with abnormal biomarker values might seem an unnatural choice for comparing biomarkers, alternative strategies also have drawbacks. Caroli and Frisoni (2010) computed Z-scores based on values of each biomarker and fitted them against Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-cog) scores, comparing linear and sigmoidal fits. Their investigation showed that a sigmoid curve fit the observed data significantly better than a linear one for most of the biomarkers, and thereby might be able to characterize the time course of those biomarkers. These results were consistent with the hypothetical model proposed by Jack et al. (2010) and Jack et al. (2013). However, the biomarker trajectories cannot be directly compared with the data-driven model developed by Jack et al. (2011), since different scales were employed in both studies. While standardization of values by converting them into Z-scores resolves this problem, it introduces a new one: by definition, the arithmetic mean of each biomarker will be 0. This makes it impossible to reasonably compare biomarker distributions based on their means using standard statistical procedures like, for example, t-tests (Jack et al., 2011; Moeller, 2015).

The arbitrariness of defining a cut-off for abnormality of a biomarker will always pose a limitation on statistical approaches relying on biomarkers. While such cut-offs simplify the interpretation of the biomarker, there is no universally correct cut-off for a given biomarker. Rather, appropriate cut-offs heavily depend on the population, and even the individual, on which a biomarker will be used. Covariates such as an individual's age, genetic risk factors, and family history of AD must be considered. For these reasons, there is no single optimal cut-off for any given biomarker (Bartlett et al., 2012; Anne and Fagan, 2014). To address this, a less rigid technique has been developed, that designates an intermediate range using two cut-offs, one permissive and the other conservative (Klunk et al., 2012; Jack et al., 2016a,b; Bzdok, 2017). The permissive point can be used for earliest detectable evidence of AD pathologic changes and the conservative one for high diagnostic certainty. Moreover, different statistical approaches, like Youden's index and the receiver operating characteristic (ROC) curve, can be applied to help determine an appropriate cut-off.

Linear traditional models are ill-equipped to handle the increasingly high-dimensional data being collected in AD studies. Thanks to recent technological advancements, the granularity of AD datasets with respect to information resolution, feature size, and complexity of meta-information have increased. For example, improved neuro-imaging techniques generate datasets with higher resolution than previously available. This information distributed over voxels, a 3D imaging unit, is hard to capture using linear models (Bzdok, 2017). Therefore, more advanced data-driven models have been developed based on machine learning. These models are generally more flexible and compatible with the complex datasets encountered in biology research (Bzdok, 2017).

Machine Learning Models

Machine learning models can be characterized as generative or discriminative. As previously mentioned, discriminative models are designed to differentiate between groups, while generative models provide better disease understanding by learning inherent properties from datasets, such as feature interdependencies.

Generative models

Generative modeling relies on the use of statistics and probability to extract patterns from data and learn the underlying distribution. In the following, three types of generative integrative AD models are reviewed: event-based models, Bayesian network learning, and autoencoders.

Event-based models. Event-based models estimate the most probable sequence of events based on the assessment of a probability density function for a particular event order. Fonteijn et al. (2012), Chen et al. (2016), and Oxtoby et al. (2018), used this method to learn the sequence of AD events based on imaging and non-imaging measurements from a clinical study. The authors first fitted simple mixture models (e.g., gaussian mixture models) to individual biomarkers in order to calculate the likelihood of the normality or abnormality status per biomarker. Given these likelihoods, by multiplication of the probabilities, the likelihoods for each possible order of events was calculated. The order with the highest probability was then selected using a greedy Markov Chain Monte Carlo algorithm to describe the temporal correlation of the biomarker trajectories over the course of AD progression.

The models developed by Fonteijn et al. (2012) and Chen et al. (2016) simplified the sequence of biomarker abnormalities over the course of the disease progression by relying on the assumption that all subjects follow a single event sequence. However, AD is highly heterogeneous and includes distinct subgroups (Ferreira et al., 2018). To account for this, Young et al. (2015) established their event-based models with two extensions: a Mallows model and a Dirichlet process mixture of generalized Mallows models. The first extension allows subjects to deviate from the main event sequence, and the latter clusters subjects according to different event sequences.

In principle, the event sequence proposed in the hypothetical model is similar to that observed using traditional and event-based models. Changes in CSF measures are the earliest events, followed by regional brain atrophies and finally succeeded by diminished cognitive scores. However, the event sequence in the hypothetical and traditional models is constructed based on predefined clinical assessments and often imprecise or subjective cut-offs. By contrast, in generative models, the sequence of events, as well as the clustering of biomarkers into normal and abnormal classes, is directly extracted from the data (e.g., the onset of a new symptom, like memory performance decline). Thus, event-based models explain the changes without a priori biases. Moreover, generative models are able to characterize uncertainty in the event ordering arising from heterogeneity in the population and thus, can address individual deviations from the generic model.

Bayesian network learning. Extensive research efforts have been made to uncover the relationships between individual biomarkers and AD. Yet the number of studies that investigated the interplay between multiple biomarkers themselves is comparably limited. Khanna et al. (2018) and Ding et al. (2018) built Bayesian network models covering different biological scales and time points to uncover the interplay amongst sets of biomarkers. Ding et al. (2018) considered the ApoE allele, PET and MRI imaging data, scores from psychological and functional tests, and the medical history of patients with respect to neurological diseases. Using a variety of feature selection metrics, they determined the most relevant features with respect to the clinical dementia rating and modeled these heterogeneous measurements using a Bayesian network to determine their probabilistic interdependencies. However, these models only capture conditional probabilities between predictor variables and clinical outcomes. They are unable to provide a causal mechanistic understanding of an observed phenomenon. Such hypothesized pathophysiological mechanisms are important for making reliable predictions and having confidence in the practical application of data-driven models. To this end, Khanna et al. (2018) employed a combination of data-driven probabilistic and knowledge-driven mechanistic approaches. They modeled clinical variables, genetic variants, pathways, and neuro-imaging readouts using Bayesian network learning to estimate dependencies between disease relevant features. Together with a cause-and-effect knowledge model derived from scientific literature, they partially reconstructed biological mechanisms that could play a role in the conversion of normal/MCI into AD pathology.

Autoencoders. The last type of generative model discussed in this review is autoencoders. In essence, an autoencoder is a neural network that aims to encode the input data into a lower dimensional representation and from that decode it again, reconstructing the original input. It has successfully been applied for different tasks on AD cohorts (Basu et al., 2019; Martinez-Murcia et al., 2019). The two main applications of this approach in the field consist of classifying patients based on AD diagnosis (Basu et al., 2019) and clustering of patient trajectories into subgroups (De Jong et al., 2019). These strategies are especially interesting for patient classification and stratification tasks in datasets where information is sparse. However, another novel and promising task for autoencoders is the generation of synthetic data from real patient level data (Gootjes-Dreesbach et al., 2019). This, in turn, could be used to circumvent legal and ethical constraints that restrict data sharing.

Discriminative models

Discriminative models are a class of models generally used for classification. Discriminative models that rely on labeled data are called supervised models, while unsupervised models use unlabeled data.

Supervised discriminative models. Diverse supervised discriminative methods such as support vector machines (SVM; Magnin et al., 2010), and multiple-kernel SVM (MKL; Hinrichs et al., 2010; Zhang et al., 2011) have been used to classify AD patients, MCI subjects, and controls. However, studies that used multiple-kernel SVM reported superior classification performance, because the use of multiple kernels facilitates the integration of multimodal biomarker data (Zhang et al., 2011). Additionally, MKL are well-suited for dealing with very high dimensional data (Young et al., 2013). MKL also enable individual weighting of biomarker modalities. This offers more flexibility for kernel combination and thus, a better integration of the data. For example Hinrichs et al. (2010), applied MKL in combination with MRI and PET imaging to differentiate between AD subjects and controls. Their method showed high classification performance, achieving 92.4% accuracy. Similarly, Zhang et al. (2018) combined MRI, PET, and CSF biomarkers to discriminate between healthy controls and AD/MCI. After integrating all biomarker data using a MKL, they deployed a linear SVM for the actual classification task, which resulted in 93.2% accuracy for classifying AD and healthy controls and 76.4% for discriminating between MCI and healthy controls. Both studies applied a similar method for classification, yet the latter one achieved a slightly higher accuracy. Comparing the approaches applied in Zhang et al. (2018) and Hinrichs et al. (2010) it becomes clear that the major reason for the difference in performance is the feature selection process. Depending on the available sample size, other methods might prove more promising (Liu et al., 2012). Moreover, Zhang et al. (2018) benefits from employing three biomarker modalities, namely, CSF measurements and two imaging modalities, compared to Hinrichs et al. (2010) who only use the two imaging modalities.

While the above kernel-based pattern recognition approaches yield categorical class decisions, Young et al. (2013) used gaussian process classification, which is a probabilistic classification algorithm. This study integrated imaging, CSF, neuropsychological, and genetic biomarkers to classify MCI subjects who remained stable and MCI patients who converted to AD within 3 years. In contrast to MKL, the probabilistic classification afforded by the gaussian process approach provides the opportunity to position the subjects according to disease stage, to stratify patients, and to model the sequence order of biomarker abnormality.

Another type of discriminative model is disease risk models. This type of supervised model can be used to predict the time to AD diagnosis for normal/MCI patients. Multiple approaches have been used to develop risk models for AD (Da et al., 2013; Li et al., 2013). Li et al. (2013) used a combination of cox regression analyses and time-dependent ROC approaches to evaluate prognostic utility and performance stability of candidate biomarkers. The authors deduced that both baseline volumetric MRI and cognitive measures can predict progression from MCI to AD. However, in participants' follow-up visits, only cognitive measurements remained predictive. Da et al. (2013) employed the cox proportional hazards models to compare the magnitudes of the relative association between predictors (patterns of brain atrophy, cognitive assessments, genetics, and CSF biomarkers) and time to conversion from MCI to AD. They concluded that brain atrophy and cognitive assessments in combination offer the highest predictive power of conversion from MCI to AD.

Although the results in both studies were similar, the time-dependent ROC curve used by Li et al. (2013) enabled them to predict disease risk as a function of time. Thus, this method provides clear benefit for a progressive disease such as AD, in which both the disease status and biomarker measurements change over time (Kamarudin et al., 2017).

The data labeling which enables supervised discriminative models to determine decision boundaries for distinguishing classes of interest can also introduce errors. Inaccurate labels will negatively affect the performance of the classifier. Such mislabeling is not uncommon in AD, due to the absence of a clear diagnostic biomarker (Fischer et al., 2017). Instead, diagnosis is currently made based on symptoms (Schott and Petersen, 2015) Furthermore, integrative data analysis is further complicated by the fact that the diagnostic criteria for MCI have changed over the years, and MCI is not consistently defined across clinical studies. While one study relies on assessing only a single cognitive domain for MCI diagnosis, such as speech or memory, others base their diagnoses on performance on cognitive tests for multiple domains. Apart from that, there are multiple pathologies for MCI; AD is just one of them. Thus, unified clear disease definitions are crucial, since the MCI classification accuracy can influence outcomes of research and clinical practice (Jak et al., 2010).

Unsupervised Discriminative Models. Unsupervised discriminative models use a variety of clustering techniques on unlabeled data, avoiding the challenges of data label accuracy. These techniques use properties of each data point to iteratively form groups, called clusters. This ultimately leads to a discrimination of the data into several clusters of highly similar data points. Given the observed biological heterogeneity among normal control subjects, Nettiksimmons et al. (2014) hypothesized that different subgroups may also be found among the MCI subjects. Using agglomerative hierarchical clustering, they sorted subjects based on MRI volumes, CSF measurements, and cognitive tests. Next, the resulting clusters were explored with regard to longitudinal atrophy, conversion time, and cognitive trajectories. Four clusters with unique biomarker patterns resulted: (i) a cluster biologically similar to normal controls. MCI patients from that cluster rarely converted to AD, (ii) one cluster with early AD pathology characteristics, (iii) another cluster of subjects with hardly any tau abnormality, but a high proportion of AD converters, and (iv) and finally one cluster with pre-AD symptoms wherein almost all subjects converted to AD. Based on these findings, they hypothesized that clusters ii and iv reflected the amyloid cascade pattern (Ricciarelli and Fedele, 2017) since both clusters presented lower CSF Aβ levels and elevated tau proteins. However, the tau level in cluster iv was higher, and more severe atrophy as well as cognitive impairment were detected. The authors concluded that more tau accumulation may lead to more cognitive decline. One of the intrinsic limitations of their clustering approach is that the number of clusters must be predefined. The maximum gap statistic is one approach to determine this number (Tibshirani et al., 2001). However, specifying the number of clusters beforehand will always bias the clustering to some extent, and choosing a reasonable number is no trivial task given the broad variety of subtypes found among AD subjects.

Toschi et al. (2019) used Density-Based Spatial Clustering of Applications with Noise (DBSCAN; Thanh et al., 2013), which does not require pre-specifying the number of clusters. They integrated five validated CSF biomarkers in order to cluster a cohort where symptomatic patients presented diagnoses ranging from self-perceived cognitive decline (Zhang et al., 2011) to MCI to AD. In contrast to the previous study, Toschi et al. (2019) adjusted all biomarker values for age, sex and their interactions to exclude them as confounders (Pourhoseingholi et al., 2012). Moreover, Toschi et al. (2019) used t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality of biomarkers space, since defining the distance between the data points in a high dimensional space of biomarkers is notoriously difficult (Domingos, 2012). Finally, they applied DBSCAN on this lower dimensional representation. DBSCAN defines a high data density region based on two parameters: (i) the radius of the neighborhood, and (ii) the minimum number of points within the radius. These values are determined by a nearest neighbor method, in which the distance of each point to their nearest n points is calculated. Afterwards, results are sorted, plotted and the value with most pronounced change is selected as the optimal value. Using DBSCAN, Toschi et al. (2019) characterized five biological clusters which were not significantly bound to the original distinct clinically phenotyped diagnostic groups. They explained that the clusters included all phenotypic groups and were not homogeneous enough to be considered as a specific AD pathophysiology. Moreover, contrary to general belief that $A\beta_{1-42}$ is linearly associated with the progression of AD and cognitive decline (Sperling et al., 2011a; Samtani et al., 2013), their findings suggest that $A\beta_{1-42}$ is less likely to contribute to phenotypic discrimination.

The dimensionality reduction technique, t-SNE, used by Toschi et al. (2019) enabled them to better separate the data and hence, to enhance cluster identification, in comparison to directly running a clustering algorithm on a high dimensional data as Nettiksimmons et al. (2014). However, their main limitation is that clustering results are highly sensitive to two parameters necessary for DBSCAN. Moreover, they did not include other biomarkers, such as imaging and genetics biomarkers, which could enhance their clustering, as previously reported by Young et al. (2013, 2018).

Unsupervised clustering algorithms are ideal for identifying subgroups and non-linear associations between individuals based on a multidimensional profile, regardless of the individual labels, in contrast to supervised algorithms. This allows the grouping of individuals based on shared pathophysiological drivers and triggers and, possibly, similar longitudinal disease trajectories. This is an advantage in the AD field due to the prevalence of unreliable labels stemming from misdiagnosis and to the biological heterogeneity of AD subjects. On the other hand, most unsupervised clustering algorithms perform better with a larger sample size than is often obtainable in AD studies (Oxtoby and Alexander, 2017). Therefore, the smaller size inherent to AD cohorts may lead to clustering instability.

To this point, we have reviewed a broad variety of data-driven integrative AD models and elaborated on their associated limitations and challenges. In the following, we enumerate more general challenges researchers encounter in the course of data-driven integrative AD modeling and suggest how these could be addressed.

Challenges of Data-Driven Modeling

Although there exists a wide range of data-driven integrative modeling approaches, not all of them are well-suited for every analytic task and each has its own strengths and weaknesses. Still, there are some challenges which affect all data-driven approaches to some degree: data collection, reproducibility of findings, and interpretability of models and results.

Data Collection

Collecting patient level data, the basis for all data-driven modeling, is a time-consuming and costly process. Additionally, it is a source of major challenges and limitations of these models. In particular, data "censoring" and "missingness," can impede modeling, bias models, or even make certain modeling techniques unfeasible.

Data censoring describes the condition in which a particular event (here AD diagnosis) is not observed for certain study participants during the study runtime. This censoring can occur in two ways: if AD diagnosis occurred before the start of the study; or if the patient drops out of the study, or the study ends without occurrence of the AD diagnosis event. A significant number of patients enrolled in clinical studies have already received a diagnosis before the beginning of the study, indicating that they are in a progressed stage of the disease (Ellis et al., 2009). It is therefore not possible to obtain indications of early disease onset in such patients. The second form of censoring arises from two sources. First, all observational cohort studies experience participant dropout for a variety of reasons, including the participation burden on caregivers or medical problems (Coley et al., 2008). Second, subjects that remain healthy throughout study runtime could still develop the disease after the study ended, meaning they were in a prodromal disease stage. It is thus impossible to know if or when the patient would eventually receive an AD diagnosis. This form of censoring is common in longitudinal AD studies, because AD is a slow-progressing disease, while the studies are typically quite short (Lawrence et al., 2017), due to limited funding (Prabhakaran and Bakshi, 2018).

Disease onset is a critical point for clinical intervention (Sperling et al., 2011b), so it is subject to extensive research efforts. It is here, however, where data censoring impedes data analysis the most. Data censoring can result in over- or under-sampling of early and advanced disease stages. This, in turn, leads to models biased toward specific disease stages (Ning et al., 2010). Various methods, such as complete data analysis (Xiang et al., 2013), imputation (Fisher et al., 2019), or analysis based on dichotomized data (Donohue et al., 2011), have been established to address censored data. Yet all of these methods may introduce error and impose complexities and biases on other integrative modeling steps, such as model interpretation, and thus need to be used with care (Prinja et al., 2010).

The complete absence of a value for variables in the observation of interest likewise poses a significant challenge to data-driven modeling. This missing data in AD cohort studies occurs for several reasons, including unwillingness of patients to undergo invasive tests like lumbar punctures, and the high cost of measuring a particular variable, such as imaging scans (Engelborghs et al., 2017). The implications of such a scenario include a loss of statistical power of the study and may bias the conclusions that can be drawn (Hughes et al., 2019). Over the past decades, novel statistical methods (Molenberghs et al., 2014) and software (Quartagno and Carpenter, 2016; Moreno-Betancur et al., 2017) have been developed for analyzing data with missing values. However, analysis restricted to individuals with complete data is generally preferred, if feasible.

Despite the challenges in collecting complete and uncensored data, the value of data in strengthening disease understanding is clear. Several large-scale AD patient datasets have been collected for use in a variety of studies (Lawrence et al., 2017) including, for example, Alzheimer's Disease Neuroimaging Initiative (ADNI; Mueller et al., 2005), Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL; Ellis et al., 2009), the Dominantly Inherited Alzheimer Network (DIAN; Moulder et al., 2013), and European Prevention of Alzheimer's Dementia (EPAD; Vermunt et al., 2018). However, these classical observational studies are subject to bias, resulting from the inclusion and exclusion criteria used to select participants (Miksad and Abernethy, 2018).

The use of electronic medical records (EMRs) has been proposed as a potential solution to reduce the bias of classical clinical trials. They provide an alternative view on patient measurements (Fröhlich et al., 2018), so, a collection of EMRs can provide a more representative view on patient measurements. However, EMRs are largely phenotypic: molecular phenomena such as genomic variants are not reflected in the data. Moreover, extracting information from EMRs requires natural language preprocessing, which itself currently remains a difficult and error-prone process.

Reproducibility

The ability to reproduce the findings of a study using different subjects is an important part of scientific research. This is particularly the case in integrative AD modeling, since the tendency of AD datasets is not to fully reflect the diversity of AD patients. Inclusion-exclusion criteria in clinical studies can lead to significant under-representation of some populations. For example, the landscape of data-driven AD models is currently dominated by only a few cohorts which are made up largely of White Caucasians, and, to a lesser extent, are constrained by geographic location (Lawrence et al., 2017). Since most observational cohorts are not representative of the general AD population (Ferreira et al., 2017), it is important to validate the resulting models with an independent cohort study. While this external validation is a necessary step to corroborate findings, it is complicated by data interoperability and sample size.

Interoperability

The ability to map the data coming from one study to data from another study is known as data interoperability[5]. Each of the major AD clinical studies was established with a specific sample and feature characterization. Since they might not be directly interoperable, extensive curation is needed before the external validation of a model can be carried out. Otherwise, the training cohort and the validation cohort would be based on different populations, and would contain different measurements. Thus, before validation, researchers must map and assess the "comparability" of both features and subjects.

Feature mapping requires specifying relationships between data elements from different data models and standardizing the terms used to represent the features in the two datasets. This is due to the fact that controlled vocabularies are not used to annotate the datasets. Thus, even if the same biomarker has been collected in two studies, it is usually referred to by different terms, impeding a direct comparison of the datasets. For example, the hippocampus is one of the earliest sites of AD pathology, and hippocampal volume is measured in ADNI and EPAD. However, ADNI identifies this biomarker as "Hippocampus," while EPAD refers to it as "lhvr" (right hemisphere) and "lhvl" (left hemisphere).

Moreover, the subject populations in each study must be comparable. For instance, if the biological sex distributions in two AD studies differ significantly, then the cognitive impairment scores of the cohorts cannot be directly compared, because female AD patients have been shown to have greater cognitive impairment than men in comparable stages of the disease (Laws et al., 2016).

There are several strategies to overcome the lack of interoperability between datasets at both feature and subject level. At the feature level, interoperability can be attained by annotating datasets according to a standard controlled vocabulary. Several such vocabularies (e.g., NIFT Iyappan et al., 2017 and PTS Iyappan et al., 2016) have been established, but significant improvements in interoperability will only come with widespread adoption (Neu et al., 2012). The most prominent example might be the AD specific standard developed by the Clinical Data Interchange Standards Consortium (CDISC; Neville et al., 2017). At the subject level, mapping between training and validation cohorts can be accomplished by identifying, in the validation cohort, a subset of subjects that is statistically comparable to the training cohort. Finally, in order to assess the comparability of subjects from different studies, techniques such as statistical matching can be used (Austin, 2011).

Sample size

The relatively small sample sizes of AD clinical studies also contributes to the challenge of reproducibility in AD integrative modeling. Many AD studies contain fewer than a thousand patients, and the longitudinal follow-up is limited. In addition, typically not all of the subjects were screened for the complete biomarker set, leading to sparse subsets of patients for whom the study contains complete data. As a result, models generated from these studies have a high margin of error and low statistical power, meaning they struggle to detect small effects.

The integration of different datasets into a larger dataset can overcome some of the challenges related to small sample sizes (Gomez-Cabrero et al., 2014). Integrated datasets provide more comprehensive data, and the resulting models have greater statistical power. However, current approaches for data integration were developed for the analysis of single-data-type datasets, and only subsequently adapted to handle datasets with multiple data types. For this reason, data integration methodologies can be ill-suited to manage the computational challenges arising from the variety of different data sizes, formats, and dimensionalities present in AD datasets, as well as their noisiness, complexity, and the level of agreement between datasets (Gomez-Cabrero et al., 2014; Gligorijević et al., 2015). Furthermore, even data acquired by analogous technologies are not necessarily integrable. For example, neuroimaging data acquired from similar scanners and similar modalities may still be stored in different formats and have different metadata content (Goble and Stevens, 2008).

Several strategies could be applied to address the interoperability challenges arising from data integration. The first strategy is to normalize and standardize data across all platforms (O'Bryant et al., 2015). However, scientific independency and freedom for innovation, as well as uniqueness of databases, must be respected. The second strategy is to collect a standardized set of biomarkers across different studies. Finally, the ideal solution would be performing a systematic longitudinal clinical and -omics follow-up of each individual in a large and rigorously characterized cohort since this would provide a statistically sufficient number of measurements in the context of subjects and variables. The Deep and Frequent Phenotyping study from Lawson et al. (2017) showed that such a cohort, in theory, is feasible. Yet, including a sufficient number of participants in such an ambitious study is costly.

Interpretability

In order for an AD model to have clinical impact, its findings must be interpretable. There are several barriers to AD model interpretability. Machine learning models often act as "black boxes"; it may be impossible to uncover the reasons for the predictions made by the model (Rudin, 2019). Indeed, as the number of features and the complexity of the computational processes used in models increases, this interpretability problem will worsen. Moreover, data-driven models are not causal and typically capture non-linear correlations between predictor and explanatory variables. While prior understanding of cause–effect relationships and detailed mechanisms might prove helpful to well-performing models, it is not necessarily required. Lack of mechanistic explanations for model prediction complicates the interpretation of data-driven findings and reduces acceptance by physicians (Fröhlich et al., 2018). Thus, the translation of data-driven models into a biomedical knowledge context is a major challenge in integrative AD modeling.

[5] https://library.ahima.org/doc?oid=65895#.Xdl-iZPYrOQ

Combining available mechanistic knowledge with machine learning-based sub-models, so-called hybrid modeling could bridge the gap between experimental biological and computational research by improving interpretability (Fröhlich et al., 2018). For example, Bayesian networks which built on causal knowledge graphs constitute such a hybrid model (Arora et al., 2019). They shed light on interdependencies across features, which can be on different scales (e.g., clinical, genetic, and molecular), and allow for predicting the outcome of purely hypothetical clinical interventions. Similarly, other recent deep learning methodologies use knowledge-derived biological networks to define the layers of neural networks in order to improve interpretability (Fortelny and Bock, 2019).

CONCLUSION

In the era of extensive biomarker profiling, big data, and artificial intelligence, integrative AD modeling comes with high promises. By integrating multi-scale, multimodal, and longitudinal patient data, such modeling approaches aim to provide a holistic picture of disease pathophysiology and progression. However, as we have discussed in this review, while integrative models have generated significant insights, and thus proved to be valuable in research, existing models do not yet fully describe critical aspects of AD.

The construction of hypothetical models simultaneously benefits and suffers from the vast amount of published knowledge. Prioritization of articles and computational text mining of literature corpora are reasonable approaches to identify a greater quantity of relevant knowledge while designing hypothetical models. In the field of data-driven integrative AD modeling, we highlighted several major ongoing challenges throughout the whole modeling process of data collection, integration of disparate data sources, data analysis, and model interpretation. Data missingness and data censoring are major bottlenecks in data collection as well as analysis and interpretation. Heterogeneity and complexity in biological data are major impediments to data integration, which is central to data-driven integrative modeling and validation. Data mapping, imprecise diagnostic stages, and biased data are barriers that hamper data analysis and interpretation. Furthermore, there is an insufficient number of subjects in studies, which restricts the statistical power of data-driven integrative AD models. Because of these challenges, to the best of our knowledge, at this point in time, there are no integrative AD models which have been used in clinical practice.

While in theory, certain existing integrative models are capable of predicting AD diagnosis and progression, they are not used in clinical practice. We see a number of steps that could bring us closer to the goal of precision medicine and that could enable patient diagnosis through integrative disease models in a clinical context. First, we, the AD research community, need to establish valid, informative biomarkers and clear criteria for AD diagnosis. This would result in reliable predictors that could be included in modeling approaches, as well as fewer diagnostic errors, which in turn reduce the effect of mislabeled data. Second, a global data schema that could support the normalization and standardization of data across measurements would ultimately facilitate improved data integration. If future cohort studies would adhere to such a schema, data integration would be straightforward and the cumulative time saved for researchers working with it would be enormous. Finally, innovative modeling approaches, such as causal inference techniques and hybrid modeling, which go beyond current state-of-the-art data-driven models by linking prior knowledge with data-driven models, need to be developed and made more robust. Overall, novel computational modeling approaches that surmount the current integrative AD modeling challenges may hold the potential to play an increasing role in the planning of medical interventions and practice.

AUTHOR CONTRIBUTIONS

SG drafted the manuscript. CR, CB, DD-F contributed to the final version of the manuscript. CH and MH-A reviewed the final version of the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support from the B-IT foundation.

REFERENCES

Aisen, P. S., Cummings, J., Jack, C. R. Jr., Morris, J. C., Sperling, R., Frölich, L., et al. (2017). On the path to 2025: understanding Alzheimer's disease continuum. *Alzheimers Res. Ther.* 9:60. doi: 10.1186/s13195-017-0283-5

Anne, M., and Fagan. (2014). CSF biomarkers of Alzheimer's disease: impact on disease concept, diagnosis, and clinical trial design. *Adv. Geriatr.* 2014:302712. doi: 10.1155/2014/302712

Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., and Druzdzel, M. J. (2019). Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Val. Health* 22, 439–445. doi: 10.1016/j.jval.2019.01.006

Atkins, S., Clear, J., and Ostler, N. (1992). Corpus design criteria. *Lit. Ling. Comput.* 7, 1–16. doi: 10.1093/llc/7.1.1

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav. Res.* 46, 399–424. doi: 10.1080/00273171.2011.568786

Bartlett, J. W., Frost, C., Mattsson, N., Skillbäck, T., Blennow, K., Zetterberg, H., et al. (2012). Determining cut-points for Alzheimer's disease biomarkers: statistical issues, methods and challenges. *Biomark. Med.* 6, 391–400. doi: 10.2217/bmm.12.49

Basu, S., Wagstyl, K., Zandifar, A., Collins, L., Romero, A., and Precup, D. (2019). "Early prediction of alzheimer's disease progression using variational autoencoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, eds D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan (Cham: Springer), 205–213. doi: 10.1007/978-3-030-32251-9_23

Bateman, R. J., Xiong, C., Benzinger, T. L., Fagan, A. M., Goate, A., Fox, N. C., et al. (2012). Clinical and biomarker changes in dominantly inherited

Alzheimer's disease. *N. Engl. J. Med.* 367, 795–804. doi: 10.1056/NEJMoa1202753

Blennow, K., and Zetterberg, H. (2018). Biomarkers for Alzheimer's disease: current status and prospects for the future. *J. Intern. Med.* 284, 643–663. doi: 10.1111/joim.12816

Boutron, I., and Ravaud, P. (2018). Misrepresentation and distortion of research in biomedical literature. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2613–2619. doi: 10.1073/pnas.1710755115

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Front. Neurosci.* 11:543. doi: 10.3389/fnins.2017.00543

Caroli, A., and Frisoni, G. B. (2010). The dynamics of Alzheimer's disease biomarkers in the Alzheimer's disease neuroimaging initiative cohort. *Neurobiol. Aging* 31, 1263–1274. doi: 10.1016/j.neurobiolaging.2010.04.024

Chen, G., Shu, H., Chen, G., Ward, B. D., Antuono, P. G., Zhang, Z., et al. (2016). Staging Alzheimer's disease risk by sequencing brain function and structure, cerebrospinal fluid, and cognition biomarkers. *J. Alzheimers Dis.* 54, 983–993. doi: 10.3233/JAD-160537

Chuang, Y. F., An, Y., Bilgel, M., Wong, D. F., Troncoso, J. C., O'Brien, R. J., et al. (2016). Midlife adiposity predicts earlier onset of Alzheimer's dementia, neuropathology and presymptomatic cerebral amyloid accumulation. *Mol. Psychiatry* 21, 910–915. doi: 10.1038/mp.2015.129

Coley, N., Gardette, V., Toulza, O., Gillette-Guyonnet, S., Cantet, C., Nourhashemi, F., et al. (2008). Predictive factors of attrition in a cohort of Alzheimer disease patients. *Neuroepidemiology* 31, 69–79. doi: 10.1159/000144087

Da, X., Toledo, J. B., Zee, J., Wolk, D. A., Xie, S. X., and Ou, Y. (2013). Integration and relative value of biomarkers for prediction of MCI to AD progression: spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. *Neuroimage Clin.* 4, 164–173. doi: 10.1016/j.nicl.2013.11.010

De Jong, J., Emon, M. A., Wu, P., Karki, R., Sood, M., Godard, P., et al. (2019). Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience* 8:giz134. doi: 10.1093/gigascience/giz134

Ding, X., Bucholc, M., Wang, H., Glass, D. H., Wang, H., Clarke, D. H., et al. (2018). A hybrid computational approach for efficient Alzheimer's disease classification based on heterogeneous data. *Sci. Rep.* 8:9774. doi: 10.1038/s41598-018-27997-8

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM* 55, 78–87. doi: 10.1145/2347736.2347755

Dong, R., Wang, H., Ye, J., Wang, M., and Bi, Y. (2019). Publication trends for Alzheimer's disease worldwide and in China: a 30-year bibliometric analysis. *Front. Hum. Neurosci.* 13:259. doi: 10.3389/fnhum.2019.00259

Donohue, M. C., Gamst, A. C., Thomas, R. G., Xu, R., Beckett, L., Petersen, R. C., et al. (2011). The relative efficiency of time-to-threshold and rate of change in longitudinal data. *Contemp. Clin. Trials* 32, 685–693. doi: 10.1016/j.cct.2011.04.007

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., et al. (2009). The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int. Psychogeriatr.* 21, 672–687. doi: 10.1017/S1041610209009405

Engelborghs, S., Niemantsverdriet, E., Struyfs, H., Blennow, K., Brouns, R., Comabella, M., et al. (2017). Consensus guidelines for lumbar puncture in patients with neurological diseases. *Alzheimers Dement.* 8:111–126. doi: 10.1016/j.dadm.2017.04.007

Ferreira, D., Hansson, O., Barroso, J., Molina, Y., Machado, A., Hernández-Cabrera, J. A., et al. (2017). The interactive effect of demographic and clinical factors on hippocampal volume: a multicohort study on 1958 cognitively normal individuals. *Hippocampus* 27, 653–667. doi: 10.1002/hipo.22721

Ferreira, D., Wahlund, L., and Westman, E. (2018). The heterogeneity within Alzheimer's disease. *Aging* 10, 3058–3060. doi: 10.18632/aging.101638

Fischer, C. E., Qian, W., Schweizer, T. A., Ismail, Z., Smith, E. E., Millikin, C. P., et al. (2017). Determining the impact of psychosis on rates of false-positive and false-negative diagnosis in Alzheimer's disease. *Alzheimers Dement*. 3, 385–392. doi: 10.1016/j.trci.2017.06.001

Fisher, C. K., Smith, A. M., and Walsh, J. R. (2019). Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci. Rep.* 9:13622. doi: 10.1038/s41598-019-49656-2

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., et al. (2012). An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage* 60, 1880–1889. doi: 10.1016/j.neuroimage.2012.01.062

Fortelny, N., and Bock, C. (2019). Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. doi: 10.1101/794503

Frisoni, G. B., Fox, N. C., Jack, C. R. Jr, Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77. doi: 10.1038/nrneurol.2009.215

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., et al. (2018). From hype to reality: data science enabling personalized medicine. *BMC Med.* 16:150. doi: 10.1186/s12916-018-1122-7

Gamberger, D., Lavrač N., Srivatsa, S., Tanzi, R. E., and Doraiswamy, P. M. (2017). Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease. *Sci. Rep.* 7:6763. doi: 10.1038/s41598-017-06624-y

Gladun, V. P. (1997). Hypothetical modeling: methodology and application. *Cybern. Syst. Anal.* 33, 7–15. doi: 10.1007/BF02665935

Gligorijević V., and PrŽulj, N. (2015). Methods for biological data integration: perspectives and challenges. *J. R. Soc. Interface* 12:20150571. doi: 10.1098/rsif.2015.0571

Goble, C., and Stevens, R. (2008). State of the nation in data integration for bioinformatics. *J. Biomed. Inform.* 41, 687–693. doi: 10.1016/j.jbi.2008.01.008

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8(Suppl. 2):I1. doi: 10.1186/1752-0509-8-S2-I1

Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M., and Fröhlich, H. (2019). Variational Autoencoder Modular Bayesian Networks (VAMBN) for simulation of heterogeneous clinical study data. *bioRxiv* 760744. doi: 10.1101/760744

Gyori, B. M., Bachman, J. A., Subramanian, K., Muhlich, J. L., Galescu, L., and Sorger, P. K. (2017). From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol.* 13:954. doi: 10.15252/msb.20177651

Hampel, H., O'Bryant, S. E., Durrleman, S., Younesi, E., Rojkova, K., Escott-Price, V., et al. (2017). A precision medicine initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling. *Climacteric* 20, 107–118. doi: 10.1080/13697137.2017.1287866

Hinrichs, C., Singh, V., Xu, G., and Johnson, S. (2010). MKL for robust multi-modality AD classification. *Med. Image Comput. Comput. Assist. Interv*. 12(Pt 2), 786–794. doi: 10.1007/978-3-642-04271-3_95

Hinrichs, C., Singh, V., Xu, G., and Johnson, S. C. (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *Neuroimage* 55, 574–589. doi: 10.1016/j.neuroimage.2010.10.081

Hughes, R. A., Heron, J., Sterne, J. A. C., and Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int. J. Epidemiol.* 48, 1294–1304. doi: 10.1093/ije/dyz032

Humayun, F., Domingo-Fernandez, D., George, A. A. P., Hopp, M. T., Syllwasschy, B. F., Detzel, M., et al. (2019). A computational approach for mapping heme biology in the context of hemolytic disorders. *bioRxiv* 804906. doi: 10.1101/804906

Iyappan, A., Gündel, M., Shahid, M., Wang, J., Li, H., Mevissen, H. T., et al. (2016). Towards a pathway inventory of the human brain for modeling disease mechanisms underlying neurodegeneration. *J. Alzheimers Dis.* 52, 1343–1360. doi: 10.3233/JAD-151178

Iyappan, A., Younesi, E., Redolfi, A., Vrooman, H., Khanna, S., Frisoni, G. B., et al. (2017). Neuroimaging feature terminology: a controlled terminology for the annotation of brain imaging features. *J. Alzheimers Dis.* 59, 1153–1169. doi: 10.3233/JAD-161148

Jack, C. R. Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Feldman, H. H., Frisoni, G. B., et al. (2016a). A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology* 2, 539–547. doi: 10.1212/WNL.0000000000002923

Jack, C. R. Jr., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* 12, 207–216. doi: 10.1016/S1474-4422(12)70291-0

Jack, C. R. Jr., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* 9, 119–128. doi: 10.1016/S1474-4422(09)70299-6

Jack, C. R. Jr., Vemuri, P., Wiste, H. J., Weigand, S. D., Aisen, P. S., Trojanowski, J. Q., et al. (2011). Evidence for ordering of Alzheimer disease biomarkers. *Arch. Neurol.* 68, 1526–1535. doi: 10.1001/archneurol.2011.183

Jack, C. R. Jr., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Lowe, V., et al. (2012). Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Arch. Neurol.* 69, 856–867. doi: 10.1001/archneurol.2011.3405

Jack, C. R. Jr., Wiste, H. J., Weigand, S. D., Therneau, T. M., Lowe, V. J., and Knopman, D. S. (2016b). Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimers Dement.* 13, 205–216. doi: 10.1016/j.jalz.2016.08.005

Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D. P., et al. (2010). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *Am. J. Geriatr. Psychiatry* 17, 368–375. doi: 10.1097/JGP.0b013e31819431d5

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* 51, 404–413. doi: 10.1038/s41588-018-0311-9

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* 17:53. doi: 10.1186/s12874-017-0332-6

Khanna, S., Domingo-Fernández, D., Iyappan, A., Emon, M. A., Hofmann-Apitius, M., and Fröhlich, H. (2018). Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci. Rep.* 8:11173. doi: 10.1038/s41598-018-29433-3

Klunk, W., Cohen, A., Bi, W., Weissfeld L,Aizenstein, H., McDade, E., et al. (2012). Why we need two cutoffs for amyloid imaging: early versus Alzheimer's-like amyloid-positivity. *Alzheimers Dement.* 8, P453–P454. doi: 10.1016/j.jalz.2012.05.1208

Kudelic R., Konecki, M., and Malekovic, M. (2011). "Mind map generator software model with text mining algorithm," in *Proceedings of the ITI 2011, 33rd International Conference on Information Technology Interfaces* (Cavtat), 487–494.

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* 51, 414–430. doi: 10.1038/s41588-019-0358-2

Lamurias, A., and Couto, F. M. (2019). "Text mining for bioinformatics using biomedical literature," in *Encyclopedia of Bioinformatics and Computational Biology*, eds S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford: Elsevier), 602–611. doi: 10.1016/B978-0-12-809633-8.20409-3

Lawrence, E., Vegvari, C., Ower, A., Hadjichrysanthou, C., De Wolf, F., and Anderson, R. M. (2017). A systematic review of longitudinal studies which measure alzheimer's disease biomarkers. *J. Alzheimers Dis.* 59, 1359–1379. doi: 10.3233/JAD-170261

Laws, K. R., Irvine, K., and Gale, T. M. (2016). Sex differences in cognitive impairment in Alzheimer's disease. *World J. Psychiatry* 6, 54–65. doi: 10.5498/wjp.v6.i1.54

Lawson, J., Murray, M., Zamboni, G., Koychev, I. G., Ritchie, C. W., Ridha, B. H., et al. (2017). Deep and frequent phenotyping: a feasibility study for experimental medicine in dementia. *J Alzheimers Dement*. 13, p1268–1269. doi: 10.1016/j.jalz.2017.06.1897

Li, S., Okonkwo, O., Albert, M., and Wang, M. C. (2013). Variation in variables that predict progression from MCI to AD dementia over duration of follow-up. *Am. J. Alzheimers Dis*. 2, 12–28. doi: 10.7726/ajad.2013.1002

Liu, M., Zhang, D., Yap, P. T., and Shen, D. (2012). Tree-guided sparse coding for brain disease classification. *Med. Image Comput. Comput. Assist. Interv.* 15(Pt 3), 239–247. doi: 10.1007/978-3-642-33454-2_30

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., and Sarazin, M. (2010). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology* 51, 73–83. doi: 10.1007/s00234-008-0463-x

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J. M., Ramirez, J., and Castillo-Barnes, D. (2019). Studying the manifold structure of Alzheimer's Disease: a deep learning approach using convolutional autoencoders. *IEEE J. Biomed. Health Inform.* 1-1. doi: 10.1109/JBHI.2019.2914970

Miksad, R. A., and Abernethy, A. P. (2018). Harnessing the Power of Real-World Evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin. Pharmacol. Ther.* 103, 202–205. doi: 10.1002/cpt.946

Moeller, J. (2015). A word on standardization in longitudinal studies: don't. *Front. Psychol.* 6:1389. doi: 10.3389/fpsyg.2015.01389

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of Missing Data Methodology*. New York, NY: Chapman and Hall/CRC. doi: 10.1201/b17622

Moreno-Betancur, M., Leacy, F. P., Tompsett, D., and White, I. (2017). *mice: The NARFCS Procedure for Sensitivity Analyses.*

Moulder, K. L., Snider, B. J., Mills, S. L., Buckles, V. D., Santacruz, A. M., Bateman, R. J., et al. (2013). Dominantly inherited Alzheimer network: facilitating research and clinical trials. *Alzheimers Res. Ther.* 5:48. doi: 10.1186/alzrt213

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., et al. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N Am.* 15, 869–877. doi: 10.1016/j.nic.2005.09.008

Nettiksimmons, J., DeCarli, C., Landau, S., and Beckett, L. (2014). Biological heterogeneity in ADNI amnestic mild cognitive impairment. *Alzheimers Dement.* 10, 511–521.e1. doi: 10.1016/j.jalz.2013.09.003

Neu, S. C., Crawford, K. L., and Toga, A. W. (2012). Practical management of heterogeneous neuroimaging metadata by global neuroimaging data repositories. *Front. Neuroinform.* 6:8. doi: 10.3389/fninf.2012.00008

Neville, J., Kopko, S., Romero, K., Corrigan, B., Stafford, B., LeRoy, E., et al. (2017). Accelerating drug development for Alzheimer's disease through the use of data standards. *Alzheimer's Dement*. 3, 273–283. doi: 10.1016/j.trci.2017.03.006

Ning, J., Qin, J., and Shen, Y. (2010). Nonparametric tests for right-censored data with biased sampling. *J. R. Stat. Soc. Series B Stat. Methodol.* 72, 609–630. doi: 10.1111/j.1467-9868.2010.00742.x

O'Bryant, S. E., Gupta, V., Henriksen, K., Edwards, M., Jeromin, A., Lista, S., et al. (2015). Guidelines for the standardization of preanalytic variables for blood-based biomarker studies in Alzheimer's disease research. *Alzheimers Dement.* 11, 549–560. doi: 10.1016/j.jalz.2014.08.099

Oxtoby, N. P., and Alexander, D. C. (2017). Imaging plus X: multimodal models of neurodegenerative disease. *Curr. Opin. Neurol.* 30, 371–379. doi: 10.1097/WCO.0000000000000460

Oxtoby, N. P., Young, A. L., Cash DM Benzinger, T. L. S., Fagan, A. M., Morris, J. C., et al. (2018). Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* 141, 1529–1544. doi: 10.1093/brain/awy050

Peng, D., Shi, Z., Xu, J., Shen, L., Xiao, S., Zhang, N., et al. (2016). Demographic and clinical characteristics related to cognitive decline in Alzheimer's disease in China: a multicenter survey from 2011 to 2014. *Medicine* 95:26. doi: 10.1097/MD.0000000000003727

Petrella, J. R., Hao, W., Rao, A., and Doraiswamy, P. M. (2019). Computational causal modeling of the dynamic biomarker cascade in Alzheimer's disease. 2019:6216530 *Comput. Math. Methods Med.* doi: 10.1155/2019/6216530

Pourhoseingholi, M. A., Baghestani, A. R., and Vahedi, M. (2012). How to control confounding effects by statistical analysis. *Gastroenterol. Hepatol. Bed Bench* 5, 79–83.

Prabhakaran, G., and Bakshi, R. (2018). Analysis of structure and cost in an American longitudinal study of Alzheimer's disease. *J. Alzheimers Dis. Parkinsonism* 8:411. doi: 10.4172/2161-0460.1000411

Prinja, S., Gupta, N., and Verma, R. (2010). Censoring in clinical trials: review of survival analysis techniques. *Indian J. Community Med.* 35, 217–221. doi: 10.4103/0970-0218.66859

Quartagno, M., and Carpenter, J. (2016). *jomo: A Package for Multilevel Joint Modelling Multiple Imputation.* R package version 2.

Rao, A., Lee, Y., Gass, A., and Monsch, A. (2011). Classification of Alzheimer's disease from structural MRI using sparse logistic regression with optional spatial regularization. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 4, 499–502. doi: 10.1109/IEMBS.2011.6091115

Reitz, C. (2016). Toward precision medicine in Alzheimer's disease. *Ann. Transl. Med.* 4:107. doi: 10.21037/atm.2016.03.05

Ricciarelli, R., and Fedele, E. (2017). The amyloid cascade hypothesis in Alzheimer's disease: it's time to change our mind. *Curr. Neuropharmacol.* 2017, 926–935. doi: 10.2174/1570159X15666170116143743

Rodriguez-Esteban, R. (2015). Biocuration with insufficient resources and fixed timelines. *Database* 2015:bav116. doi: 10.1093/database/bav116

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Samtani, M. N., Raghavan, N., Shi, Y., Novak, G., Farnum, M., and Lobanov, V. (2013). Disease progression model in subjects with mild cognitive impairment from the Alzheimer's disease neuroimaging initiative: CSF biomarkers predict population subtypes. *Br. J. Clin. Pharmacol.* 75, 146–161. doi: 10.1111/j.1365-2125.2012.04308.x

Schott, J. M., and Petersen, R. C. (2015). New criteria for Alzheimer's disease: which, when and why? *Brain* 138(Pt 5), 1134–1137. doi: 10.1093/brain/awv055

Simon, C., Davidsen, K., Hansen, C., Seymour, E., Barnkob, M. B., and Olsen, L. R. (2018). BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* 19:57. doi: 10.1186/s12859-019-2607-x

Singh, M., Murthy, A., and Singh, S. (2015). Prioritization of free-text clinical documents: a novel use of a bayesian classifier. *JMIR Med. Inform.* 3:e17. doi: 10.2196/medinform.3793

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011a). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 280–292. doi: 10.1016/j.jalz.2011.03.003

Sperling, R. A., Jack, C. R. Jr., and Aisen, P. S. (2011b). Testing the right target and right drug at the right stage. *Sci. Transl. Med.* 3:111cm33. doi: 10.1126/scitranslmed.3002609

Thanh, N.,Tran, T., Drab, K., and Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *J. Chemom. Intell. Lab. Syst.* 120, 92–96. doi: 10.1016/j.chemolab.2012.11.006

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. Royal Stat. Soc.* 63, 411–423. doi: 10.1111/1467-9868.00293

Tombaugh, T. N. (2005). Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Arch. Clin. Neuropsychol.* 20, 485–503. doi: 10.1016/j.acn.2004.11.004

Toschi, N., Lista, S., Baldacci, F., Cavedo, E., Zetterberg, H., Blennow, K., et al. (2019). Biomarker-guided clustering of Alzheimer's disease clinical syndromes. *Neurobiol. Aging* 83, 42–53. doi: 10.1016/j.neurobiolaging.2019.08.032

Vermunt, L., Veal, C. D., Ter Meulen, L., Chrysostomou, C., van der Flier, W., Frisoni, G. B., et al. (2018). European prevention of Alzheimer's dementia registry: recruitment and pre screening approach for a longitudinal cohort and prevention trials. *Alzheimers. Dement.* 14, 837–842. doi: 10.1016/j.jalz.2018.02.010

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., and Ye, J. (2013). "Multi-source learning with block-wise missing data for Alzheimer's disease prediction,". in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 185–193. doi: 10.1145/2487575.2487594

Younesi, E., and Hofmann-Apitius, M. (2013). From integrative disease modeling to predictive, preventive, personalized and participatory (P4) medicine. *EPMA J.* 4:23. doi: 10.1186/1878-5085-4-23

Young, A. L., Marinescu, R. V., Oxtoby, N. P., Bocchetta, M., Yong, K., Firth, N. C., et al. (2018). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat. Commun.* 9:4273. doi: 10.1038/s41467-018-05892-0

Young, A. L., Oxtoby, N. P., Huang, J., Marinescu, R. V., Daga, P., Cash, D. M., et al. (2015). Multiple orderings of events in disease progression. *Inf. Process. Med. Imagin*g 24, 711–722. doi: 10.1007/978-3-319-19992-4_56

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., and Ourselin, S. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *Neuroimage Clin.* 2, 735–745. doi: 10.1016/j.nicl.2013.05.004

Zhang, D., Wang, Y., Zhou, L., Yuan, H., and Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 55, 856–867. doi: 10.1016/j.neuroimage.2011.01.008

Zhang, J., Zhou, W., Cassidy, R. M., Su, H., Su, Y., Zhang, X. (2018). Risk factors for amyloid positivity in older people reporting significant memory concern. *Comprehensive Psychiatry* 80, 126–131. doi: 10.1016/j.comppsych.2017.09.015

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 13 February 2020
doi: 10.3389/fmolb.2020.00013

Development of Supervised Learning Predictive Models for Highly Non-linear Biological, Biomedical, and General Datasets

David Medina-Ortiz[1,2], *Sebastián Contreras*[2], *Cristofer Quiroz*[3] *and Álvaro Olivera-Nappa*[1,2]*

[1] Departamento de Ingeniería Química, Biotecnología y Materiales, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile, [2] Centre for Biotechnology and Bioengineering, Universidad de Chile, Santiago, Chile, [3] Facultad de Ingeniería, Universidad Autónoma de Chile, Talca, Chile

In highly non-linear datasets, attributes or features do not allow readily finding visual patterns for identifying common underlying behaviors. Therefore, it is not possible to achieve classification or regression using linear or mildly non-linear hyperspace partition functions. Hence, supervised learning models based on the application of most existing algorithms are limited, and their performance metrics are low. Linear transformations of variables, such as principal components analysis, cannot avoid the problem, and even models based on artificial neural networks and deep learning are unable to improve the metrics. Sometimes, even when features allow classification or regression in reported cases, performance metrics of supervised learning algorithms remain unsatisfyingly low. This problem is recurrent in many areas of study as, per example, the clinical, biotechnological, and protein engineering areas, where many of the attributes are correlated in an unknown and very non-linear fashion or are categorical and difficult to relate to a target response variable. In such areas, being able to create predictive models would dramatically impact the quality of their outcomes, generating an immediate added value for both the scientific and general public. In this manuscript, we present RV-Clustering, a library of unsupervised learning algorithms, and a new methodology designed to find optimum partitions within highly non-linear datasets that allow deconvoluting variables and notoriously improving performance metrics in supervised learning classification or regression models. The partitions obtained are statistically cross-validated, ensuring correct representativity and no over-fitting. We have successfully tested RV-Clustering in several highly non-linear datasets with different origins. The approach herein proposed has generated classification and regression models with high-performance metrics, which further supports its ability to generate predictive models for highly non-linear datasets. Advantageously, the method does not require significant human input, which guarantees a higher usability in the biological, biomedical, and protein engineering community with no specific knowledge in the machine learning area.

Keywords: highly non-linear datasets, supervised learning algorithms, clustering, statistical techniques, recursive binary methods

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda,
Institute of Applied Mathematics and Systems Research, National Autonomous University of Mexico, Mexico

Reviewed by:
Sotiris Kotsiantis,
University of Patras, Greece
Georgios Kostopoulos,
Department of Mathematics, School of Science, University of Patras, Greece

****Correspondence:***
Álvaro Olivera-Nappa
aolivera@ing.uchile.cl

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *09 September 2019*
Accepted: *22 January 2020*
Published: *13 February 2020*

Citation:
Medina-Ortiz D, Contreras S, Quiroz C and Olivera-Nappa Á (2020) Development of Supervised Learning Predictive Models for Highly Non-linear Biological, Biomedical, and General Datasets. Front. Mol. Biosci. 7:13. doi: 10.3389/fmolb.2020.00013

INTRODUCTION

In the so-called era of Data, Big Data seems to be a common term. As the name suggests, its determining characteristic is the amount of information, a quantity so large that it has required the development of new technologies and algorithms to obtain useful information from them (Katal et al., 2013; Sagiroglu and Sinanc, 2013; Gandomi and Haider, 2015). The above has attracted the interest of various actors, and among them, the field finds enthusiasts, detractors, and skeptics. In recent times, academic interest in Big Data revealed by the number of journals, conferences, and initiatives dedicated to the subject, has shown a consistently growing trend (Ekbia et al., 2015; Gandomi and Haider, 2015). From this increase, we can infer that, in addition to introducing new study directions and fields, Big Data has changed how research is carried out (Abbasi et al., 2016). The proliferation of information generators has created gigantic volumes and great diversity of data, and the evolution of the methods to analyze, store, transmit, and use them are radically reforming the scientific computing scenario (Hu et al., 2014; Asch et al., 2018; Oussous et al., 2018). Machine Learning (ML) techniques are an example of such methods (Al-Jarrah et al., 2015; Qiu et al., 2016; Zhou et al., 2017).

ML operates under the premise that it is possible to learn from the data and to generate predictions from the trends it may exhibit. ML, and any learning process in general, first involves a pattern discrimination stage, which is subsequently used for conjecturing predictions for new examples. Among the best-known ML methods, two separate groups can be drawn: supervised learning (Singh et al., 2016) and unsupervised learning (Ghahramani, 2003) methods. The first group of methods, usually associated with the classification and regression tasks, requires knowledge about a response variable, which is assumed to be related to and inferred from it. The second group of methods, generally related to clustering or pattern recognition tasks, does not require a previously known response variable since the output is clusters of behaviors that naturally emerge from the data (Witten et al., 2005). Examples of widely-used ML techniques are Artificial Neural Networks (ANN), Decision Trees (DT), Support Vector Machines (SVM), Naïve Bayes, *k*-nearest neighbors (KNN), and ensemble methods such as Boosting or Bagging, among others (Witten et al., 2005; Kourou et al., 2015). A general weakness of ML techniques, reported in different tenors, is an intrinsic part of their core: as they train from limited data, their results depend on their limited experience and, lacking a theoretical background, they frequently fail to cast predictions over exotic examples not present in the training set (Kourou et al., 2015; Michael et al., 2018). Some researchers commonly classify ML-trained models as "black boxes," a term that results quite accurate for the ANN's applications (Olden and Jackson, 2002; Qiu and Jensen, 2004). However, models as DT, SVM, and KNN, for example, actually do rescue information about the decision-making workflow in their architecture, giving some insights about the reasons behind their results. In the area of biomedicine, where the applications are wide and very promising (Costa, 2014; Greene et al., 2014; Lee and Yoon, 2017), researchers call for a new era in the application of ML (Camacho et al., 2018), where the incorporation of information will be a key feature for success (Auffray et al., 2016; Michael et al., 2018). For instance, applications of ML may be found in studies related to cancer diagnosis and treatment (Kourou et al., 2015; Hinkson et al., 2017), diabetes research (Kavakiotis et al., 2017), decision support in critical care (Johnson et al., 2016), genomic medicine (Leung et al., 2015), among others.

Many times, the datasets do not have information about how their features interact to generate responses or clusters, which, added to the noise that datasets usually have, complicates its treatment. Researchers have pointed out this fact, emphasizing that it is difficult to bridge the gap between prediction and reality if the mechanistic background of the phenomenon to be predicted is not evident (Coveney et al., 2016). Depending on how complex the underlying relationships between the features are, classification or prediction models would be trained more or less smoothly. However, that complexity could also represent a prohibitive constraint, resulting in unacceptable performances of the trained models. Consequently, we may find natural that the success of ML techniques when training predictive models strongly rely on the data. In this work, we will call *linear datasets* those in which ML methods based on linearity assumptions generate models with *outstanding* performance measures. We will refer those datasets in which this does not happen as *non-linear datasets*. Some datasets result too complicated for linear models but may be suitable for applying mildly non-linear algorithms, such as non-linear Functional Data Analysis (FDA), Random Forest, AdaBoost, Gradient Tree Boosting, among others, or after performing a data pretreatment stage (Kourou et al., 2015). For this work, we will focus on those datasets in which, even after attempting to apply non-linear techniques, trained models do not reach acceptable performance. We will refer to these sets as *highly non-linear datasets*.

Previous works handling non-linear biological and biomedical datasets have used different Machine Learning-driven approaches to obtain predictors. Some of them use artificial neural networks (ANN) because of the high-performance metrics that these methods might achieve (Almeida, 2002; Rani, 2011; Shaikhina and Khovanova, 2017). Nevertheless, such performances can be altered by modifying the network hyperparameters (such as the number of layers or neuron units), often on the cost of overfitting the data. Other works have applied distance-based methods such as KNN (Ahmad et al., 2017), kernel-driven spatial transforms as SVM (Shi et al., 2013; Xiang et al., 2017), and variations of Partial Least Squares PLS (Sun et al., 2017), all after performing a specially tailored data pretreatment. This non-standard pretreatment results in the loss of generality of such approaches. Examples of the used data pretreatment techniques are classical Principal Components Analysis (PCA) and its variants, Factor Analysis (FA), and non-linear approaches as the t-distributed Stochastic Neighbor Embedding (t-SNE), Laplacian Eigenmaps and Locally Linear Embedding (LEM), and isometric mapping Isomaps (ISO), among others (Lee et al., 2008; Pandit et al., 2016; Rydzewski and Nowak, 2016; Doerr et al., 2017; Tribello and Gasparotto, 2019).

Since highly non-linear datasets are usually obtained while gathering scientific data, attempts have been performed using them to somehow develop predictive or interpretative models. However, these approaches lack generality as they have usually been developed for particular applications and used bare algorithms, which were combined with data pretreatment techniques, as described above, to increase performance metrics. Some of the examples we will use as study subjects in this manuscript relate to the fields of protein engineering, specifically stability assessment on point mutations (Capriotti et al., 2005; Masso and Vaisman, 2008; Getov et al., 2016) and protein localization in *E. coli* (Horton and Nakai, 1997; Zhang and Ling, 2001; Deshpande and Karypis, 2002; Ratanamahatana and Gunopulos, 2002), and clinical medicine, such as mammographic mass evolution (Elter et al., 2007) and thoracic surgery. Yet, the generation of a general methodology to treat these (highly) non-linear datasets in order to get predictive models is still an open problem, which we intend to tackle in the present manuscript.

Aiming to solve the model training underperformance issue over highly non-linear datasets, we present RV-Clustering, a library programmed in Python language, optimized for the development of predictive models for these datasets. In the following sections, the different modules implemented in the library and a new methodology to adequately obtain models in a highly non-linear dataset are described in detail. Following the workflow proposed by our methodology, the library implements different stages of data pretreatment and linearity assessment. In case the dataset is proven to be highly non-linear, the recursive binary partition, which is the central point of the algorithm, is carried out. The idea behind the method is the following: first, using unsupervised learning methods, a partition of the input dataset is generated. Afterward, different predictive models are locally trained in each subset, taking advantage of similarities among subset members to reach better performance metrics. After the local models are trained, they are validated and combined to form a meta-model. Before casting predictions on new cases, a global classification model is created to assign them to the subset where they belong, according to their features. The predictions result from applying the local meta-model on the new examples. We have successfully tested the proposed methodology in several highly non-linear datasets from a broad spectrum of origins, such as from the biomedical, biotechnology, and protein engineering areas. The versatility introduced by the proposed methodology highlights its potential benefits for users from all areas of knowledge, not only limited only to the fields mentioned above.

METHODS

Both the source code and the executable elements of RV-Clustering were implemented under the Python 2.7 programming language (Oliphant, 2007), mainly using the Scikit-learn (Pedregosa et al., 2011), Python Data Analysis (Pandas) (McKinney, 2011), and NumPy (Van Der Walt et al., 2011) libraries. The RV-Clustering library was designed under the Object-Oriented Programming paradigm (Wegner, 1990), aiming to provide the modularity required to perform actions separately in the proposed workflow. We tested the different functionalities of the library through the analysis of diverse datasets, mainly extracted from bibliographic reports of specific mutations in proteins and the effect they have on their properties and stability, and from open databases, such as BRENDA (Jeske et al., 2018), ProTherm (Bava et al., 2004), and the UCI Machine Learning repository (Dua and Graff, 2017).

OVERVIEW OF THE RV-CLUSTERING METHODOLOGY

RV-Clustering is a Python library, optimized for the creation and validation of predictive models for highly non-linear datasets. Its functionalities range from the typical data pretreatment techniques to the generation of predictive models for highly non-linear datasets. Our library stands out from others because of its ease of use, its modularity, the robustness of the implemented algorithms, and its open-source access. The details about the different commands and instructions for installing RV-Clustering in a local computer are available in the authors' Github repository (https://github.com/dMedinaO/nonlinearModels). Without being specific, RV-Clustering consists of different modules aiming to:

- Provide data pretreatment techniques.
- Assess the degree of non-linearity of the dataset.
- Create predictive models based on both supervised and unsupervised learning algorithms.
- Build and train meta-models.
- Generate partitions of the dataset, where models reach high performances more efficiently while being trained.
- Evaluate performance metrics of the implemented models.

To highlight the motivation behind the proposed library and methodology, we will explain its different modules as they appear in the proposed workflow. Briefly, RV-Clustering modules for the treatment of highly non-linear datasets are based on a recursive binary partition of the initial dataset and subsequent training of the predictive models for assigning new examples to the constitutive subsets. Afterward, RV-Clustering generates different predictive models within the resulting partition, generating a battery of local models that predicts examples inside the subset. When the user wants to evaluate a new example, RV-Clustering assigns it to one of the subsets within the partition, and then the local models cast the predictions to form the output. RV-Clustering also reports the performance metrics and statistical analyses of the resulting classification model, the within-the-partition local models, and the general meta-model.

Algorithm 1: RV-Clustering methodology

```
Result: Predictive meta-model for a (highly) non-linear dataset
$x_{\mathrm{user}}$ : User defined linearity threshold for a performance metric $x$;
$x_{\mathrm{mod}}$ : Model/meta-model performance metric $x$;
Assess linearity of the dataset, $x_{\mathrm{linear}}$;
if $x_{linear} \leq x_{user}$ then
    Explore linear and mildly non-linear models within the dataset, $x_{mod}$;
    if $x_{mod} \leq x_{user}$ then
        Generate a partition of the dataset and a classification model within it;
        Generate local meta-models in subsets of the partition;
        Couple the classification model with the local meta-models to create a general model;
        Validate the general model, $x^{\mathrm{gen}}_{\mathrm{mod}}$;
        if $x^{gen}_{mod} \leq x_{user}$ then
            Suggest corrections and restart the algorithm;
        else
            Accept model;
        end
    else
        Accept model;
    end
else
    Accept model;
end
```

RESULTS

RV-Clustering Modules Through the Proposed Methodology

This section comprises the description of the different modules implemented in the RV-Clustering command library and the proposed methodology. **Figure 1** represents the workflow of our method. As an input, RV-Clustering receives the dataset and configuration parameters for the evaluation of different criteria such as the minimum percentage of elements in each group, the kind of model to be trained, and the minimum ratio accepted for the detection of class imbalance, in the case of classification models. At this stage, the user also must declare thresholds to evaluate whether the dataset is considered as linear or non-linear, and minimum expected performance metrics in the exploratory stage of predictive models.

Data Preprocessing

RV-Clustering incorporates a dataset preprocessing stage that allows encoding categorical variables using One Hot Encoder and assessing the class imbalance, if applicable. Finally, RV-Clustering standardizes the dataset and divides it into two groups: a training subset (80% of the original dataset) and a validation subset (the remaining 20%).

Evaluation of Dataset Linearity

In the first instance, RV-Clustering evaluates whether the dataset is non-linear according to our definition. To do this, the user must indicate if the desired model is for or classification. If the models to be trained are regression models, the tool applies a linear regression on the dataset based on ordinal least squares and obtains the coefficient of determination value of the result (R^2). Otherwise, it applies a variation of the Ho-Kashyap algorithm (Serpico and Moser, 2006), in which different linear classification methods, based on Support Vector Machines (SVM) and its variants, are implemented. Finally, we compare the accuracy of the obtained models with the minimum acceptance threshold defined by the user. Thus, any dataset that does not meet this criterion is classified as non-linear and is a candidate to undergo the process of recursive binary partitions.

Initial Exploration of Predictive Models

RV-Clustering allows the user to perform an exploratory stage for testing the performance metrics of predictive models based on supervised learning algorithms. This evaluation receives as input: (i) the dataset, (ii) the performance measure of interest, (iii) the minimum performance threshold, (iv) the type of response (categorical or numerical), and (v) the response column identifier.

To perform the exploration, the model training module of our tool applies different supervised learning algorithms to the dataset, depending on the type of response. After training the models, we obtain distributions of performance metrics, selecting the model with the highest performance according to the user-input metric. If the performance is higher than the threshold declared by the user, the tool reports as output the respective model and all its performance metrics. Otherwise, a message informing that no model meets the desired requirements will appear. If that were the case, there are two different actions to take that may help to reverse the result: (i) reducing the dimensionality of the dataset by selecting the most informative attributes or, on the contrary, (ii) adding further information to the dataset. The first requires knowledge about the available techniques for dimensionality reduction, while the addition of information may not be favorable if it is not informative enough and only serves to increase the noise in the dataset. Finally, if none of the options works, it is recommended to submit the dataset to the recursive binary partition stage proposed in this work.

It is essential to mention that this stage is complementary to the evaluation of the linearity of the datasets since the contemplated algorithms are not linear regressions or hyperplane generation-based. Alternatively, we instead employ probability distributions (Naïve Bayes and derivatives), evaluation of characteristics (Decision Trees), or boosting methods (Random Forest, Adaboost, Bagging, Gradient Tree Boosting) for model training.

Recursive Binary Partitions

The main objective of the recursive binary partition process is the generation of subsets from the initial dataset, wherein we could increase the performance metrics previously obtained in

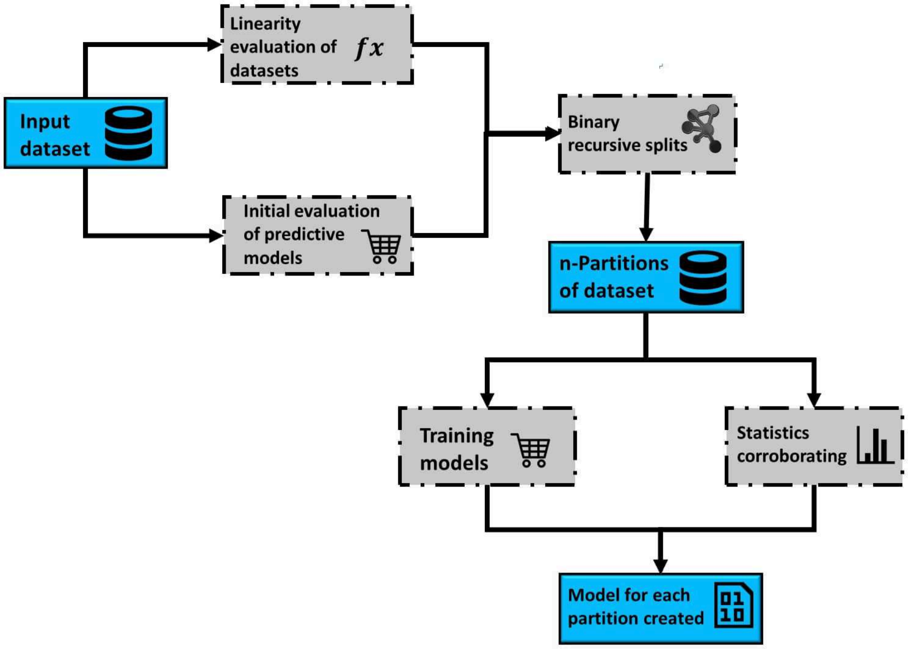
Linearity evaluation of datasets
fx
Input dataset
Initial evaluation of predictive models
Binary recursive splits
n-Partitions of dataset
Training models
Statistics corroborating
Model for each partition created

FIGURE 1 | Representative scheme of the workflow associated with the methodology proposed to develop predictive models for highly non-linear datasets, based on the use of the RV-Clustering library.

the exploratory stage of supervised learning models. A binary search trees-inspired algorithm (Bentley, 1975), where the search is optimized in the tree path, generate the partitions. In each iteration, the initial dataset is subjected to an exploration of different unsupervised learning clustering methods, such as the Birch, k-Means, and Agglomerative algorithms, conditioned to the generation of two elements. In the cases of k-Means and Birch, our algorithm automatically tests different distance metrics, while for Agglomerative Clustering, the affinity parameter and linkage methods are automatically varied. Each proposed partition is evaluated using the silhouette coefficient and the Calinski-Harabasz index. Subsequently, we evaluate the number of subset elements of those partitions that have the highest clustering performance indexes. The number of elements in each subset should be equal or higher than the minimum threshold previously selected by the user. Class imbalance generated by the partition is assessed according to a user-determined threshold for classification models. Finally, if the partition in a given iteration meets all the mentioned criteria, it is accepted, and the recursive division continues for each tree branch. At the end of the execution, we will have n subsets, which will be statistically studied to evaluate if each generated partition is significantly different from the others, if each element effectively belongs to its corresponding subset, and if all the features are informative for all subsets, in order to avoid any redundancy that could affect the model training stage.

Creation of Models to Classify New Examples in the Generated Partition

In order to classify examples within the generated partition, different classification models are created, using supervised learning algorithms. For this, the training dataset, which is already a subset of the input dataset, is divided into two sets for training (80%) and validating (20%) the classification models. The first subset undergoes a model exploratory stage training with k-cross-validation, with k-values varying depending on the size of the set. We obtain the accuracy, recall, precision, and F1 scores for each model, and also their statistical distributions. From these four distributions of performance metrics, the models with the maximum values in these distributions are selected, forming a set of at most four independent models (one per each performance metric). These four models are used to generate a weighted meta-model with a classification criterion obtained by the votation of the individual models, assigning each element to the subset pointed by the majority of the individual models. Finally, we compare the classifications generated by the meta-model with the actual values of the validation set to obtain the overall performance metrics.

Model Training

Each subset A_i within the partition generated in the binary recursive division undergoes a predictive model exploration stage, and the best j models are selected and combined to form a local meta-model. The selection criterion is associated with the maximum value of each metric of interest selected by the user, which may be accuracy, recall, precision, or F1 for classification models, or R^2, Pearson, Kendall τ, or Spearman rank coefficients for regression models, hence $j \leq 4$. RV-Clustering estimates an overall performance for the models over the entire dataset, weighting the individual metrics in the generated partition. Let x_i be a metric of the models' performance over A_i. The corresponding i-weighted performance is given by

$$\hat{x}_i = x_i \cdot \frac{|A_i|}{\left|\bigcup_{i=1}^{n} A_i\right|}, \tag{1}$$

and the final measure is obtained from the summation of the $\hat{x}_i$, which corresponds to the probabilistic expected value of x, $\mathbb{E}(x)$ assigning a probability $\mathbb{P}(A_i) = \frac{|A_i|}{|\bigcup_{i=1}^{n} A_i|}$ to the subset A_i,

$$\hat{x} = \sum_{i=1}^{n} \hat{x}_i = \sum_{i=1}^{n} x_i \mathbb{P}(A_i) = \mathbb{E}(x). \tag{2}$$

We compare the obtained weighted measure with the performance values obtained in the initial stage, reporting them both. Finally, the tool uses the validation set to obtain the real metrics $x^{\text{gen}}_{\text{mod}}$ of the general model created, and report the results associated with the classification or prediction of new examples. To do this, RV-Clustering uses the classification model to assign each example to the subset in the partition where it should belong, and then, using the local meta-model corresponding to that subset, obtain the predicted value. We compare this value with the real value and generate the performance metrics corresponding to the type of model.

An index for assessing over-fitting local meta-models within the partitions *IOF* is presented in Equation (3), defined as the difference between the expected (via Equation 2) and the real performance metric.

$$IOF = \frac{\hat{x} - x^{\text{gen}}_{\text{mod}}}{x^{\text{gen}}_{\text{mod}}}, \qquad IOF_i = \frac{x_i - x^{\text{gen}}_{\text{mod}}}{x^{\text{gen}}_{\text{mod}}} \tag{3}$$

Similarly to Equation (1), it is possible to obtain a local *IOF* for subset i, IOF_i. If the *IOF* or any of the local IOF_i values are >5% or another user-customizable value, the recursive binary partition algorithm should be repeated, conditioned to producing subsets with more elements. Negative values of *IOF* do not have any implications, as they only show that the performance of the global model is greater than the expected value, accounting for a synergy between individual meta-models.

Predicting New Examples

The proposed method creates a partition splitting the input dataset into n subsets. Hence, as we work independently in each subset, we obtain n independent meta-models. In order to classify new examples within the subsets of the obtained partition, we train a classification model, which assigns every new example to the subset where it should belong. For this, RV-Clustering classifies the new example into a particular subset in the partition, applying the predictions of local meta-models. We can directly calculate the improvement of the original result I from the linearity assessment index and the final performance metric,

$$I = \frac{x^{\text{gen}}_{\text{mod}} - x_{\text{linear}}}{x_{\text{linear}}} \tag{4}$$

CASES OF STUDY

The proposed methodology and library modules were tested with different highly non-linear datasets according to our previous definition, related to clinical diagnosis, biotechnology, and protein engineering. Each one of the proposed scenarios is presented below in three different case studies.

Case Study I: Use of RV-Clustering in Clinical Datasets

The prediction of the clinical risk associated with mutations in proteins, the probability of having a disease, or the need to carry out an invasive or dangerous exam, among others, are activities of high interest in the biomedical area. Taking this into consideration, the different points of the methodology proposed in this article were applied to three highly non-linear datasets, which represent Mammographic Mass, Heart-Disease, and Thoracic Surgery. The datasets were extracted from the UCI-Machine Learning (Dua and Graff, 2017) repository and, in all cases, the required models are of the classification type, since their response is categorical.

When performing the linearity assessment, all the datasets turned out to be highly non-linear, considering a minimum threshold of 0.8 for the linearity metrics. This stringent criterion was selected to impose a high quality of the classification since false positive and false negative errors should be as low as possible for a clinical test. The performances obtained in the model exploration stage using mildly non-linear methods did not reach the minimum threshold values, so RV-Clustering proceeded to apply the binary partition methods proposed in this work. **Figure 2** shows the partition generated for each dataset. In each case, the cardinality of the generated subsets varies as the depth of the resulting binary tree increases. The performance metrics obtained for Mammographic Mass and Thoracic Surgery models applying the proposed methodology is considerably greater than those obtained in the exploratory stage since accuracy is improved from 54 to 87% in the first case, and from 71 to 83% in the second case. In the Heart-Disease Cleveland dataset, no considerable improvement was achieved. We consider this to be due to the large number of classes presented by this dataset. Given this result, as RV-Clustering ensures class balance in each subset within the partition, the recursive binary partition method should not be used with datasets whose response categories are >5, especially when the number of examples is limited, because it may lead to detriments on the performance metrics initially achieved. This limitation arises from the lack of information in the dataset itself, as the generation of regressions or predictions of high-dimensional responses based on few data examples remains an open problem.

Case Study II: Use of RV-Clustering in Biotechnological Datasets

Another approach of a broad interest in the use of data mining and ML techniques is the development of predictive models for the optimisation of experimental plans in biotechnological applications. Through the generated predictive models, it is possible to reduce the use of economic and human resources

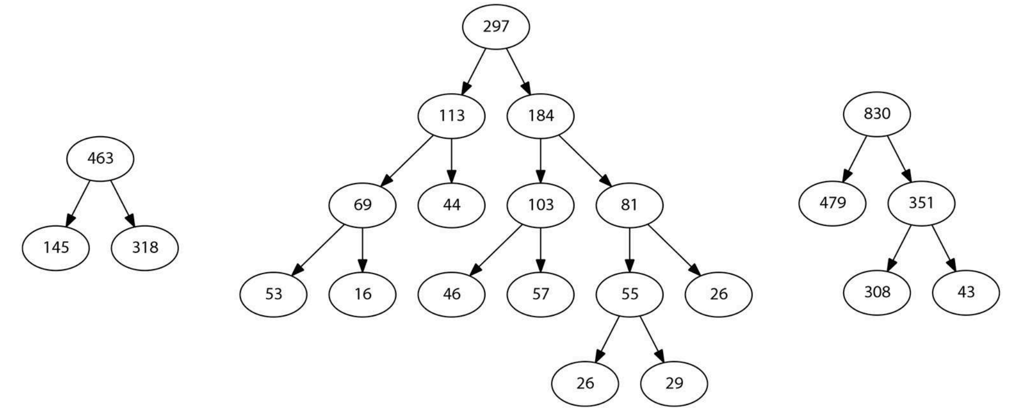

FIGURE 2 | Representative schemes of the partitions and the flows of divisions generated for the example datasets associated with case study I: Thoracic Surgery dataset **(left)**, Heart Disease Cleveland dataset **(center)**, and Mamographic Mass dataset **(right)**. The number of final partitions, their cardinality and the performance measures achieved by the models trained in each case are also presented.

and the duration of the experimental projects dramatically. As an example, a dataset with information on the classification of protein localisation sites in *E. coli*, extracted from the UCI Machine Learning repository (Dua and Graff, 2017), will be used. This dataset was subjected to the linearity assessment, contemplating a minimum acceptance threshold of 0.7 in linearity metrics. As the highest accuracy achieved was 56%, RV-Clustering classified this dataset as non-linear. However, when applying the model exploration module, satisfactory results were obtained. The distributions presented in **Figure 3** show a set of models that have performance measures greater than those of the threshold imposed. Hence, it is not necessary to proceed to the binary recursive partition stage. The best models trained in the exploration stage are selected to create a weighted meta-model, whose accuracy and precision reached 88.1 %.

In particular, given the properties of the input dataset, it was possible to obtain a meta-model with performance metrics above those imposed as an experimental requirement, only by applying the exploratory module. This fact highlights the efficiency of RV-Clustering, always aiming to satisfy the user requirements to obtain as-good-as-required models as fast as possible and without incurring in greater trade-offs in quality-time. Using the modules implemented in RV-Clustering, it was possible to improve the initial accuracy of 56% to a value of 88.1%, confirming that the proposed workflow is appropriated. It is crucial to know which algorithms are the most suitable for a given application, and it is a great advantage of RV-Clustering to test them in such a way that all the possibilities are evaluated, without requiring any specific knowledge on algorithms for getting high-quality results.

Case Study III: Use of RV-Clustering for the Evaluation of Protein Stability Given Point Mutations

The evaluation of the effect that point mutations have in protein stability is one of the most visited topics in protein engineering. Different approaches have been proposed, considering methods based on electrostatic potentials, statistics, ML techniques, among others. The methods mentioned above allow a mutation to be classified as stable or non-stable or to generate stability predictions based on the difference in free energy ($\Delta\Delta G$) caused by the replacement of the residue. Applying the approach proposed by Capriotti et al. (2005) for describing mutations and considering three independent descriptors, thermodynamic, structural and residue-environmental, a dataset comprising 11 proteins and 2,247 mutations associated was generated (see **Figure 4**, left). In the created dataset, the response column represents the $\Delta\Delta G$ values, associated with the difference between mutated residue and wild residue. These values were obtained from the ProTherm (Bava et al., 2004) database.

The application of the linearity assessment module classified the dataset as non-linear, since the performance metrics obtained by applying linear methods did not exceed the threshold of 0.6 for predictive models. Furthermore, as it was not possible to achieve significantly higher performance measures in the model exploration stage, the dataset was classified as highly non-linear. By applying the proposed methodology for binary recursive partition, nine subsets were obtained (see **Figure 4**, right), and different meta-models were developed locally. Intra-partition over-adjustment was avoided by applying a k-cross-validation, with $k = 10$. Subsequently, a meta-model for the classification of new examples to the different partitions was generated. Finally, the general metrics of the model were obtained for the validation set (see **Figure 5**, left). By comparing the resulting performance metrics and the initial values obtained in the exploration stage of predictive models, an average improvement of 40% was achieved in each measure of interest. For example, the initial Pearson's coefficient of 0.58 was improved to 0.92 after applying the methodology here presented. A scatter plot of the real and predicted values for the effect of point mutations shows that the error distribution has a random and bounded behavior (see **Figure 5**, right), which corroborates the quality of the obtained results.

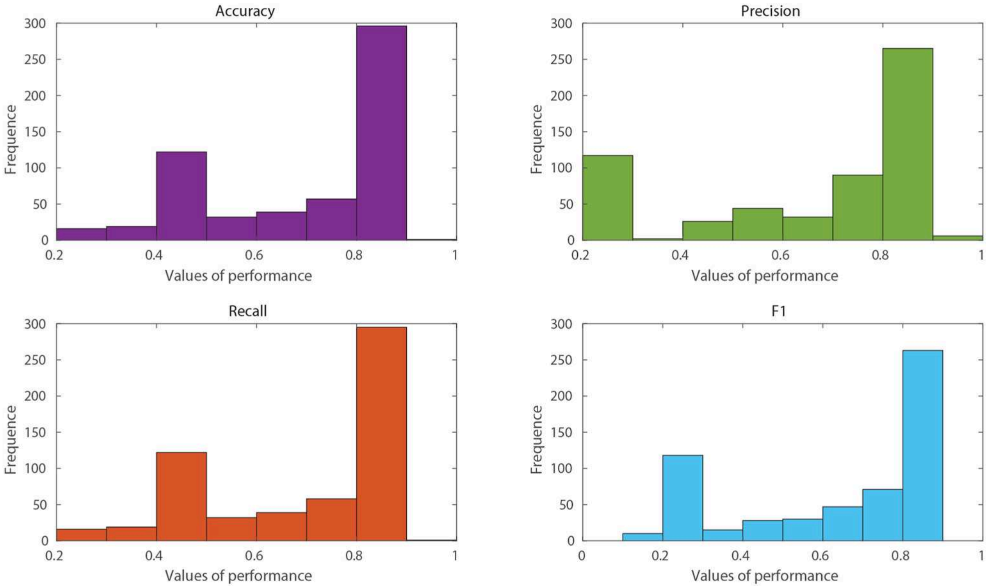

FIGURE 3 | Histograms of performance metrics obtained in the exploration stage by the RV-Clustering library for the protein location in an *E. coli* dataset. The highest values were obtained by methods based on Bagging or Boosting algorithms, accounting for the non-linearity of the dataset.

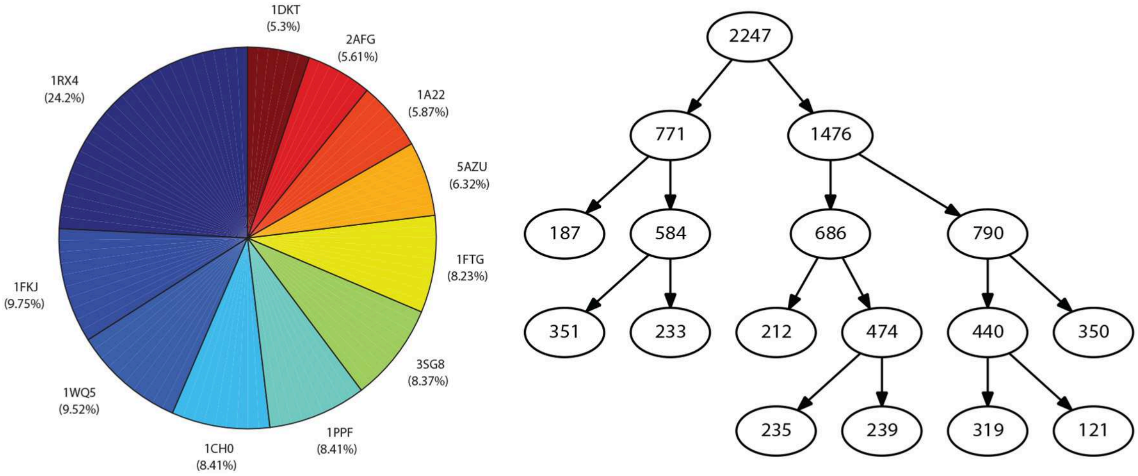

FIGURE 4 | Representation of the dataset associated with case study III: Distribution of mutations for the considered proteins **(left)**, and Resulting partition after applying the methodology proposed in this work **(right)**.

DISCUSSION

Improvements on Performance Metrics

The different datasets tested in the cases studies serve to illustrate the great capacities of the proposed method since it not only improves the performance measures, but it does so efficiently from a computational point of view, generating as-good-as-required models in the shortest time possible. This result is achieved thanks to the RV-Clustering library modularity and the structure of the presented methodology, which considers advancing to the next complexity level only when models generated so far do not meet user requirements.

Another advantage of this new approach is the transparency of the results. Model performance metrics, by themselves, may not be sufficiently informative and mislead to wrong conclusions about the quality of the predictive outcome; they should always be analyzed in context. In our work, the different metrics associated with different elements (models, meta-models, global model) are analyzed together and combined using the proposed indexes. This combination of metrics is used both for improvement evaluation between the initial linear assessment stage and the final performance and for the evaluation of over-fitting in local meta-models within the partition. **Table 1** presents the results of the considered cases of study, all of which show a significant

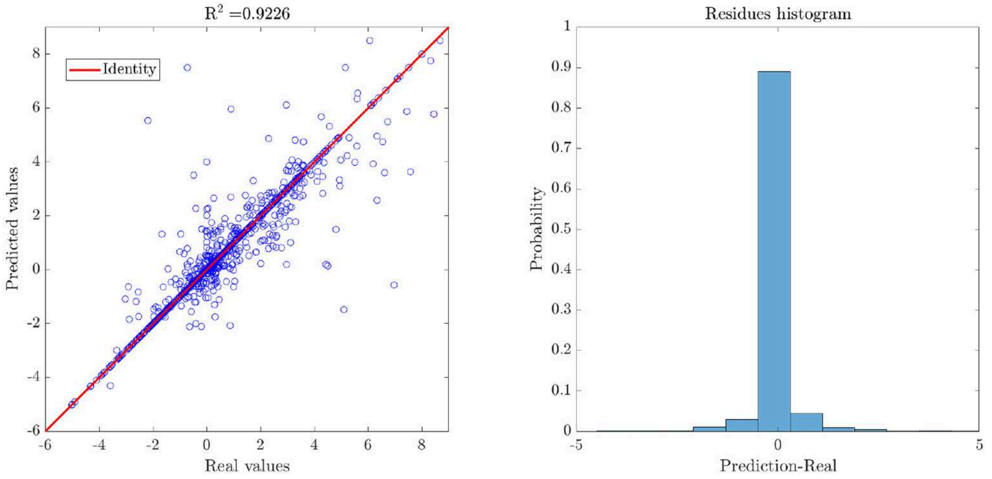

FIGURE 5 | Results of the generated weighted meta-model, where the predicted values are obtained from the average of the predictions of the individual methods. As the error seems to have a random distribution around zero, $\Delta\Delta G$ values predicted by the meta-model do not present considerable biases.

TABLE 1 | Evolution of performance (accuracy) of the models generated in different progressive steps of the proposed methodology.

Dataset	x_{linear}	$\hat{x}$	$x^{\text{gen}}_{\text{mod}}$	Improvement after applying RV-Clustering's methodology (%)	*IOF*(%)
Mammografic mass	0.54	0.85	0.87	61.1	−2.3
Thoracic surgery	0.71	0.78	0.87	22.5	−10.3
Protein location in *E. coli*	0.56	–	0.88	57.1	–
Protein stability*	0.58	0.82	0.92	48.3	−4.7

*The performance of the final RV-Clustering generated model is represented by x^{gen}_{mod}, while x_{linear} and $\hat{x}$ are the results of intermediate steps of the method (linear assessment step and model exploration step, respectively). * Pearson's coefficient.*

TABLE 2 | Comparison of reported performance metrics for the studied experimental datasets.

Dataset	Reported by	Reported performance	RV-Clustering performance
Protein stability (point mutations)	Capriotti et al., 2005	0.71	0.92
	Deshpande and Karypis, 2002	0.73	
Classification of protein location in *E. coli*	Zhang and Ling, 2001	0.84	0.88
	Horton and Nakai, 1997	0.68	
	Ratanamahatana and Gunopulos, 2002	0.84	
Mammographic mass	Elter et al., 2007	0.87	0.87
Thoracic surgery	None	None	0.87

improvement in their metrics. No over-fitting of the local meta-models was observed in the different subgroups of the partition since all *IOF* values were negative. The previous discussion also accounts for synergistic effects between the classification model and the different meta-models within the partition, since overall performance metrics are higher than weighted individual ones. All of the above translates into an average percentage increase of 47.3% in the performance metrics of the predictive models for the highly non-linear biological datasets considered, as presented in **Table 1**. As the performance metrics increase as the methodology proceeds, the best model will always be the latest delivered (except in cases where $IOF > 0$). To stop at early stages by imposing lower values of x_{user} is a decision based on a time-quality trade-off, as our methodology was thought for delivering as-good-as-required models.

Table 2 presents the overall improvement in the performance metrics after applying our methodology, compared to the values reported in the original works. As our methodology incorporates most of the best state-of-the-art available algorithms and progressively applies them, the worst scenario would always be better than the original one.

Testing on Artificial Datasets

In order to test the proposed methodology and the robustness of our library, we generated different artificial datasets with tailored properties, aiming to evaluate its response against (a) noise intensity, (b) presence of outliers, (c) degree of non-linearity of the input dataset, and (d) maximum dimension of the input dataset, with further recommendations based on the fitting procedure.

Given that our methodology is very intuitive to understand when applied to regression models (as discussed in section 3), all models trained in this section were of the regression type. We explain each of the cases in the subsections below.

Noise Intensity

To show the influence of noise intensity or experimental errors, we tested our methodology with two different datasets: an artificial dataset containing a linear ground truth function, and the dataset of Case Study III. We introduced an additive proportional error to the response variable, characterized by a variable amplitude α. Adding this error to the experimental values y_{exp} resulted in the following expression for y_i^{noise}:

$$y_i^{\text{noise}} = y_{\text{exp}}(x_i)(1 + \alpha(2\mathcal{U} - 1)) \tag{5}$$

where $\mathcal{U}$ is a random variable with uniform values in $[0, 1]$. Equation (5) was selected because of its statistical properties, given that the expected value of the noisy random variable is its corresponding experimental value:

$$\begin{aligned} \mathbb{E}\left(y_i^{\text{noise}}\right) &= y_{\text{exp}}(x_i)\,(1 + \alpha\,(2\mathbb{E}\,(\mathcal{U}) - 1)) \\ &= y_{\text{exp}}(x_i)\left(1 + \alpha\left(\cancel{2}\frac{1}{\cancel{2}} - 1\right)\right) \\ &= y_{\text{exp}}(x_i). \end{aligned}$$

Aiming to test how heavily the increasing noise impacts the performance metrics, we considered two scenarios: (a) adding α-noise to the experimental $\Delta\Delta G$ data (Case of Study III), classified as highly non-linear, with an unknown ground truth function, and (b) adding α-noise to numerical experiments with known ground truth $y = x$, which included a white Gaussian noise with $\sigma = 5\%$, in order to resemble real-world measurements. For both scenarios, we considered $\alpha \in [5, 10, 20, 30, 40, 50\%]$, as shown in **Figure 6**.

In the first case, as the ground truth function is linear, we set $x_{\text{user}} = 0.95$ to force our algorithm to move forward into the second step of our methodology. However, even when the noise intensity was $\alpha = 20\%$, models generated in the first step of our methodology (linear assessment stage) still reached performance metrics over the threshold $x_{\text{linear}} > x_{\text{user}}$. When the noise intensity was higher, linear models did not meet the required performance, but those generated by DTs and RF did, preventing the algorithm from entering into the binary splitting stage. Despite the generated models reaching high-performance measures at every α−noise scenario (see the left plot in **Figure 8**), a bifurcation in the quality of the predictive outcome appears when the nature of the training algorithms shifts from linear regressions to DTs. As shown in **Figure 7**, the scatter plot of predictions and ground truth (original data without noise) present high dispersion when $\alpha \geq 40\%$, even though models reached high performance metrics, which accounts for models fitting the noisy data rather than the original trend. The above highlights the need for a preliminary analysis of the data, as moderate to high noise can mislead the results and affect the quality of the produced models. However, a 40% or higher noise level is large by any measure, and would not be usually considered as simple noise but rather as a composition of signals. In this sense, the fitting given by our

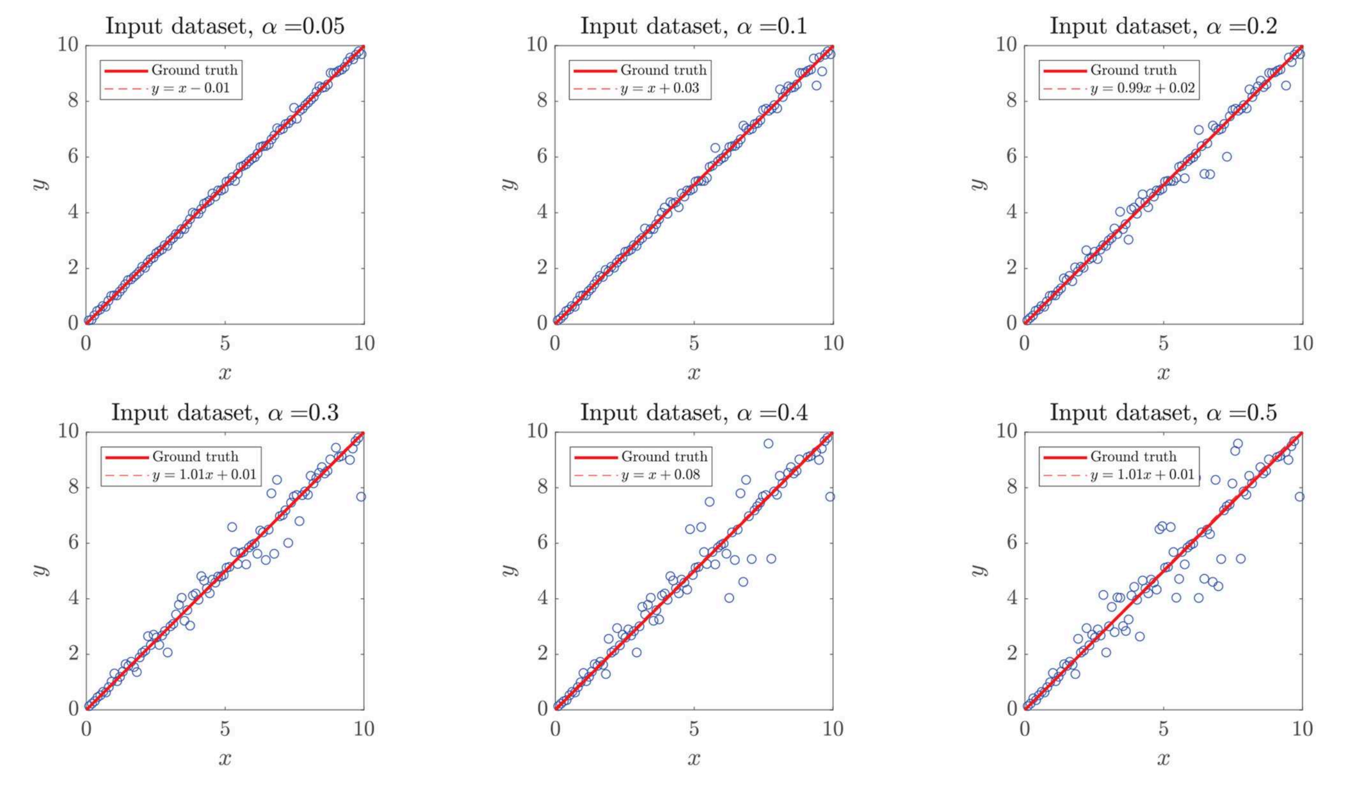

FIGURE 6 | Simulated dataset with added white noise α. The plots represent simulated (y) vs. ground truth (x) data points (circles), the identity line (continuous line), and the crude statistical regression of the resulting dataset (discontinuous line). Added noise followed a Gaussian distribution around the expected value $y = x$, not affecting the expected value of the distribution, which translates to regression lines very similar to the identity.

algorithms in the presence of high "noise" points into the right direction by identifying the data points as coming from a model different from the linear ground truth function. The outstanding predictions of the models generated at low α can be explained by the linear nature of the ground truth.

When the considered dataset was classified as highly non-linear, added noise had a stronger impact on performance metrics, as shown in the center graph of **Figure 8**. In this case, the range of the $y-$axis is much wider than in other cases. Noise levels over $\alpha = 20\%$ have a more significant impact over

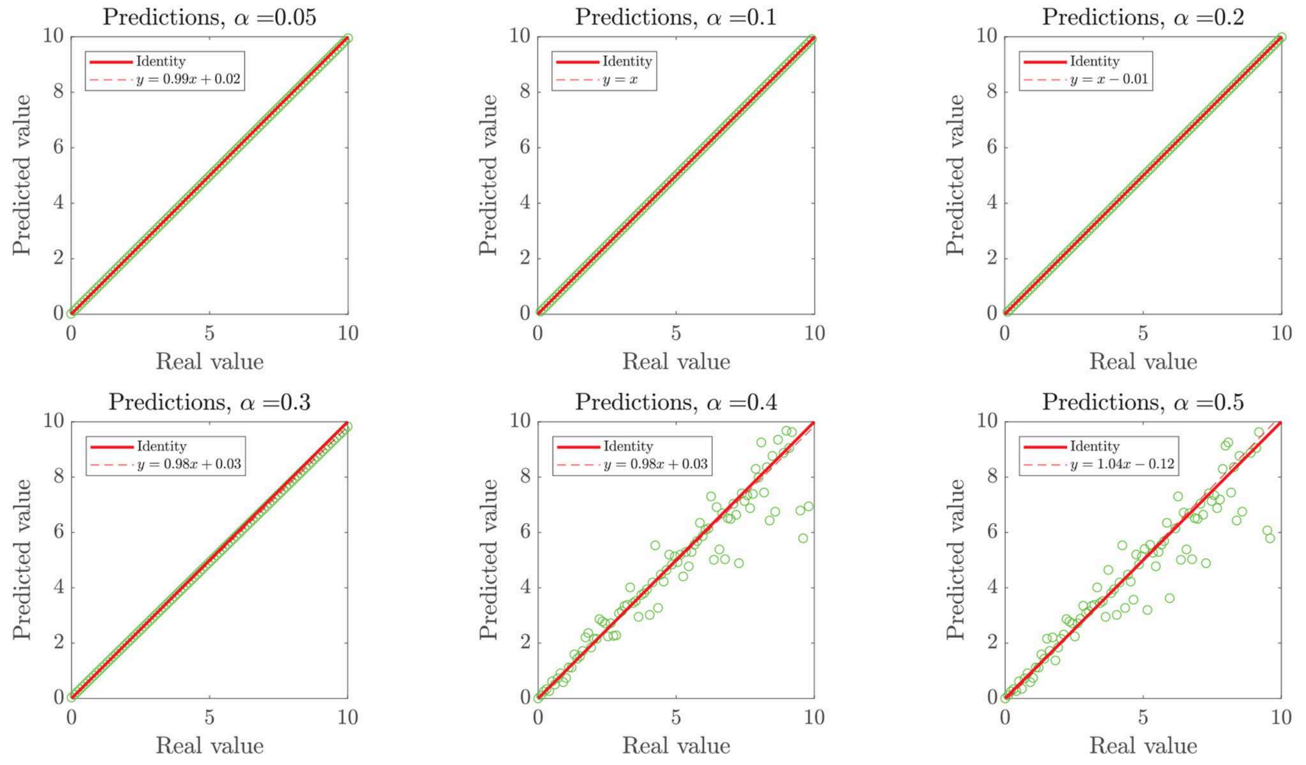

FIGURE 7 | Model-prediction of the simulated linear dataset with $\alpha-$ induced noise in 100 data points. The plots represent predicted (y) vs. real (ground truth without noise, x) data points (circles), the identity line (continuous line), and the crude linear statistical regression of the scatter (discontinuous line). Since training datasets for models included noise, we expect particular discordance between the dispersion of high α scenarios and the predictive outcome of noise-fitting models trained therein, when compared to the original noise-free dataset.

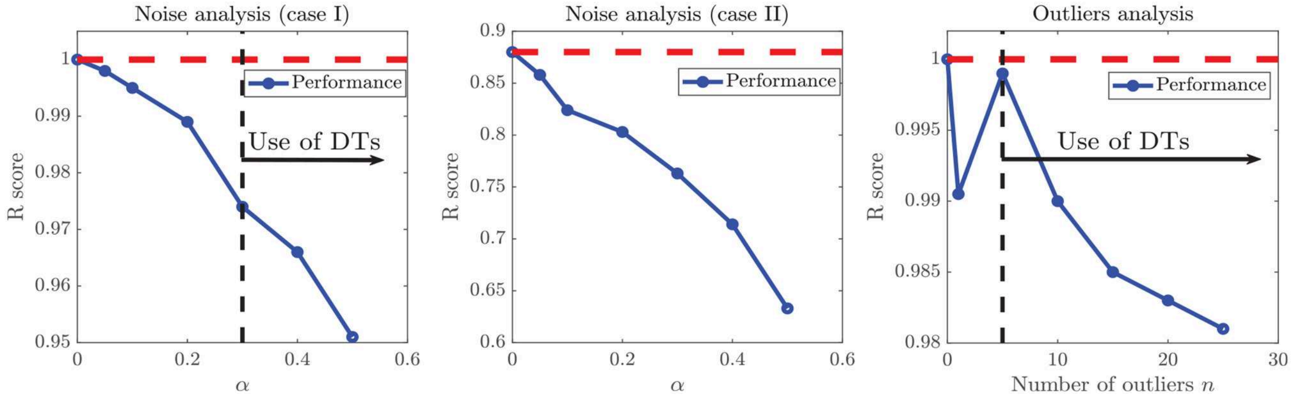

FIGURE 8 | Evolution of model performance metrics against noise. **(Left)** Model performance on a linear ground truth function with white noise. **(Center)** Model performance on experimental data (Case Study III) with white noise. **(Right)** Model performance on a linear ground truth function with different number of outliers. In artificial datasets with linear ground truth functions (left and right images) x_{user} was set equal to 0.95 to force the algorithm to continue further in the proposed methodology. When the linear model performance fell under the selected threshold, the algorithm swap to DT models, which rose the performance metrics again, generating a break in the sloping trends.

performance metrics, since the slope of the α vs. performance curve is always decreasing. Such impact can be assessed from the decrease in the improvement after applying the RV-Clustering methodology (see the sixth and seventh row of **Table 3**). Given that noise levels under $\alpha = 20\%$ do not have a severe impact on the performance metrics of the generated models, we show our methodology to be robust against low to moderate white noise.

Presence of Outliers

To evaluate the robustness of our methodology and command library against the presence of outliers in the dataset, we performed the following numerical experiment. Starting with data with a known ground truth function, $y(x) = x$, we added a white Gaussian noise $N_1 \sim \mathcal{N}(0, \sigma_1 = 0.25)$. Hence, our "experimental" dataset was the collection of random variables $y_i^{\text{noise}} \sim \mathcal{N}(x_i, \sigma)$. To simulate the existence of n outliers, we superposed a flat Gaussian distribution $N_2 \sim \mathcal{N}(0, \sigma_2 \gg \sigma_1)$, as depicted in **Figure 9**, and applied the method described in the Algorithm 2.

We simulated different datasets of $N = 100$ examples, and turned n of them into "outliers," with $n = \{1, 5, 10, 15, 20, 25\}$, as shown in **Figure 10**. Noticeably, the added outliers modify the nature of the original Gaussian distribution, which is demonstrated by the drift between the identity and the purely statistical regression of the data points as more outliers are added to the dataset. In such sense, those outliers drift considerably from the expected values of the original distribution. Nevertheless, the presence of less than ~10% outliers does not affect the performance of the final model. Even when outliers are not symmetric (see examples with 5, 10, and 15 outliers in **Figure 10**).

TABLE 3 | Evolution of performance (accuracy) of the models generated in different progressive steps of the proposed methodology, applied to noisy variations of the dataset used in Case Study 3.

Induced noise α[%]	x_{linear}	$\hat{x}$	$x^{\text{gen}}_{\text{mod}}$	Improvement after applying RV-Clustering's methodology (%)	*IOF*(%)
0	0.58	0.82	0.92	58.62	−10.87
5	0.57	0.79	0.86	51.86	−7.58
10	0.55	0.78	0.82	48.74	−5.22
20	0.55	0.77	0.80	47.34	−3.86
30	0.54	0.75	0.76	41.04	−1.57
40	0.53	0.69	0.71	33.96	−3.64
50	0.52	0.59	0.63	22.44	−6.32

As shown in the right plot of **Figure 8**, the presence of outliers negatively affects the linearity of the dataset as perceived by the methodology, since linear models do not meet the required performance and the RV-clustering workflow would move forward to DTs and non-linear algorithms. Nevertheless, and once again because of the linearity of the ground truth function, DTs would produce models with outstanding performance, producing a clear break in the sloping trend of the n vs. performance curve of **Figure 8** and preventing the algorithm

Algorithm 2: Numerical experiment with simulated outliers.

```
Result: Simulated outliers for the numerical experiment
$p_0$ : cumulative probability of $\mathcal{N} \sim (0, \sigma_2)$, at $x = \sigma_1$.;
while $j \leq n$ do
    $k$ : random integer between 1 and $N$;
    $s = \mathcal{U}$, and $p = (1 - p_0) \cdot \mathcal{U}'$, where $\mathcal{U}$ and $\mathcal{U}'$ take
    uniform values in $[0, 1]$;
    if $s > 0.5$ then
        $s = 1$;
        $p = 1 - p$;
    else
        $s = -1$;
    end
    $y_k^{\text{noise}}$ = inverse of the cumulative probability function of
    $\mathcal{N}' \sim (x_k + 2s\sigma_1, \sigma_2)$, at probability $p$.;
    $j = j + 1$;
end
```

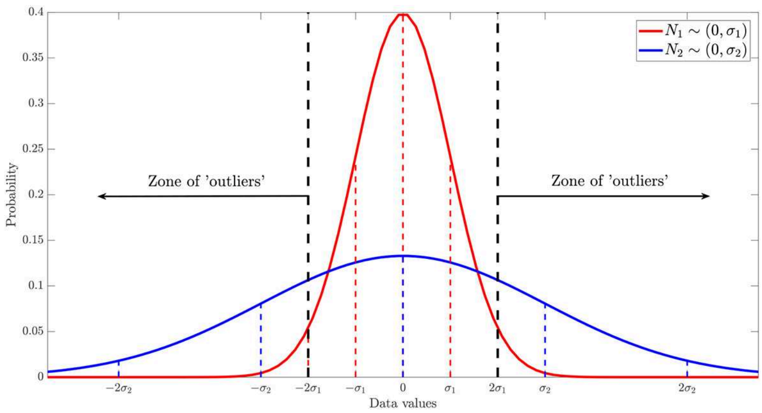

FIGURE 9 | Scheme of the non-arbitrary and statistical methodology proposed to generate outliers, given a known dataset with random error.

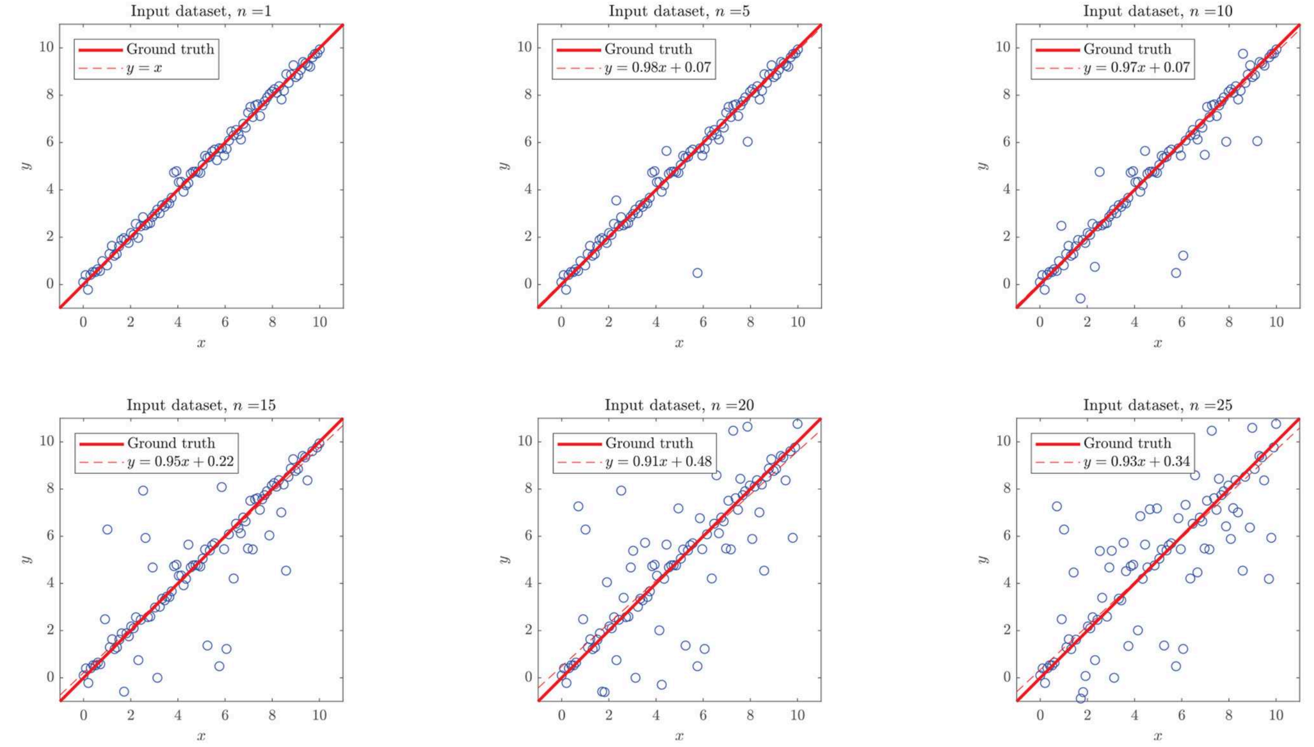

FIGURE 10 | Simulated linear dataset with n outliers in 100 data points. The added outliers modify the nature of the distribution, as they make the regression considerably drift from the identity. In such sense, those outliers might not be "pure" in a strict statistical definition (as they depend on the distribution), they drift considerably from the expected values of the original distribution.

to proceed to the recursive binary splitting stage. When a high number of outliers are expected within the dataset, we recommend to directly proceed to probability-based methods by setting a high x_{user} threshold.

Degree of Non-linearity of the Input Dataset

To evaluate the robustness of our methodology and command library against the degree of non-linearity of the ground truth function, we simulated different points from the 2-D Rosenbrock function (Rosenbrock, 1960), with parameters $a = 5$ and $b = 2$, over the $[0, 3]^2$ rectangle. Data for the numerical experiment were randomly extracted from the $[0, 3]^2$ rectangle, and a proportional white Gaussian noise was added to resemble experimental conditions. When setting a threshold $x_{\text{user}} = 0.9$ the dataset would be classified as non-linear, and the methodology would proceed to explore non-linear algorithms for training models. Among the algorithms that produced models with outstanding performance metrics, we found DTs (0.998), Bagging (0.995), Random Forest (0.995), KNN (0.98), and Adaboost (0.95), with an over-fitting assessment of k-cross-validation, $k = 10$. As expected, given the non-linear nature of the ground truth function of the dataset, the best performing algorithms mentioned above are based on feature analysis, bagging, or boosting. In particular, we expected KNN to be within the outstanding algorithms, given its distance-based generation of predictions, although it occupies only the fourth place among the best predictors.

Visually, we can corroborate that the best models were those based in DTs, Bagging and Random Forest algorithms (see **Figure 11**). All these models are able to predict extreme values of the function, the local maximum at $(0, 3)$, the valley of minimum values at (x, x^2) and the extreme values around $(3, 0)$. Random Forest and Bagging model predictions are smoother than other models and are good to predict function values in sectors with higher slopes and variability. Smoothness in this frame can be interpreted as a measure of the model insensitivity to noise, which points to Random Forest models as the best ones in this respect.

Maximum Problem Size, Properties of the Input Dataset, and Further Recommendations

We tested different cases where the dimensions of the input dataset were progressively increasing, aiming to determine a size threshold for the datasets RV-Clustering may process in a reasonable time. Our exploration found special cases where the input datasets may produce errors. The maximum dataset size that can be processed is less than $10,000 \times 1,000$, i.e., 10,000 examples with 1,000 features. In the current implementation of RV-Clustering, when submitting a dataset with such dimensions, more than 16 GB of RAM are used, which results in process abortion. To prevent the situation mentioned above, we suggest applying a dimension reduction technique prior to using our methodology, and taking the resulting dataset with fewer features as the input dataset for RV-Clustering. As maximum execution time, a dataset with 10,000 examples and 500 features would take

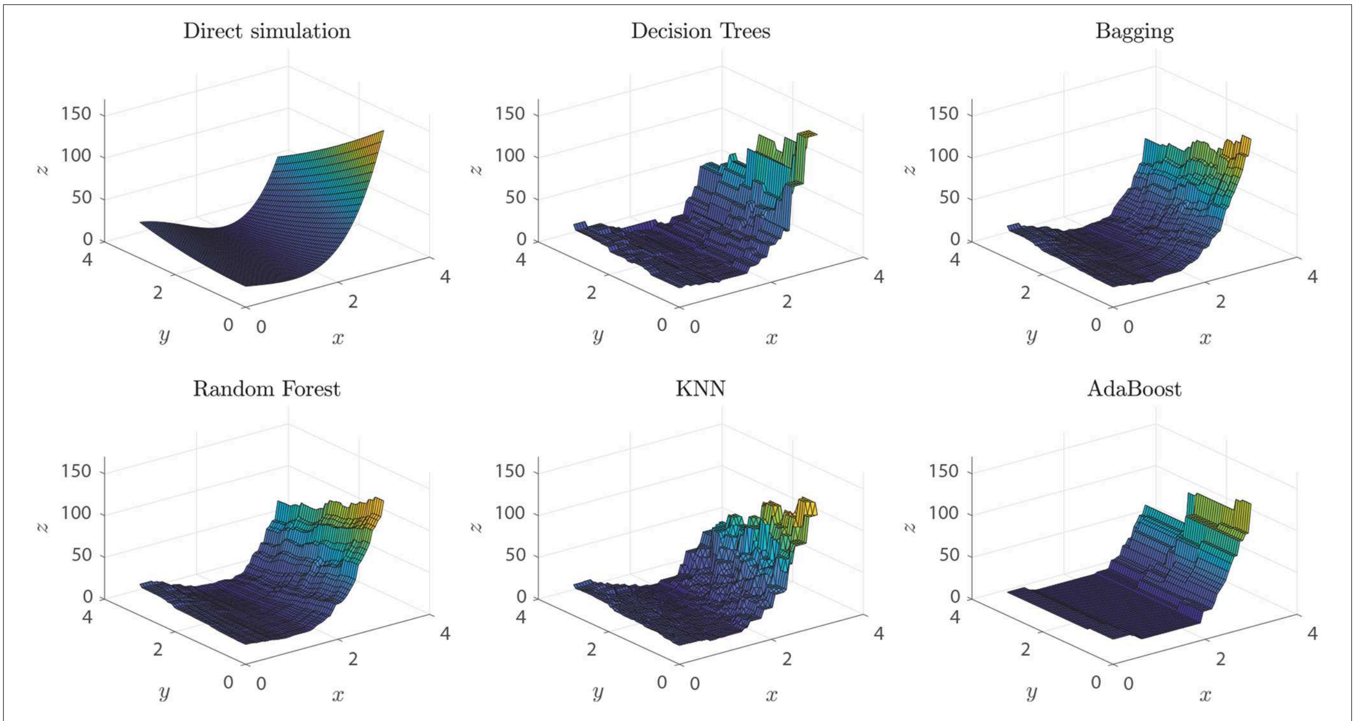

FIGURE 11 | Rosenbrock function predictions of each of the different models generated. Direct simulation represents the real ground truth Rosenbrock function.

6 days and 2 h to be processed by a seventh generation Intel Core i5 processor.

As further recommendations and good practices for using the RV-Clustering tool, we suggest:

- Standardizing numerical datasets with float size less or equal than 64.
- Keeping in mind that categorical datasets where the number of features is >20% the number of examples would be coded using One Hot Encoder, hence consuming more resources and taking much more time to be processed.
- Carefully "refining" user datasets before submitting a job to RV-Clustering. For example, numerical datasets with alphanumerical entrances would stop the process, and a warning message would pop-out.
- Especially in the case of regression models (which are not "protected" with a class balance assessment), procuring that data is well-distributed and there are no information gaps in the predictor variables. Not taking care of this situation may lead to poor fitting of the un-populated zones or filling-in with erroneous predictions if unattended, respectively. The first point can be corrected by pre-processing the data to collapse the populated zones into fewer data points to balance their weights, or selecting a different performance metric as the control variable. For the second point, unfortunately, it is not possible to find an always-working solution: as we do not know *a priori* the real values of the data in the unpopulated zone, the errors in the predictions are unbounded. We can avoid this fact being a problem for the algorithm by splitting the dataset in parts, and processing each subset separately, or forcing the algorithm to proceed to the binary splitting stage. However, this solution will not give any model prediction for the unpopulated gap zone in the original data.

CONCLUSIONS

We presented a new methodology for the design and implementation of classification or regression models for highly non-linear datasets, together with the RV-Clustering library, which corresponds to a set of modules implemented in Python that allow the manipulation of these datasets and the training of predictive models through supervised learning algorithms. This new methodology is based on a binary recursive division of the dataset, in order to generate subsets in which it would be possible to train predictive models with higher final performances, taking advantage of similarities between members. In each subset of the generated partition, models are trained, and the best ones are combined to form a meta-model. Separately, a model to classify new examples within the subsets in the partition is created. Finally, we generate a global model that assigns new examples to a particular subset using the classification mentioned above model, and predicts their value using the local meta-model for each case.

We successfully tested this new method in different non-linear datasets from different origins in the clinical, biomedical, biotechnological, and protein engineering fields. On those datasets, predictive meta-models were created, and high performance metrics were achieved, far above those obtained with other methods. The use of numerical experiments helped us to test the boundaries of our methodology, controlling the predictive outcome and the ground truth of the datasets. A natural relationship appears regarding the metrics for the linearity assessment: if the number of dimensions is high, the dataset would likely be classified as non-linear, at least in one of its dimensions. This does not necessarily imply that mildly non-linear methods will fail, but if so, our method would recommend directly applying the binary recursive division method to increase the performance measures of predictive models, despite the higher computational cost.

Our method applies state-of-the-art algorithms in a special order and following a novel strategy to optimize the results, which allows generating classification or regression models in general datasets, especially those addressed in this manuscript as highly-non linear. However, since our method uses previously developed ML methods, we are bound by their own limitations, in the sense that many of the flaws of our method are but a legacy of the ML algorithms used. Taking this into account, we recommend the use of the library and the proposed methodology in datasets with a reduced number of categories in their categorical variables since the library encodes them using One Hot Encoder. The recursive binary partition methodology should not be used when the number of classes is much larger than the available examples, as it may lead to detriments on the performance metrics because of the class balance buffer incorporated in the algorithm.

Future work contemplates the development of a web-based computational tool implementing our methodology, allowing non-specific users to enjoy the advantages of RV-Clustering, without the need to invest time gaining the knowledge that would be required by command-line execution. As the development of predictive models is common to different areas of application, we expect our methodology, library, and the future web-based service, to become a useful tool for the scientific community and a significant contribution to state of the art.

DATA AVAILABILITY STATEMENT

The https://github.com/dMedinaO/nonlinearModels repository contents the datasets generated and analyzed for this study.

AUTHOR CONTRIBUTIONS

DM-O, SC, and ÁO-N: conceptualization. DM-O and ÁO-N: methodology and project administration. DM-O and CQ: validation. DM-O, SC, CQ, and ÁO-N: investigation. DM-O and SC: writing and original draft preparation. SC and ÁO-N: writing, review, and editing. ÁO-N: supervision and funding resources.

FUNDING

This research has been financed mainly by the Centre for Biotechnology and Bioengineering—CeBiB (PIA project FB0001, Conicyt, Chile).

ACKNOWLEDGMENTS

DM-O gratefully acknowledges Conicyt, Chile, for Ph.D. fellowship 21181435, and Gonzalo Muñoz Rojas, for his support and computer guidance during the implementation of the algorithm.

REFERENCES

Abbasi, A., Sarker, S., and Chiang, R. H. (2016). Big data research in information systems: toward an inclusive research agenda. *J. Assoc. Inform. Syst.* 17, 1–32. doi: 10.17705/1jais.00423

Ahmad, J., Javed, F., and Hayat, M. (2017). Intelligent computational model for classification of sub-golgi protein using oversampling and fisher feature selection methods. *Artif. Intell. Med.* 78, 14–22. doi: 10.1016/j.artmed.2017.05.001

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., and Taha, K. (2015). Efficient machine learning for big data: a review. *Big Data Res.* 2, 87–93. doi: 10.1016/j.bdr.2015.04.001

Almeida, J. S. (2002). Predictive non-linear modeling of complex data by artificial neural networks. *Curr. Opin. Biotechnol.* 13, 72–76. doi: 10.1016/S0958-1669(02)00288-4

Asch, M., Moore, T., Badia, R., Beck, M., Beckman, P., Bidot, T., et al. (2018). Big data and extreme-scale computing: pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry. *Int. J. High Perform. Comput. Appl.* 32, 435–479. doi: 10.1177/1094342018778123

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., et al. (2016). Making sense of big data in health research: towards an eu action plan. *Genome Med.* 8:71. doi: 10.1186/s13073-016-0376-y

Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., and Sarai, A. (2004). Protherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 32(Suppl. 1), D120–D121. doi: 10.1093/nar/gkh082

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Commun. ACM* 18, 509–517. doi: 10.1145/361002.361007

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell* 173, 1581–1592. doi: 10.1016/j.cell.2018.05.015

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306–W310. doi: 10.1093/nar/gki375

Costa, F. F. (2014). Big data in biomedicine. *Drug Discov. Today* 19, 433–440. doi: 10.1016/j.drudis.2013.10.012

Coveney, P. V., Dougherty, E. R., and Highfield, R. R. (2016). Big data need big theory too. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 374:20160153. doi: 10.1098/rsta.2016.0153

Deshpande, M., and Karypis, G. (2002). "Evaluation of techniques for classifying biological sequences," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Taipei: Springer), 417–431.

Doerr, S., Ariz-Extreme, I., Harvey, M. J., and De Fabritiis, G. (2017). Dimensionality reduction methods for molecular simulations. *arXiv:1710.10629*. Available online at: https://arxiv.org/abs/1710.10629

Dua, D., and Graff, C. (2017). *Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. Available online at: http://archive.ics.uci.edu/ml

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., et al. (2015). Big data, bigger dilemmas: a critical review. *J. Assoc. Inform. Sci. Technol.* 66, 1523–1545. doi: 10.1002/asi.23294

Elter, M., Schulz-Wendtland, R., and Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. *Med. Phys.* 34, 4164–4172. doi: 10.1118/1.2786864

Gandomi, A., and Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. *Int. J. Inform. Manage.* 35, 137–144. doi: 10.1016/j.ijinfomgt.2014.10.007

Getov, I., Petukh, M., and Alexov, E. (2016). SAAFEC: predicting the effect of single point mutations on protein folding free energy using a knowledge-modified MM/PBSA approach. *Int. J. Mol. Sci.* 17:512. doi: 10.3390/ijms17040512

Ghahramani, Z. (2003). "Unsupervised learning," in *Summer School on Machine Learning* (Berlin: Springer), 72–112.

Greene, C. S., Tan, J., Ung, M., Moore, J. H., and Cheng, C. (2014). Big data bioinformatics. *J. Cell. Physiol.* 229, 1896–1900. doi: 10.1002/jcp.24662

Hinkson, I. V., Davidsen, T. M., Klemm, J. D., Chandramouliswaran, I., Kerlavage, A. R., and Kibbe, W. A. (2017). A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Front. Cell Dev. Biol.* 5:83. doi: 10.3389/fcell.2017.00083

Horton, P., and Nakai, K. (1997). "Better prediction of protein cellular localization sites with the it k nearest neighbors classifier," in *ISMB, Vol. 5* (Halkidiki), 147–152.

Hu, H., Wen, Y., Chua, T.-S., and Li, X. (2014). Toward scalable systems for big data analytics: a technology tutorial. *IEEE Access* 2, 652–687. doi: 10.1109/ACCESS.2014.2332453

Jeske, L., Placzek, S., Schomburg, I., Chang, A., and Schomburg, D. (2018). Brenda in 2019: a european elixir core data resource. *Nucleic Acids Res.* 47, D542–D549. doi: 10.1093/nar/gky1048

Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., and Clifford, G. D. (2016). Machine learning and decision support in critical care. *Proc. IEEE Inst. Electr. Electron. Eng.* 104:444. doi: 10.1109/JPROC.2015.2501978

Katal, A., Wazid, M., and Goudar, R. (2013). "Big data: issues, challenges, tools and good practices," in *2013 Sixth International Conference on Contemporary Computing (IC3)* (Noida: IEEE), 404–409.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* 15, 104–116. doi: 10.1016/j.csbj.2016.12.005

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Lee, C. H., and Yoon, H.-J. (2017). Medical big data: promise and challenges. *Kidney Res. Clin. Pract.* 36:3. doi: 10.23876/j.krcp.2017.36.1.3

Lee, G., Rodriguez, C., and Madabhushi, A. (2008). Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 5, 368–384. doi: 10.1109/TCBB.2008.36

Leung, M. K., Delong, A., Alipanahi, B., and Frey, B. J. (2015). Machine learning in genomic medicine: a review of computational problems and data sets. *Proc. IEEE* 104, 176–197. doi: 10.1109/JPROC.2015.2494198

Masso, M., and Vaisman, I. I. (2008). Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics* 24, 2002–2009. doi: 10.1093/bioinformatics/btn353

McKinney, W. (2011). "Pandas: a foundational python library for data analysis and statistics," in *Python for High Performance and Scientific Computing* (Dallas, TX), 14.

Michael, K. Y., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J., and Ideker, T. (2018). Visible machine learning for biomedicine. *Cell* 173, 1562–1565. doi: 10.1016/j.cell.2018.05.056

Olden, J. D., and Jackson, D. A. (2002). Illuminating the "black box" : a randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Modell.* 154, 135–150. doi: 10.1016/S0304-3800(02)00064-9

Oliphant, T. E. (2007). Python for scientific computing. *Comput. Sci. Eng.* 9, 10–20. doi: 10.1109/MCSE.2007.58

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., and Belfkih, S. (2018). Big data technologies: a survey. *J. King Saud Univer. Comput. Inform. Sci.* 30, 431–448. doi: 10.1016/j.jksuci.2017.06.001

Pandit, R., Shehu, A., Ioerger, T., and Haspel, N. (2016). "A principled comparative analysis of dimensionality reduction techniques on protein structure decoy data," in *Proceedings of the International Conference on Bioinformatics and Computational Biology* (Las Vegas, NV), 4–6.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *arXiv:1201.0490*. Available online at: https://arxiv.org/abs/1201.0490

Qiu, F., and Jensen, J. R. (2004). Opening the black box of neural networks for remote sensing image classification. *Int. J. Remote Sens.* 25, 1749–1768. doi: 10.1080/01431160310001618798

Qiu, J., Wu, Q., Ding, G., Xu, Y., and Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP J. Adv. Signal Process.* 2016:67. doi: 10.1186/s13634-016-0382-7

Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach. *arXiv:1110.2626*. doi: 10.5121/ijdkp.2011.1501

Ratanamahatana, C. A., and Gunopulos, D. (2002). *Scaling Up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection.* CiteSeerX. Available online at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.704

Rosenbrock, H. (1960). An automatic method for finding the greatest or least value of a function. *Comput. J.* 3, 175–184. doi: 10.1093/comjnl/3.3.175

Rydzewski, J., and Nowak, W. (2016). Machine learning based dimensionality reduction facilitates ligand diffusion paths assessment: a case of cytochrome p450cam. *J. Chem. Theory Comput.* 12, 2110–2120. doi: 10.1021/acs.jctc.6b00212

Sagiroglu, S., and Sinanc, D. (2013). "Big data: a review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)* (San Diego, CA: IEEE), 42–47.

Serpico, S. B., and Moser, G. (2006). Weight parameter optimization by the Ho–kashyap algorithm in MRF models for supervised image classification. *IEEE Trans. Geosci. Remote Sens.* 44, 3695–3705. doi: 10.1109/TGRS.2006.881118

Shaikhina, T., and Khovanova, N. A. (2017). Handling limited datasets with neural networks in medical applications: a small-data approach. *Artif. Intell. Med.* 75, 51–63. doi: 10.1016/j.artmed.2016.12.003

Shi, Z.-X., Dai, Q., He, P.-A., Yao, Y.-H., and Liao, B. (2013). "Subcellular localization prediction of apoptosis proteins based on the data mining for amino acid index database," in *2013 7th International Conference on Systems Biology (ISB)* (Huangshan: IEEE), 43–48.

Singh, A., Thakur, N., and Sharma, A. (2016). "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (Delhi: IEEE), 1310–1315.

Sun, S., Zhang, X., and Peng, Q. (2017). A high-order representation and classification method for transcription factor binding sites recognition in *Escherichia coli. Artif. Intell. Med.* 75, 16–23. doi: 10.1016/j.artmed.2016.11.004

Tribello, G. A., and Gasparotto, P. (2019). Using dimensionality reduction to analyze protein trajectories. *Front. Mol. Biosci.* 6:46. doi: 10.3389/fmolb.2019.00046

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* 13:22. doi: 10.1109/MCSE.2011.37

Wegner, P. (1990). Concepts and paradigms of object-oriented programming. *ACM Sigplan Oops Messeng.* 1, 7–87.

Witten, I. H., Frank, E., and Hall, M. A. (2005). *Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Xiang, Q., Liao, B., Li, X., Xu, H., Chen, J., Shi, Z., et al. (2017). Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine. *Artif. Intell. Med.* 78, 41–46. doi: 10.1016/j.artmed.2017.05.007

Zhang, H., and Ling, C. X. (2001). "An improved learning algorithm for augmented naive bayes," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Hong Kong: Springer), 581–586.

Zhou, L., Pan, S., Wang, J., and Vasilakos, A. V. (2017). Machine learning on big data: opportunities and challenges. *Neurocomputing* 237, 350–361. doi: 10.1016/j.neucom.2017.01.026

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 09 October 2020
doi: 10.3389/fmolb.2020.502573

A Middle-Out Modeling Strategy to Extend a Colon Cancer Logical Model Improves Drug Synergy Predictions in Epithelial-Derived Cancer Cell Lines

Eirini Tsirvouli[1]*, *Vasundra Touré*[1], *Barbara Niederdorfer*[2], *Miguel Vázquez*[2†], *Åsmund Flobak*[2,3] *and Martin Kuiper*[1]

[1] Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway, [2] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway, [3] The Cancer Clinic, St. Olav's University Hospital, Trondheim, Norway

OPEN ACCESS

Edited by:
Alberto Jesus Martin,
Universidad Mayor, Chile

Reviewed by:
Jie Zheng,
ShanghaiTech University, China
Tero Aittokallio,
University of Oslo, Norway
Jing Tang,
University of Helsinki, Finland

***Correspondence:**
Eirini Tsirvouli
eirini.tsirvouli@ntnu.no

†Present address:
Miguel Vázquez,
Barcelona Supercomputing Center,
Barcelona, Spain

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: 03 October 2019
Accepted: 22 September 2020
Published: 09 October 2020

Citation:
Tsirvouli E, Touré V,
Niederdorfer B, Vázquez M, Flobak Å
and Kuiper M (2020) A Middle-Out
Modeling Strategy to Extend a Colon
Cancer Logical Model Improves Drug
Synergy Predictions
in Epithelial-Derived Cancer Cell
Lines. Front. Mol. Biosci. 7:502573.
doi: 10.3389/fmolb.2020.502573

Cancer is a heterogeneous and complex disease and one of the leading causes of death worldwide. The high tumor heterogeneity between individuals affected by the same cancer type is accompanied by distinct molecular and phenotypic tumor profiles and variation in drug treatment response. *In silico* modeling of cancer as an aberrantly regulated system of interacting signaling molecules provides a basis to enhance our biological understanding of disease progression, and it offers the means to use computer simulations to test and optimize drug therapy designs on particular cancer types and subtypes. This sets the stage for precision medicine: the design of treatments tailored to individuals or groups of patients based on their tumor-specific molecular cancer profiles. Here, we show how a relatively large manually curated logical model can be efficiently enhanced further by including components highlighted by a multi-omics data analysis of data from Consensus Molecular Subtypes covering colorectal cancer. The model expansion was performed in a pathway-centric manner, following a partitioning of the model into functional subsystems, named modules. The resulting approach constitutes a middle-out modeling strategy enabling a data-driven expansion of a model from a generic and intermediate level of molecular detail to a model better covering relevant processes that are affected in specific cancer subtypes, comprising 183 biological entities and 603 interactions between them, partitioned in 25 functional modules of varying size and structure. We tested this model for its ability to correctly predict drug combination synergies, against a dataset of experimentally determined cell growth responses with 18 drugs in all combinations, on eight cancer cell lines. The results indicate that the extended model had an improved accuracy for drug synergy prediction for the majority of the experimentally tested cancer cell lines, although significant improvements of the model's predictive performance are still needed. Our study demonstrates how a tumor-data driven middle-out approach toward refining a logical model of a biological system can further customize a computer model to represent specific cancer cell lines and provide a basis for identifying synergistic effects of drugs targeting specific regulatory proteins. This approach bridges between preclinical cancer model data and clinical patient data and may thereby ultimately be of help to develop patient-specific *in silico* models that can steer treatment decisions in the clinic.

Keywords: logical model simulations, drug synergy prediction, systems medicine, model validation, middle-out modeling, model curation, cancer cell fate decisions

INTRODUCTION

Computational models that describe biological systems can help to provide insight into how these systems control regulatory events at the molecular level (Smolen et al., 2000; Le Novère, 2015). The ability to correctly predict the effects of systems perturbations by *in silico* simulations is a good indicator of how well the computational model represents biological reality. Indeed, computer models for diseased systems are being used to simulate drug perturbations and to develop, evaluate and prioritize putative drugs *in silico* (Flobak et al., 2015; Rubio-Perez et al., 2015). Approaches that use quantitative modeling rely on information including kinetic rate constants for regulatory components and their interactions, but this type of detailed quantitative data is only available for a small fraction of the regulatory interactions that underlie cell fate decision mechanisms. The much more abundant availability of binary molecular interactions, also defined as 'causal statements' (Touré et al., 2020) allows the use of the Boolean formalism as a powerful alternative mathematical framework for *in silico* simulation studies. The ability of Boolean models to represent discrete levels of a system furthermore complies well with the need for basic representations of cellular states, as these equate to stable states of regulatory networks that are interconnected through logical rules that may reach new stable states under different conditions, e.g., normal, diseased, and drug-perturbed. In systems medicine efforts to understand cancer, Boolean networks have been used previously to model biological systems driving cancer and were found useful for studying tumor progression and understand cancer signaling mechanisms (Srihari et al., 2014; Pirkl et al., 2016), predict tumor metastatic capabilities and therapy resistance (Srinivas, 2015), identify cancer-specific biomarkers, driver genes, drug targets (Irurzun-Arana et al., 2017; Sahoo et al., 2018; Qiu et al., 2019), and predict drug effects (Fumiã and Martins, 2013; Azuaje, 2017), including the possible synergistic effect of combinations of drugs (Flobak et al., 2015). Depending on the purpose of computational simulations, Boolean models can describe either a very specific process, such as a specific cancer-related signaling pathway (Grieco et al., 2013), or a collection of processes that together result in a biological phenomenon, such as the signaling pathways involved in metastasis. These models can vary in size, but they rarely comprise more than some tens of components.

In regulatory models based on the Boolean mathematical framework, a model component, also called 'node,' can either be active or inactive, which in Boolean algebra can be represented as 1 and 0, respectively. The state of a particular node (referred to as local state) is updated according to logical rules that capture the regulatory effects (activation or inhibition) of all the regulators of that node in the network, taking into account their activity state (Glass and Kauffman, 1973; Thomas, 1973). Logical rules in Boolean models follow the logical formalism and employ the operators AND, OR and NOT. Each Boolean model can be represented as a graph of nodes connected by a set of directed and signed edges, representing the causal interactions between the nodes. The same network graph can support multiple Boolean models, with logical rules specific for the system that the model should represent. Starting from an initial state, Boolean models that adequately represent biological systems are able to reach only a limited set of stable states (often only one), called attractors (Naldi et al., 2009; Helikar et al., 2012; Naldi et al., 2018), which can be considered as the mathematical equivalent of cellular states. Attractors can refer to a single stable state (singleton attractor), a set of stable states that repeat themselves in sequence (simple or complex cyclic attractor), or a set of stable states in which the system randomly oscillates (Wang et al., 2012; Irurzun-Arana et al., 2017). If a Boolean model can reach a stable state in which its node activities match experimentally observed activity states of their biological counterparts (e.g., the results of biomarker analysis), it indicates that the model captures to some extent relevant aspects of the biological system.

Boolean modeling toolkits (Gonzalez et al., 2006; Naldi et al., 2018) provide for a variety of analyses that can be further used to test, validate and enhance a model. Apart from being descriptive of a biological system and identifying attractors that comply with a particular state of a cell, Boolean models can also be predictive and be explored to simulate cellular behavior under perturbed conditions (Joo et al., 2018). Perturbation analysis allows the simulation of a system under different conditions, similar to knock-out, over-expression, or chemically induced perturbations in laboratory experiments. Such simulations can be designed for a variety of purposes, e.g., to analyze the regulatory system *per se* and identify critical nodes whose perturbation leads to significant functional changes in the system, thereby generating hypotheses as to their biological function in the system. Attractor analysis is also important to identify trajectories (a series of states that the network traverses through while reaching a stable state) in the system's behavior (Huang et al., 2005). In the case of gene regulatory networks, attractors are usually associated with specific phenotypes (Cho et al., 2016; Yang J. M. et al., 2018). Furthermore, a disruption of the balance found in these stable states of normal cellular systems can many times be associated with specific diseases, including cancers, allowing the mechanistic understanding of cancer development and progression (Bachmann et al., 2012), which can provide an important advantage when designing cancer therapies.

This paper focuses on a specific use of Boolean modeling, namely its use for computer simulations to identify effective combinations of targeted drugs that act synergistically in growth inhibition of a set of cancer cell lines. It is well known that a combination of drugs can have a higher effect on treated cells than the individual drugs alone would suggest if their effect were only additive (Roell et al., 2017). This effect, called drug synergy, results from systems interactions between the drugs and may yield a higher treatment efficacy. The use of synergistic drug combinations may address some of the current treatment limitations in cancer by reducing the emergence of drug resistance, which is frequently observed with single-agent therapies (Al-Lazikani et al., 2012; Gottesman et al., 2016), and lowering the chance of potential side effects and toxicity because the individual drugs can be used at lower dosages (Crystal et al., 2014; Trairatphisan et al., 2016). The use of drug combinations may serve as a stepping stone toward precision medicine, in which limitations of single-agent treatment, such as low response rates and acquired drug resistance, may be overcome by treatment regimes that use drug combination therapy optimized for the individual patient (Madani Tonekaboni et al., 2018). Combinatorial treatment refers to the targeting of multiple molecular components of a tumor cell-fate decision network, either by the combination of two or more targeted drugs or by combining targeted drugs with other therapies like immunotherapy, antibody-based therapy, and chemotherapy, with the aim to exploit synthetic lethality and tumor vulnerabilities and dependencies to treat cancer (Al-Lazikani et al., 2012).

With the availability of a relatively large number of targeted drugs (Yu et al., 2019), this may provide for a substantial number of potentially powerful combinations of drugs, but despite the availability of automated screening platforms using efficient high throughput technologies, the testing of combinatorial drug effects in the laboratory depends on vast amounts of large-scale dose-response data that is extremely time and resource-demanding (Pirkl et al., 2016; Joshi and Durden, 2019). The collection of all possible combinations of the large repertoire of targeted drugs presents a vast search space, and the number of possible interactions that need to be screened quickly becomes unmanageable, especially when taking into account different drug doses, combinations with more than 2 drugs, timing effects of drug administration, and the high intratumor, interpatient and cancer type variability that needs to be replicated in assays. Consequently, the screening for potential synergy is currently conducted mainly on compounds with an already known effect and/or where the combination of specific drugs makes sense based on empirical observations, significantly limiting the subspace of possible combinations that are actually tested (Cheng et al., 2019). *In silico* methods, therefore, pose an attractive pre-screening possibility, provided that the computer predictions can reliably limit the experimental search space (Crystal et al., 2014; Tolcher et al., 2018). More specifically, this means that predictive models must accurately predict the cellular response to medication, reveal the potential synergy between different drugs, produce few or no false-negative predictions (potentially powerful drug combinations that would be excluded from further testing) and preferably also few false positives (drug combinations that in further testing prove to be ineffective). Computational models that meet these criteria can alleviate the screening burden and create insight in the molecular mechanisms that lead to perturbational synergy (Flobak et al., 2015; Jeon et al., 2018; Madani Tonekaboni et al., 2018; Cheng et al., 2019; Tang et al., 2019).

Therefore, it is of high importance to develop high-quality logical models for predicting drug synergies and validating them by testing against experimental observations. The construction of computational models of biological systems can either involve a top-down approach that uses genome-wide omics data analysis to reveal the underlying regulatory network structure, or a bottom-up approach, in which a regulatory network is built from single entities and their interactions, often based on literature that describes their detailed analysis in various experimental settings (Shahzad and Loor, 2012). Bottom-up approaches are usually based on the manual curation of models, focusing on entities of interest, such as biological entities that are also drug targets, or driver genes for cancer. During this manual curation, the modeler many times confronts a series of subjective decisions on the relevance of entities, interactions and, more generally, the specific cellular processes to incorporate in a model, to properly represent a biological system. For the purpose of assessing the effect of particular perturbations, there is the additional challenge to identify and encode multi-level nodes that can be directly associated with a phenotype and, thus, serve as phenotypic readouts in the model. These phenotypic output nodes provide a convenient way to assess and quantify the effect of the drugs *in silico* simulations.

Here it is presented how a top-down multi-omics data analysis can identify candidate genes that should be considered for addition to an existing model, serving as seed genes that provide guidance for additional bottom-up modeling. The cell signaling components used were highlighted by the analysis of multi-omics data from the Consensus Molecular Subtypes (CMSs) (Guinney et al., 2015) study of colorectal cancer (CRC), to expand the generic cancer cell fate decision network CASCADE 2.0 that was built previously by our group (Niederdorfer et al., 2020). This approach effectively constitutes a middle-out strategy that allowed the expansion in a pathway-centric manner, capturing processes that were highlighted as possibly important for CRC subtype development. Furthermore, the model was manually partitioned into functional subsystems, named modules, allowing a continuous switching between top-down (finding modules and seed genes) and bottom-up modeling (module completion) during the manual curation of each module, in order to comprehensively capture cell fate decision mechanisms. Additionally, modules served as a "binning principle" of nodes and regulatory relationships, providing an intermediate network level, placed between the individual binary interactions and a fully connected network. This allowed for a multilevel assessment of the system, focusing on the modular regulatory effect on output nodes, and their perturbational response to targeted drugs. The evaluation of the performance of the model in predicting experimentally validated synergies of combinations of 18 established cancer drugs in a panel of eight cancer cell lines revealed that the model performs similarly well for a majority of carcinoma cell lines in the panel, and not only for colorectal cancer that it was originally specialized for. Our results suggest that a middle-out modeling approach may be appropriate for optimizing the representation of specific cancers or cancer cell lines, or indeed other disease types for which multi-omics data is available.

MATERIALS AND METHODS

Tools, Data Standards, and Exchange Formats

An overview of the software tools and their versions used in this study can be found in the **Supplementary Material**. Genes and proteins were represented with the standard identifier nomenclature for each entity type, namely HUGO Gene Nomenclature Committee (HGNC) symbols and UniProt IDs, respectively. Several files are available at are publicly available at https://github.com/druglogics/cascade, with information about the CASCADE 3.0 model: the model's interactions as a Simple Interaction File (SIF); a file containing the supporting evidence for each of the interactions in the model; a file with information about node translation and module assignment; and a cytoscape.cys file of the network and its topology as shown in **Figure 3**. The github repository also contains information about other CASCADE models, including the CASCADE 2.0 model that was used as the basis for this work.

Model Assembly by Manual Curation

Logical models are usually created manually by carefully screening the literature for evidence that supports the linking of components and their regulatory relationships in a Prior Knowledge Network (PKN) that represents a biological system. The CASCADE 2.0 model is a manually curated logical model, representative for the most prevalent cancer types (Niederdorfer et al., 2020). The CASCADE 2.0 model consists of 144 nodes and 366 interactions, including two output nodes called *Prosurvival* and *Antisurvival*. Each node was annotated with its HUGO gene symbol. In the case of several isoforms, a family-node representative of all isoforms was used. Family nodes are notated with a _f in their name, while protein complexes and genes are notated with _c and _g, respectively.

Niederdorfer et al. (2020) describe several model versions, including a version in which the model topology was refined so that it better recapitulates the biological mechanisms of the analyzed cell lines. In the current analyses, the more generic cancer model was used. These different models were all constructed according to the following design principles: (1) include targets of specific drugs for which the effects should be simulated; (2) contain entities that are known to be involved in specific or more general oncogenic processes, and (3) contain specific nodes that will allow a read-out of the state of the cell fate (phenotype output nodes): actively dividing (*Prosurvival*) or growth-inhibited/apoptotic (*Antisurvival*). In this study, the CASCADE 2.0 model was taken as a basis for extending into a logical model that contains the major components and processes that can be identified as significantly perturbed in one or more of the colorectal cancer Consensus Molecular Subtype datasets (see below), which was named CASCADE 3.0.

Identification of Affected Processes in Consensus Molecular Subtypes of Colorectal Cancer

An expression-based classification of the patients in the TCGA-COAD cohort was performed according to the Consensus Molecular Subtypes (CMS) classification for colorectal cancer (CRC), as described in the **Supplementary Material**. This patient classification aimed to identify commonalities and differences between the four subtypes at a genomic, transcriptional and functional (i.e., pathway) level. All the subsequent analyses were conducted separately for each CMS class of patients unless stated otherwise, and *p*-values were adjusted using the Benjamini–Hochberg method, to correct for the false discovery rate (FDR) across multiple tests (Benjamini and Hochberg, 1995).

The omics data used in the current project (i.e., mRNA expression, somatic copy number variation and mutation data) were publicly available data published as part of the TCGA-COAD project (Cancer Genome Atlas Network, 2012). Data from patients that were not classified into one of the CMS classes were not used in the analyses, while non-tumorous data from adjacent tissues of the classified patients were used when needed (further discussed in the following sections).

Differential Expression Analysis

Using the RNA-sequencing data of TCGA-COAD, a statistical analysis of differential expression was performed on the transcriptomes of the tumor samples using the edgeR RNA-Seq expression analysis package (Robinson et al., 2010). Data from the same patient, but originating from different vials, portions, analytes or aliquots, were averaged. RNAs with very low counts across all libraries (*fewer than 6–7 counts*) and genes that were expressed in only one sample were discarded, as they were deemed not significant. Since the high expression of some genes in a sample can lead to the under-sampling of the others, a normalization step to correct for differences in the library sizes was performed. The same filtering and normalization steps were performed in available normal tissue samples of TCGA. The differential expression analysis (DEA) was carried out against this collection of normal samples, for all the subtypes. Protein-coding genes with an FDR-adjusted *p*-value of less than 0.05 and a logarithmic fold change (logFC) greater than 2 or lower than −2 were deemed significantly differentially expressed.

Somatic Copy-Number Alterations (SCNV) Analysis

The GISTIC 2.0 tool (Mermel et al., 2011) in the GenePattern platform (Reich et al., 2006) was used to identify genomic regions that were significantly amplified or deleted across the different subtypes, based on the amplitude of the aberrations as well as their frequency of occurrence across the tumor samples. For this analysis, masked segment copy number variation data of TCGA-COAD were retrieved and used. In masked data, segments with probes known to contain germline mutations are removed allowing the identification of the cancer-associated, somatic copy number variation. The recurrently aberrant regions and their containing genes were identified with a threshold FDR < 0.01.

Recurrent Somatic Mutation Analysis

The MutSigCV tool (Lawrence et al., 2013) in the GenePattern platform (Reich et al., 2006) was used to identify recurrent mutations in the cancer genome of TCGA-COAD patients. The mutational profile of TCGA-COAD patients containing information on mutation type, category and its effect, was used to. Recurrent mutations are identified by calculating the probability of a non-silent mutation to have happened by chance compared to the background mutation rate estimated by silent mutations and other patient-specific and position-based confounding covariates. A threshold FDR < 0.05 was used.

Functional Analysis by Enrichment

Initially, genes that were found to be either differentially expressed, located in recurrently aberrant chromosomal regions or recurrently mutated were considered important for colorectal cancer cells. To further investigate the functional role of the affected genes in each subtype, independent enrichment analyses were performed against the Reactome (Fabregat et al., 2018), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), and Atlas of Cancer Signaling Networks (ACSN) (Kuperstein et al., 2015) databases. For KEGG and Reactome databases the clusterProfiler package was used (Yu et al., 2012) while the ACSNmineR package (Deveau et al., 2016) was used for ACSN. Results with FDR-adjusted *p*-values lower than 0.05 were considered significant. In parallel, similar enrichment analyses were performed with the CASCADE 2.0 components (Niederdorfer et al., 2020), to identify the main pathways and processes represented in this model. A comparison of the results of these analyses revealed the processes that were affected in the CMS classes but not significantly represented in the topology of the CASCADE 2.0 model.

The CASCADE 3.0 Model

Middle-Out Expansion of the Initial Model

The middle-out modeling process was characterized by a combination of a top-down and bottom-up approach. As already described, the first steps were governed by the genome-wide analysis of relevant omics data, a typical approach in top-down modeling where correlations between genes or proteins are investigated by deploying various statistical and bioinformatics analyses. More specifically, the top-down step and overrepresentation analysis highlighted the affected processes in the different CMSs, after which the nodes of CASCADE 2.0 network could be annotated and mapped to these overrepresented signaling pathways and biological processes. However, as additional missing process and pathway components were added during the construction of CASCADE 3.0, the module assignment for some nodes had to be further refined, in ways that it better represented the role of these entities in the modeled system.

When most of the nodes were assigned to modules, four of the initial modules were divided into two segments: one containing entities involved in the core signaling pathway and the other containing the negative regulators of that pathway. The core signaling pathway included proteins involved in signal transduction, starting mostly from receptors sensing a signal and all the signaling proteins (i.e., the positive regulators of the response and the main effector of the pathway) that enable the regulatory response to the signal. The negative regulators were placed in the other module, including entities involved in negatively regulating the pathway's main effector, meaning that they are involved in potential regulatory feedback loops, as seen for example in the WNT and MAPKs modules.

As a next step, a bottom-up approach was employed to expand the modules so that they comprehensively represent relevant pathways. As is common practice in bottom-up approaches, this step was focused on individual biological entities and their interactions, using a variety of databases, knowledge bases and sometimes literature. The expansion of the modules and the construction of the extended CASCADE 3.0 model was an iterative process of manual curation: Each module was manually checked against existing knowledge to comprehensively capture its intra- and inter-modular regulatory, causal interactions that would likely contribute to the overall cell fate decision mechanism that the model should represent. A detailed list of all the knowledge resources used during the curation processes is presented in the **Supplementary Material**. Most of the initial curation work drew on the cell signaling pathway database Signor (Perfetto et al., 2016), which, in combination with the primary modules from the original CASCADE 2.0, guided the addition of new nodes in each of the pre-existing modules. An important part of this curation process was the identification of the context under which an interaction was observed. In order to retain high confidence in the accuracy with which the model describes the biology of colorectal cancer cells, only regulatory interactions relevant for cells of tissues from which CRC subtypes originate were selected. In case interactions were reported in other tissues, additional literature was checked to decide whether to include or discard the interaction. Furthermore, interactions that were reported for specific biological processes not relevant to the biological system that the models should capture, for example, cardiac development, were omitted.

Taking into account the possible cross-talk of signaling pathways and the multi-functionality of many biological molecules, all nodes were examined for their potential participation in several pathways. Because of this, some nodes, including entities such as adaptor proteins or cytoplasmic kinases, were functionally attributed to multiple modules, but are presented and analyzed in CASCADE 3.0 only as members of one main module. The assignment to these modules was based on the available knowledge on the functional role of the nodes and the number of interactions it shared with the other members of that module.

Logical Modeling

The transformation of the expanded PKN into a Boolean model was done by the definition of logical rules that describe the overall regulatory input that each node receives: for this, the regulatory effect of each of the input nodes (activating or inhibiting) was combined with the logical AND, OR and NOT operators. The local state of each entity depends on the state of the combined set of nodes that regulate it. As described above, those regulations are captured in the logical rules that govern the update of the state of each node. As a point of departure, general logical rules that assume that an entity is active if *any* of its activators is active and *none* of its inhibitors is active were defined. According to the general rules, a protein's activity will be downregulated by any active inhibitor, regardless of the upregulation input of one or more activators (Shmulevich et al., 2002).

According to the notation of the logical formalism, the rules of the nodes' activities are of the form:

Node X = (Activator A OR Activator B . . . Activator n) AND NOT (Inhibitor A OR Inhibitor B OR . . . Inhibitor n)

Two additional "phenotype" nodes were added to the model, representing the two cellular states *Prosurvival* and *Antisurvival.* These nodes were implemented as multivalued nodes (with possible local states 0, 1, 2, or 3) which serve as cellular state readout and allow to assess the overall proliferative state of the system. The global state of the system is computed as the overall sum of the negative value of *Antisurvival* and the positive value of *Prosurvival*, ranging between −3 and +3.

Drug Synergy Prediction

Drug synergy predictions were performed with a custom-built modeling pipeline that combines several software modules that together provide a highly automated computational framework for logical model assemble and simulations (Flobak et al., manuscript in preparation)[1]. The pipeline can customize a general logical model to a specific cell line, after which it uses a collection of models (ensemble) each equally fit to represent a cell type to predict the effect of a drug perturbation, as well as their potential synergies in case of drug combinations. Initially, omics data of a specific cell type are translated into entity steady state activities (1 or 0). Such omics data can be either genome-wide or biomarker-specific, and can include among others proteomics, genomics, and transcriptomics data, either separately or in combination. The omics data serves to produce a training set of steady state activities that the network nodes should display when the logical model reaches a stable state. As this is dependent on the exact configuration of the overall logical rules of the model, these logical rules are optimized with the help of a genetic algorithm that changes sets of logical rules and analyses the steady state values from the altered model against this training set.

The genetic algorithm iteratively "mutates" the logical rules of some nodes each time, by randomly switching between AND and NOT and then the global stable states of the mutated models are calculated using the BNReduction tool (Veliz-Cuba et al., 2014). The mutated models that show the highest fitness (their stable state node activities better resemble the data in the training set) are further mutated for a selected number of iterations.

This optimization process results in an ensemble of models, all having the same topology but with slight differences in their logical rule structures, each model of an ensemble complying more or less equally well with the regulatory system that should be represented. The logical model ensemble is next systematically subjected to a set of *in silico* drug perturbations by assessing the combinatorial effects of drugs on the models as observed by the combination of states of the phenotype output nodes. These output nodes are multi-valued (global state ranging from −3 to +3) and the state of these nodes is defined by the predicted local states of key nodes that provide 'regulatory input' to these phenotype nodes, such as the cyclins, MYC and other survival factors that add additively to Prosurvival and the caspases and other pro-apoptotic entities that add to Antisurvival. With the global state ranging from −3 (only activity from anti-survival nodes) to +3 (only activity from pro-survival nodes), the quantification of the effect of the drugs to the viability of a system after single and combinatorial drug simulations was possible. For example, a drug that results in a global state of −3 has a more prominent effect than a drug that results in a global state of −2 or −1. The global state of the combinatorial treatment was then compared to the global states of the treatment of each individual drug. If the drugs that together result in a viability (i.e., output nodes' state) smaller than the minimum of the viability of each individual drug, they were scored as synergistic. These predictions were then compared to observed synergies validated by experimental data produced by the combination of 19 small molecule inhibitors and their 171 combinations (Flobak et al., 2019). As discussed in Niederdorfer et al. (2020), an inhibitor of PTEN (SF-1670) that was found to be under characterized regarding its off-target effects and was involved in the majority of synergies was not included in the analysis, reducing the data used to 18 small molecule inhibitors and their 153 combinations. The inhibitors were targeting various modules of the models and were tested in all eight cell lines used in the simulations. Furthermore, all drugs used in the screen were subjected to in depth characterization including an extensive target profiling, in associating the drugs and their targets with the model's nodes. The overall performance of the model with respect to true positive, false positive, true negative and false negative drug synergy predictions was assessed using AUC-ROC curves as performance measurement (Sammut and Webb, 2017).

In this project, three different sets of inferred entity states were used as training data to the genetic algorithm. Two of the data sets include activity states inferred from two distinct sets of omics data, using the Paradigm tool (Vaske et al., 2010), while the third set contains protein activities inferred from transcription factor activity information, using the Viper tool (Alvarez et al., 2016). The first set of activity states from Paradigm, referred as *Combination-based*, makes use of cell line specific copy number variation, gene expression, RPPA for total protein abundance and RPPA for phosphosites to infer entity states. The second set of states from Paradigm, referred as *mRNA expression-based*, uses only the cell-line specific mRNA expression data. For the *TF activity based*, data from "Genomics of Drug Sensitivity in Cancer" (GDSC) project were used as an input for Viper.

[1] https://druglogics.github.io/druglogics-doc/index.html

The model optimization work indicated that larger sized training sets not necessarily correlated with higher AUC values. This might be explained by imperfections in the training data: inferred data, so not experimentally confirmed data, may have errors in it that limit the "freedom" available to the genetic algorithm to adequately fit model steady states to the real biological state of the system. For that reason, the *Combination-based* and *mRNA expression-based* datasets were reduced to only those nodes for which the smaller data sets also contained a predicted state. That also allowed a more direct assessment of training data sets with respect to their ability to correctly serve as local states that would be observed in biological reality. More details about the reduction of the training data can be found in the **Supplementary Material**. All three training sets were subsequently used to evaluate how the model performs for eight cancer cell lines (see **Table 1**).

RESULTS

Identification of Affected Processes

Omics Data Analysis

The candidates for regulatory network inclusion were identified through a multi-omics data analysis that included transcriptomic (i.e., gene expression data) and genomic (i.e., somatic mutations and copy number alterations) profiles of CRC patients, effectively identifying affected processes and pathways in these patients' tumors. The differential gene expression analysis identified the highest number of affected genes and displayed a significant overlap between the differentially expressed genes in the subtypes, all involved in fundamentally dysregulated processes in cancer, such as DNA repair, cell adhesion, and cell cycle control. The identification of somatic copy number alterations (SCNAs) corroborated the profiles of the molecular subtypes and revealed both known and novel aberrant chromosomal regions. CMS2 and CMS4 displayed the highest number of SCNAs, whereas the two remaining subtypes had a low number of aberrant regions. Interestingly, a much higher number of genes was found correlating with deleted peaks than with amplified peaks, for all the subtypes. Among the 114 unique aberrant regions across all subtypes, five regions were altered in all subtypes. Four of those regions (16p13.2, 20p12.1, 5q12.1, and 4q22.1) showed deletions, while 8p11.21 was amplified in all the subtypes. The 20p12.1 region has been previously reported as recurrent in CRC (Davison et al., 2005), but there are no reports for the presence of known cancer genes in any of the regions. Some of the identified SCNAs have been previously reported for their involvement in other cancer types, but not in CRC. A number of genes located in these chromosomal regions have been associated with clinical characteristics of cancer patients and could potentially be investigated as biomarkers or drug targets for CRC (Coppedè et al., 2014). The somatic mutation analysis did not show any association between the mutation of specific pathways and specific subtypes, as the major signaling pathways known to be altered in CRC tend to be mutated in all the subtypes. Given its Microsatellite Instability (MSI) status resulting from a defective DNA mismatch repair machinery, CMS1 patients are expected to have a predisposition to hypermutability (Yu et al., 2019). For that reason, CMS1 patients had the highest number of recurrently mutated genes.

A list of the affected genes was produced for each subtype and classified as either differentially expressed in comparison to normal tissue, recurrently mutated, or located in a recurrently amplified or deleted region with respect to normal copy number. The total number of affected genes per category in each subtype is presented in **Table 2**, and their overlap in **Figure 1**.

TABLE 1 | Description of the eight cancer cell lines used in the synergy prediction analysis.

Cell line	ID	Tissue	Disease
AGS	RRID:CVCL_0139	Stomach	Gastric adenocarcinoma
Colo205	RRID:CVCL_0218	Colon; derived from metastatic site: ascites	Colon adenocarcinoma
DU145	RRID:CVCL_0105	Prostate; derived from metastatic site: brain	Prostate carcinoma
SW620	RRID:CVCL_0547	Colon; derived from metastatic site: lymph node	Colon adenocarcinoma
MDA-MB-468	RRID:CVCL_0419	Breast; derived from metastatic site: pleural effusion	Breast adenocarcinoma
A498	RRID:CVCL_1056	Kidney	Renal cell carcinoma
SF295	RRID:CVCL_1690	Brain	Glioblastoma
UACC62	RRID:CVCL_1780	Skin	Melanoma

TABLE 2 | Total number of genes that were found to be affected in the omics data analysis.

Subtype	Recurrently mutated genes	Amplified genes	Deleted genes	Upregulated genes	Downregulated genes
CMS1	55	541	135	1625	1658
CMS2	6	438	2508	1793	1789
CMS3	11	12	2054	1185	1501
CMS4	10	587	3461	2003	919

Affected genes are defined as genes that were either differentially expressed in comparison to normal tissue, amplified or deleted with respect to normal copy number or recurrently mutated.

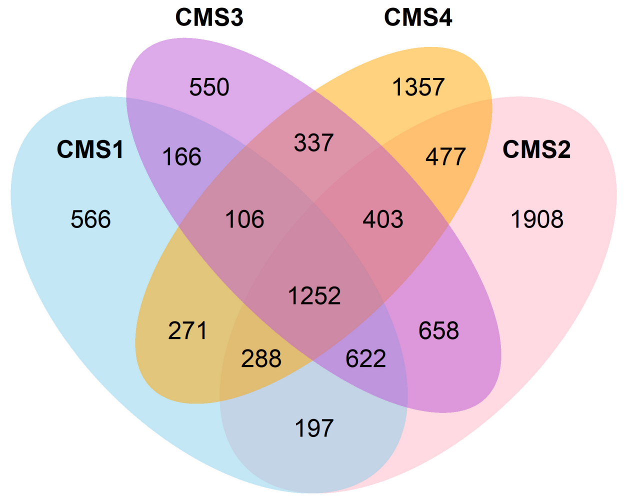

FIGURE 1 | Venn diagram showing the overlap between genes affected in the four CSMs.

Functional Analysis of CMS Genes by Enrichment

An analysis of the affected genes for their functional enrichment was performed against the ACSN, Reactome, and KEGG databases (Kanehisa and Goto, 2000; Kuperstein et al., 2015; Fabregat et al., 2018). The results are shown in **Figure 2** and are represented as individual, non-redundant, cancer-related pathways. Non-cancer related pathways and processes mainly found in cancer-associated cells in the tumor microenvironment, such as cancer-associated fibroblasts and immune cells, were not included in the results as the model does not account for inter-cellular interaction. Since the model aims to represent regulatory interactions involved in signaling pathways, metabolic pathways that were found to be deregulated, especially in the case of the metabolic subtype, could not be represented in the model and thus were also excluded from the results. A similar enrichment and aggregation analysis was done for the nodes of the CASCADE 2.0 model and a comparison with the affected CMS pathways (see **Figure 3**) highlights the signaling pathways that are affected in CRC but were not included in CASCADE 2.0. The identified missing processes included the Hippo, Hedgehog, and Notch signaling pathways, as well as DNA repair and cell adhesion, all with well documented involvement in both CRC and cancer in general (Wierzbicki and Rybarczyk, 2015; Vinson et al., 2016; Wu et al., 2017; Boesch et al., 2018; Mirza-Aghazadeh-Attari et al., 2018).

Construction of the CASCADE 3.0 Model

Manual Extension of the CASCADE 2.0 Model

Guided by the results of the top-down analysis and founded on prior knowledge from several databases and the literature, the CASCADE 3.0 model was constructed through the addition of nodes and edges to describe fundamentally dysregulated processes in all CRC subtypes. Those processes involved cell fate controlling processes such as cell cycle progression, regulation of apoptosis and response to DNA damage, as well as signaling pathways that were identified to be missing from the CASCADE 2.0. The final network consists of 183 nodes and 605 edges (see **Figure 3**). In addition to the inclusion of new nodes and interactions, small refinements were performed in the model. Nodes representing genes (*notated with _g in CASCADE 2.0*) were removed from the model, and replaced with their gene product node, including their regulatory interactions. Additionally, the CK1_f node, containing CSNK1A1, CSNK1D, and CSNK1E isoforms was split into two nodes, due to the involvement of the two latter isoforms in a newly added Hippo pathway module. Finally, in order to more accurately represent the regulation of the cell cycle by MYC (Bretones et al., 2015), the edge representing the direct interaction of MYC with the *Prosurvival* output node was replaced by an edge representing the promotion of proliferation through the activation of CCND1. Finally, in addition to the *Prosurvival* and *Antisurvival* output nodes, a new output node representing *Metastasis* could be included, based on the observation that several pathways were involved in metastasis-related processes (e.g., Epithelial-to-Mesenchymal transition and cell motility). However, due to the lack of appropriate screening data for this effect, it was omitted from the model, but it could be considered in future extensions of the model.

Topological Comparison of Original and Extended Models - Final Modules

The 183 nodes of the extended model were grouped into 25 manually curated pathways modules representing altered pathways or functions in the CMSs of colorectal cancer. Additionally, four of those modules (WNT, PI3K/AKT, TGFβ, and JAK/STAT) represent the negative regulators of a specific pathway and its respective main effector. An example of such a set of negative regulators is the β-catenin destruction complex. The components of the complex (i.e., APC, AXIN1, CK1, and GSK3B), are involved in the WNT pathway, but they negatively regulate its main effector (β-catenin), so they were assigned to a separate module (WNT negative regulators module) than the core signaling pathway (WNT module). The resulting modules vary in size and structure, and nodes grouped in a module do not necessarily share interactions with each other. This is specifically the case for modules with entities exerting similar regulatory functions (e.g., the Anti-apoptotic module), but do not necessarily interact with each other to achieve that function. The modules share numerous interactions with each other, a reflection of the fact that biological pathways cannot be delineated as completely independent groups, and perturbations in one module are likely to affect the behavior of other modules. In biological systems, module cross-talk can give rise to emerging functions that differ from their original functions (Lorenz et al., 2011). This is especially true when cells execute more complex behaviors, such as invasion in cancer systems, which are often controlled by many processes and a result of the interaction of many modules (Koutsogiannouli et al., 2013).

A side-by-side comparison of the topologies of the two networks is shown in **Figure 3**, allowing an easy identification of the added or expanded modules in CASCADE 3.0. Of the 144 nodes of the CASCADE 2.0, 36 were among the

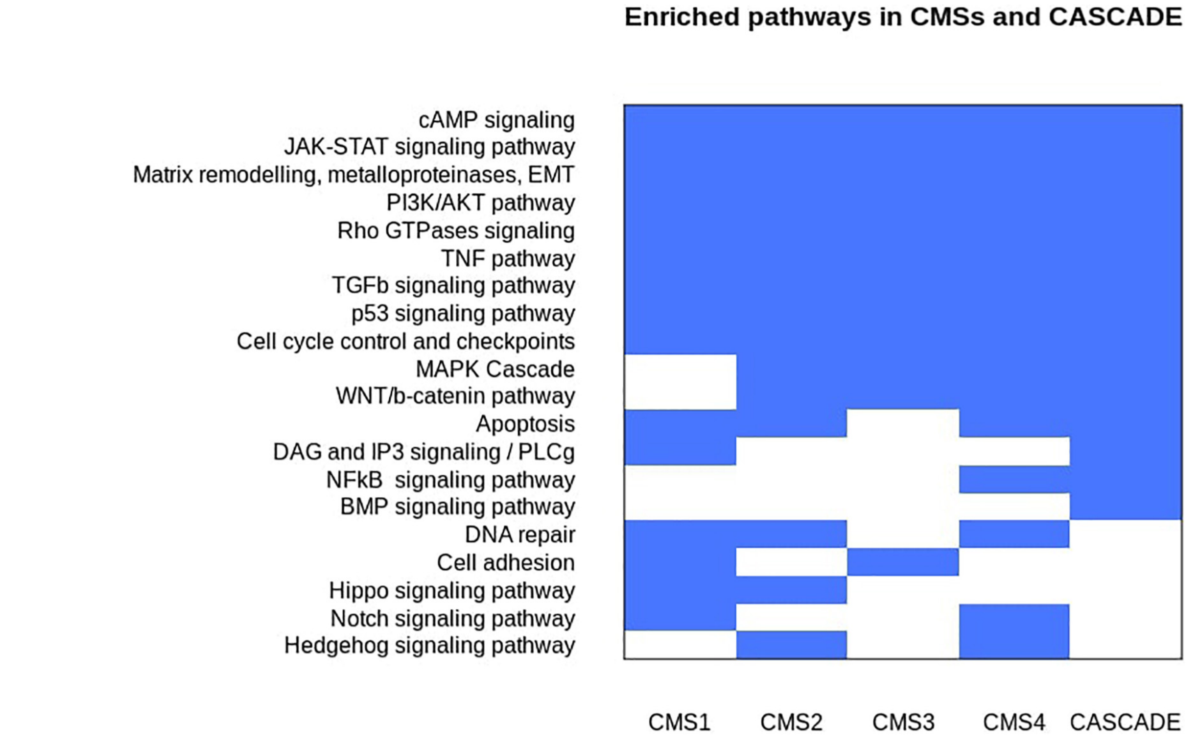

FIGURE 2 | Discrete heatmap of the presence or absence of signaling pathways and biological processes that were identified as affected in the Consensus Molecular Subtypes of colorectal cancer and/or represented in the CASCADE 2.0 model. Blue colored cells represent the enrichment of a process in a subtype and/or the model.

genes affected in at least one of the molecular subtypes of CRC and these were assigned to 16 different modules. These modules represented key oncogenic processes, such as proliferation-promoting transcription factors, apoptosis, the JAK/STAT signaling pathway, and MAPK cascades. The majority of the affected genes present in CASCADE 2.0 was found to be part of signaling pathways whose dysregulation is considered to be a driver event in colorectal cancer: the PI3K/AKT, WNT and Transforming Growth Factor-β signaling pathways (Tiwari et al., 2018). While only five modules representing missing pathways identified in the enrichment analysis step were added, all other modules were expanded either with a few entities or, in some cases, a substantial number of them. The modules that had the most nodes added represented processes such as adhesion and EMT, negative regulation of apoptosis, cell cycle control and checkpoints, and DNA damage response and repair.

Drug Synergy Prediction

Evaluation of the Model's Overall Performance

The drug synergy predictions were performed with three sets of models, each trained with a different set of inferred node activity states: TF-, Combination-, and mRNA expression-based. Model optimization and drug synergy scoring was as described in Methods, and prediction performance was benchmarked against experimental data obtained for eight cancer cell lines (Flobak et al., 2019) and evaluated using AUC values that define the ability of a model to distinguish experimentally validated synergies and non-synergies (Sammut and Webb, 2017).

The distribution of AUC values between 0.5 (no prediction efficiency) and 1.0 (optimal model predictions) shows that the model's performance depends both on the training data that was used and the cell line for which predictions were produced. As shown in **Figure 4**, models tend to perform better when trained to the *Combination-based* training set, and model performance can be very high for some cell lines, while for other cell lines drug synergies prove to be difficult to predict with any training set. The model displayed a (relatively) good performance for both the colorectal adenocarcinoma (Colo205) and gastric adenocarcinoma (AGS) cell lines. Synergy predictions for the prostate carcinoma cell line DU145 were consistently of moderate accuracy (AUC values ∼0.6–0.7), while predictions for the melanoma cell line (UACC62) was consistently the poorest, with an AUC value lower than 0.5 for the *TF activity* training data set. Prediction performance for the other cell lines range from moderate to very high, depending on the training set used. In order to ensure that the performance of the *Combination-based* training set was significantly improved when compared to the other training data sets, a one-sided t-test was performed. The comparison of the performance between the *Combination-* and *TF activity-based* training data showed significant improvement (p-value $= 0.024$). At the same time, the difference between the Combination- and mRNA-based training data sets was not significant (p-value of 0.2671). However, all but two cell lines have

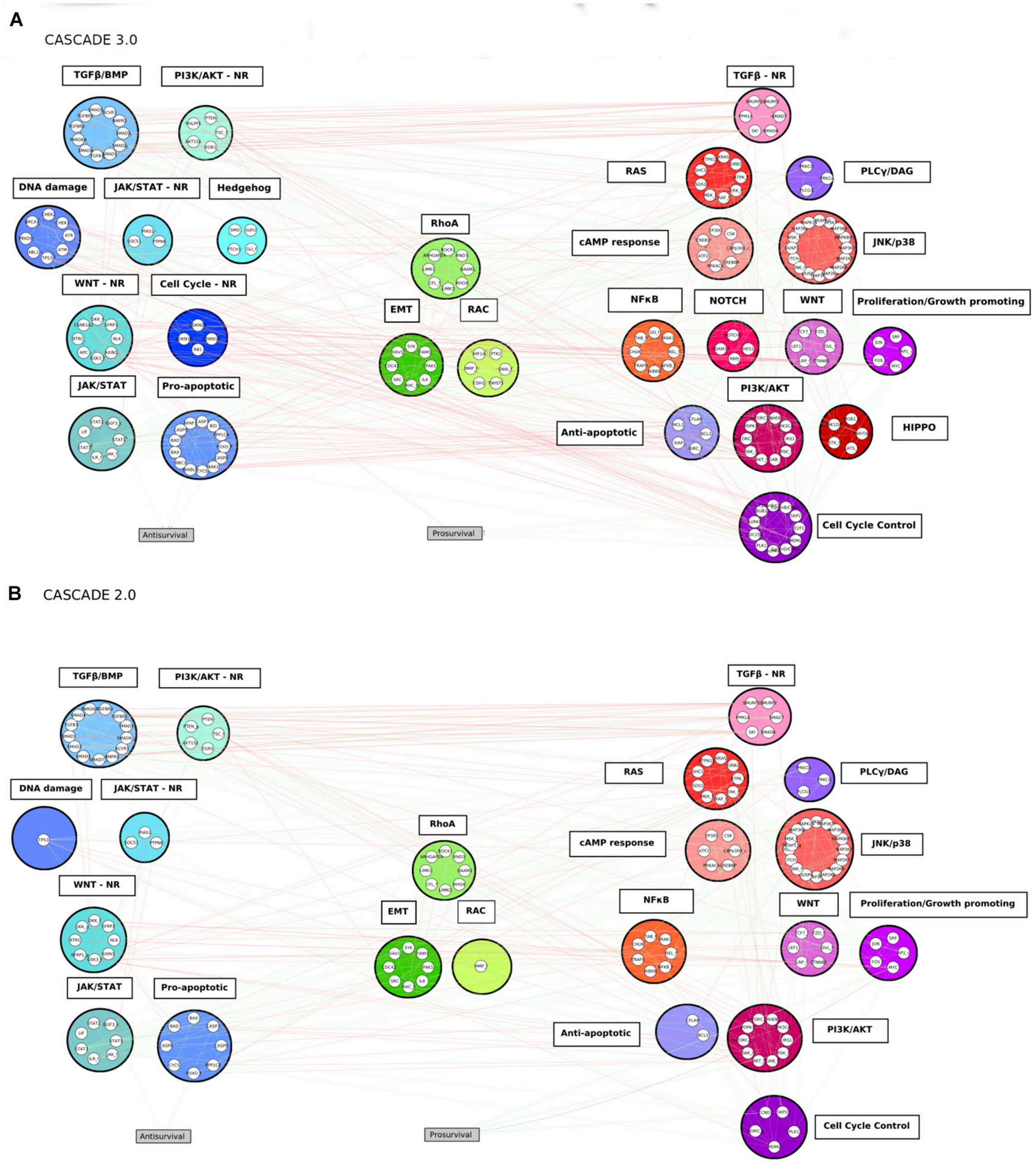

FIGURE 3 | The CASCADE 3.0 **(A)** and CASCADE 2.0 **(B)** models. The nodes are grouped according to pathway modules. The modules are grouped based on their promotion of apoptosis (blue colored modules), metastasis (green colored modules) or proliferation (red colored modules) when the pathways they represent are active. The position of a module in the figure displays its proximity to the output node: the smaller the average shortest path of the module to the output node it is related to, the closer to the output it is placed. Empty slots in the CASCADE 2.0 topology show the individual nodes or complete pathway modules that were added in the CASCADE 3.0 model. Gray rectangular nodes represent the output nodes.

an improved performance with the Combination-based data and these data were therefore selected as the one yielding the highest performance. This variance in performance could indicate the inability of specific computational tools to correctly infer node activity states for specific cell lines, the importance of these states when training the model, or that models for these cell lines need specific topology optimization in addition to the logical rule optimization, to make them more stable with respect to the training data.

As the *Combination-based* training data set was the most informative one, this set was used as the basis for a comparison of the performance between the initial (i.e., CASCADE 2.0) and the updated (CASCADE 3.0) model. The obtained AUC scores for each cell line are shown in **Figure 5**, with the model

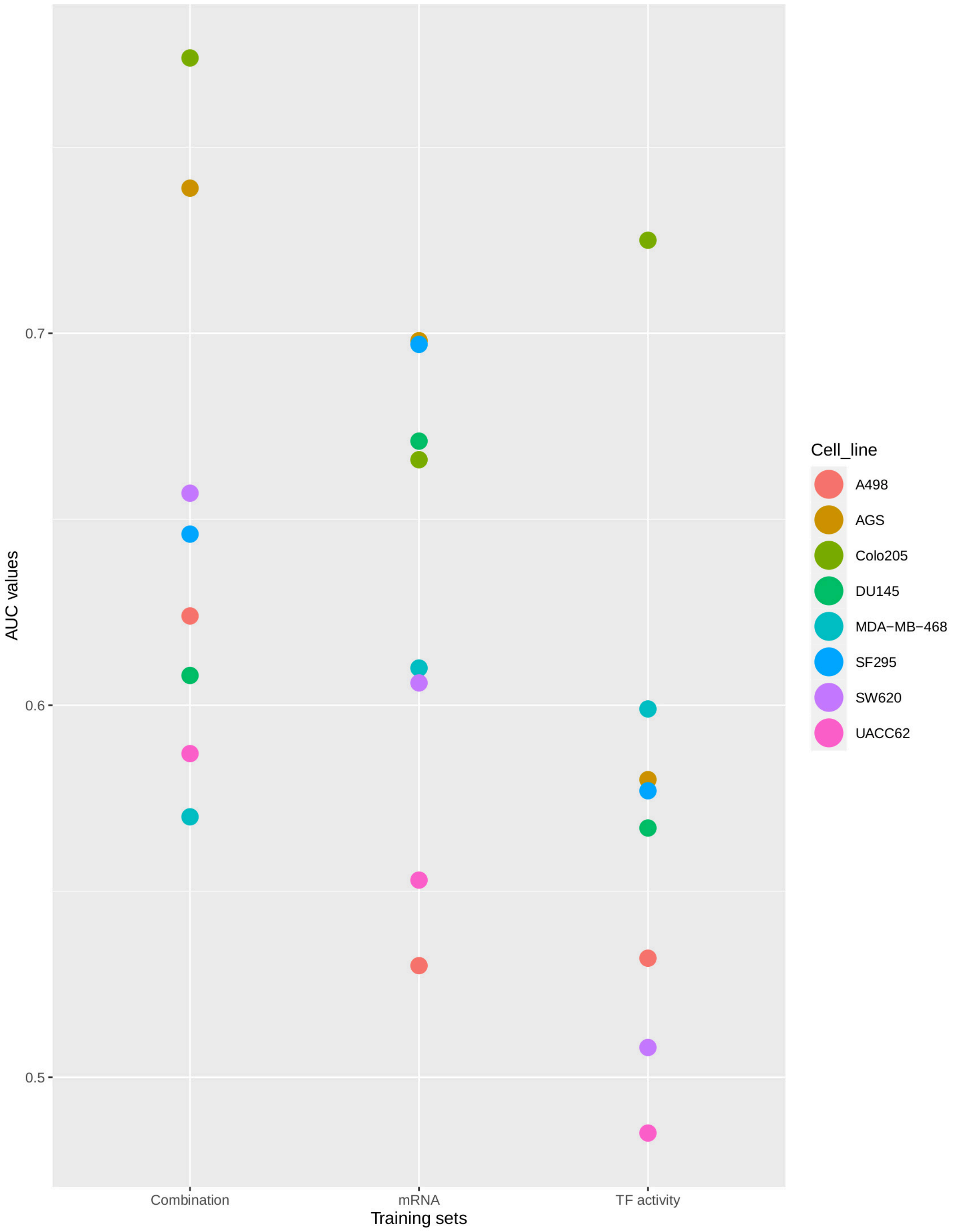

FIGURE 4 | Synergy prediction AUCs per cell line. The plot shows the AUC values of the ROC curves produced for drug synergy prediction performance using CASCADE 3.0 optimized to three different training sets. Colors represent different cell lines. The AUC values are plotted on the *X*-axis. The training sets are shown on the *Y*-axis.

performances shown side by side. The statistical significance of the difference between the performances of the two models was computed by a paired Wilcoxon signed-rank test. The median AUC of CASCADE 2.0 was found to be significantly less than the median AUC of CASCADE 3.0 (*p*-value of 0.03). As seen in **Figure 5**, with CASCADE 3.0, there was a considerable improvement of the performance in all cell lines except the kidney carcinoma cell line, A498. As mentioned above, CASCADE 3.0's improvement was most conspicuous for the Colo205 cell line. While CASCADE 3.0 overall seems to perform better for almost

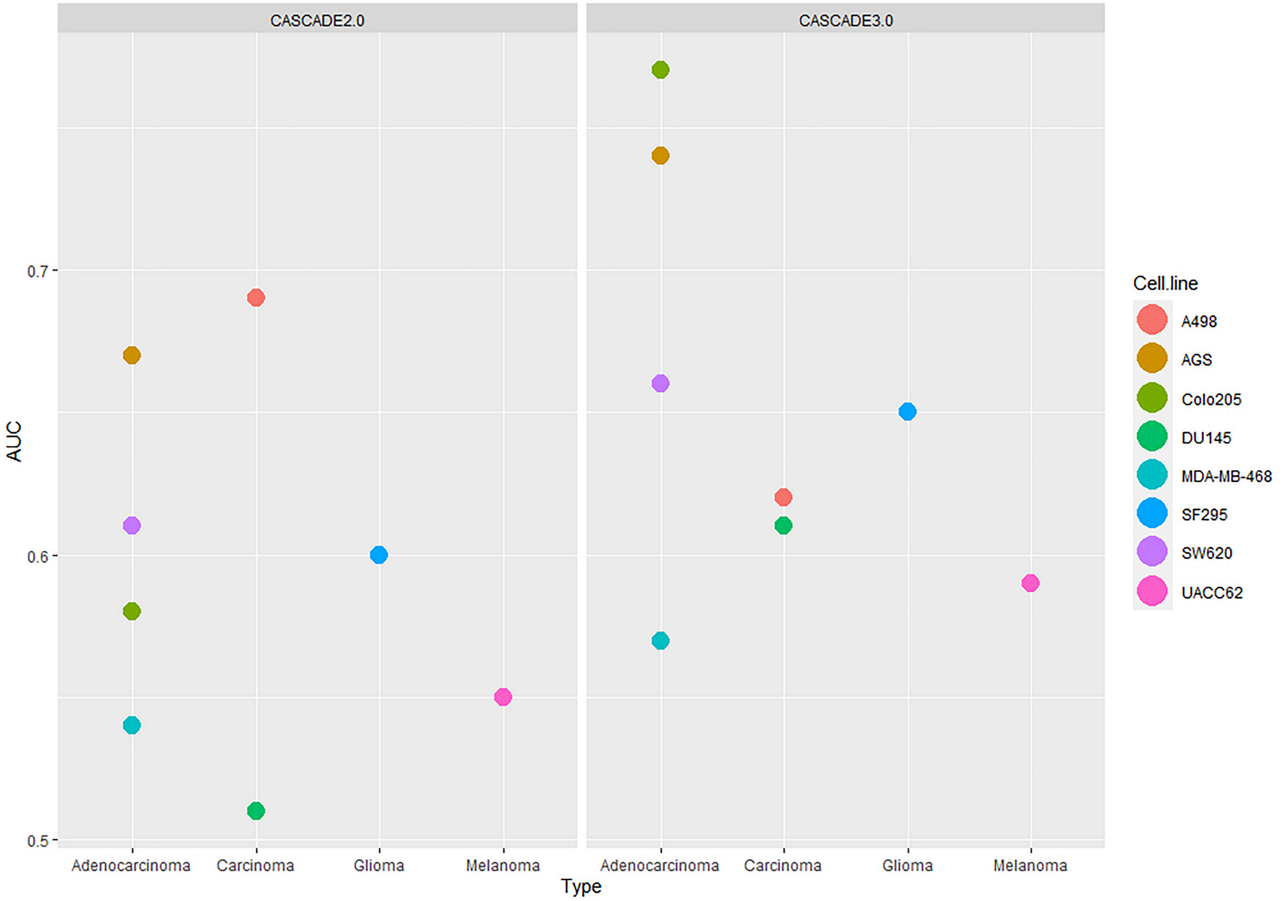

FIGURE 5 | Comparison of the performances of CASCADE 2.0 **(left)** and CASCADE 3.0 **(right)** with respect to their ability to predict drug synergies. Performance is measured in AUC values and presented in the *y*-axis. On the *x*-axis, cell lines are grouped based on their tissue of origin and each point of the plot represents a different cell line.

all cell lines, the range of improvement is most noticeable for cell lines originating from adenocarcinomas. This may indicate that the model extensions may better capture processes relevant to adenocarcinomas in general, rather than those specific for colorectal cancer and its subtypes.

Analysis of the Individual Predicted Synergies in Different Cell Lines

The mapping of the interactions between the 18 drugs used in this project and their target-entities revealed that the 20 entities of the model that serve as a target to those drugs are members of only 11 of the 25 modules. Two of the 18 drugs had no experimentally observed involvement in any synergy, reducing the number of modules involved in drug synergies to ten. Multiple drugs included in the screening and simulations were found targeting entities belonging to the same module. Specifically, four, three and two different drugs were targeting the PI3K/AKT, JNK/p38, and JAK/STAT modules, respectively. As many times cancer therapies take advantage of the dependency of cancer cells on an oncogene and/or loss of a tumor suppressor, and with the aforementioned pathways being among the most frequently altered pathways in several types of cancer, it is expected that multiple drugs have been designed to target these specific pathways (Thomas et al., 2015; Mayer and Arteaga, 2016; Martínez-Limón et al., 2020). The remaining seven modules included only one drug target each.

To visualize potential patterns in the ability of the model to correctly predict experimentally observed synergies, **Figure 6** displays the synergies in a module, represented as connecting edges between the drug targets, and in a cell line-specific manner. The **Figure 6A** shows all experimentally observed synergies, and **Figure 6B** shows the observed synergies that were also predicted. Only predictions obtained with the best performing training data set (*Combination-based*) are shown.

Three modules, PI3K/AKT, JNK/p38, and RAC, appear to be involved in the majority of the observed synergies, with most of the synergies observed in at least four cell lines. Additionally, the PI3K/AKT and JNK/p38 modules presented cases of intra-module synergies, with targeting of two entities in these individual modules resulting in synergistic response. Some modules, such as the one composed of the WNT negative regulators, were involved in synergies observed in

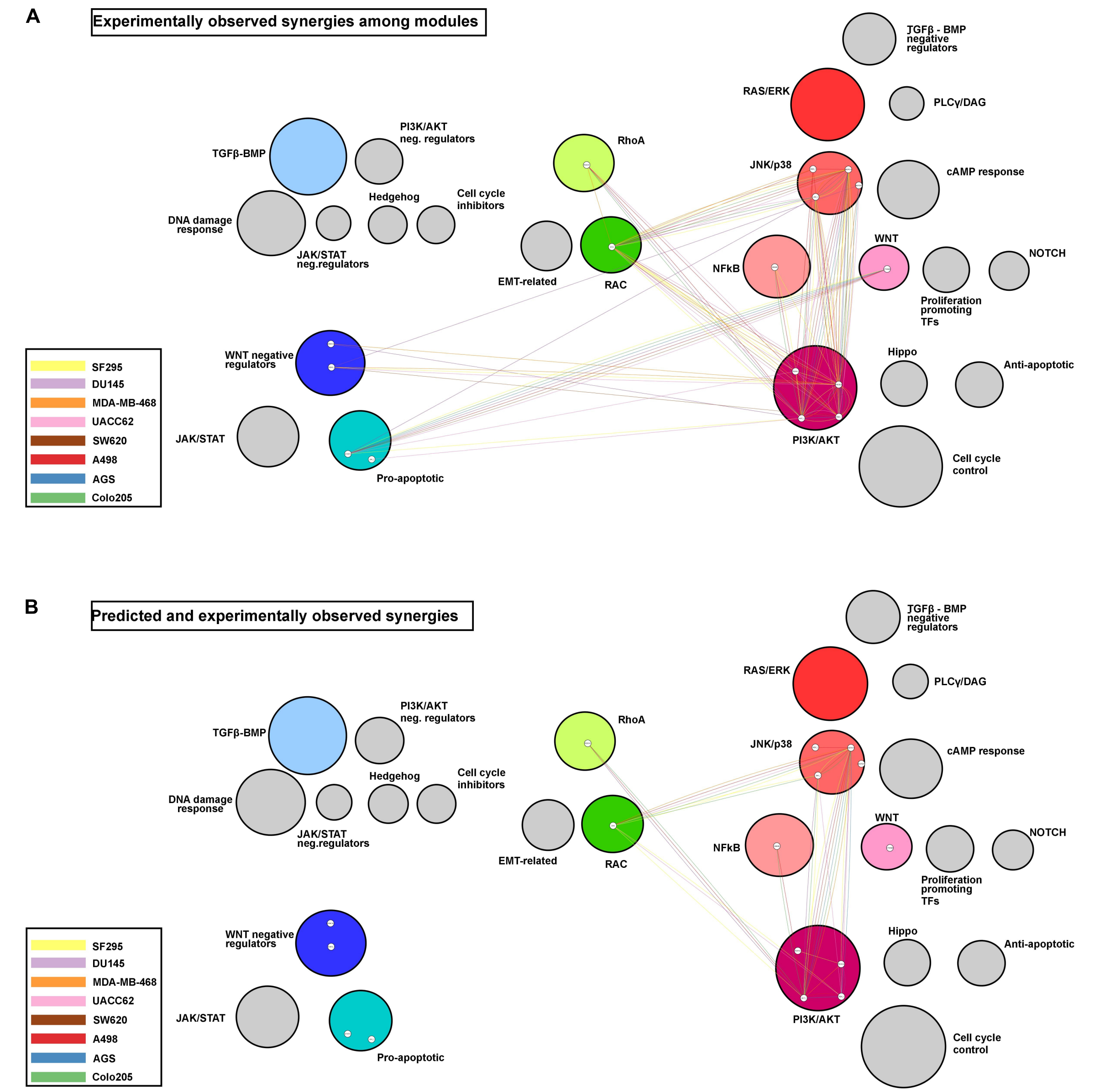

FIGURE 6 | Network modules and synergy prediction. Network representing the experimentally observed **(A)** and the predicted and experimentally observed **(B)** synergies as edges connecting the drug targets involved. Colored circles represent modules with targeted nodes, while circle size reflects the number of nodes in each module. The colors of the edges represent the cell line for which a synergy was observed and/or predicted.

only a few cell lines. It is interesting to note that the model fails to correctly predict drug synergies for the two modules with drug targets displayed toward the left in **Figure 6** (i.e., the modules that promote apoptosis when active), for any of the cell lines.

DISCUSSION

As standard treatment plans for cancer patients are often thwarted by acquired drug resistance of tumors, or the adverse effects and toxicities of monoregimen therapies (DeVita et al., 1975) combinatorial treatment with multiple chemical agents is being proposed as a solution (Kummar et al., 2010; Al-Lazikani et al., 2012; Madani Tonekaboni et al., 2018; Goldman et al., 2019). *In silico* screening of drug combinations may be particularly helpful in the pre-clinical stage, as it may serve to identify large numbers of combinations that need not be tested because they are unlikely to exhibit synergy (Celebi et al., 2019). *In silico* pre-screening may therefore solve many of the logistical and financial challenges that testing the enormous combinatorial drug compound space poses for screening facilities, provided that the computations predictions are of sufficient quality. Since the testing in the laboratory will only include a subset of the possible combinations (Cheng et al., 2019), it is critical to have a pre-screening procedure that produces low numbers of false negatives, as any potential blockbuster combination among them would not be tested.

Several approaches for *in silico* identification of drug synergies have been explored, some of them employing the computational modeling framework (Klinger et al., 2013; Miller et al., 2013; Flobak et al., 2015; Vitali et al., 2016; Eduati et al., 2017; Niederdorfer et al., 2020), as the current paper. In addition, machine learning (Jeon et al., 2018; Preuer et al., 2018; Sidorov et al., 2019; Yang et al., 2020), graph theory (Li et al., 2018), and multi-omics integration and analysis (Celebi et al., 2019; John et al., 2020) have been used. While machine learning approaches can be both highly flexible and accurate, there are certain limitations that should be acknowledged, such as the need to include expensive, hard to obtain training datasets, and for certain approaches (e.g., neural networks) they offer limited insights to what features confer predictability. On the other side, with logical models that make use of the abundantly available interaction data, synergy predictions can be successfully obtained from combining a prior knowledge network with observations, without the need for actual drug synergy training data, as demonstrated in the current manuscript. However, in a community effort to assess the computational prediction approaches (Menden et al., 2019), it was underlined that *in silico* synergy prediction remains a challenge even with using training data, and before such applications reach the clinic certain obstacles have to be overcome. One of these, the ability to tailor a computer model to the unique patient-specific molecular profiles, is key for the development of personalized therapies (Menden et al., 2019). To overcome this bottleneck, several methods to integrate patient-specific molecular characteristics have been proposed, with most of them exploring the use of multi-omics and perturbation data. As also demonstrated by our paper as well as others, logical models can be trained to both omics (Silverbush et al., 2017; Béal et al., 2019) and/or perturbation data (Fey et al., 2015; Eduati et al., 2020) in order to be further specified to specific cell lines or even patients.

The main aim of the project was to explore the use of multi-omics data to further extend and enhance a logical model that was produced by a manual curation effort. Analysis of colorectal tumor-derived omics data was used to define pathway modules representing functionally-related groups of proteins. Modules relevant for colorectal cancer were obtained through a workflow that combined multiple omics data to identify pathways and processes affected in the consensus molecular subtypes (CMS) of colorectal cancer (CRC). This top-down approach efficiently revealed CRC-specific processes, all with well-documented roles either in general tumor formation, or specific colorectal and/or general adenocarcinoma tumor ontogenesis. Next, a bottom-up approach was performed to extend an existing cancer model (CASCADE 2.0) with the new network nodes together with additional functionally relevant pathway and module components, to produce CASCADE 3.0. These approaches together exemplify an efficient middle-out workflow for cell fate decision network building, combining the best of well established top-down and bottom-up modeling approaches (Xavier et al., 2014). The top level results (affected pathways) were used to set the boundaries regarding the processes that should be present in the model, while the study of the individual entities involved in these pathways (bottom level) was guiding the curation and integration of these entities and their interacting partners in the system. The main advantage of this approach is that it provides a direct link between a collection of clinically relevant molecular phenotypes for very specific cancer (colorectal cancer subtypes) and a general model scaffold for cancer-related cell fate decisions. More specifically, it provides a modeler with very direct guidance for model refinement, essentially a blueprint of the modules whose inclusion should be considered. Similar workflows should allow model refinements for essentially any cellular system, provided that ample genome-wide information of that biological system is available. The modeler, however, will still face the responsibility to critically assess each model extension and guarantee the overall quality of the final model.

The broad availability of curated pathway resources and the definition of condition- and context specific modules could alleviate this workload, but it would be even better if a collection of reusable and interchangeable modular structures would be available that could be added or removed according to the different modeling purposes for different biological systems. The capacity of modules as building blocks has indeed been investigated in various types of biological networks (Segal et al., 2004; Schroeder, 2015), and the interest in building models in a modular manner is increasing.

To assess the quality of the CASCADE 3.0 model to predict drug synergies, simulations were performed for eight different cancer cell lines from various tissue origins, using three training sets for model configuration. The expectation was that the CMS extensions to the CASCADE 2.0 model would enhance the model performance for colorectal cancer. Model predictions were tested against experimentally observed synergies, and the AUC values indicated that CASCADE 3.0 had an improved prediction for Colo205, a colorectal adenocarcinoma cell line. However, the second adenocarcinoma cell line, SW620, displayed a more variable performance across the training data, with AUC values ranging from almost random (∼0.5) to 0.7. Interestingly, a multi-omics analysis of 34 colorectal adenocarcinoma cell lines (Berg et al., 2017) classified Colo205 and SW620 to different colorectal consensus molecular subtypes, as they have significant molecular differences. Among others, their CNV and gene expression profiles are quite distinct, causing Colo205 to be classified as a colon-like cell line, and SW620 as an undifferentiated cell line. These molecular differences and different subgroup classifications may indicate different underlying cellular signaling network activities or even different network topologies of these seemingly similar colorectal cancer cell lines, which in turn may explain the difference in CASCADE 3.0 model performance. In addition to Colo205, other well-performing cell lines include the gastric and prostate cancer cell lines. Interestingly, Colo205, AGS and DU145 all originate from the same tissue type, the epithelial, hinting to a pattern in the model's performance. By grouping the cell lines by their tissues of origin, it became evident that the model had a tendency to perform considerably better for the epithelial cancers (i.e., adenocarcinomas and carcinomas), and not only for the colorectal adenocarcinoma that it was specified for. Tumors are often classified by the organ they arise in. However, the molecular profiling of major cancer types has revealed surprising similarities between the molecular profiles of tumors arising from the same tissue type, but in different organs (Lin et al., 2017). For instance, the oncogenic role of the newly added Hippo, Hedgehog, and Notch modules is well reported in both prostate cancer (Zhang et al., 2015; Su and Xin, 2016; Buttyan et al., 2018) and gastric cancer (Kang et al., 2016; Yao et al., 2017; Akyala and Peppelenbosch, 2018). Together with the notion that targeted therapy based on molecular features is more effective (Senft et al., 2017), as practiced in precision medicine, the observation that CASCADE 3.0 has an overall better performance on cell lines displaying similar molecular phenotypes, may provide a handle on further optimizing logical modules for cancer cell line sets with shared other molecular profiles. This hypothesis could be further investigated in larger scale datasets, where the predictions of the model can be tested against additional drug combination data, as for example the drug synergy data reported in DrugComb (Zagidullin et al., 2019) and SYNERGxDB (Seo et al., 2020), and potentially in a broader set of cell lines.

The combination of proteomics with genomics data has been proposed as the most effective way to infer the state of an entity (Senft et al., 2017), corroborating our observation that models trained to the combined data set tend to perform generally better. The noticeable variation of the performance with different training data even for a specific cell line underlines the importance of correctly assessing the entities' states before training a model, which would need careful, high-quality assays for all proteins represented in the logical model. In most biological systems, however, it is assumed that the state of only a specific subset of its nodes is rather sufficient to control the global state of the system (Gao et al., 2014; Dnyane et al., 2018; Yang J. M. et al., 2018). Based on this, the accurate identification of the states of a well-chosen subset of nodes in the model rather than the majority of its nodes can be an attractive alternative (Niederdorfer et al., 2020). However, since the behavior of Boolean networks depends on multiple node and network features (Kauffman et al., 2004; Kochi et al., 2014), and often on the combined effect of those individual features (Kochi et al., 2014), it is essential to identify which of those features can be best used to assess the importance of a node for the global state of a system. Several features, including well-established or novel topology metrics (e.g., in-degree, out-degree, various path lengths, and centrality measures) and dynamical characteristics (e.g., bias and sensitivity of Boolean functions, presence of feedback loops), have been proposed to identify those nodes (Kochi et al., 2014; Sheikhahmadi et al., 2015; Wang et al., 2017). These findings suggest that further work on identifying such 'high leverage' nodes, or even complete modules that are critical for a model's performance and whose state therefore should be accurately assessed, is much needed.

Most of the observed synergies that could be predicted involve one of the PI3K/AKT, JNK/p38, or RAC modules. These modules play a central regulatory role in both normal and malignant cells, and many studies have already investigated and supported the effectiveness of combinatorial over single-agent treatment targeting these pathways, either in combination with each other or together with other pathways (Jain et al., 2017; Pons-Tostivint et al., 2017; Rocca et al., 2018). Alternatively, the apparent higher success rate for these modules may also be a consequence of the bias of this study toward drugs targeting the PI3K/AKT and JNK/p38 modules (seven of the 18 drugs). The classification of the modules (see **Figure 3**) based on whether they promote apoptosis, metastasis or proliferation, when the pathway they represent is active, revealed that the model fails to predict synergies for drugs targeting module combinations from different functional classes (apoptosis and proliferation), while it could predict most synergies that involved a combinatorial targeting of proliferation-associated modules. This observation may indicate a lack of regulatory detail in specific subparts, namely the apoptosis-related modules, or their cross-talk with the other parts of the network, especially given their direct interaction with the *Antisurvival* phenotype. To test this, additional curation efforts could be performed in an iterative way while testing model performance. Additional reasons that might affect the performance of the model in drug synergy prediction may be found in the lack of knowledge about the specificity of some cancer drugs (Rázga and Némethová, 2017). They may have unforeseen off-target effects that for a variety of reasons cannot be taken into account in the perturbed model simulations, which could seriously affect the model's performance (Saginc et al., 2017). For the moment, there are additional frontiers that need to be crossed before logical model-based therapy design can become relevant for the clinic.

In summary, this paper illustrates that middle-out model building provides an efficient approach to extend and optimize a logical model for specific cancer cell lines, or even individual patients, for more accurate drug effect simulations. The results illustrate that guided extensions of models to optimize their representation of a disease system can provide important insights and guide experimental design toward the identification of effective drug combinations. This approach allows the prioritization of the proposed synergies in a pre-clinical setting, to facilitate the selection of candidate drugs combinations that should be experimentally tested on cell lines.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

ET and MK designed the approach. MK supervised the project. ET carried out the multi-omics data analysis and model extensions and analyses. VT provided help with the logical model simulations. BN made available and provided guidance in the use of the CASCADE 2.0 model for model extensions.

ÅF and MV constructed and made available the NTNU logical model simulation pipeline. All authors contributed to the manuscript.

FUNDING

This project received funding from the Norwegian Research Council (project number 259842), under the frame of ERACoSysMed-1, the ERA-Net for Systems Medicine in clinical research and medical practice, The Liaison Committee between the Central Norway Regional Health Authority (Samarbeidsorganet) and the Norwegian University of Science and Technology (NTNU), and it received support from the NTNU Strategic Research Area NTNU Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.502573/full#supplementary-material

REFERENCES

Akyala, A. I., and Peppelenbosch, M. P. (2018). Gastric cancer and Hedgehog signaling pathway: emerging new paradigms. *Genes cancer* 9, 1–10. doi: 10.18632/genesandcancer.168

Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* 30, 679–692. doi: 10.1038/nbt.2284

Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., et al. (2016). Network-based inference of protein activity helps functionalize the genetic landscape of cancer. *Nat. Genet.* 48, 838–847. doi: 10.1038/ng.3593

Azuaje, F. (2017). Computational models for predicting drug responses in cancer research. *Brief. Bioinform.* 18, 820–829.

Bachmann, J., Raue, A., Schilling, M., Becker, V., Timmer, J., and Klingmüller, U. (2012). Predictive mathematical models of cancer signalling pathways. *J. Intern. Med.* 271, 155–165. doi: 10.1111/j.1365-2796.2011.02492.x

Béal, J., Montagud, A., Traynard, P., Barillot, E., and Calzone, L. (2019). Personalization of logical models with multi-omics data allows clinical stratification of patients. *Front. Physiol.* 9:1965. doi: 10.3389/fphys.2018.01965

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berg, K. C., Eide, P. W., Eilertsen, I. A., Johannessen, B., Bruun, J., Danielsen, S. A., et al. (2017). Multi-omics of 34 colorectal cancer cell lines-a resource for biomedical studies. *Mol. Cancer* 16, 1–16. doi: 10.1186/s12943-017-0691-y

Boesch, M., Spizzo, G., and Seeber, A. (2018). Concise review: aggressive colorectal cancer: role of epithelial cell adhesion molecule in cancer stem cells and epithelial-to-mesenchymal transition. *Stem Cells Transl. Med.* 7, 495–501. doi: 10.1002/sctm.17-0289

Bretones, G., Delgado Villar, M. D., and León Serrano, J. (2015). Myc and cell cycle control. *Biochim. Biophys. Acta* 1849, 506–516. doi: 10.1016/j.bbagrm.2014.03.013

Buttyan, R., Li, N., and Massah, S. (2018). Hedgehog in prostate cancer explained. *Oncoscience* 5, 67–68. doi: 10.18632/oncoscience.405

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337. doi: 10.1038/nature11252

Celebi, R., Don't Walk, O. B., Movva, R., Alpsoy, S., and Dumontier, M. (2019). In-silico prediction of synergistic anti-cancer drug combinations using multi-omics data. *Sci. Rep.* 9, 1–10. doi: 10.1038/s41598-019-45236-6

Cheng, F., Kovács, I. A., and Barabási, A. L. (2019). Network-based prediction of drug combinations. *Nat. Commun.* 10, 1–11. doi: 10.1038/s41467-019-09692-y

Cho, S. H., Park, S. M., Lee, H. S., Lee, H. Y., and Cho, K. H. (2016). Attractor landscape analysis of colorectal tumorigenesis and its reversion. *BMC Syst. Biol.* 10:96. doi: 10.1186/s12918-016-0341-9

Coppedè, F., Lopomo, A., Spisni, R., and Migliore, L. (2014). Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer. *World J. Gastroenterol.* 20, 943. doi: 10.3748/wjg.v20.i4.943

Crystal, A. S., Shaw, A. T., Sequist, L. V., Friboulet, L., Niederst, M. J., Lockerman, E. L., et al. (2014). Patient-derived models of acquired resistance can identify effective drug combinations for cancer. *Science* 346, 1480–1486. doi: 10.1126/science.1254721

Davison, E. J., Tarpey, P. S., Fiegler, H., Tomlinson, I. P., and Carter, N. P. (2005). Deletion at chromosome band 20p12. 1 in colorectal cancer revealed by high resolution array comparative genomic hybridization. *Genes Chromosomes Cancer* 44, 384–391. doi: 10.1002/gcc.20252

Deveau, P., Barillot, E., Boeva, V., Zinovyev, A., and Bonnet, E. (2016). Calculating biological module enrichment or depletion and visualizing data on large-scale molecular maps with ACSNMineR and RNaviCell R packages. *bioRxiv [Preprint]*. doi: 10.1101/064469

DeVita, V. T. Jr., Young, R. C., and Canellos, G. P. (1975). Combination versus single agent chemotherapy: a review of the basis for selection of drug treatment of cancer. *Cancer* 35, 98–110. doi: 10.1002/1097-0142(197501)35:1<98::aid-cncr2820350115>3.0.co;2-b

Dnyane, P. A., Puntambekar, S. S., and Gadgil, C. J. (2018). Method for identification of sensitive nodes in Boolean models of biological networks. *IET Syst. Biol.* 12, 1–6. doi: 10.1049/iet-syb.2017.0039

Eduati, F., Doldàn-Martelli, V., Klinger, B., Cokelaer, T., Sieber, A., Kogera, F., et al. (2017). Drug resistance mechanisms in colorectal cancer dissected with cell type–specific dynamic logic models. *Cancer Res.* 77, 3364–3375. doi: 10.1158/0008-5472.can-17-0078

Eduati, F., Jaaks, P., Wappler, J., Cramer, T., Merten, C. A., Garnett, M. J., et al. (2020). Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies. *Mol. Syst. Biolo.* 16:e8664. doi: 10.15252/msb.209690

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 46, D649–D655. doi: 10.1093/nar/gkx1132

Fey, D., Halasz, M., Dreidax, D., Kennedy, S. P., Hastings, J. F., Rauch, N., et al. (2015). Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci. Signal.* 8:ra130. doi: 10.1126/scisignal.aab0990

Flobak, Å, Baudot, A., Remy, E., Thommesen, L., Thieffry, D., Kuiper, M., et al. (2015). Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS Comput. Biol.* 11:e1004426. doi: 10.1371/journal.pcbi.1004426

Flobak, Å, Niederdorfer, B., Nakstad, V. T., Thommesen, L., Klinkenberg, G., and Lægreid, A. (2019). A high-throughput drug combination screen of targeted small molecule inhibitors in cancer cell lines. *Sci. Data* 6, 1–10. doi: 10.1038/s41597-019-0255-7

Fumiã, H. F., and Martins, M. L. (2013). Boolean network model for cancer pathways: predicting carcinogenesis and targeted therapy outcomes. *PLoS One* 8:e69008. doi: 10.1371/journal.pone.0069008

Gao, J., Liu, Y. Y., and D'souza, R. M. (2014). Barabá si AL. Target control of complex networks. *Nat. Commun.* 5:5415. doi: 10.1038/ncomms6415

Glass, L., and Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.* 39, 103–129. doi: 10.1016/0022-5193(73)90208-7

Goldman, A., Khiste, S., Freinkman, E., Dhawan, A., Majumder, B., Mondal, J., et al. (2019). Targeting tumor phenotypic plasticity and metabolic remodeling in adaptive cross-drug tolerance. *Sci. Signal.* 12:eaas8779. doi: 10.1126/scisignal.aas8779

Gonzalez, A. G., Naldi, A., Sanchez, L., Thieffry, D., and Chaouiya, C. (2006). GINsim: a software suite for the qualitative modelling, simulation and analysis of regulatory networks. *Biosystems* 84, 91–100. doi: 10.1016/j.biosystems.2005.10.003

Gottesman, M. M., Lavi, O., Hall, M. D., and Gillet, J. P. (2016). Toward a better understanding of the complexity of cancer drug resistance. *Annu. Rev. Pharmacol. Toxicol.* 56, 85–102. doi: 10.1146/annurev-pharmtox-010715-103111

Grieco, L., Calzone, L., Bernard-Pierrot, I., Radvanyi, F., Kahn-Perles, B., and Thieffry, D. (2013). Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS Comput. Biol.* 9:e1003286. doi: 10.1371/journal.pcbi.1003286

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., et al. (2015). The consensus molecular subtypes of colorectal cancer. *Nat. Med.* 21, 1350–1356. doi: 10.1038/nm.3967

Helikar, T., Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., et al. (2012). The cell collective: toward an open and collaborative approach to systems biology. *BMC Syst. Biol.* 6:96. doi: 10.1186/1752-0509-6-96

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. E. (2005). Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* 94:128701. doi: 10.1103/PhysRevLett.94.128701

Irurzun-Arana, I., Pastor, J. M., Trocóniz, I. F., and Gómez-Mantilla, J. D. (2017). Advanced Boolean modeling of biological networks applied to systems pharmacology. *Bioinformatics* 33, 1040–1048. doi: 10.1093/bioinformatics/btw747

Jain, P., Silva, A., Han, H. J., Lang, S. S., Zhu, Y., Boucher, K., et al. (2017). Overcoming resistance to single-agent therapy for oncogenic BRAF gene fusions via combinatorial targeting of MAPK and PI3K/mTOR signaling pathways. *Oncotarget* 8:84697. doi: 10.18632/oncotarget.20949

Jeon, M., Kim, S., Park, S., Lee, H., and Kang, J. (2018). In silico drug combination discovery for personalized cancer therapy. *BMC Syst. Biol.* 12:16. doi: 10.1186/s12918-018-0546-1

John, A., Qin, B., Kalari, K. R., Wang, L., and Yu, J. (2020). Patient-specific multi-omics models and the application in personalized combination therapy. *Fut. Oncol.* 16, 1737–1750. doi: 10.2217/fon-2020-0119

Joo, J. I., Zhou, J. X., Huang, S., and Cho, K. H. (2018). Determining relative dynamic stability of cell states using boolean network model. *Sci. Rep.* 8, 1–14. doi: 10.1038/s41598-018-30544-0

Joshi, S., and Durden, D. L. (2019). Combinatorial approach to improve cancer immunotherapy: rational drug design strategy to simultaneously hit multiple targets to kill tumor cells and to activate the immune system. *J. Oncol.* 2019:5245034. doi: 10.1155/2019/5245034

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kang, W., Cheng, A. S., Yu, J., and To, K. F. (2016). Emerging role of Hippo pathway in gastric and other gastrointestinal cancers. *World J. Gastroenterol.* 22, 1279–1288. doi: 10.3748/wjg.v22.i3.1279

Kauffman, S., Peterson, C., Samuelsson, B., and Troein, C. (2004). Genetic networks with canalyzing Boolean rules are always stable. *Proc. Natl. Acad. Sci. U.S.A.* 101, 17102–17107. doi: 10.1073/pnas.0407783101

Klinger, B., Sieber, A., Fritsche-Guenther, R., Witzel, F., Berry, L., Schumacher, D., et al. (2013). Network quantification of EGFR signaling unveils potential for targeted combination therapy. *Mol. Syst. Biol.* 9:673. doi: 10.1038/msb.2013.29

Kochi, N., Helikar, T., Allen, L., Rogers, J. A., Wang, Z., and Matache, M. T. (2014). Sensitivity analysis of biological Boolean networks using information fusion based on nonadditive set functions. *BMC Syst. Biol.* 8:92. doi: 10.1186/s12918-014-0092-4

Koutsogiannouli, E., Papavassiliou, A. G., and Papanikolaou, N. A. (2013). Complexity in cancer biology: is systems biology the answer? *Cancer Med.* 2, 164–177. doi: 10.1002/cam4.62

Kummar, S., Chen, H. X., Wright, J., Holbeck, S., Millin, M. D., Tomaszewski, J., et al. (2010). Utilizing targeted cancer therapeutic agents in combination: novel approaches and urgent requirements. *Nat. Rev. Drug Discov.* 9, 843–856. doi: 10.1038/nrd3216

Kuperstein, I., Bonnet, E., Nguyen, H. A., Cohen, D., Viara, E., Grieco, L., et al. (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* 4:e160. doi: 10.1038/oncsis.2015.19

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.* 16, 146–158. doi: 10.1038/nrg3885

Li, H., Li, T., Quang, D., and Guan, Y. (2018). Network propagation predicts drug synergy in cancers. *Cancer Res.* 78, 5446–5457. doi: 10.1158/0008-5472.CAN-18-0740

Lin, E. W., Karakasheva, T. A., Lee, D. J., Lee, J. S., Long, Q., Bass, A. J., et al. (2017). Comparative transcriptomes of adenocarcinomas and squamous cell carcinomas reveal molecular similarities that span classical anatomic boundaries. *PLoS Genet.* 13:e1006938. doi: 10.1371/journal.pgen.1006938

Lorenz, D. M., Jeng, A., and Deem, M. W. (2011). The emergence of modularity in biological systems. *Phys. Life Rev.* 8, 129–160. doi: 10.1016/j.plrev.2011.02.003

Madani Tonekaboni, S. A., Soltan Ghoraie, L., Manem, V. S. K., and Haibe-Kains, B. (2018). Predictive approaches for drug combination discovery in cancer. *Brief. Bioinform.* 19, 263–276. doi: 10.1093/bib/bbw104

Martínez-Limón, A., Joaquin, M., Caballero, M., Posas, F., and de Nadal, E. (2020). The p38 pathway: from biology to cancer therapy. *Int. J. Mol. Sci* 21:1913. doi: 10.3390/ijms21061913

Mayer, I. A., and Arteaga, C. L. (2016). The PI3K/AKT pathway as a target for cancer treatment. *Annu. Rev. Med.* 67, 11–28. doi: 10.1146/annurev-med-062913-051343

Menden, M. P., Wang, D., Mason, M. J., Szalai, B., Bulusu, K. C., Guan, Y., et al. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* 10, 1–17. doi: 10.1038/s41467-019-09799-2

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12:R41. doi: 10.1186/gb-2011-12-4-r41

Miller, M. L., Molinelli, E. J., Nair, J. S., Sheikh, T., Samy, R., Jing, X., et al. (2013). Drug synergy screen and network modeling in dedifferentiated liposarcoma identifies CDK4 and IGF1R as synergistic drug targets. *Sci. Signal.* 6:ra85. doi: 10.1126/scisignal.2004014

Mirza-Aghazadeh-Attari, M., Darband, S. G., Kaviani, M., Mihanfar, A., Attari, J. A., Yousefi, B., et al. (2018). DNA damage response and repair in colorectal cancer: defects, regulation and therapeutic implications. *DNA Repair.* 69, 34–52. doi: 10.1016/j.dnarep.2018.07.005

Naldi, A., Berenguier, D., Fauré, A., Lopez, F., Thieffry, D., and Chaouiya, C. (2009). Logical modelling of regulatory networks with GINsim 2.3. *Biosystems* 97, 134–139. doi: 10.1016/j.biosystems.2009.04.008

Naldi, A., Hernandez, C., Levy, N., Stoll, G., Monteiro, P. T., Chaouiya, C., et al. (2018). The CoLoMoTo interactive notebook: accessible and reproducible computational analyses for qualitative biological networks. *Front. Physiol.* 9:680. doi: 10.3389/fphys.2018.00680

Niederdorfer, B., Touré, V., Vazquez, M., Thommesen, L., Kuiper, M., Lægreid, A., et al. (2020). Strategies to enhance logic modeling-based cell line-specific drug synergy prediction. *Front. Physiol.* 11:862. doi: 10.3389/fphys.2020.00862

Perfetto, L., Briganti, L., Calderone, A., Cerquone Perpetuini, A., Iannuccelli, M., Langone, F., et al. (2016). SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.* 44, D548–D554. doi: 10.1093/nar/gkv1048

Pirkl, M., Hand, E., Kube, D., and Spang, R. (2016). Analyzing synergistic and non-synergistic interactions in signalling pathways using Boolean Nested Effect Models. *Bioinformatics* 32, 893–900. doi: 10.1093/bioinformatics/btv680

Pons-Tostivint, E., Thibault, B., and Guillermet-Guibert, J. (2017). Targeting PI3K signaling in combination cancer therapy. *Trends Cancer* 3, 454–469. doi: 10.1016/j.trecan.2017.04.002

Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C., and Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* 34, 1538–1546. doi: 10.1093/bioinformatics/btx806

Qiu, Y., Huang, Y., Tan, S., Li, D., Van Der Zijp-Tan, A. C., Borchert, G. M., et al. (2019). Exploring observability of attractor cycles in Boolean networks for biomarker detection. *IEEE Access* 7, 127745–127753. doi: 10.1109/access.2019.2937133

Rázga, F., and Némethová, V. (2017). Selective therapeutic intervention: a challenge against off-target effects. *Trends Mol. Med.* 23, 671–674. doi: 10.1016/j.molmed.2017.06.007

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J. P. (2006). GenePattern 2.0. *Nat. Genet.* 38, 500–501. doi: 10.1038/ng0506-500

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rocca, S., Carrà, G., Poggio, P., Morotti, A., and Brancaccio, M. (2018). Targeting few to help hundreds: JAK, MAPK and ROCK pathways as druggable targets in atypical chronic myeloid leukemia. *Mol. Cancer* 17:40. doi: 10.1186/s12943-018-0774-4

Roell, K. R., Reif, D. M., and Motsinger-Reif, A. A. (2017). An introduction to terminology and methodology of chemical synergy—perspectives from across disciplines. *Front. Pharmacol.* 8:158. doi: 10.3389/fphar.2017.00158

Rubio-Perez, C., Tamborero, D., Schroeder, M. P., Antolín, A. A., Deu-Pons, J., Perez-Llamas, C., et al. (2015). *In silico* prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* 27, 382–396. doi: 10.1016/j.ccell.2015.02.007

Saginc, G., Voellmy, F., and Linding, R. (2017). Harnessing off-target effects. *Nat. Chem. Biol.* 13, 1204–1205. doi: 10.1038/nchembio.2519

Sahoo, D., Wei, W., Auman, H., Hurtado-Coll, A., Carroll, P. R., Fazli, L., et al. (2018). Boolean analysis identifies CD38 as a biomarker of aggressive localized prostate cancer. *Oncotarget* 9:6550. doi: 10.18632/oncotarget.23973

Sammut, C., and Webb, G. I. (2017). *Encyclopedia of Machine Learning and Data Mining*. Berlin: Springer. doi: 10.1007/978-1-4899-7687-1

Schroeder, F. C. (2015). Modular assembly of primary metabolic building blocks: a chemical language in *C. elegans*. *Chem. Biol.* 22, 7–16. doi: 10.1016/j.chembiol.2014.10.012

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* 36, 1090–1098. doi: 10.1038/ng1434

Senft, D., Leiserson, M. D., Ruppin, E., and Ze'ev, A. R. (2017). Precision oncology: the road ahead. *Trends Mol. Med.* 23, 874–898. doi: 10.1016/j.molmed.2017.08.003

Seo, H., Tkachuk, D., Ho, C., Mammoliti, A., Rezaie, A., Madani Tonekaboni, S. A., et al. (2020). SYNERGxDB: an integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology. *Nucleic Acids Res.* 46, W494–W501. doi: 10.1093/nar/gkaa421

Shahzad, K., and Loor, J. (2012). Application of top-down and bottom-up systems approaches in ruminant physiology and metabolism. *Curr. Genom.* 13, 379–394. doi: 10.2174/138920212801619269

Sheikhahmadi, A., Nematbakhsh, M. A., and Shokrollahi, A. (2015). Improving detection of influential nodes in complex networks. *Physica A Stat. Mech. Appl.* 436, 833–845. doi: 10.1016/j.physa.2015.04.035

Shmulevich, I., Dougherty, E. R., and Zhang, W. (2002). From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proc. IEEE* 90, 1778–1792. doi: 10.1109/jproc.2002.804686

Sidorov, P., Naulaerts, S., Ariey-Bonnet, J., Pasquier, E., and Ballester, P. (2019). Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Front. Chem.* 7:509. doi: 10.3389/fchem.2019.00509

Silverbush, D., Grosskurth, S., Wang, D., Powell, F., Gottgens, B., Dry, J., et al. (2017). Cell-specific computational modeling of the PIM pathway in acute myeloid leukemia. *Cancer Res.* 77, 827–838. doi: 10.1158/0008-5472.can-16-1578

Smolen, P., Baxter, D. A., and Byrne, J. H. (2000). Mathematical modeling of gene networks. *Neuron* 26, 567–580. doi: 10.1016/s0896-6273(00)81194-0

Srihari, S., Raman, V., Leong, H. W., and Ragan, M. A. (2014). Evolution and controllability of cancer networks: a boolean perspective. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 83–94. doi: 10.1109/tcbb.2013.128

Srinivas, P. (2015). "Boolean network modeling for systematic identification of anticancer drug resistance in colorectal cancer," in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, New York, NY, 514–514. doi: 10.1145/2808719.2811436

Su, Q., and Xin, L. (2016). Notch signaling in prostate cancer: refining a therapeutic opportunity. *Histol. Histopathol.* 31, 149–157. doi: 10.14670/HH-11-685

Tang, J., Gautam, P., Gupta, A., He, L., Timonen, S., Akimov, Y., et al. (2019). Network pharmacology modeling identifies synergistic Aurora B and ZAK interaction in triple-negative breast cancer. *NPJ Syst. Biol. Appl.* 5, 1–11. doi: 10.1007/978-3-319-69980-6_1

Thomas, R. (1973). Boolean formalization of genetic control circuits. *J. Theor. Biol.* 42, 563–585. doi: 10.1016/0022-5193(73)90247-6

Thomas, S. J., Snowden, J. A., Zeidler, M. P., and Danson, S. J. (2015). The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. *Br. J. Cancer* 113, 365–371. doi: 10.1038/bjc.2015.233

Tiwari, A., Saraf, S., Verma, A., Panda, P. K., and Jain, S. K. (2018). Novel targeting approaches and signaling pathways of colorectal cancer: an insight. *World J. Gastroenterol.* 24:4428. doi: 10.3748/wjg.v24.i39.4428

Trairatphisan, P., Wiesinger, M., Bahlawane, C., Haan, S., and Sauter, T. (2016). A probabilistic boolean network approach for the analysis of cancer-specific signalling: A case study of deregulated PDGF signalling in GIST. *PLoS One* 11:e0156223. doi: 10.1371/journal.pone.0156223

Tolcher, A. W., Peng, W., and Calvo, E. (2018). Rational approaches for combination therapy strategies targeting the MAP kinase pathway in solid tumors. *Mol. Cancer Ther.* 17, 3–16. doi: 10.1158/1535-7163.mct-17-0349

Touré, V., Vercruysse, S., Acencio, M. L., Lovering, R. C., Orchard, S., Bradley, G., et al. (2020). The Minimum Information about a Molecular Interaction Causal Statement (MI2CAST). *Bioinformatics* doi: 10.1093/bioinformatics/btaa622 [Epub ahead of print].

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26, i237–i245. doi: 10.1093/bioinformatics/btq182

Veliz-Cuba, A., Aguilar, B., Hinkelmann, F., and Laubenbacher, R. (2014). Steady state analysis of Boolean molecular network models via model reduction and computational algebra. *BMC Bioinformatics* 15:221. doi: 10.1186/1471-2105-15-221

Vinson, K. E., George, D. C., Fender, A. W., Bertrand, F. E., and Sigounas, G. (2016). The Notch pathway in colorectal cancer. *Int. J. Cancer* 138, 1835–1842. doi: 10.1002/ijc.29800

Vitali, F., Cohen, L. D., Demartini, A., Amato, A., Eterno, V., Zambelli, A., et al. (2016). A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer. *PLoS One* 11:e0162407. doi: 10.1371/journal.pone.0162407

Wang, R. S., Saadatpour, A., and Albert, R. (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Phys. Biol.* 9:055001. doi: 10.1088/1478-3975/9/5/055001

Wang, S., Du, Y., and Deng, Y. (2017). A new measure of identifying influential nodes: efficiency centrality. *Commun. Nonlinear Sci. Numerical Simulat.* 47, 151–163. doi: 10.1016/j.cnsns.2016.11.008

Wierzbicki, P. M., and Rybarczyk, A. (2015). The Hippo pathway in colorectal cancer. *Folia Histochem. Cytobiol.* 53, 105–119. doi: 10.5603/fhc.a2015.0015

Wu, C., Zhu, X., Liu, W., Ruan, T., and Tao, K. (2017). Hedgehog signaling pathway in colorectal cancer: function, mechanism, and therapy. *Onco Targets Ther.* 10, 3249–3259. doi: 10.2147/ott.s139639

Xavier, J. C., Patil, K. R., and Rocha, I. (2014). Systems biology perspectives on minimal and simpler cells. *Microbiol. Mol. Biol. Rev.* 78, 487–509. doi: 10.1128/MMBR.00050-13

Yang, G., Gómez Tejeda Zañudo, J., and Albert, R. (2018). Target control in logical models using the domain of influence of nodes. *Front. Physiol.* 9:454. doi: 10.3389/fphys.2018.00454

Yang, J. M., Lee, C. K., and Cho, K. H. (2018). Global stabilization of boolean networks to control the heterogeneity of cellular responses. *Front. Physiol.* 9:774. doi: 10.3389/fphys.2018.00774

Yang, M., Jaaks, P., Dry, J., Garnett, M., Menden, M. P., and Saez-Rodriguez, J. (2020). Stratification and prediction of drug synergy based on target functional similarity. *npj Syst. Biol. Appl.* 6, 1–10. doi: 10.1038/s41540-020-0136-x

Yao, Y., Ni, Y., Zhang, J., Wang, H., and Shao, S. (2017). The role of Notch signaling in gastric carcinoma: molecular pathogenesis and novel therapeutic targets. *Oncotarget* 8, 53839–53853. doi: 10.18632/oncotarget.17809

Yu, C., Hong, H., Zhang, S., Zong, Y., Ma, J., Lu, A., et al. (2019). Identification of key genes and pathways involved in microsatellite instability in colorectal cancer. *Mol. Med. Rep.* 19, 2065–2076. doi: 10.3892/mmr.2019.9849

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., et al. (2019). DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res.* 47, W43–W51. doi: 10.1093/nar/gkz337

Zhang, L., Yang, S., Chen, X., Stauffer, S., Yu, F., Lele, S. M., et al. (2015). The hippo pathway effector YAP regulates motility, invasion, and castration-resistant growth of prostate cancer cells. *Mol. Cell. Biol.* 35, 1350–1362. doi: 10.1128/MCB.00102-15

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership