Exploring synergies between plant metabolic modelling and machine learning

Marta Sampaio\textsuperscript{a,b,\*}, Miguel Rocha\textsuperscript{a,b}, Oscar Dias\textsuperscript{a,b,\*}

\textsuperscript{a}Centre of Biological Engineering, University of Minho, Campus of Gualtar, 4710-057 Braga, Portugal
\textsuperscript{b}LABBELS, Associate Laboratory, Braga, Guimarães, Portugal

Abstract

As plants produce an enormous diversity of metabolites to help them adapt to the environment, the study of plant metabolism is of utmost importance to understand different plant phenotypes. Omics data have been generated at an unprecedented rate for several organisms, including plants, and are widely used to study the central dogma of molecular biology, connecting the genome to phenotypes. Constraint-based modelling (CBM) methods, working over genome-scale metabolic models (GSMMs), have been crucial for organising and analysing omics data by integrating them with biochemical knowledge. In 2009, the first plant GSMM was reconstructed and, since then, several advances have been made, including the creation of context- and multi-tissue models that have supported the study of plant metabolism. Nevertheless, plant metabolic modelling remains very challenging. In parallel, as omics datasets are complex and heterogeneous, machine learning (ML) models have been applied in their interpretation to foster knowledge discovery. Recently, the first studies combining both CBM and ML approaches have emerged and have shown promising results. Here, we present the major advances in plant metabolic modelling and review the main CBM-ML hybrid studies. Finally, we discuss the application of machine learning to address the unique challenges of plant metabolic modelling.

© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

1. Introduction .................................................................................................................. 1886
2. Plant metabolic modelling ......................................................................................... 1886
   2.1. Arabidopsis thaliana ......................................................................................... 1888
   2.2. Zea mays ................................................................................................. 1890
   2.3. Oryza sativa ............................................................................................. 1890
   2.4. Other organisms ......................................................................................... 1890
3. Major challenges and limitations ............................................................................. 1891
4. Machine learning and constraint-based modelling .............................................. 1892
   4.1. Fluxomics analysis ....................................................................................... 1893
   4.2. Multiomics analysis .................................................................................... 1894
   4.3. Generation of constraint-based models and fluxomics ...................... 1895
   4.4. Other applications ..................................................................................... 1896
5. Perspective: Machine learning and plant metabolic modelling ...................... 1896
6. Conclusion .............................................................................................................. 1897

Funding ...................................................................................................................... 1897
CRediT authorship contribution statement .......................................................... 1897
Declaration of Competing Interest ........................................................................ 1897
References ............................................................................................................... 1897

\* Corresponding authors at: Centre of Biological Engineering, University of Minho, Campus of Gualtar, 4710-057 Braga, Portugal.
E-mail addresses: msampaio@ceb.uminho.pt (M. Sampaio), odias@ceb.uminho.pt (O. Dias).

https://doi.org/10.1016/j.csbj.2022.04.016
2001-0370/© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

Plants are multicellular eukaryotic photosynthetic organisms indispensable for human life. They are the ultimate food source for almost all animals, including humans (legumes, fruits, cereals, among others), maintain the atmosphere balance by consuming carbon dioxide and releasing oxygen, and provide many materials for human use such as wood, fibres for clothing, drugs, pesticides, oils, and fuels [1]. Plants are sessile organisms, unable to escape from environmental stresses or pathogens. Consequently, plants face a wide range of adverse environmental conditions and interact with several pathogenic or beneficial organisms. As a result, plants have the most complex metabolic networks that produce an enormous diversity of metabolites to help them grow, adapt to the environment, and defend against pathogens [2]. Since plants’ growth and survival are intrinsically linked to metabolism, its study is essential for understanding the mechanisms of fruit production and metabolic responses to different environmental stresses.

Metabolism has been studied by Systems Biology approaches, like Constraint-based Modelling (CBM), which use computational and mathematical models to analyze biological systems as a whole, modelling the inner components and their respective interactions [3]. The rise of next-generation technologies enabled the sequencing of complete genomes and later the reconstruction of Genome-Scale Metabolic Models (GSMMs), which are in silico metabolic flux models derived from genome annotation, representing all metabolic reactions taking place within an organism. These models allow performing in silico simulations of metabolic phenotypes under different environmental or genetic conditions [3]. Although GSMMs have been reconstructed mainly for unicellular organisms, several models are available for plants [4]. These models have a wide range of applications, such as understanding photosynthesis and analysing metabolic behaviour under different conditions. In addition to providing a better understanding of cellular phenotypes, GSMMs can also help design new strategies to improve the production of relevant metabolites. Currently, the reconstruction of plant GSMMs is still very challenging and time-consuming due to the large diversity of metabolites and extensive compartmentalisation of plant cells [5,6].

Recently, vast amounts of omics data have been generated from high-throughput technologies, leading to the development of several methods for integrating context-specific omics data as constraints in metabolic models, which are especially valuable for complex organisms like plants [7–12]. Omics data have been widely used in molecular biology to understand the underlying mechanisms leading to an organism’s phenotype, bridging the gap between genotype and phenotype.

Although genome-scale metabolic modelling has been crucial for organising and analysing omics data, integrating different omics (genomics, transcriptomics, proteomics, metabolomics) is hitherto an inefficient task [13]. Omics datasets are large, complex, and heterogeneous; hence, Machine Learning (ML) has been extensively used to process, analyse, and integrate different types of omics and extract biological knowledge from data [14]. CBM and ML have been mainly used independently in molecular biology, but integrating these approaches has improved predictions’ accuracy and increased the interpretability of the results. Recently, several reviews of CBM–ML hybrid studies have been published, suggesting the growth potential of this area [13,15–19].

In this article, we review the state of the art in plant metabolic modelling and the recent studies integrating CBM and ML. First, we introduce the data resources used to reconstruct and improve GSMMs and describe existing plant GSMMs and their application in the study of plant phenotypes, highlighting these models’ major advances and limitations. Then, we describe the main studies combining ML and CBM approaches, including their strengths and conclusions to elucidate how these studies can be applied or adapted to tackle the unique features of plant metabolism. We address this subject with a different perspective from existing reviews [13,15–19], focusing on the systematic application of ML to solve unique problems of plant metabolic modelling, and therefore conclude our review by underlining the main challenges and benefits of combining these approaches.

2. Plant metabolic modelling

During the reconstruction of GSMMs, different biochemical databases allow obtaining up-to-date information on the organism, which support the development and refinement of the metabolic network, namely genome annotations, biochemical data of metabolic reactions, and functional information on enzymes [20]. Table 1 describes the most important databases containing plant metabolic data. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [21] and MetaCyc [22] are the most used generic databases for the analysis of metabolic pathways. The National Center for Biotechnology Information (NCBI) [23], Universal Protein Resource (UniProt) [24], BRaunschweig Enzyme Database (BRENDA) [25], Transporter Classification Database (TCDB) [26] and PubChem [27] are generic databases used for extracting detailed information on genomes, proteins, enzymes, transporters, and chemical compounds, respectively. PlantCyc [28], Plant Reactome [29] and MetaCrop [30] are databases with metabolic data for several plants species, whereas SolCyc [31] only includes information for the Solanaceae family and The Arabidopsis Information Resource (TAIR) [32] is specific for A. thaliana. Species-specific plant databases have been created from MetaCyc and are available at the Plant Metabolic Network (PMN) resource [28].

The assembled metabolic network is then converted to a mathematical representation, involving the formulation of the biomass equation and definition of organism-specific constraints. Therefore, the model consists of a set of ordinary differential equations, representing the changes in metabolites’ concentrations over time. Usually, a pseudo-steady state assumption is applied to simplify the model to linear equations, assuming that the metabolite’s concentration is constant throughout time. Equation (1) represents this steady-state’s mass balancing, where S is the stoichiometric matrix and v is the flux vector. In S, rows represent metabolites and columns represent reactions. $S_i$ is the stoichiometric coefficient of metabolite i in reaction j [20].

$$S \cdot v = 0$$  \hspace{1cm} (1)

After reconstruction, the GSMMs can be simulated with constraint-based approaches, like Flux Balance Analysis (FBA) [34], to predict the metabolic phenotypes of an organism under different conditions. These methods require the definition of a relevant objective function, representing the metabolic goal of the organism, which can be defined as the maximisation or minimisation of a metabolic flux during the simulation, usually biomass maximisation. Another constraint-based method is Flux Variability Analysis (FVA), which calculates each reaction’s minimum and maximum flux for a defined set of constraints [35]. The FBA approach was extended to Dynamic Flux Balance Analysis (dFBA) [36], which assumes that intracellular metabolites are at steady state, but exchange metabolites and total biomass are constrained with dynamic equations, representing the rates of uptake or excretion.

Other methods have been developed to improve flux predictions through the integration of context-specific omics data, mainly transcriptomics, within metabolic models [7–12]. Omics
Table 1
Description of the most relevant databases of plant metabolic data.

| Database         | Ref. | Description                                                                 | Data                        |
|------------------|------|------------------------------------------------------------------------------|-----------------------------|
| KEGG             | [21] | Generic database resource that comprises genomes, metabolic pathways, chemical compounds, diseases, and drugs. | Metabolic data              |
| Metacyc          | [22] | Comprehensive database of extensively curated metabolic pathways, containing information on reactions, enzymes, genes, and compounds for several organisms. | Curated metabolic data      |
| BioCyc           | [22] | Collection of organism-specific pathway genome databases (PGDBs), each containing the complete genome and predicted metabolic network of an organism. | Organism-specific predicted metabolic data |
| NCBI             | [23] | Online repository containing several databases for genomics and biomedical information and tools for extracting and analysing the data. | Reviewed and unreviewed sequence data |
| UNIPROT          | [24] | Resource for protein sequence and related information, including manually reviewed data (Swiss-Prot) and automatic, non-reviewed protein annotations (TrEMBL). | Curated and predicted protein data |
| BRENDA           | [25] | Main database of manually annotated enzyme functional data, which uses the Enzyme Commission (EC) classification system. | Curated enzyme data          |
| Tcdb             | [26] | Curated database containing information on transport systems from several organisms, including sequence, structure, and function, and uses the Transport Classification (TC) system to classify transport proteins. | Curated transport data      |
| PubChem          | [27] | The largest database of chemical information, including molecular structure, physical properties, and biological activities of compounds. | Unreviewed chemical data    |
| PlantCyc and PMN | [28] | PlantCyc contains more than 1000 curated metabolic pathways, for at least 350 plant species. This database is the centre of PMN, a resource of plant metabolic databases, and is used as a reference to create plant-specific PGDBs. The current version of PMN (15.0) comprises 126 plant-specific metabolic databases, including curated and predicted databases. | Curated and predicted plant metabolic data |
| PlantReactome    | [29] | Manually curated and comparative pathway database for plants, being part of the Gramene, which is a resource for comparative functional genomics [33]. Plant Reactome used O. sativa as a reference species to manually curate metabolic and regulatory pathway data for 97 plant species, also providing a suite of tools for the analysis of large-scale omics datasets. | Curated plant metabolic data |
| MetaCrop         | [30] | Repository of detailed and manually curated metabolic information for six major crop plants with agronomic importance. It allows to export the data automatically for the creation of metabolic models. | Curated plant metabolic data |
| SolCyc           | [31] | Collection of PGDBs for Solanaceae species, including databases for S. lycopersicum (tomato), Solanum tuberosum (potato), Nicotiana tabacum (tobacco), Capsicum annum (pepper), and Petunia × hybrida (petunia). | Curated metabolic data of Solanaceae species |
| TAIR             | [32] | Database of genetic and molecular data for A. thaliana, including genome sequence and gene structures, products, and expression datasets as well as tools for data visualisation and analysis. | Curated genetic and metabolic data of A. thaliana |

Table 2
Description of the most relevant databases of plant omics data.

| Database         | Ref. | Description                                                                 | Data                        |
|------------------|------|------------------------------------------------------------------------------|-----------------------------|
| SRA              | [29] | Archive for next-generation raw sequence data.                              | Sequences                   |
| GenBank          | [40] | Comprehensive collection of all publicly available DNA sequences and respective annotations. | Sequences                   |
| RefSeq           | [41] | A comprehensive, curated, and non-redundant collection of sequences, including genomes, transcripts, and proteins. | Sequences                   |
| Nucleotide       | [42] | A collection of sequences from different sources including GenBank and RefSeq. | Sequences                   |
| GEO              | [45] | Repository of functional genomics data, including raw and processed data with descriptive metadata. | Genomics and transcriptomics |
| DDBJ             | [43] | Public database of nucleotide sequences at National Institute of Genetics.   | Sequences                   |
| ENA              | [44] | A comprehensive nucleotide sequence resource, including raw sequencing data, assembly information and functional annotations. | Sequences                   |
| ArrayExpress     | [46] | Database of functional genomics data and respective metadata.               | Genomics and transcriptomics |
| Expression Atlas | [47] | A resource for gene and protein expression data for multiple organisms and across different biological conditions. | Transcriptomics              |
| RODC             | [48] | Database of mRNA-sequencing expression data for plants.                     | Transscriptomics             |
| PlantExpress     | [49] | Database of gene expression data from microarrays for O. sativa and A. thaliana. | Transscriptomics             |
| ProteomicsDB     | [50] | Database for quantitative Mass Spectrometry (MS)-based proteomics data. Currently, it also includes RNA-Seq expression datasets, drug-target interactions, and protein turnover data. | Proteomics                  |
| PRIDE            | [51] | Repository of MS-based proteomics data, including protein identification and quantification, post-translational modifications, analysed mass spectra and technical metadata. | Proteomics                  |
| PeptideAtlas     | [52] | Database of peptides identified in MS proteomics experiments. It provides tools for processing and analysing raw MS output data. | Proteomics                  |
| GPMDM            | [53] | Database for analysis, validation, and storage of MS proteomics data.        | Proteomics                  |
| Massive          | [54] | Community resource for raw MS data, including proteomics datasets.           | Proteomics                  |
| PPDB             | [55] | Database for integrating MS-based proteomics data of Z. mays and A. thaliana. | Proteomics                  |
| Metabolights     | [56] | Repository for metabolomics data and associated metadata, covering metabolite structures, reference spectra, concentrations, and functions. | Metabolomics                |
| MetabolomeExpress| [57] | Online server for processing, interpreting, and storing MS metabolomics data | Metabolomics                |
| MetabolomicsWorkbench | [58] | Repository for metabolomics data and associated metadata from MS and nuclear magnetic resonance studies. | Metabolomics                |
| GDM              | [59] | Collection of reference mass spectra and retention times for metabolites.    | Metabolomics                |
data have allowed the study of the central dogma by detecting and quantifying genes, transcripts, proteins, and metabolites in biological samples. Hence, omics data offer insights into the metabolism, allowing the detection and analysis of differential expression patterns across varied environmental conditions [37,38].

The most popular databases of omics data are presented in Table 2. The main databases of sequence data and annotations include Sequence Read Archive (SRA) [39], GenBank [40], Reference Sequence Database (RefSeq) [41], and Nucleotide [42] from NCBI, DNA DataBank of Japan (DDBJ) [43] and European Nucleotide Archive (ENA) [44]. Gene Expression Omnibus (GEO) [45] and ArrayExpress [46] contain functional genomics data and respective metadata and Expression Atlas database [47] holds gene expression data. Other databases, such as the Plant Omics Data Center (PODC) [48] and Plant Express [49] only contain transcriptomics data for plants. Proteomics data can be retrieved from sources like ProteomicsDB [50], PRoteomics IDentifications (PRIDE) [51], PeptideAtlas [52], Global Proteome Machine Database (GPMDB) [53], Mass Spectrometry Interactive Virtual Environment (MassIVE) [54] and Plant Proteomics Database (PPDB) [55]. Metabolomics data can be found at MetaboLights [56], MetabolomeExpress [57], Metabolomics WorkBench [58] and Colim Metabolome Database (GDM) [59].

The integration of omics in metabolic models is especially important in higher organisms, like plants and mammals, as they are complex organisms composed of different cells and tissues. Therefore, generic models may lead to wrong interpretations, as certain reactions or pathways are only active in specific tissues or conditions. This is even more challenging in the case of non-model organisms, whose metabolism is poorly characterised.

Additionally, the metabolic behaviour of higher organisms involves interactions between multiple cells or tissues. Hence, multi-tissue models have been reconstructed to understand such complex behaviour [60–68]. A multi-tissue model is usually composed of several copies of a GSMM, connected by inter-tissue exchange reactions. Moreover, tissue-specific omics can define the constraints for each tissue model to improve the flux predictions [69,70].

In 2009, Poolman published the first plant GSMM for Arabidopsis thaliana [71]. Several models have been developed since, not just for model plants like A. thaliana, but also for more complex plants [72], such as Zea mays (maize) [73–75] and Oryza sativa (rice) [76–79], Fig. 1 summarises the plant GSMMs published to date. Generally, these models have proven to be robust and accurately predict specific aspects of central carbon metabolism [72]. The existing plant GSMMs are described below, grouped by organism, and ordered by publication date.

### 2.1 Arabidopsis thaliana

Poolman et al. [71] reconstructed the first plant GSMM for A. thaliana heterotrophic cell suspension culture. This model was mainly derived from the AraCyc database (version 4.5) [80] and produces biomass components in the proportion observed experimentally in heterotrophic suspension cultures. In 2013, Cheung et al. [81] extended this model to include the subcellular localisation of central metabolic reactions across five compartments (cytosol, plastid, mitochondrion, peroxisome, and vacuole) and to account for growth, transport, and cell maintenance energy costs, including ATP and reductive costs. They simulated the model under different environmental conditions and discovered that accounting for energy costs of transport and maintenance substantially improves flux predictions, regardless of the objective function used in the simulation.

AraGEM [82] was the first plant GSMM to represent the metabolism of a compartmentalised photosynthetic cell (same five compartments as in Cheung’s model [81]), describing photosynthesis, photorespiration and respiration while identifying metabolic changes between them. This model was updated by Saha et al. [73] and later by Chung et al. [83] to include terpenoid biosynthesis reactions. Recently, Siriwach et al. [84] have combined the AraGEM model with time-series gene expression data, creating condition-specific models of A. thaliana under drought and control conditions to gain insights for the development of tolerant plants.

Mintz-Oron et al. [85] have reconstructed the fully compartmentalised GSMM for A. thaliana, which encompasses the subcellular localisation of all reactions, across the five compartments of the AraGEM model, plus the Golgi Complex and Endoplasmatic Reticulum. They extracted ten tissue-specific models from this generic GSMM by integrating protein expression data of eight tissues and cell cultures in light and dark conditions. The authors then used the seed-specific model to predict the genetic knockouts that result in vitamin E overproduction. Töpfer et al. [86,87] have combined this generic model with time-resolved transcriptomics data from different temperature and light conditions to understand the metabolic acclimation of A. thaliana to stressful environments.

An evidence-based model for A. thaliana was reconstructed by Seaver et al. [88] from the generic model available on PlantSEED [89], including the seven compartments mentioned plus the nucleous and the cell wall, and combined with transcriptomics and metabolomics data to extract specific models for eight root tissues at different developmental stages [63]. The tissue-specific models were used to build a multi-tissue model of the root for analysing the flux distribution of hormones indole-3-acetate and trans-Zeatin through the root.

A more complex model of A. thaliana was developed to represent the leaf metabolism over a day-night cycle [90]. This diel model was reconstructed by duplicating the previous model [81] into two modules, day, and night, and manually adding the trans-
porters between these two phases. The authors simulated both phases in a single optimisation problem by applying specific constraints specifying that photon influx is allowed in the day (photoautotrophic metabolism) and is set to zero at night (heterotrophic metabolism). This model simulates and clarifies the interactions between the two phases by allowing storage metabolites synthesised during the day to be used at night and vice-versa.

More recently, multi-tissue models for *A. thaliana* were reconstructed to represent different tissues and their interactions. Dal’Molin et al. [61] have developed a framework to create a multi-tissue model comprising leaf, stem, and root of *A. thaliana* across the diurnal cycle. In this approach, the tissues exchange metabolites through a shared compartment (common pool) rather than directly transported between two tissues, which can reduce redundancy when more than two tissues are interconnected. Additionally, a storage pool manages storage and retrieval of metabolites. Therefore, in this framework, a multi-tissue model is defined by a stoichiometric matrix representing the internal reactions and three matrices for the exchange reactions with the environment, the transport reactions through the common pool and the accumulation of metabolites in the storage pool. The multi-tissue model consisted of three replicates of the AraGEM model (representing root, stem, and leaf) and two common pools, one for exchanges between leaf and stem and another for exchanges between stem and root. To simulate the diurnal cycle, the multi-tissue model was duplicated to represent each state (light and dark) and a storage pool was created, with starch being the only stored metabolite. The model was used to study carbon and nitrogen translocation between tissues.

Following this strategy, the diel model was used to build a dynamic multi-tissue diel model (Fig. 2) to study metabolic changes across multiple growth stages under different nutrient availability [62]. All reactions of the diel GSMM were replicated to represent the leaf and root model, and the transport between root and leaf was performed through a common pool representing the phloem. This multi-tissue model was simulated by dFBA [36] to explore carbon and nitrogen partitioning between root and leaf over different developmental stages.

A different approach has followed by Schroeder et al. [65] to study the evolution of metabolism across the lifecycle of *A. thaliana*. While previous studies have only considered metabolism at a single point (growth or a single diurnal cycle), this optimisation framework takes a series of “snapshots” of core-carbon metabolism. These snapshots comprise the plant mass, growth rate, and fluxes at one-hour intervals across 61 days of growth, including the stages of seed germination, leaf development, flower production and silique ripening. In this study, a core multi-tissue metabolic model (referred to as *p-ath780*) comprising leaf, root, seed, and stem tissues was reconstructed. The core model only includes the central metabolic pathways of *A. thaliana*. The tissue-specific models were built based on the available literature and experiemen-
2.2. Zea mays

Dal’Molin et al. [92] made the first efforts towards reconstructing a Z. mays GSMM, combining biochemical information of Sorghum bicolor (sorghum), Z. mays and Saccharum officinarum to build the C4GEM model. This model represents the two leaf tissues, mesophyll (M) and bundle sheath (BS) cells, where photosynthesis of C4 plants takes place, and the interactions between them. It also includes the main five compartments (cytosol, plastid, mitochondrion, peroxisome, and vacuole).

Since then, three more GSMMs were reconstructed for Z. mays leaf. The first model, referred to as iRS1563 [73], was based on Arabidopsis (with the same five compartments) and Z. mays genome and was used to predict metabolic phenotypes for two natural brown midribs’ (bm) mutants with defective lignin biosynthesis.

Another Z. mays leaf model was reconstructed by Simons et al. [74], with a significant increase in the number of genes and reactions, representing secondary metabolism. It comprises the two tissues of leaf, M and BS cells, and gene expression data was used to identify the active reactions in each tissue. Regarding the compartments, it includes the five of C4GEM plus the plasmatic, thylakoid, and inner mitochondrial membranes. This model was used to assess the assimilation of nitrogen within the leaf under different nitrogen conditions and later was constrained by incorporating enzyme activity data to detect metabolic differences between nineteen Z. mays lines [93].

Seaver et al. [88] have also reconstructed an evidence-based model for Z. mays, with the same compartments as the A. thaliana model [88], and it was used to extract tissue-specific models for leaf, embryo, and endosperm of Z. mays by integrating gene expression data within the model.

The most recent Z. mays leaf GSMM (iRB5204) was developed by Bogart et al. [75], based on the CornCyc database (version 4.0) [94] and previous models. They reconstructed a high-confidence model named iRB2140, including curated reactions only and the same compartments of Simons’ model [74], except for the plasmatic membrane. Then, the authors created a two-tissue model to represent M and BS cells of the leaf by duplicating the iRB2140 model and adding transport reactions between the two tissues (Fig. 3). They incorporated known nonlinear kinetic constraints and transcriptomics data from more and less differentiated cells to reconstruct a whole-leaf model identified as iEB2140x2x15, representing 15 developmental stages of the maize leaf.

2.3. Oryza sativa

The first GSMM was reconstructed from the RiceCyc database [95] by Poolman et al. [76] and included three compartments: cytosol, chloroplast and mitochondrion. It was analysed to identify metabolic responses to different light intensities. This model was curated and extended by Chatterjee et al. [77] to encompass the peroxisome compartment and reactions involved in chlorophyll synthesis.

Another O. sativa leaf model named iOS2164 was developed by Lakshmanan et al. [78] by adding the vacuole, the endoplasmic reticulum and the thylakoid as compartments, and all electron-transport reactions. The authors have integrated transcriptomics data within this model to evaluate the metabolic responses to different light conditions. Later, this model was combined with gene expression data of different tissues at different developmental stages to generate tissue-specific models and highlight the metabolic differences between tissues [96].

All these rice models describing the metabolism of O. sativa japonica were reviewed in [97]. Chatterjee et al. [79] reconstructed a GSMM for O. sativa indica, which included cytosol, mitochondrion, peroxisome and chloroplast compartments, and used this model to characterise the metabolic responses to variations in Rubisco activity and light intensity and under different enzymatic costs constraints.

2.4. Other organisms

Other organisms, like Solanum lycopersicum (tomato), Solanum tuberosum (potato), Medicago truncatula (barrelclover), Glycine max (soybean), Setaria viridis (green foxtail) and Quercus suber (Cork oak), only have one GSMM available. For Hordeum vulgare (barley), a dynamic multi-tissue model was created, and stoichiometric multi-tissue models were also created from the GSMMs of M. truncatula, G. max, S. viridis and Q. suber.

- Hordeum vulgare. A framework for analysing metabolic dynamics of H. vulgare on a whole-plant scale was developed by integrating a steady-state multi-organ model with dynamic constraints from a functional plant model [60]. Organ-specific

![Fig. 3. Schematic overview of the two-tissue model of Z. mays, representing the M and BS leaf cells of C4 plants [75]. Each cell includes 4 compartments: cytoplasm, chloroplast, peroxisome and mitochondrion, and small molecules are directly exchanged between the two tissues by transport reactions. The M cells exchange carbon dioxide and oxygen with the intercellular air space while BS cells exchange sucrose, glutathione and glycine with phloem and import water and inorganic nutrients from xylem. This type of model representation is used to understand the photosynthesis in C4 plants, mainly the interactions between M and BS cells.](image-url)
models for leaf, stem and seed were reconstructed by collecting primary metabolism data from literature and databases and combined into one multi-organ model. Next, dFBA was applied and exchange fluxes predicted by the functional plant model were used to constrain FBA at each time interval. This framework allowed studying metabolic interactions between source and sink organs of *H. vulgare*, accounting for temporal and environmental changes.

- **Solanum lycopersicum.** The only model of *S. lycopersicum*, referred to as iHY3410, represents the leaf and enables the analysis of metabolic flux distributions on photorespiration pathways under drought stress [98].

- **Solanum tuberosum.** Botero et al. [99] reconstructed a GSMM of *S. tuberosum* late blight to study the effect of this disease on the leaf metabolism, suggesting the suppression of photosynthesis. This model encompases the metabolic pathways of the leaf and the interaction between the plant and *Phytophthora infestans* through the integration of gene expression data of infected *S. tuberosum*.

- **Medicago truncatula.** A fully compartmentalised model for *M. truncatula* was developed by Pfau et al. [64] and allowed the analysis of their rhizobial symbiosis for nitrogen fixation by connecting the plant model to a model of its symbiont and evaluating the effects of the symbiosis in plant growth. Then, a multi-tissue model representing the root and shoot of *M. truncatula* was reconstructed by integrating tissue-specific gene expression data and connecting the resulting root- and shoot-specific models with a combined biomass reaction and inter-tissue transporters derived from literature.

- **Glycine max.** Moreira et al. [66] reconstructed a GSMM of *G. max* and duplicated this model to create a multi-tissue model representing two tissues of *G. max* seedlings: the cotyledons and hypocotyl/root axis (HRA). The multi-tissue model was constrained with the biomass compositions observed experimentally over four days of seedling growth to simulate the mobilisation of seed reserves during this period and detect metabolic differences between the two tissues, as well as inter-actions between them.

- **Setaria viridis.** Similarly, a model of *S. viridis* [67] was reconstructed and used to create a multi-tissue model representing the C4 leaf (including both M and BS cell types) and stem. These models have identified implications of proton balancing on flux distributions during photosynthesis of C4 leaves and reactions involved in the biosynthesis of cellulose, hemicellulose, and lignin in the stem.

- **Quercus suber.** Recently, a reconciled GSMM for *Q. suber* was semi-automatically reconstructed by Cunha et al. [68] using merlin [100] and performing extensive manual curation. merlin is a user-friendly framework developed for reconstructing draft GSMMs automatically and assisting manual curation efforts in these tasks. This is the first model reconstructed for a woody tree. Transcriptomics data was integrated with the model to obtain tissue-specific models for the leaf, inner bark and phellogen, which were merged into a diel multi-tissue model to predict interactions among tissues at light and dark phases and study the synthesis of suberin monomers. In the future, this model can be extended and used to explore the metabolic patterns associated with high-quality cork, which is economically relevant for Portugal.

Overall, the most significant advances in plant metabolic modelling were made first for *A. thaliana*. AraGEM [82] is the most used plant model and was the first to allow the simulation of photosynthesis and photorespiration metabolic processes, including compartmentalization. Next, a relevant advance was the diel model of Cheung et al. [90], which allows simulating the leaf metabolism over the diurnal cycle in a single problem. Then, the creation of the multi-tissue model by Dal’Molin et al. [61] represented a significant improvement of these models. This model allows to analyse the metabolism across different tissues (leaf, stem, and root) and also the different tissues’ metabolic interactions. Although the multi-tissue model of Shaw et al. [62] only comprises two tissues, leaf and root, its novelty was to include dynamic constraints for the exchange metabolites.

Finally, it is important to highlight the integration of omics into models to create tissue- or condition-specific models to originate more realistic flux predictions. Although, transcriptomics data are usually used to reconstruct specific models [63,84,86,87], Mintz-Oron et al. have constrained the models with proteomics [85]. Regarding *Z. mays*, the models were helpful for studying the photosynthesis of C4 plants, which occurs between the two leaf tissues, M and BS. Of these models, the one from Simons et al. [74] stands out as it includes more secondary metabolism reactions, more compartments, and constraints based on gene expression. Likewise, iOS2164 [78] is the most complete model for *O. sativa*, as it contains more compartments and transcriptomics-based constraints. Meanwhile, the advances made for *A. thaliana* were applied in the reconstruction of complex multi-tissue models for other organisms, such as *M. truncatula* [64], *G. max* [66], *S. viridis* [67], and a woody tree, *Q. suber* [68].

### 3. Major challenges and limitations

Although several plant GSMMs and studies that successfully use them to understand plant metabolic processes are available, the existing approaches still have limitations as plant metabolic modelling is very challenging [72]. Annotation of plant genomes is incomplete, and database information on plant enzymatic reactions and metabolites is limited, especially for secondary metabolism, resulting in an inaccurate model with network gaps, requiring extensive and time-consuming validation. Most plant metabolic models have been validated to predict changes in plant central carbon metabolism, though generally neglecting secondary metabolism. Therefore, these models cannot correctly predict plant adaptation to the environment and interactions with pathogens [6]. An exception is the *Z. mays* model [74], which presents extensive coverage of the secondary metabolism.

Another challenge in plant modelling is to place reactions in the correct compartment. Plant cells are composed of multiple compartments, and little is known about the subcellular localisation of reactions and metabolites. Most enzymes of the central metabolism are known to be present in more than one compartment, which makes the modelling process even more difficult. Adding compartments to models raises other problems, such as the lack of information about transport reactions, substrate specificity, and energetic costs [5]. The assignment of reactions to compartments in plant models is typically performed by searching databases and using subcellular localisation prediction tools.

As plants are exposed and adapted to several environmental stresses, their cellular objectives are surely much more complex than maximising cell growth. For instance, during environmental changes or pathogen interactions, the fluxes are redirected from the primary metabolism to secondary metabolic pathways to produce the metabolites for the plant’s adaptation and defence [6,72]. The most used CBM’s objective functions working over plant models are minimising the total flux, minimising the photon uptake, and maximising biomass. Although these objective functions have been successfully applied for simulating the metabolism of plant tissues at specific developmental stages or under certain environmental conditions, they do not apply to all possible scenarios [72]. Therefore, defining an appropriate objective function in plant models remains exceptionally challenging.
Another challenge in plant modelling is the definition of constraints affecting plants. Most plant models use the biomass synthesis at a defined growth rate as constraint when the objective function is the minimisation of total reaction fluxes [64,66,71,79,90,98]. However, this may not be enough to correctly predict the fluxes, as net biomass synthesis uses only a fraction of the cell’s total energy [72]. Indeed, Cheung et al. [81] proved that accounting for transport and maintenance energy costs increases phenotype predictions’ accuracy, showing that the definition of appropriate constraints is essential for obtaining realistic metabolic predictions. Moreover, plants’ photosynthesis, photospiration and respiration add complexity to their metabolic networks and complicate the modelling process [6]. Other factors affecting plant metabolism include complex interactions with symbionts and pathogens, competition mechanisms, and changes in available nutrients.

Finally, one of the main problems of plant modelling is that most models are generic and include all reactions known to take place in that plant, regardless of cell type or environmental conditions. As plants contain a wide variety of cell types, each with its specific active metabolism, generic models may lead to wrong interpretations as certain reactions or pathways are inactive in a specific cell type, even though being strongly active in others. Similarly, environmental conditions may influence the expression of metabolic genes; thus, enzymes may be active in specific conditions while inactive in others. Therefore, the reconstruction of context-specific models is crucial to obtain more realistic metabolic flux predictions. Several studies have integrated transcriptomics data with generic plant GSMMs to improve flux predictions and create plant tissue- and condition-specific models [78,84–88,93,96]. However, as gene expression levels generally do not strongly correlate with reaction fluxes [8], the use of omics to improve the GSMMs is still challenging and inaccurate.

Tissue-specific models can be merged to form a multi-tissue model that simulates metabolic interactions between tissues [69,70]. However, reconstructing multi-tissue models raises the challenge of defining the metabolites transported between tissues, highlighting the lack of information regarding this topic. As depicted above, there are already several multi-tissue models of plants [60–68]. In most models, the different tissues replicate the original model, with few metabolic differences, and are connected by inter-tissue reactions or a common compartment. Light availability is a common constraint to differentiate context-specific models, for instance, leaves and roots, or diurnal and nocturnal leaves. The huge advantage of these multi-tissue models is that they simulate metabolic interactions between different tissues and organs, providing insights into complex resource allocation processes occurring in plants [101].

Despite the several challenges of plant metabolic modelling, significant advances have been made in the last years, which have allowed the reconstruction of more accurate plant generic, context-specific and multi-tissue metabolic models. These have been successfully applied for simulating photosynthetic and heterotrophic metabolism, improving the production of metabolites of interest, and understanding metabolic phenotypes under different environmental conditions or at different developmental stages. Therefore, metabolic modelling approaches have proven to be a relevant tool for understanding plant metabolism, and through the integration of omics data, the fluxes predicted by these models became more accurate and adjusted to environmental conditions.

An improvement to the current studies with plant GSMMs would be integrating more than one type of omics data to increase the accuracy of the simulations and contribute to a better understanding of complex biological processes across the whole plant. However, the challenge remains on how to integrate multiple heterogeneous data into predictive multi-scale models [15].

4. Machine learning and constraint-based modelling

In the last decades, the development of high-throughput technologies has led to the generation of large amounts of omics data that are complex and heterogeneous, making their analysis and the extraction of knowledge very challenging. Hence, the processing and interpretation of omics data require the use of appropriate tools, such as ML algorithms, which can identify patterns, select relevant features, and make inferences from the observed data without defining biological assumptions [15,16].

ML has been defined as the study of algorithms that can automatically learn and improve by experience and adapt to new data input without being explicitly programmed [102]. ML algorithms have been applied in interpreting large metabolic datasets and developing tools to study cellular metabolism [13,103,104].

An important distinction in ML is between “supervised” and “unsupervised” learning methods. In supervised learning, the model learns from a training dataset with both inputs and desired outputs, so it can later make predictions on the output of unseen...
observations. In contrast, unsupervised models are trained with unlabelled data to identify the underlying structure, patterns, or data distribution [102]. Principal Component Analysis (PCA) and clustering are the most used unsupervised methods for dimensionality reduction and identifying sub-groups in the datasets. Some of the most used supervised and unsupervised algorithms are described in Table 3.

Despite the many benefits of applying ML methods to omics data, this task is still challenging. Omics datasets are scattered and noisy, with missing values and technical errors, making it difficult for the model to differentiate between true data patterns and error profiles. In addition, as the process for data acquisition is intricate and expensive, omics datasets usually have few samples and show class imbalance, where the class representing the control group generally has more instances than the other. Together with the high number of features, which is characteristic of omics datasets, these issues lead to the development of complex, overfitted ML models and poor generalisation. Furthermore, as omics data are very heterogenous and have many applications, there is no ML algorithm or pipeline suitable for all problems. Hence, choosing the best ML algorithm, model parameters, and feature selection method requires deep knowledge of ML methods and the area of application. Also, this knowledge is essential for the proper interpretation of results generated by the models, which can be difficult to analyse. Therefore, sharing large-scale high-quality omics datasets is crucial for developing good predictive models [105–107].

Recently, the first studies combining ML and CBM approaches have emerged and were previously reviewed in [13,15–19]. The integration of ML and CBM comprises three main cases: fluxomics analysis, multi-omics analysis and generation of constraint-based models and fluxomics (Fig. 4). In fluxomics analysis, the flux distribution predicted by CBM methods is analysed by ML methods. In multi-omics analysis, omics data can be integrated within metabolic models to generate context-specific fluxomics data, which ML can analyse in combination with omics data from high-throughput technologies. In alternative, ML can be trained with omics datasets to produce or improve metabolic models or fluxomics data.

4.1. Fluxomics analysis

In the following examples, fluxomics data is generated by CBM methods and analysed by ML.

Although PCA has been widely applied to simplify and identify patterns in metabolic data, its results are complex and difficult to interpret biologically. Therefore, two approaches, named Principal Elementary mode Analysis (PEMA) [108] and Principal Metabolic Flux Mode Analysis (PFMA) [109], have combined PCA and CBM to identify the flux modes that explained most flux variance and have minimum deviations from a steady-state condition. The PEMA approach was extended to dynamic conditions using supervised ML (dynEMR-DA) in a subsequent work [110]. Here, dynamic EFMs were defined as partially activated EFMs at each time point of the simulation. A small kinetic model of _S. cerevisiae_ was simulated under different conditions. The non-steady-state flux distributions were decomposed into a set of dynamic EFMs, which were examined by discriminant analysis to identify the pathways that best differentiated between conditions.

Hierarchical clustering was used by Magnusdottir et al. [111] to predict ecological interactions between human gut bacteria. The models were simulated alone and paired with every other model to represent co-growth under different fibre diets and oxygen conditions. The relative fitness was calculated for each pair of organisms and used to define the type of interaction. Lastly, the ratio of pairwise interaction types was clustered per condition and per taxonomy, which has resulted in three main clusters comprising microbes with different carbohydrate fermentation capabilities.
suggesting that this capability may define the types of interactions between microbes.

Interactions of human gut bacteria were also predicted by DiMucci et al. [112] but using supervised ML and dFBA. Single and pairwise simulations were performed using dFBA, and the relative final biomass was used to classify the interactions as negative or nonnegative. A random forest (RF) classifier was trained with vectors representing the presence or absence of exchange reactions in each organism to predicted potential interactions between two microbes and identify relevant fluxes for the prediction.

Another example is the study of Shaked et al. [113] that has used ensemble learning to predict drug side effects. A human GSMM was used to calculate the flux bounds of the reactions through FVA after knocking out the genes that represent drug targets. These reaction bounds were used as features to an ensemble of Support Vector Machines (SVMs), where each model represented a side effect, resulting in a list of all potential side effects of the metabolically acting drug.

Other studies have trained ML models with fluxomics data obtained by CBM methods to predict microbial production rate under different bioprocess settings [114,115]. Recently, this strategy was used to predict amino acids concentrations in a fed-batch Chinese hamster ovary (CHO) cell culture. The metabolic model was constrained with experimental measurements and predicted the initial amino acid consumption rates. Then, the flux predictions were refined and extended by linear regressions to a time-course profile [116].

Another study tried to clarify the glycosylation process by training Artificial Neural Networks (ANNs) with the fluxes of the reactions involved in nucleotide sugar donor synthesis, which were calculated by a stoichiometric model of CHO cells, to predict the glycan distribution of the antibodies produced [117].

4.2. Multiomics analysis

Multiomics analysis involves the integration of predicted fluxomics with experimental data using ML. For instance, this approach was used to predict essential genes [118] and reactions [119] and yielded better results than using only CBM methods.

Li et al. [120] have created condition-specific models by integrating gene expression data of cancer cell lines under different environments within a human GSMM. These models were simu-
lated through FBA, and a K-nearest neighbours (KNN) model used the resulting fluxes to predict new targets for cancer drugs.

Following a two-stage data integration strategy, Kim et al. [121] have developed a normalised, well-annotated multi-omics database for *E. coli* to provide high-quality data for data-driven predictive analysis. Hence, ML models were trained with experimental data, including transcriptomics, proteomics, metabolomics, and growth rates, and with fluxomics data obtained by condition-specific models, which resulted from the integration of proteomics and transcriptomics with a GSMM. This work was the first to integrate a comprehensive set of omics to train an ML model.

Recently, Culley et al. [122] have integrated transcriptomics and fluxomics data of different *S. cerevisiae* mutants to predict growth, using three different strategies: early, intermediate, and late integration. The strain-specific fluxomics data were obtained by simulating the models constrained with the transcriptomics data. Gene expression and fluxomics were analysed separately and as a single dataset by ML models and feature selection was applied to reduce dimensionality for the early integration. In the intermediate integration, two multi-view methods were applied: Bayesian Efficient Multiple-Kernel Learning (BEMKL), which creates and combines different kernel matrices for each dataset, and a multimodal artificial neural network (MMANN), that contains a layer for each dataset fused via additional layers. Finally, RFs trained with each dataset independently were combined in an ensemble model for the late integration. The authors concluded that adding fluxomics to gene expression make the results more accurate and biologically interpretable. This study was extended by Magazzù et al. [123] who showed that regularised linear models could outperform MMANN in multi-omics analyses and highlighted the relevance of using fluxomics for better understanding the interactions among genes and phenotypes.

Lewis et al. [124] have integrated predicted fluxomics with experimental omics to overcome the lack of metabolomics in cancer datasets and created a ML classifier to identify biomarkers for radiation resistance. Context-specific models were generated by integrating transcriptomics and mutation data from radiation-resistant and non-resistant tumours. The FBA-predicted fluxomics were integrated with experimental omics using the late integration approach: multiple classifiers were trained on an individual dataset and combined in a meta classifier that calculates the final probability of radiation resistance.

Regarding drug side effects, Guebila et al. [125] have integrated drug-induced gene expression data within a metabolic model of the small intestine epithelial cells to generate drug-specific fluxomics data. Gene expression and the predicted fluxomics were trained by a multilabel SVM to predict gastrointestinal drug effects and the drugs were clustered according to their metabolic and transcriptomic profiles which give new insights into the adverse reactions in the gut.

In another study, Vijayakumar et al. [126] proposed a pipeline that applies FBA and ML to improve phenotypic prediction in cyanobacteria. Condition-specific models were constrained by transcriptomics data and simulated using multi-objective FBA with three goals: biomass, photosystems I & II, and ATP maintenance reactions. Then, ML was trained with transcriptomics and predicted fluxomics and allowed the identification of key genes and reactions related to each condition. This pipeline clarified the mechanisms used by cyanobacteria to deal with variations in light intensity and salinity that could not be detected using transcriptomics alone.

Kavvas et al. [127] presented the Metabolic Allele Classifier (MAC), a metabolic model-based ML classifier that uses FBA to predict allele-specific antimicrobial resistance. MAC was formulated within the FBA structure but using allele-specific flux capacity constraints and antibiotic-specific objective functions. This method takes the genome sequence of a *Mycobacterium tuberculosis* strain as input and classifies the strain as either resistant or susceptible to a specific antibiotic by optimising the antibiotic-specific objective function. Thus, MAC uses linear programming that acts as an ML classifier, in which the predicted flux state corresponds to each class, elucidating the biochemical processes leading to antibiotic resistance.

Lastly, a deep-learning approach, entitled DeepMetabolism, was developed to predict *E. coli* phenotypes, using CBM, biological knowledge and gene expression data to define the neural network structure [128]. The first step is unsupervised pre-training and consists of an autoencoder with five layers: the first three represent gene expressions, protein abundances and phenotypes, respectively, while the fourth and fifth layers are decoders representing reconstructed protein and gene layers, respectively. The layers were connected using biological knowledge: gene-protein rules (GPRs) from the GSMM connected the first and second layers, and GSMM’s simulations identified essential reactions for each phenotype to connect the second and third layers. The second step consists of supervised training, using the same autoencoder model, though only with the first three layers trained to predict phenotypes.

### 4.3. Generation of constraint-based models and fluxomics

Instead of using ML to analyse fluxomics data, ML models can be trained with experimental data, and the results can be used to create models or improve flux predictions [129–132].

For instance, a web-based platform called Mflux was developed to predict the central bacterial metabolism fluxes using ML models [129]. The models, namely SVM, KNN and decision trees, were trained with experimental fluxomics data under different conditions to predict the flux of the central metabolism reactions and associate metabolic fluxes with the conditions. As the fluxes predicted by ML may not follow the stoichiometry of metabolic networks, quadratic programming was applied to adjust the predicted fluxes to satisfy the stoichiometric constraints.

ML can also be used to pre-process omics data and define additional constraints for CBM. For instance, Brunk et al. [130] presented a workflow that combines ML, metabolomics and a GSMM to characterise *E. coli* strain variation. PCA is applied to reduce the dimensionality of metabolomics data and identify key metabolites driving strain variation. Then, the pre-processed metabolomics data were used to adjust the flux bounds of the GSMM of *E. coli* to achieve a better characterisation of the flux network of each strain.

Another example is the unsteady-state flux balance analysis (uFBA) workflow for integrating time-course metabolomics to predict metabolic fluxes in dynamic conditions [131]. The first step is to discretise nonlinear metabolomics data into time intervals of linear metabolic states using PCA. Then, the authors perform a linear regression to estimate the change rate for each metabolite for a specific state, using a 95% confidence interval of the rate as reaction flux bounds in a constraint-based model. As metabolomics data can be incomplete due to experimental errors, uFBA also implements a relaxation algorithm to determine the minimum number of unmeasured metabolites whose concentration needs to vary for the model to be feasible.

Finally, another study has trained ANN models with experimental enzyme concentrations to predict the fluxes for the NADH consumption by glycerol-3-phosphate dehydrogenase in the upper part of glycolysis. In this case, no kinetic parameters or stoichiometric constraints were considered but a large and diverse dataset of enzyme concentrations is needed to obtain accurate flux predictions [132].
4.4. Other applications

ML has also been indirectly applied to other steps of GSMM reconstruction, namely genome annotation and gap-filling [16]. For instance, an approach used several multiclassification ML models to classify enzymatic reactions using a dataset of hydrolysis and redox reactions. The ML models predicted whether oxidoreductases or hydrolases catalysed specific reactions and the subclasses for each type [133]. Another method named DeepAnnotator applies deep learning models trained with DNA embeddings to identify genes and annotate prokaryotic genome sequences [134]. Regarding gap-filling, a set of ML models predicted the pathways present in an organism [135]. The models were trained with curated information on which pathways are present and absent in six organisms, achieving performance similar to other pathway prediction algorithms. Another approach uses association rule mining trained with UniProt entries to predict metabolic pathways in prokaryotes [136].

As mentioned above, using datasets with a large number of samples is crucial to obtain good ML models. The studies cited in Table 4 show that the number of samples used varies greatly depending on the type of ML and the application, ranging from a few dozens to thousands of samples. Generally, unsupervised studies use datasets with fewer samples, except for the study of Magnusdottir et al. [111], which includes over 200,000 samples. Furthermore, the studies for bacteria and yeast usually have more samples than those using human data, which is expected as data acquisition is cheaper and more accessible for smaller organisms. The study of Lewis et al. was the human study that used the most extensive dataset, including data for 915 patient tumours.

The problem of having datasets with few samples is even more evident for plants due to the lack of efforts to collect, integrate and standardise plant omics datasets and experimental conditions. Usually, some procedures are adopted to deal with small datasets. First, its common to choose simpler models with few parameters, such as logistic regression, to avoid overfitting. In addition, the use of regularization techniques and the creation of ensemble models enhances the power of generalisation. Second, it is also important to remove outlier observations and to select the relevant features to decrease the bias in the dataset. Another strategy that can improve the results is to extend the dataset by creating synthetic observations or integrating data from other sources, which is still very challenging. If the dataset is unbalanced, one solution is to perform an oversampling, which consists of increasing the number of observations of the minority class [137]. Finally, choosing the method for model validation is also important to get realistic performance metrics. For instance, the Nested Cross Validation approach was proven to make an unbiased performance evaluation [138]. Furthermore, other strategies have been developed to overcome the low-quality data problem in diverse applications, such as decomposition methods to generate extra data samples and impute missing features [139].

5. Perspective: Machine learning and plant metabolic modelling

As integrating regulatory networks in GSMMs is still very challenging, ML models trained with transcriptomics data allow detecting and rectifying systematic errors associated with the GSMMs’ flux predictions [16].

Multi-omics analyses seem to be the most promising application of combining ML and CBM, as it involves integrating traditional omics with fluxomics data predicted by CBM methods, which can provide meaningful insights about complex biological processes. Rana et al. [16] proposed an iterative scheme to combine these approaches. In such an approach, ML is initially used to analyse the data that will define the input constraints in a GSMM and later to analyse the predicted fluxes combined with experimental omics data. This process iteratively refines a GSMM until reaching consistency between CBM simulations, ML predictions and experimental data.

The integration of omics from high-throughput technologies with fluxomics data provided by GSMMs is advantageous, as it seeks to overcome the specific limitations of each data type [13,15]. Firstly, experimental omics data covers several areas, such as genomics, transcriptomics, proteomics, or metabolomics, while CBM is usually limited to fluxomics. Secondly, generating omics data does not require prior knowledge of the underlying networks, whereas GSMMs are based on extensive prior knowledge of metabolic networks, making their reconstruction time-consuming. Thirdly, although omics can be obtained promptly, these may contain intrinsic noise and experimental errors, requiring preprocessing and potentially leading to ambiguous interpretations. In contrast, GSMMs are curated and have straightforward interpretation, though relying on strong assumptions and the accuracy of flux predictions limited by the model quality and available knowledge [15]. Therefore, integrating omics data with metabolic models or the predicted fluxomics data can reduce ambiguity, generate accurate predictions, and provide more comprehensive analyses. Challenges remain in combining heterogeneous omics datasets with GSMMs. In the future, the increase in the number of omics layers will lead to the development of new multi-view algorithms. Hence, their combination with CBM is expected to grow as well [13,15].

As plants are very complex, the studies of plant metabolism and physiology will significantly benefit from multi-omics analysis and the combination of ML and CBM approaches. Plant growth, responses to biotic and abiotic stresses, fruit composition, and emerging phenotypes involve complex mechanisms and multiple interactions between system components; thus, they must be studied as a system, comprising information of all levels.

Currently, no work combining ML and CBM methods to study plant metabolism is available. However, several studies have integrated omics data with plant metabolic models and built context-specific and multi-tissue models. In addition, ML models have already been used to analyse plants’ omics data, but most plant multi-omics’ studies analyse the different omics types separately [140]. Given the large amount of plant omics data generated, ML models can combine plant multi-omics and integrate them into CBM models. The resulting fluxomics data and experimental omics can then be jointly interpreted with ML models to identify, for instance, key genes or reactions associated with specific phenotypes.

If enough data is available, part of the hybrid studies described in the previous section can be applied to analyse plant metabolism. For instance, the study of Vijayakumar et al. [126] can be used to identify key genes or reactions that best differentiate between conditions, elucidating the mechanisms of plants to adapt to different environmental conditions, such as variations in water and salt levels, and light intensities. Also, using multi-objective FBA for simulating the condition-specific models can be suitable for plants as their cellular objectives are complex and might differ between
conditions. Similarly, the work of Lewis et al. [124] can be applied to identify biomarkers for plant tolerance to adverse environmental conditions, such as drought- or salt-tolerance, or diseases. Both examples can be applied to analyse the metabolic differences between plant varieties.

Another possible application of ML to CBM is predicting interactions between plants and microbes, pathogens or symbionts. In the first case, the goal is to understand the mechanisms leading to disease and disease resistance and identify new drug targets. The latter aims to predict the plant-symbiont interaction network and analyse the effect of symbionts in metabolites or fruits production. For instance, the previous hybrid studies that predict interactions between human gut bacteria [111,112] could be adapted to predict interactions between plants and microbes. This will rely on phenotype predictions from metabolic models of both plant and microbes, creating models encompassing both organisms and their metabolic interactions. The plant-pathogen models can also be used to predict drug side effects on plants, using approaches similar to the studies [113,125].

In addition, as plant GSMMs present extensive metabolic gaps, ML models can be used for gap-filling and for predicting interactions between different tissues to generate better multi-tissue models, which will allow studying complex mechanisms related to plant responses to the environment and fruit quality. As depicted above, one of the challenges in plant metabolic modelling is the definition of constraints. Based on the outputs of ML models trained with experimental omics, approaches like uFBA [131] and the work of Nagaraja et al. [132] can be adapted to define the appropriate constraints for specific metabolic processes, such as photosynthesis and photosynthesis. Also, the best objective functions for describing a particular condition could be inferred from the context-specific omics data available.

Hence, we believe that most CBM-ML hybrid approaches can be applied to plants, including supervised and unsupervised methods. One main challenge will be collecting suitable plant omics data for the analysis of interest. Although large amounts of plant omics datasets have been generated, most of these are scattered and non-standardised, which hampers their analysis. Choosing the best ML method to use will depend on the available data and the purpose of each analysis. For unsupervised learning, studies like the ones of Folch-Fortuny et al. [108], Bharda et al. [109], Magnusdottir et al. [111], and Brunk et al. [130] were developed to explore data variation, identify metabolic groups and characterise metabolic patterns, using PCA or clustering methods and unlabelled data. The other studies have used supervised learning models, such as SVMs, ANNs, LASSO regressions and RFS, and created predictors that can be applied to new data. The applications of supervised models included the prediction of drug effects and essential genes, the identification of biomarks and novel drug targets and the estimation of microbial growth rate. The use of deep learning models with CBM is very limited, as omics datasets usually contain few samples, which is even more evident in plant omics. The number of samples in plant omics datasets is still low for traditional ML methods, hampering the development of good predictive models.

Therefore, ML complements CBM methods by defining the input constraints to metabolic models and improving interpretation of the results. Given the complexity of plant metabolic networks, CBM-ML hybrid studies will provide a more comprehensive and accurate view of the metabolic processes and variations in plants.

6. Conclusion

In this article, we described the main developments in plant metabolic modelling, underlining the current challenges and limitations hindering the study of plant metabolism. Although there is little knowledge about the metabolic pathways of plants, many advances have been made in this field, including the reconstruction of complex, context-specific, and multi-tissue models that generate more realistic predictions. Even so, challenges remain in defining constraints affecting plants, choosing appropriate objective functions, and characterising the metabolic differences across different tissues and conditions.

With the rapid generation of large amounts of omics datasets, the use of ML in systems biology will continue to increase. ML is a valuable tool for reducing the dimensionality of omics datasets and extracting knowledge from data. Here, we have also described the main hybrid studies combining CBM and ML developed for other organisms showing promising results for several applications, such as predicting essential genes and reactions, phenotypes of interest, genetic and microbial interactions, and new drug targets. Although these studies were mainly applied to microbes and human cells, some can be adapted to plants, for instance, to predict plant-symbiont interactions and identify key molecules to characterise each phenotype.

Therefore, we believe that using ML in plant metabolic modelling will fill the gaps in plant biochemical knowledge with insights retrieved from the experimental omics analysis. The integration of fluxomics with experimental omics will allow to better understand complex biological processes and interactions occurring in plants.

Funding

This work was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2020 unit and the PhD scholarship (SFRH/BD/144643/2019) to Marta Sampaio. Oscar Dias also acknowledges FCT for the Assistant Research contract obtained under CEEC Individual 2018.

CRediT authorship contribution statement

Marta Sampaio: Conceptualization, Writing – original draft. Miguel Rocha: Conceptualization, Writing – review & editing. Oscar Dias: Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Argueso CT, Assmann SM, Birnbaum KD, Chen S, Dinneyen JR, Doherty CJ, et al. Directions for research and training in plant omics: Big Questions and Big Data. Plant Direct 2019;3:1–16. https://doi.org/10.1002/pld.1333.
[2] Verpoort R. Plant secondary metabolism. Metab Eng Plant Second Metab, Kluwer Academic Publishers 2000. https://doi.org/10.1002/0470869143. kc025.
[3] Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BD. Reconstruction of Biochemical Networks in Microbial Organisms. Nat Rev Microbiol 2008. https://doi.org/10.1038/nrmicro1849.
[4] Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biol 2019;20:1–18. https://doi.org/10.1186/s13059-019-1730-3.
[5] Sweetlove LJ, George Ratcliffe R. Flux-balance modeling of plant metabolism. Front Plant Sci 2011;2:1–10. https://doi.org/10.3389/fpls.2011.00018.
[6] Collakova E, Yen JY, Senger RS. Are we ready for genome-scale modeling in plants? Plant Sci 2012;191–192:53–70. https://doi.org/10.1016/ j.plantsci.2012.04.016.
[7] Robaina Estévez S, Nikolaou Z. Generalized framework for context-specific metabolic model extraction methods. Front Plant Sci 2014;5:491. https://doi.org/10.3389/fpls.2014.00491.
Bhadra S, Blomberg P, Castillo S, Rousu J. Principal metabolic flux mode analysis. Bioinformatics 2018;34:2409–17. https://doi.org/10.1093/bioinformatics/bty494.

Folch-Fortuny A, Teusink B, Hoefsloot HCJ, Smilde AK, Ferrer A. Dynamic elementary mode modelling of non-steady state flux data. BMC Syst Biol 2018;12:1–15. https://doi.org/10.1186/s12918-018-0598-3.

Magnúsdóttir S, Heinzen A, Kutt L, Ravcheev DA, Bauer E, Norosha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol 2017;35:81–9. https://doi.org/10.1038/nbt.7073.

DiMucci D, Kon M, Segrè D. Machine learning reveals missing edges and putative interaction mechanisms in microbial ecosystem networks. MSystems 2018;3:1–13. https://doi.org/10.1128/mSystems.00181-18.

Shaked I, Oberhardt MA, Atias N, Shiran R, Ruppin E. Metabolic network prediction of drug side effects. Cell Syst 2016;2:209–13. https://doi.org/10.1016/j.cels.2016.03.001.

Oyetunde T, Liu D, Martin HG, Tang YJ. Machine learning framework for assessment of microbial factory performance. PLoS ONE 2019;14:1–15. https://doi.org/10.1371/journal.pone.0210558.

Czajka JJ, Oyetunde T, Tang YJ. Integrated knowledge mining, genome-scale modeling, and machine learning for predicting Yarrowia lipolytica bioproduction. Metab Eng 2021;67:227–36. https://doi.org/10.1016/j.ymben.2021.07.003.

Schimm SM, Morrison C, Wei W, Zhang L, Lewis NE. A genome-scale metabolic network model and machine learning predict amino acid concentrations in Chinese Hamster Ovary cell cultures. Biotechnol Bioeng 2021;118:2118–23. https://doi.org/10.1002/bit.27714.

Antonakoudis A, Strain B, Barbosa R, Jimenez del Val I, Kontoravdi C. Synergising stoichiometric modelling with artificial neural networks to predict antibiotic glycosylation patterns in Chinese hamster ovary cells. Comput Chem Eng 2021;154. https://doi.org/10.1016/j.compchemeng.2021.107471.

Nandi S, Subramanian A, Sarkar RR. An integrative machine learning strategy for improved prediction of essential genes in Escherichia coli metabolism using flux-coupled features. Mol Bioysyt 2017;13:1584–96. https://doi.org/10.1039/c7nb00234d.

Plaïmas K, Mallim JP, Oswald M, Svara F, Sourjik V, Eils R, et al. Machine learning based analyses on metabolic networks supports high-throughput knockout screens. BMC Syst Biol 2008;2:1–11. https://doi.org/10.1186/1752-0509-2-67.

Li L, Zhou X, Ching WK, Wang P. Predicting enzyme targets for cancer drugs by profiling human Metabolic reactions in NCI-60 cell lines. BMC Bioinf 2010;11. https://doi.org/10.1186/1471-2105-11-501.

Kim M, Rai N, Zorraquino V, Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nat Commun 2016;7:1–12. https://doi.org/10.1038/ncomms13090.

Culley C, Vijayakumar S, Rahman PKSM, Angione C. A hybrid flux balance analysis and machine learning pipeline elucidates metabolic adaptation in cyanobacteria. iScience 2020;23. https://doi.org/10.1016/j.isci.2020.101818.

Kavvas ES, Yang L, Monk JM, Heckmann D, Palsson BO. A biochemically-interpretable machine learning classifier for microbial GWAS. Nat Commun 2020;11:1–11. https://doi.org/10.1038/s41467-020-16310-9.

Guo W, Xu Y, Feng X. DeepMetabolism: A Deep Learning System to Predict Phenotype from Genome Sequencing 2017;1:7–.

Wu SG, Wang Y, Jiang W, Oyetunde T, Yao R, Zhang X, et al. Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. PLoS Comput Biol 2016;12. https://doi.org/10.1371/journal.pcbi.1004838.

Brunk E, George KW, Alonso-Gutierrez J, Thompson M, Baidoo E, et al. Characterizing strain variation in engineered E. coli using a multi-omics based workflow. Physiol Behav 2017;176:139–48. https://doi.org/10.1016/j.physbeh.2017.03.001.

Bordbar A, Yurkovich JT, Paglia G, Rolfsen S, Sigurjónsson Ö, Palsson BO. Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. Sci Rep 2017;7:1–12. https://doi.org/10.1038/srep46249.

Nagaraja AA, Fontaine N, Delsaut M, Charton P, Damour C, Offmann B, et al. Meta-Flux prediction using artificial neural network (ANN) for the upper part of glycolysis. PLoS ONE 2019;14:10–2. https://doi.org/10.1371/journal.pone.0216128.

Cai Y, Yang H, Li W, Liu G, Lee PW, Tang Y. Multiclassification prediction of enzymatic reactions for oxidoreductases and hydrolases using reaction fingerprints and machine learning methods. J Chem Inf Model 2018;58:1169–81. https://doi.org/10.1021/acs.jcim.7b00656.

Amin MR, Yurovsky A, Tian Y, Skiena S. DeepAnnotator: genome annotation with deep learning. ACM-BCB 2018. In: Proc 2018 ACM Int Conf Bioinformatics Comput Biol Heal Informatics. p. 254–9. https://doi.org/10.1145/3233547.3233777.

Dale JM, Popescu L, Karp PD. Machine learning methods for metabolic pathway prediction. BMC Bioinf 2010;11:15. https://doi.org/10.1186/1471-2105-11-15.

Boudellioua I, Said R, Hoehndorf R, Martin MJ, Solovyev V. Prediction of metabolic pathway involvement in prokaryotic uniprotkb data by association rule mining. PLoS ONE 2016;11:1–16. https://doi.org/10.1371/journal.pone.0158896.

Maheswari JP. Breaking the curse of small datasets in Machine Learning. Towards Data Science n.d. https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-362980c044d (accessed March 25, 2022).

Valabas A, Gowen E, Polakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS ONE 2019;14:. https://doi.org/10.1371/journal.pone.0224365.

Cañada CF, Sun Z, Tanaka T, Mari-Puig P, Solé-Casals J. Machine learning methods with noisy, incomplete or small datasets. Appl Sci 2021;11. https://doi.org/10.3390/app11091532.

Jamil IN, Remali J, Azzan KA, Nor Muhammad NA, Arita M, Goh HH, et al. Systematic Multi-omics integration (M0I) approach in plant systems biology. Front Plant Sci 2020;11. https://doi.org/10.3389/fpls.2020.00944.