merlin v4.0: an updated platform for the reconstruction of high-quality genome-scale metabolic models

João Capela[1], Davide Lagoa[1], Ruben Rodrigues[1], Emanuel Cunha[1],Fernando Cruz[1] Ana Barbosa[1], José Bastos[1], Diogo Lima[1], Eugénio C. Ferreira[1], Miguel Rocha[1], Oscar Dias[1,*]

[1] Centre of Biological Engineering, University of Minho, 4704-553, Braga, Portugal

* To whom correspondence should be addressed.

Tel: 00351 253 604 422

Email: odias@deb.uminho.pt

ABSTRACT

Genome-scale metabolic models have been recognized as useful tools for better understanding living organism's metabolism. *Merlin* (https://merlin-sysbio.org/) is an open-source and user-friendly resource that hastens these models' reconstruction process, conjugating manual, and automatic procedures, while leveraging user's expertise with a curation-oriented graphical interface. An updated and redesigned version of *merlin* is herein presented. Since 2015, several features were implemented in *merlin,* along with profound changes in the software architecture, operating flow, and graphical interface. The current version (4.0) includes the implementation of novel algorithms and third-party tools for genome functional annotation, draft assembly, model refinement, and curation. Such updates led to an increase in the user-base, resulting in multiple published works including genome metabolic (re-)annotation and model reconstruction of multiple (lower and higher) eukaryotes and prokaryotes.

INTRODUCTION

Genome-scale Metabolic Models (GSMM) are genome-wide representations of a given organism's metabolism. Accordingly, the metabolic information inferred from the genome is integrated with biochemical data, commonly retrieved from reference databases. Within their broad spectrum of applications (1), high-quality GSMM can be used to predict phenotypes under different genetic and environmental conditions. Such models have been guiding metabolic engineering towards maximising cell factories' efficiency, predicting the most suitable conditions for driving flux into the production of compounds of interest.

During the past two decades, an increasing number of genome sequences have become available (2). Correspondingly, the number of curated GSMM has taken the pace, as it overtook the mark of the six thousand since 1999's *Haemophilus influenzae* model was published (1). Nevertheless, such models' reconstruction process is often time-consuming and laborious, as it could last from a few months to years (3).

Given the usefulness of GSMM, several endeavours were devised to accelerate their reconstruction's extensive tasks. State-of-the-art platforms can integrate fully and semi-automatic methods and graphical interfaces capable of assisting the model's manual curation (4). Automatic methods provide useful clues and resources to hasten the reconstruction of metabolic networks. Nonetheless, manual curation is recognised as a relevant task to ensure the high quality of genome-wide metabolic reconstructions (3). Such fact derives from the often absent or incomplete genome annotations (5) and biochemical data. As far as this is concerned, a balance between automatic processes and manual curation is desirable.

Merlin (6) is a comprehensive open-source platform, regularly updated, initially released in 2010 and published in 2015, aiming to assist and accelerate the main tasks of the GSMM' reconstruction. Multiple tools for genome functional annotation, draft assembly, model refinement, and validation have been implemented as plugins since the last major version.

Over the last years, *merlin's* user-base has grown considerably, resulting in the reconstruction of GSMM of multiple organisms from all domains of life, ranging from small-sized genome bacteria to the complexity of higher eukaryotes such as the oak tree.

Herein, we present the newest version of *merlin* (version 4), which includes profound software architecture changes, database management, and significant updates in already implemented operations. Furthermore, new valuable features were developed and integrated into the framework, mainly as *plugins*. Additionally, the graphical interface suffered profound alterations, as it was adjusted to enhance the user-friendliness and the assistance of the manual evaluation and curation. Boosted by all these improvements, *merlin* version 4 is an oriented and powerful tool with great scalability that ensures the trade-off between automatic and manual reconstruction workflows.

SOFTWARE ARCHITECTURE AND OVERVIEW

Merlin version 4 is implemented on top of AIBench, a Java application framework for scientific software development (7), and includes four main functional modules: Genome Functional Annotation, Model Reconstruction, Curation, and Data Import/Export. The software architecture is divided into layers of software dependencies. Herein, two main layers can be considered: *plugins* and *project* (outer and grey shaded layer in Figure 1, respectively).

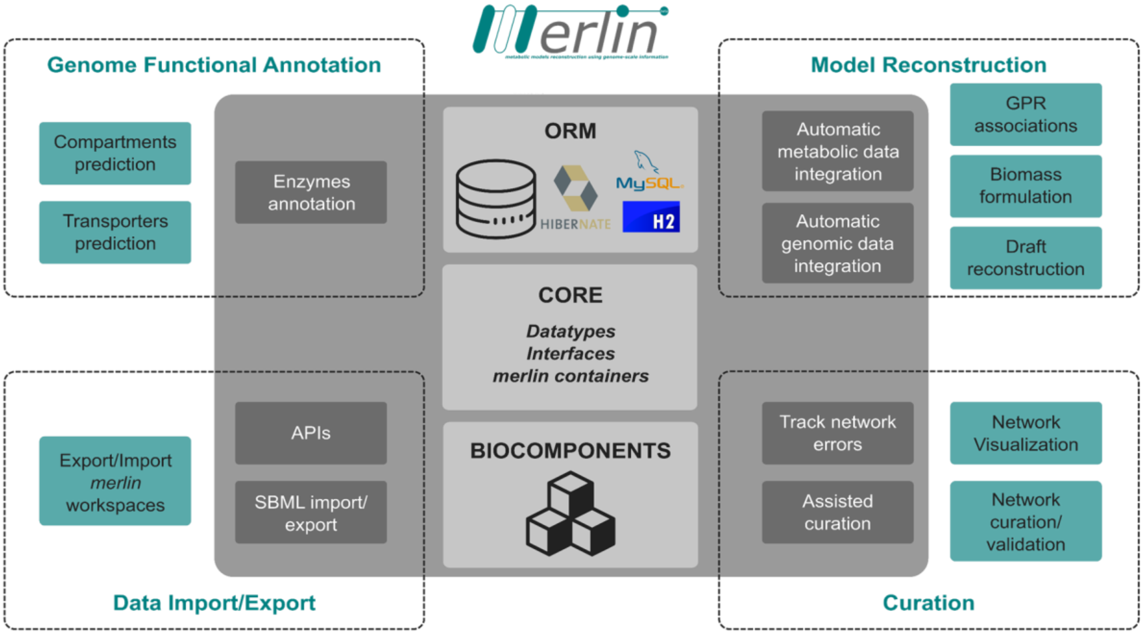

Figure 1 – *Merlin*'s holistic software architecture. This figure captures the functional modules covered by *merlin* and the layers of software dependencies. The modules are divided into Genome Functional Annotation, Model Reconstruction, Curation and Data Import/Export. Each of these modules comprises several tools associated to a given stage of reconstruction or data acquisition. These tools are either in the *plugins* form (outer, green-shaded boxes) or are included in the software *project* (inner, dark-grey shaded boxes). The *project,* which includes the main components of the software, is also composed by the BIOCOMPONENTS, CORE and ORM modules (central, light-grey boxes). Such components are the cornerstone of the software, as all the operations depend on the classes contained here.

As shown in Figure 1, the *project*'s *CORE* represents the software's core, where datatypes, interfaces, and containers are implemented to manipulate relevant data (e.g. genes, metabolites). Moreover, the database access and Object-relational mapping (ORM) modules were integrated into the *project* layer and are discussed in the next subsection.

The *project*'s *BIOCOMPONENTS* represent the internal library wrappers that allow handling and manipulating a given model without depending on the internal database. Accordingly, these computational objects can be used for independent operations.

Finally, Application Programming Interfaces (APIs) for external data access and other essential operations are contained in the *merlin'*s *project* layer. Table S2.2 (Supplementary data) provides a brief description of each database used by *merlin* and enumerates the operations that use data retrieved from those sources.

The *plugins'*s layer represents optional features that can be both installed and updated at any time. Moreover, a *plugins* management system was implemented to allow users to update the software without downloading and installing a completely new version. Each plugin and their implementation are enumerated and described in Table S2.1 (Supplementary data).

This new architecture is more flexible than the previous, as it allows the implementation and release of new features more easily and effectively.

ORM

The new features available in the newest version require the storage and manipulation of complex biological data. In the previous version, the interaction between the MySQL (https://www.mysql.com/) database and the biological data involved prepared statements. Database maintenance and management using prepared statements can be a time-consuming task that requires technical skills and may hinder the development of new features.

To address this problem, two frameworks, Hibernate (https://hibernate.org/) and Spring Data (https://spring.io/projects/spring-data) were integrated into *merlin*'s engine. Hibernate is an open-source ORM framework created to simplify development, allowing the conversion of Plain Old Java Objects (POJO) classes into database tables by using mapping configurations. The data present in a Java object is automatically converted into Structured Query Language (SQL) data without the development of SQL queries, required for the prepared statements. The adoption of an ORM framework improved the database maintenance, as any modification of Java entities can be automatically reflected into the database schema, facilitating the development of database migration patches between versions of *merlin* and its plugins. Another improvement prompted by the ORM adoption was the option to use the H2 database (http://h2database.com/), a Java embedded database, which overrides the technical requirement of installing MySQL in the machine that will run *merlin*.

Spring data was integrated into *merlin's* engine to foster Hibernate framework functionalities and further improve *merlin*'s database management. In the new version of *merlin*, Spring-based data accesses manage the database operations performed by Hibernate ORM and *merlin* CORE data structures. The data accesses split the database operations from *merlin*'s operations, enabling the development of improved data storage schemas, transparent to *merlin*'s operations, which allows integrating different types of database, e.g. non-relational databases like MongoDB (https://www.mongodb.com/) or Neo4J (https://neo4j.com/).

NEW FEATURES AND UPDATES

Workspaces management

In the previous version, *merlin* allowed deploying several "Projects" simultaneously to reconstruct different GSMM independently. In *merlin* version 4.0, "Projects" are renamed to "Workspaces", which characterises better the reconstruction environment. A new feature offered by *merlin* is that "Workspaces" can be exported, imported, and cloned at any moment or stage of the reconstruction process. This useful feature allows backing up and recovering "Workspaces" at any time, or just change machines without losing any work. Moreover, backward compatibility is guaranteed by a *plugin* that allows importing previous versions of merlin's "Workspaces".

New features for genome's functional annotations

*G*enome enzymatic annotation routines in *merlin* include the selection of both gene products and Enzyme Commission (EC) numbers from Basic Local Alignment Search Tool (BLAST) (8) or HMMER (9) results. For this selection, the scoring algorithm that accounts for the frequency and taxonomy described elsewhere (6, 10) is applied. In previous versions of *merlin*, the parameterization of the scoring algorithm was manually and empirically determined, often being a time-consuming process or not providing the optimal parameters for the genome being annotated.

SamPler (11) is a semi-automatic method that relies on statistical metrics and the manual annotation of a genome sample to determine the optimal parameters. Such tool is available in the *plugin* form to configure *merlin*'s annotation routine, semi-automatically. Though useful and offering optimum results, an automatic procedure is also available for this task. *Merlin* provides the *automatic workflow* operation that annotates genes according to a list of taxonomically related genus or species. Such list is used to prioritize the gene product and EC number from entries associated with the utmost organisms or genus. Hence, homologous genes from taxonomically related organisms will be chosen as the best candidates for the gene product and EC number selection.

TRIAGE (12) a tool that performs the annotation of transport proteins and the generation of transport reactions was published just after *merlin*. However, due to its strict rules for the Transport Candidate Gene (TCG) identification, the Transporter Systems Tracker (*TranSyT*) was developed (13). *TranSyT* generates a set of system-specific transport reactions with gene-protein-reaction (GPR) rules associated. Afterwards, these reactions are integrated into the model, automatically.

The compartmentalisation of proteins and metabolites requires loading reports from *WolfPSORT* (14), *PSORTb3* (15), or *LocTree3* (16). Compared to older versions, *merlin* can no longer use a remote Java API for accessing *WolfPSORT*'s functionalities. Instead, it offers operations to integrate each tool's prediction report rapidly. For *WolfPSORT* and *LocTree3* reports, *merlin* web-scraps the prediction results directly on the web, requiring only the Uniform Resource Locator (URL) of the report webpage, whereas, for *PSORTb3*, *merlin* parses a prediction file ("Long Format"), and integrates the results in the database. *LocTree3* reports allow widening the range of options for subcellular localisation prediction of enzymes and metabolites, providing a valuable tool whose performance is equal or better than other state-of-the-art software (16). These annotations can be integrated into the model, defining thresholds for the subcellular localisation prediction scores.

Biomass formulation

The most common approach towards the biomass formulation includes the adoption of biomass equations from taxonomically related organisms. Although recurrent and assumed not to propagate significant errors (17), this method has been suggested as incorrect by a more recent study (18). Instead, estimating the average protein and (deoxy)ribonucleotide contents from the genome seems to provide better predictions (18). Hence, *merlin* provides an operation that estimates the contents as mentioned above, automatically. Furthermore, gene expression data can be used to adjust the protein contents to experimentally data.

Moreover, *merlin* provides templates for biomass equations. These templates include averaged contents of each biomass-related macromolecule for different types of organisms. These values were retrieved from the literature, from the Model SEED database (19), or can be experimentally determined. Nevertheless, it is worth noting that these templates can only serve as baselines, requiring further curation and adjustments.

New features for network curation

Debugging large networks can be a time-consuming task, even for the most experienced curator. Both *merlin* graphical interface and services provide means to hasten such a laborious task.

The network topology can be assessed using the "*Draw in Browser*" functionality, which allows visualising KEGG pathways in the default browser. Two new plugins, namely *MetExploreViz* (20) and *Escher Maps* (21) complement the network topology analysis package, allowing to visualise more than one pathway simultaneously and highlighting network characteristics such as compartments, shared pathways, and network gaps.

Although these tools ease manual curation, they cannot evaluate which reactions are carrying flux, which metabolites are not being produced, or assess the model's consistency. Therefore, new *plugins* were implemented, namely Biological networks In Silico Optimization (*BioISO*) (22) and *MEMOTE* (23). *BioISO* highlights whether a set of reactions are carrying flux when maximised or minimised. Also, it enables tracking errors that impair the synthesis of a given reaction's metabolites. This tool assists the manual curation and gap-filling along with *merlin*'s user-friendly graphical interface that permits adding, editing, and removing reactions. Lastly, *MEMOTE* (23) constitutes a suite of standardised tests proposed by the modelling community to assess the model's quality. The quality of the models reconstructed in *merlin* can be verified without leaving *merlin's* graphical interface.

Changes in the Graphical Interface

Merlin's GUI has significantly changed since the previously published version (6). Besides alterations in the colours and graphical components, the workspace's entities have changed as well.

In version 4, *model*, *annotation*, and *validation* are the main modules. As shown in Figure S1.1 (Supplementary material), the *model* module is subdivided into five main entities: *genes, proteins, metabolites, reactions,* and *pathways*. The information associated with each of these entities is enumerated in comprehensive tables where users are allowed to edit, insert, and remove elements at any moment during the reconstruction process. These entities represent the metabolic information present in the model. The *annotation* module is subdivided into *enzymes* and *compartments*, in which results from the genome functional annotation are enumerated. The inclusion of the *compartments* entity in *merlin* 4 allows users to curate the subcellular localisation prediction results. Lastly, *validation* can include tables with results retrieved from curation and validation tools such as *BioISO* and *MEMOTE*.

Other features

Merlin allows generating a draft reconstruction from other models. This feature uses BLAST (8) or Smith-Waterman (24) alignments' algorithms to decide which reactions should be inherited from the input model. The ultimate output is a draft reconstruction ready for refinement and curation.

Furthermore, several databases and external GSMM' cross-references are assigned to all metabolites and reactions. These are displayed in a comprehensive table after pressing the magnifier button in the corresponding row of the respective entity (either *metabolites* or *reactions*).

Merlin 4 also allows to import and export the genome, annotation results in GenBank (25) file format, and the GSM model in several levels and versions of the Systems Biology Markup Language (SBML) (26) format (level 2 versions 1 to 4 and level 3 versions 1 and 2) at any stage of the reconstruction.

RESULTS AND DISCUSSION

Updated workflow

Merlin's *modus operandi* has changed since 2015's published version. The general workflow is depicted in Figure 2. The whole operating flow is assisted by an enhanced user-friendly interface that leverages the user expertise and curation's quality.

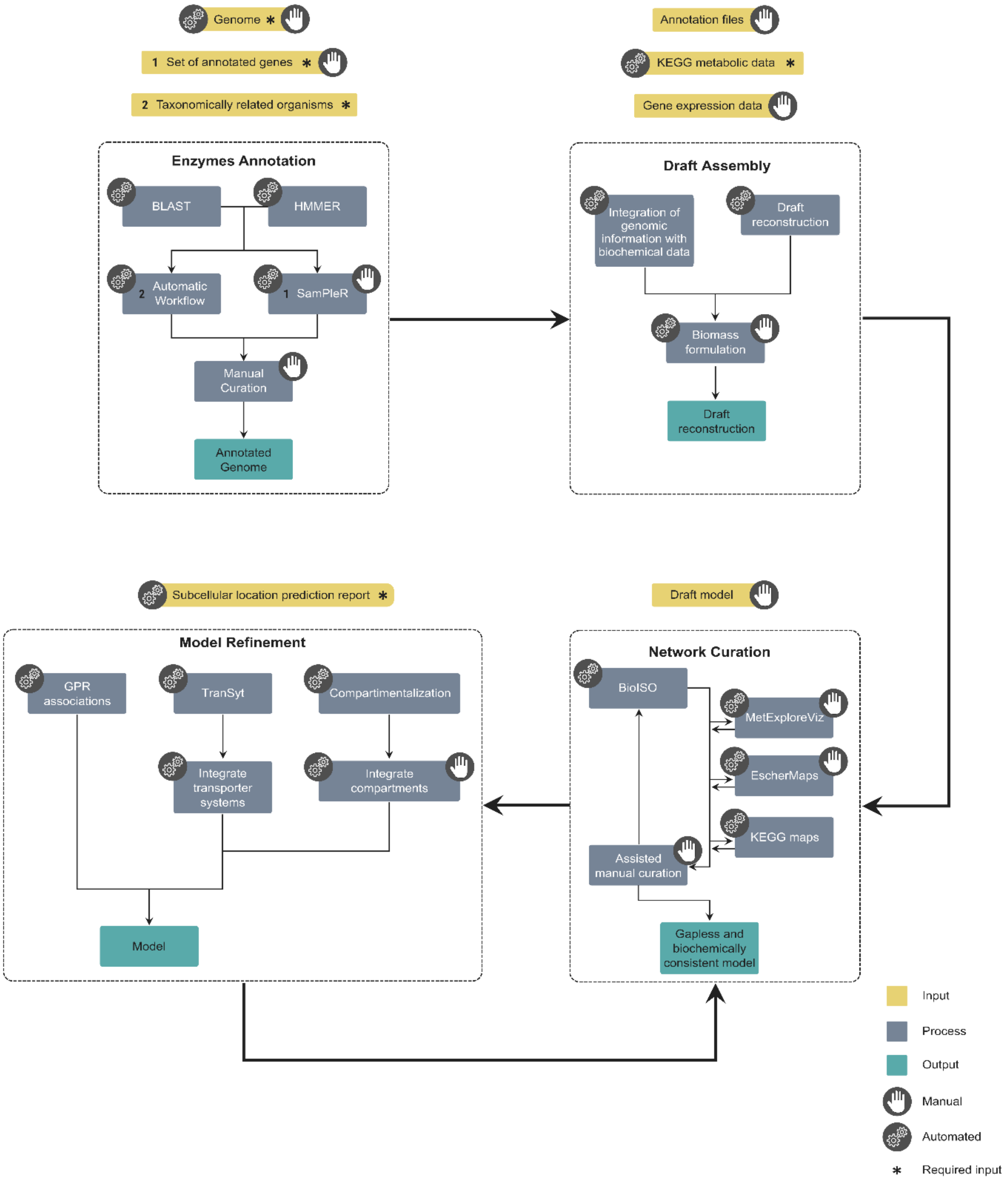

Figure 2 – *Merlin* updated workflow. The reconstruction steps, the possible inputs and, outputs and processes. The "hand" symbol indicates manual processes, the "engine" symbols signals the completely automatic ones, while the semi-automatic ones are marked by both. Each stage's possible inputs are shown in yellow boxes, processes in blue, and outputs in green. Finally, the asterisk corresponds to the processes' required inputs. The workflow encompasses four main stages: enzymes annotation, draft assembly, model curation, and model refinement. The output of each stage is not necessarily required for the next, as *merlin* provides operations to import external genome annotation files and GSMM. Hence, users are able to start reconstructing a model from scratch or from externally obtained annotations and models. If the process of reconstruction is entirely performed in *merlin*, the output of each stage is required for the following one.

The starting point is either selecting or creating a new *Workspace* using the menu *workspace* -> *open* at the top bar illustrated in Figure S1.1. When creating a new *Workspace*, the user may import an organism's genome automatically; however, as highlighted in Figure 2, manually importing the genome is also possible.

The next step is to automatically annotate the enzymes with either BLAST (8) or HMMER (9), going through the *annotation* -> *enzymes* -> *BLAST* or *HMMER*. On the other hand, annotation reports may be imported, using *annotation* -> *enzymes* -> *load* menu.

The following stage is applying the scoring system described in (6) to the annotation output. Setting the best score parameters from the previous annotation is facilitated using new *plugins* such as *SamPler* at *annotation* -> *enzymes* -> *SamPler*. Alternatively, the scoring may be overruled with the *automatic workflow* at *annotation* -> *enzymes* -> *automatic workflow*. The output of these semi- and automatic methods will be a set of annotated enzymes. The described processes' results will be available at the *enzyme's* view, as shown in Figure S1.2 in the supplementary material.

Nevertheless, the manual curation of each one of these results is advisable. Such process is highly facilitated by the graphical interface provided by *merlin,* as shown in Figure S1.2. Furthermore, the *enzymes* view (Figure S1.2) provides information about the state of the annotation, gene products, EC numbers, and scores. Each row's magnifier button provides information on the BLAST (8) or HMMER (9) operations results. Moreover, the *status* column highlights the enzyme's revision state, following the pattern described in (27), and redirecting users to the UniProtKB (28) database site. The candidate gene products and EC numbers are available in the dropdown boxes, being easily updated and integrated into the model if necessary.

The next step is integrating the curated enzymes annotation with metabolic information (metabolites and reactions), retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (29) (second stage in Figure 2). Additionally, finalising the assembly of the so-called draft metabolic network will require the biomass formulation, which can be performed using the *plugin* at *model* -> *create* -> *e-biomass equation*. The biomass pseudo-reactions and KEGG information will be enumerated in the *reactions* view (Figure S1.3), the *metabolites,* and *pathways* view. The integrated information is available and can be updated at any stage of the reconstruction. Moreover, in this second stage of the workflow, a draft model can be automatically reconstructed using the *model* -> *draft model reconstruction* operation.

Gaps and inconsistencies are likely to be found in the draft reconstruction. Hence, as highlighted in the third stage of the workflow (Figure 2), *merlin* provides several tools to highlight blocked and unbalanced reactions, and correct reactions' reversibility, among others to assist in the network manual curation. Additionally, the *reactions* view (Figure S1.3 in Supplementary material) allows users to rapidly insert, edit, duplicate, and remove existing reactions through explicit buttons and checkboxes. This view also enables users to visualise and include reactions from the universal internal database, facilitating manual gap-filling.

BioISO analysis can be performed in *validation -> BioISO -> execute BioISO,* providing useful insights into the network's state. After running *BioISO*, a table is docked in the dashboard's *validation* module, rendering the results as in Figure S1.4 (Supplementary material). Likewise, other tools such as *MetExploreViz*, *Escher,* accessed through *validation -> network visualisation,* and *Draw in Browser* operation (Figure S1.3) can be used to get insights into the network topology.

The transporter systems annotation and transport reactions generation can be performed with *TranSyT,* by executing *model -> create -> transport reactions TranSyT*. These reactions will be then listed in the *reactions* view*,* having associated a surrogated pathway designated *Transporters Pathway*. The model compartmentalisation is then advisable, either through *PSort3b, LocTree3* or *WolfPSort,* by executing *annotation -> compartments -> load reports*. The results will be rendered at the *compartments* view docked in the dashboard's *enzymes* module as shown in Figure S1.5 in the Supplementary material. Further integration through *annotation -> compartments -> integrate to model* is mandatory, allowing the user to define thresholds over the obtained prediction scores and ignore possible erroneous compartments. Finally, the GPR associations are automatically generated using BLAST or Smith-Waterman alignments' algorithms through *merlin*'s operation *model -> create -> gene-protein-reaction associations*.

MEMOTE test suite can be then applied to assess the general state of the model. This *plugin* is available through *validation -> memote* and the tests' report will be rendered in a comprehensive table that will be docked under the *validation* module of the dashboard (Figure S1.6. in the Supplementary material).

Lastly, the model can be exported (*workspace -> export -> model*) in the SBML format with valid identifiers and cross-references for metabolites and reactions.

Merlin provides an optimized workflow that allows users to start the reconstruction process from scratch to a high-quality GSM model. Nevertheless, the workflow's flexibility, complemented by import/export operations, enables users to begin or continue the reconstruction process from external genome annotations or external GSMM. Such feature also enables the curation, reannotation, and refinement of already reconstructed models and annotated genomes using *merlin*'s internal tools. Moreover, each of the obtained results within *merlin* can be exported in GenBank, Excel Workbook, and SBML formats.

Validation

Since 2015, tenths of high-quality GSMM were reconstructed using *merlin* and at least 11 were published (30–36). A list of models completely developed in *merlin* is provided in Table S2.3 (Supplementary material). The applications of these models include synergies with food biotechnology, being integrative tools for drug targeting, and efficient biomaterials production. Noticeably, the first genome-wide metabolic model of a ligneous tree was developed using *merlin*. Moreover, *merlin*'s curation tools and interface have been used to (re-)annotate several organisms' genomes (37–40).

The metabolic annotation of *Pythium irregulare* CBS 494.86's genome has been employed in *merlin* (37, 41), delivering interesting results towards the better comprehension of Eicosapentaenoic acid production. EC numbers were linked to 3809 genes, whereas 945 to membrane transporter proteins. Genes associated with amino acid and lipid production, as well as with the consumption of carbon and nitrogen sources present in wastewater were identified. Such result provides important insights into the metabolism of *P. irregulare* CBS 494.86 and the possible applications in the production of value-added lipids for industrial purposes.

The continuous development of *merlin* provided valuable tools for the improvement of a *Kluyveromyces. lactis* model (42). Such model is recognized as a reference among the yeast community, as it serves as baseline for both experimental studies on metabolism and regulation towards biotechnological applications (43–49) and to build models of other yeast (50, 51). The *Streptococcus pneumoniae* R6 (31) model has been reconstructed in *merlin,* having exciting results regarding the genome annotation and predictive capability. In this work, 67 essential genes that were not listed in the Online GEne Essentiality (OGEE) database (52) were proposed as critical for certain environmental conditions, guiding the discovery of novel drug targets. Moreover, the phenotype predictions under different conditions were assessed and corroborated by five different studies.

Likewise, other pathogens were modelled using *merlin*. A remarkable example was *Candida albicans* (32) that now provides an accurate platform for drug targeting. The model correctly predicted 78% of the essential genes (84 out of 108 validated experimentally) under anaerobic growth for different carbon and nitrogen sources.

Recently, four Lactic Acid Bacteria (LAB) GSMM were reconstructed with *merlin;* namely, *Streptococcus thermophilus LMD-9*, *Lactobacillus acidophilus La-14, Lactobacillus helveticus CNRZ32,* and *Lactobacillus rhamnosus GG*. The growth rate under different carbon sources, amino acid auxotrophies, and minimal medium were in good agreement with the experimental data. These GSMM were compared to other LAB models in different environmental conditions with the aim on food biotechnology applications (33).

The *Quercus suber* GSM model (*Cunha et al. 2021*) epitomises the first genome-scale metabolic model of a ligneous tree. This model comprises 6475 metabolites, 6248 reactions, 7901 genes, and 8 different compartments. The authors replicated and simulated growth under autotrophic and heterotrophic conditions, covering the photorespiration process. Furthermore, although not straightforward, this model includes plant secondary metabolism and complete pathways for the bioproduction of suberin monomers, compounds of paramount importance in cork production. Finally, the Quantum Yield (QA) and Assimilation Quotient (AQ) predicted by the model were validated and are in accordance with values reported in the literature. Another highlight of this work is the conversion of the generic model, reconstructed in *merlin,* into tissue-specific and multi-tissue models, corroborating the scalability and usability of *merlin* models using Troppo (53).

These models confirm the reliability and robustness of *merlin.* Moreover, the models' applications range from food biotechnology, drug targeting to biomaterials' production. The large extent and relevance of such applications are particularly endorsed by several published research articles (43–47) and different collaborations. Finally, the reconstruction of large-sized genome organisms such as *Quercus suber* highlights, unequivocally, *merlin*'s scalability and quality.

Since its launch in 2016, *merlin*'s website was visited more than eleven thousand times and over three thousand downloads were performed. *Merlin* has gained notoriety over the years and is nowadays recognised as a reference software for the reconstruction of high-quality GSMM by the scientific community through multiple research articles (54–62), reviews, and book chapters (1, 4, 63–80).

CONCLUSION

Merlin 4.0 is an updated and robust open-source software developed in Java to reconstruct high-quality GSMM. Several improvements and new features were included, when compared with the last published version (*merlin* 2.0), these updates are related not only to the reconstruction process but also the software architecture and graphical interface. Accordingly, new semi- and completely automatic plugins were added, establishing synergies with the improved user-friendly graphical interface to facilitate the model's curation. Moreover, *merlin's* software architecture is currently much more modular, allowing the easy insertion of both in-house and third-party computational tools.

Genome functional annotation, draft assembly, model curation, and refinement represent tasks covered by *merlin*. The genome functional annotation, BLAST (8) or HMMER (9) results annotation can be accelerated by either *SamPler* or the *automatic workflow*. *TranSyT* performs the transporter systems annotation and integration the subcellular localisation prediction is also ensured. Besides hastening the draft assembly, it can be generated from existing models. The model curation is leveraged by multiple tools for tracking network errors and inconsistencies, complemented by network visualisation add-ons.

The scientific community has extensively used *merlin*, as corroborated by the multiple published works (32–35, 42, 81, 82) and considerable user-base expansion. The scalability and reliability of *merlin* 4.0 have been showcased with the reconstruction of multiple models, but particularly with the *Quercus suber's*, which is the first GSMM of a ligneous tree.

In conclusion, *merlin* is an integrated, open-source, and updated platform designed to anchor GSMM' reconstruction. The trade-off between automatic processes and assisted manual curation ensured by *merlin*, allows users to leverage their genomics and biochemistry expertise towards high-quality reconstructions.

DATA AVAILABILITY

merlin is an open-source application, currently available for Linux, Windows, and MacOS. It is distributed under the GNU General Public License at the website https://www.merlin-sysbio.org. Moreover, *merlin* source code, including plugins, is fully available at https://github.com/merlin4-sysbio.

Comprehensive documentation is provided at https://merlin-sysbio.org/tutorial.html. Animated snapshots of *merlin's* interface and clear descriptions of each step in the reconstruction are therein provided.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS AND

This work is a result of the project 22231/01/SAICT/2016: Biodata.pt Infraestrutura Portuguesa de Dados Biológicos, supported by the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). The authors would also like to acknowledge Portuguese Foundation for Science and Technology (FCT) for providing PhD scholarships to E. Cunha (DFA/BD/8076/2020) and F. Cruz (SFRH/BD/139198/2018). Oscar Dias acknowledges FCT for the Assistant Research contract obtained under CEEC Individual 2018.

FUNDING

The authors would like to acknowledge the Centre of Biological Engineering (CEB, UMinho) for financial and equipment support and the FCT under the scope of the strategic funding of UIDB/04469/2020 unit.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest

REFERENCES

1. Gu,C., Kim,G.B., Kim,W.J., Kim,H.U. and Lee,S.Y. (2019) Current status and applications of genome-scale metabolic models. *Genome Biol.*, **20**.
2. Mukherjee,S., Stamatis,D., Bertsch,J., Ovchinnikova,G., Katta,H.Y., Mojica,A., Chen,I.M.A., Kyrpides,N.C. and Reddy,T.B.K. (2019) Genomes OnLine database (GOLD) v.7: Updates and new features. *Nucleic Acids Res.*, **47**, D649–D659.
3. Thiele,I. and Palsson,B. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.
4. Mendoza,S.N., Olivier,B.G., Molenaar,D. and Teusink,B. (2019) A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.*, **20**, 158.
5. Salzberg,S.L. (2019) Next-generation genome annotation: We still struggle to get it right. *Genome Biol.*, 10.1186/s13059-019-1715-2.
6. Dias,O., Rocha,M., Ferreira,E.C. and Rocha,I. (2015) Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Res.*, 10.1093/nar/gkv294.
7. Fdez-Riverola,F., Glez-Peña,D., Lõpez-Fernández,H., Reboiro-Jato,M. and Méndez,J.R. (2012) A JAVA application framework for scientific software development. *Softw. - Pract. Exp.*, 10.1002/spe.1108.
8. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 10.1016/S0022-2836(05)80360-2.
9. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, 10.1093/bioinformatics/14.9.755.
10. Dias,O., Rocha,M., Ferreira,E.C. and Rocha,I. (2018) Reconstructing high-quality large-scale metabolic models with merlin. In *Methods in Molecular Biology*. Humana Press Inc., Vol. 1716, pp. 1–36.
11. Cruz,F., Lagoa,D., Mendes,J., Rocha,I., Ferreira,E.C., Rocha,M. and Dias,O. (2019) SamPler – a novel method for selecting parameters for gene functional annotation routines. *BMC Bioinformatics*, 10.1186/s12859-019-3038-4.
12. Dias,O., Gomes,D., Vilaca,P., Cardoso,J., Rocha,M., Ferreira,E.C. and Rocha,I. (2017) Genome-Wide Semi-Automated Annotation of Transporter Systems. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, **14**, 443–456.
13. Lagoa,D., Faria,J.P., Liu,F., Cunha,E., Ferreira,E.C., Henry,C.S. and Dias,O. (2021) TranSyT: the Transport Systems Tracker.
14. Horton,P., Park,K.-J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585-7.
15. Yu,N.Y., Wagner,J.R., Laird,M.R., Melli,G., Rey,S., Lo,R., Dao,P., Cenk Sahinalp,S., Ester,M., Foster,L.J., *et al.* (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 10.1093/bioinformatics/btq249.
16. Goldberg,T., Hecht,M., Hamp,T., Karl,T., Yachdav,G., Ahmed,N., Altermann,U., Angerer,P., Ansorge,S., Balasz,K., *et al.* (2014) LocTree3 prediction of localization. *Nucleic Acids Res.*, **42**, W350-5.

17. Varma,A. and Palsson,B.O. (1993) Metabolic capabilities of escherichia coli. II. Optimal growth patterns. *J. Theor. Biol.*, 10.1006/jtbi.1993.1203.

18. Santos,S. and Rocha,I. (2016) Estimation of biomass composition from genomic and transcriptomic information. *J. Integr. Bioinform.*, **13**, 161–169.

19. Seaver,S.M.D., Liu,F., Zhang,Q., Jeffryes,J., Faria,J.P., Edirisinghe,J.N., Mundy,M., Chia,N., Noor,E., Beber,M.E., *et al.* (2021) The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res.*, 10.1093/nar/gkaa746.

20. Chazalviel,M., Frainay,C., Poupin,N., Vinson,F., Merlet,B., Gloaguen,Y., Cottret,L. and Jourdan,F. (2018) MetExploreViz: Web component for interactive metabolic network visualization. *Bioinformatics*, 10.1093/bioinformatics/btx588.

21. King,Z.A., Dräger,A., Ebrahim,A., Sonnenschein,N., Lewis,N.E. and Palsson,B.O. (2015) Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput. Biol.*, **11**.

22. Cruz,F., Capela,J., Ferreira,E.C., Rocha,M. and Dias,O. (2021) BioISO: an objective-oriented application for assisting the curation of genome-scale metabolic models.

23. Lieven,C., Beber,M.E., Olivier,B.G., Bergmann,F.T., Ataman,M., Babaei,P., Bartell,J.A., Blank,L.M., Chauhan,S., Correia,K., *et al.* (2020) MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.*, **38**, 272–276.

24. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, 10.1016/0022-2836(81)90087-5.

25. Sayers,E.W., Cavanaugh,M., Clark,K., Pruitt,K.D., Schoch,C.L., Sherry,S.T. and Karsch-Mizrachi,I. (2021) GenBank. *Nucleic Acids Res.*, 10.1093/nar/gkaa1023.

26. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., and the rest of the SBML Forum:, Arkin,A.P., Bornstein,B.J., Bray,D., *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

27. Models,R.H.L.M., Ferreira,C., Rocha,I., Dias,O. and Rocha,M. (2018) Chapter 1 with merlin.

28. The UniProt Consortium (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.*, 10.1093/nar/gky1049.

29. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, 10.1093/nar/gkw1092.

30. Dias,O., Pereira,R., Gombert,A.K., Ferreira,E.C. and Rocha,I. (2014) iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis. *Biotechnol. J.*, **9**, 776–790.

31. Dias,O., Saraiva,J., Faria,C., Ramirez,M., Pinto,F. and Rocha,I. (2019) iDS372, a Phenotypically Reconciled Model for the Metabolism of Streptococcus pneumoniae Strain R6. *Front. Microbiol.*, **10**, 1283.

32. Viana,R., Dias,O., Lagoa,D., Galocha,M., Rocha,I. and Teixeira,M.C. (2020) Genome-Scale Metabolic Model of the Human Pathogen Candida albicans: A Promising Platform for Drug

Target Prediction. *J. Fungi*, **6**, 171.

33. Cunha,E., Cruz,F., Silva,M., Ribeiro,C., Rocha,I., Zeidan,A. and Dias,O. (2021) Unveiling metabolic traits of dairy bacteria through in silico genome-scale metabolic modelling.
34. Pereira,B., Miguel,J., Vilaça,P., Soares,S., Rocha,I. and Carneiro,S. (2018) Reconstruction of a genome-scale metabolic model for Actinobacillus succinogenes 130Z. *BMC Syst. Biol.*, **12**, 61.
35. Wang,Y., Tong,Z. and Xie,J. (2019) Reconstruction and in Silico Simulation Towards Electricigens Metabolic Network of Electronic Mediator. In.pp. 217–218.
36. Mienda,B.S. and Dräger,A. (2021) Genome-Scale Metabolic Modeling of Escherichia coli and Its Chassis Design for Synthetic Biology Applications. In.pp. 217–229.
37. Fernandes,B.S., Dias,O., Costa,G., Kaupert Neto,A.A., Resende,T.F.C., Oliveira,J.V.C., Riaño-Pachón,D.M., Zaiat,M., Pradella,J.G.C. and Rocha,I. (2019) Genome-wide sequencing and metabolic annotation of Pythium irregulare CBS 494.86: Understanding Eicosapentaenoic acid production. *BMC Biotechnol.*, 10.1186/s12896-019-0529-3.
38. Resende,T., Correia,D.M., Rocha,M. and Rocha,I. (2013) Re-annotation of the genome sequence of Helicobacter pylori 26695. *J. Integr. Bioinform.*, 10.1515/jib-2013-233.
39. Gomes,D., Aguiar,T.Q., Dias,O., Ferreira,E.C., Domingues,L. and Rocha,I. (2014) Genome-wide metabolic re-annotation of Ashbya gossypii: New insights into its metabolism through a comparative analysis with Saccharomyces cerevisiae and Kluyveromyces lactis. *BMC Genomics*, 10.1186/1471-2164-15-810.
40. Dias,O., Gombert,A.K., Ferreira,E.C. and Rocha,I. (2012) Genome-wide metabolic (re-) annotation of Kluyveromyces lactis. *BMC Genomics*, 10.1186/1471-2164-13-517.
41. Fernandes,B.S., Dias,O., Zaiat,M., Pradella,J.G.C. and Rocha,I. (2017) Eicosapentaenoic acid (EPA) production in silico by Pythium irregulare, as value-added product, using sugarcane vinasse as carbon source.
42. Dias,O., Pereira,R., Gombert,A.K., Ferreira,E.C. and Rocha,I. (2014) iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis. *Biotechnol. J.*, **9**, 776–790.
43. Pentjuss,A., Stalidzans,E., Liepins,J., Kokina,A., Martynova,J., Zikmanis,P., Mozga,I., Scherbaka,R., Hartman,H., Poolman,M.G., *et al.* (2017) Model-based biotechnological potential analysis of Kluyveromyces marxianus central metabolism. *J. Ind. Microbiol. Biotechnol.*, 10.1007/s10295-017-1946-8.
44. Jin,Y.S. and Cate,J.H. (2017) Metabolic engineering of yeast for lignocellulosic biofuel production. *Curr. Opin. Chem. Biol.*, 10.1016/j.cbpa.2017.10.025.
45. Gorietti,D., Zanni,E., Palleschi,C., Delfini,M., Uccelletti,D., Saliola,M., Puccetti,C., Sobolev,A.P., Mannina,L. and Miccheli,A. (2015) 13C NMR based profiling unveils different α-ketoglutarate pools involved into glutamate and lysine synthesis in the milk yeast Kluyveromyces lactis. *Biochim. Biophys. Acta - Gen. Subj.*, 10.1016/j.bbagen.2015.07.008.
46. Schabort,D.T.W.P., Letebele,P., Steyn,L., Kilian,S.G. and Du Preez,J.C. (2016) Differential RNA-seq, multi-network analysis and metabolic regulation analysis of kluyveromyces marxianus reveals a compartmentalised response to xylose. *PLoS One*, 10.1371/journal.pone.0156242.
47. Weiner,M., Tröndle,J., Albermann,C., Sprenger,G.A. and Weuster-Botz,D. (2016) Perturbation

experiments: Approaches for metabolic pathway analysis in bioreactors. In *Advances in Biochemical Engineering/Biotechnology*.

48. Ortiz-Merino,R.A., Varela,J.A., Coughlan,A.Y., Hoshida,H., da Silveira,W.B., Wilde,C., Kuijpers,N.G.A., Geertman,J.M., Wolfe,K.H. and Morrissey,J.P. (2018) Ploidy variation in Kluyveromyces marxianus separates dairy and non-dairy isolates. *Front. Genet.*, 10.3389/fgene.2018.00094.
49. Nurcholis,M., Lertwattanasakul,N., Rodrussamee,N., Kosaka,T., Murata,M. and Yamada,M. (2020) Integration of comprehensive data and biotechnological tools for industrial applications of Kluyveromyces marxianus. *Appl. Microbiol. Biotechnol.*, 10.1007/s00253-019-10224-3.
50. Tomàs-Gamisans,M., Ferrer,P. and Albiol,J. (2016) Integration and validation of the genome-scale metabolic models of Pichia pastoris: A comprehensive update of protein glycosylation pathways, lipid and energy metabolism. *PLoS One*, 10.1371/journal.pone.0148031.
51. Marcišauskas,S., Ji,B. and Nielsen,J. (2019) Reconstruction and analysis of a Kluyveromyces marxianus genome-scale metabolic model. *BMC Bioinformatics*, 10.1186/s12859-019-3134-5.
52. Chen,W.H., Minguez,P., Lercher,M.J. and Bork,P. (2012) OGEE: An online gene essentiality database. *Nucleic Acids Res.*, 10.1093/nar/gkr986.
53. Ferreira,J., Vieira,V., Gomes,J., Correia,S. and Rocha,M. (2020) Troppo - A Python Framework for the Reconstruction of Context-Specific Metabolic Models. In *Advances in Intelligent Systems and Computing*. Springer Verlag, Vol. 1005, pp. 146–153.
54. Machado,D., Herrgård,M.J. and Rocha,I. (2016) Stoichiometric Representation of Gene–Protein–Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction. *PLoS Comput. Biol.*, **12**.
55. Hädicke,O. and Klamt,S. (2017) EColiCore2: a reference network model of the central metabolism of Escherichia coli and relationships to its genome-scale parent model. *Sci. Rep.*, **7**, 39647.
56. Jouhten,P., Boruta,T., Andrejev,S., Pereira,F., Rocha,I. and Patil,K.R. (2016) Yeast metabolic chassis designs for diverse biotechnological products. *Sci. Rep.*, **6**, 29694.
57. Pfau,T., Christian,N., Masakapalli,S.K., Sweetlove,L.J., Poolman,M.G. and Ebenhöh,O. (2016) The intertwined metabolism of Medicago truncatula and its nitrogen fixing symbiont Sinorhizobium meliloti elucidated by genome-scale metabolic models. *bioRxiv*, 10.1101/067348.
58. Karlsen,E., Schulz,C. and Almaas,E. (2018) Automated generation of genome-scale metabolic draft reconstructions based on KEGG. *BMC Bioinformatics*, **19**.
59. Santibáñez,R., Garrido,D. and Martin,A.J.M. (2020) Atlas : automatic modeling of regulation of bacterial gene expression and metabolism using rule-based languages. *Bioinformatics*, 10.1093/bioinformatics/btaa1040.
60. Alballa,M., Aplop,F. and Butler,G. (2020) TranCEP: Predicting the substrate class of transmembrane transport proteins using compositional, evolutionary, and positional information. *PLoS One*, **15**, e0227683.
61. Lam,T.J., Stamboulian,M., Han,W. and Ye,Y. (2020) Model-based and phylogenetically adjusted quantification of metabolic interaction between microbial species. *PLOS Comput. Biol.*, **16**, e1007951.

62. Simensen,V., Voigt,A. and Almaas,E. (2020) High-quality genome-scale metabolic model of Aurantiochytrium sp. T66. *bioRxiv*, 10.1101/2020.11.06.371070.

63. Angione,C. (2019) Human Systems Biology and Metabolic Modelling: A Review—From Disease Metabolism to Precision Medicine. *Biomed Res. Int.*, **2019**, 1–16.

64. Thor,S., Peterson,J.R. and Luthey-Schulten,Z. (2017) Genome-Scale Metabolic Modeling of Archaea Lends Insight into Diversity of Metabolic Function. *Archaea*, **2017**, 1–18.

65. Rau,M.H. and Zeidan,A.A. (2018) Constraint-based modeling in microbial food biotechnology. *Biochem. Soc. Trans.*, **46**, 249–260.

66. Chiappino-Pepe,A., Pandey,V., Ataman,M. and Hatzimanikatis,V. (2017) Integration of metabolic, regulatory and signaling networks towards analysis of perturbation and dynamic responses. *Curr. Opin. Syst. Biol.*, **2**, 59–66.

67. Castillo,S., Patil,K.R. and Jouhten,P. (2019) Yeast Genome-Scale Metabolic Models for Simulating Genotype–Phenotype Relations. In.pp. 111–133.

68. Mendes-Soares,H. and Chia,N. (2017) Community metabolic modeling approaches to understanding the gut microbiome: Bridging biochemistry and ecology. *Free Radic. Biol. Med.*, **105**, 102–109.

69. Little,B.J., Blackwood,D.J., Hinks,J., Lauro,F.M., Marsili,E., Okamoto,A., Rice,S.A., Wade,S.A. and Flemming,H.-C. (2020) Microbially influenced corrosion—Any progress? *Corros. Sci.*, **170**, 108641.

70. Patra,P., Das,M., Kundu,P. and Ghosh,A. (2021) Recent advances in systems and synthetic biology approaches for developing novel cell-factories in non-conventional yeasts. *Biotechnol. Adv.*, **47**, 107695.

71. Fondi,M. and Fani,R. (2017) Constraint-based metabolic modelling of marine microbes and communities. *Mar. Genomics*, **34**, 1–10.

72. Xu,X., Liu,Y., Du,G. and Liu,L. (2020) Systems biology, synthetic biology, and metabolic engineering. In *Systems and Synthetic Metabolic Engineering*. Elsevier, pp. 1–31.

73. Choi,K.R., Jang,W.D., Yang,D., Cho,J.S., Park,D. and Lee,S.Y. (2019) Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. *Trends Biotechnol.*, **37**, 817–837.

74. Kim,W.J., Kim,H.U. and Lee,S.Y. (2017) Current state and applications of microbial genome-scale metabolic models. *Curr. Opin. Syst. Biol.*, **2**, 10–18.

75. Machado,D., Andrejev,S., Tramontano,M. and Patil,K.R. (2018) Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.*, **46**, 7542–7553.

76. Lakshmanan,M., Koduru,L. and Lee,D.-Y. (2018) Software Applications for Phenotype Analysis and Strain Design of Cellular Systems. In *Emerging Areas in Bioengineering*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, pp. 771–791.

77. Chung,W.Y., Zhu,Y., Mahamad Maifiah,M.H., Shivashekaregowda,N.K.H., Wong,E.H. and Abdul Rahim,N. (2021) Novel antimicrobial development using genome-scale metabolic model of Gram-negative pathogens: a review. *J. Antibiot. (Tokyo).*, **74**, 95–104.

78. Weber,T. and Kim,H.U. (2016) The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synth. Syst. Biotechnol.*, **1**, 69–79.

79. Lopes,H. and Rocha,I. (2017) Genome-scale modeling of yeast: chronology, applications and critical perspectives. *FEMS Yeast Res.*, **17**.

80. Muller,E.E.L., Faust,K., Widder,S., Herold,M., Martínez Arbas,S. and Wilmes,P. (2018) Using metabolic networks to resolve ecological properties of microbiomes. *Curr. Opin. Syst. Biol.*, **8**, 73–80.

81. Cunha,E., Zeidan,A. and Dias,O. (2021) Towards the Reconstruction of the Genome-Scale Metabolic Model of Lactobacillus acidophilus La-14. In.pp. 205–214.

82. Dias,O., Saraiva,J., Ferreira,E.C. and Rocha,I. (2015) Towards a genome-scale metabolic model of the Streptococcus pneumoniae R6 avirulent strain.