# Combined mechanistic modeling and machine-learning approaches in systems biology – A systematic literature review

**Article** *in* Computer Methods and Programs in Biomedicine · June 2023

DOI: 10.1016/j.cmpb.2023.107681

CITATIONS
2

READS
95

**6 authors**, including:

Anna Procopio
Universita' degli Studi "Magna Græcia" di Catanzaro
**18** PUBLICATIONS **92** CITATIONS
SEE PROFILE

Leandro Donisi
Università degli Studi della Campania "Luigi Vanvitelli"
**60** PUBLICATIONS **549** CITATIONS
SEE PROFILE

Francesco Amato
Universita' degli Studi di Napoli Federico II
**424** PUBLICATIONS **7,742** CITATIONS
SEE PROFILE

Carlo Cosentino
Universita' degli Studi "Magna Græcia" di Catanzaro
**176** PUBLICATIONS **3,943** CITATIONS
SEE PROFILE

Computer Methods and Programs in Biomedicine 240 (2023) 107681

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

# Combined mechanistic modeling and machine-learning approaches in systems biology – A systematic literature review

Anna Procopio[a,1], Giuseppe Cesarelli[b,1], Leandro Donisi[c], Alessio Merola[a], Francesco Amato[b,2,*], Carlo Cosentino[a,2,*]

[a] Department of Experimental and Clinical Medicine, Università degli Studi Magna Græcia, Catanzaro, 88100, Italia
[b] Department of Electrical Engineering and Information Technology, Università degli Studi di Napoli Federico II, Napoli, 80125, Italy
[c] Department of Advanced Medical and Surgical Sciences, Università della Campania Luigi Vanvitelli, Napoli, 80138, Italy

ARTICLE INFO

*Article history:*
Received 17 May 2023
Revised 14 June 2023
Accepted 14 June 2023

*Keywords:*
Mathematical modeling
Machine learning
Reinforcement learning
Systems biology
Simulation
Systematic literature review

ABSTRACT

*Background and objective:* Mechanistic-based Model simulations (MM) are an effective approach commonly employed, for research and learning purposes, to better investigate and understand the inherent behavior of biological systems. Recent advancements in modern technologies and the large availability of omics data allowed the application of Machine Learning (ML) techniques to different research fields, including systems biology. However, the availability of information regarding the analyzed biological context, sufficient experimental data, as well as the degree of computational complexity, represent some of the issues that both MMs and ML techniques could present individually. For this reason, recently, several studies suggest overcoming or significantly reducing these drawbacks by combining the above-mentioned two methods. In the wake of the growing interest in this hybrid analysis approach, with the present review, we want to systematically investigate the studies available in the scientific literature in which both MMs and ML have been combined to explain biological processes at genomics, proteomics, and metabolomics levels, or the behavior of entire cellular populations.

*Methods:* Elsevier Scopus®, Clarivate Web of Science™ and National Library of Medicine PubMed® databases were enquired using the queries reported in Table 1, resulting in 350 scientific articles.

*Results:* Only 14 of the 350 documents returned by the comprehensive search conducted on the three major online databases met our search criteria, i.e. present a hybrid approach consisting of the synergistic combination of MMs and ML to treat a particular aspect of systems biology.

*Conclusions:* Despite the recent interest in this methodology, from a careful analysis of the selected papers, it emerged how examples of integration between MMs and ML are already present in systems biology, highlighting the great potential of this hybrid approach to both at micro and macro biological scales.



## 1. Introduction

Systems biology (SB) is a multidisciplinary field that aims to understand the behavior of biological systems through the computational analysis of biological mechanisms. In this context, mathematical modeling represents a crucial aspect of SB, as it allows researchers to make predictions about how biological systems will behave under different conditions, and to test these predictions using experimental data. Models can range from simple mathematical equations to more complex systems, often based on differential equations or logical models, obtained by incorporating information provided by multiple sources. By creating and refining models of biological systems, researchers can gain a deeper understanding of how these systems function, and how they can be manipulated for a variety of applications, including drug discovery and biotechnology. Significant advances in fields such as genome sequencing and high-throughput measurements have enabled the collection of exhaustive datasets on the function of biological systems and their constituent molecules. In this context, SB, in combination with bioinformatics analysis tools, aims to pro-

* Corresponding authors.
*E-mail addresses:* francesco.amato@unina.it (F. Amato), carlo.cosentino@unicz.it (C. Cosentino).
[1] Equal contribution.
[2] Co-last.

https://doi.org/10.1016/j.cmpb.2023.107681

vide appropriate tools for the analysis, interpretation, and integration of all omics data, as well as for the investigation and formulation of new theoretical hypotheses, [1,2]. Different modeling approaches, more specialized towards genomics and metabolomics [3–5] to proteomics [6,7], have been proposed to describe the biological systems from different perspectives, [8]. However, although these mechanistic models provide detailed information on the observed process, their complexity increases exponentially with the degree of detail, i.e. with the number of species and unknown parameters to be estimated, [9].

At the same time, in recent years artificial intelligence (AI) strategies, particularly machine learning (ML) techniques, have been increasingly applied in the biological and biotechnological fields, thanks to their ability to introduce more automated analysis and decision-making, [10]. However, also ML techniques present several limitations, such as their inability to deal with sparse or biased data, leading to ill-posed problems and non-physical predictions. Additionally, since ML techniques are solely dependent on data, they are unable to provide a mechanistic basis for explaining complex behaviors and making reliable predictions, [4].

The limitations of mechanistic models (MMs) and ML methods have prompted the scientific community to investigate novel approaches based, for example, on the combination of these two methodologies, as documented in [9–11]. In support of these hybrid approaches, Yeo and Selvarajoo [4] proposed instead a combination of modeling/simulation and ML to address several biological issues, not explainable by only ML. Furthermore, examples of the application of these hybrid approaches are documented in clinical, [12,13], and industrial sectors, [14–16]. Recently, in healthcare, Sharafutdinov e al. proposed a hybrid pipeline in which mechanistic models are used to generate virtual patients' synthetic data, which are exploited to improve the performance of unsupervised machine learning methods for the identification of patients with suspected acute respiratory distress syndrome, [17].

Since the combination of MMs with ML represents a relatively new area of interest in healthcare, to date, no systematic reviews have been proposed about the combined applications of this hybrid approach in SB. To this end, in this article, we shall review the contributions related to the SB field where the hybrid use of MMs (especially the deterministic models) and ML has been applied. To this aim, we follow the 2020 update of the "Preferred Reporting Items for Systematic Reviews and Meta-Analyses" (PRISMA) [18] guidelines, that have been shown to overcome many of the drawbacks of traditional [19] critical reviews.

This systematic review is articulated as follows: in Section 2 the selection and exclusion criteria of the articles coming from the online databases are reported, as well as the main methodologies for both MMs and ML. Section 3 introduces the hybrid approach and its application at the genomics, proteomics, metabolomics, and cellular dynamics level, while Section 4 focuses on the main conclusions.

## 2. Materials and methods

### 2.1. Systematic literature review

As previously anticipated in the Introduction section, the reporting of this systematic review is guided by the PRISMA guidelines, [20]. A literature search was carried out seeking documents on three academic search engines: Elsevier Scopus®, Clarivate Web of Science™and National Library of Medicine PubMed®. The queries employed to enquire the three above-mentioned databases are reported in Table 1. Especially, these queries allow to search documents that showed the following specific features: firstly, documents had to be focused on the SB field and/or on SB specific applications; secondly, documents should have presented data processing workflows based on the combining use of both MMs and ML.

As depicted in Fig. 1, the identification stage was based on a) setting up the research on the engines, b) exporting the results (date of the last check: 10/05/2023), and c) eliminating the duplicated records. The next screening stage is carried out in three steps; firstly i) the documents that are neither original/research articles nor conference papers are excluded; then ii) the remaining documents are screened initially by title and Abstract and, finally, iii) by the screening of the full text, to check for coherence with the review objective/s. As Fig. 1 reports, only 14 out of the initial 350 documents found met the criteria established for this systematic review.

### 2.2. Simulation in systems biology

Mathematical models are the cornerstone of the ever-growing field of SB. In this context, the mechanistic models, based on the fundamental rules of physics and biochemistry, represent the most appealing approach to provide an accurate and comprehensive description of the researched biological process, [21]. This peculiarity makes mathematical models effective tools for analyzing and investigating the reality of interest: a mathematical model may be used to replicate alternative experimental scenarios, hence directing the development of novel biological hypotheses [22,23].

The complexity of mathematical models and their simulations increases linearly with the level of detail chosen to represent the observed reality: each model consists of i) dependent and independent variables, i.e., state variables and variables with respect to which the dependent variables vary, respectively, and ii) a certain number of parameters. In particular, the parameters need a calibration phase, during which their value is determined experimentally or estimated from experimental data. The process of identifying such values for model parameters represents one of the main drawbacks of mathematical models: due to the microscopic nature of the biological processes and the limited capability of direct measurements of the physicochemical quantities, parameter estimation is usually based on the fitting of limited data.

These models could be either static or dynamic. Specifically, dynamical models account for time, allowing them to replicate the evolution of the entire system across time. They are frequently referred to as simulation models precisely for this reason. Fig. 2 shows a schematic representation of the types of mechanistic models investigated in this review, from the most complex, detailed, and accurate, i.e., dynamical models based on ordinary differential equations, to the simplest algebraic ones, [9].

#### 2.2.1. Ordinary differential equations

Ordinary Differential Equations (ODEs) allow modeling the variation over time of a variable of interest, which generally in SB coincides with the concentration of particular biochemical species. For research purposes, the models based on ODEs are primarily exploited as tools for simulation and analysis to investigate the dynamical systems in several biological contexts, ranging from the study of specific cellular pathways, [24–26], to the study of the dynamics underlying specific diseases, [27–31]. In the development of drug delivery systems, the Food and Drug Administration (FDA) encourages the standardization of nanoparticle design and development processes through the use of tools based on mathematical models, such as ODE systems, which enables an understanding of how formulation variables might affect the quality of the final product, [32].

#### 2.2.2. Constraint-based models

Constraint-based models (CBMs) are mainly employed to represent metabolic interaction networks. Starting from the hypothe-

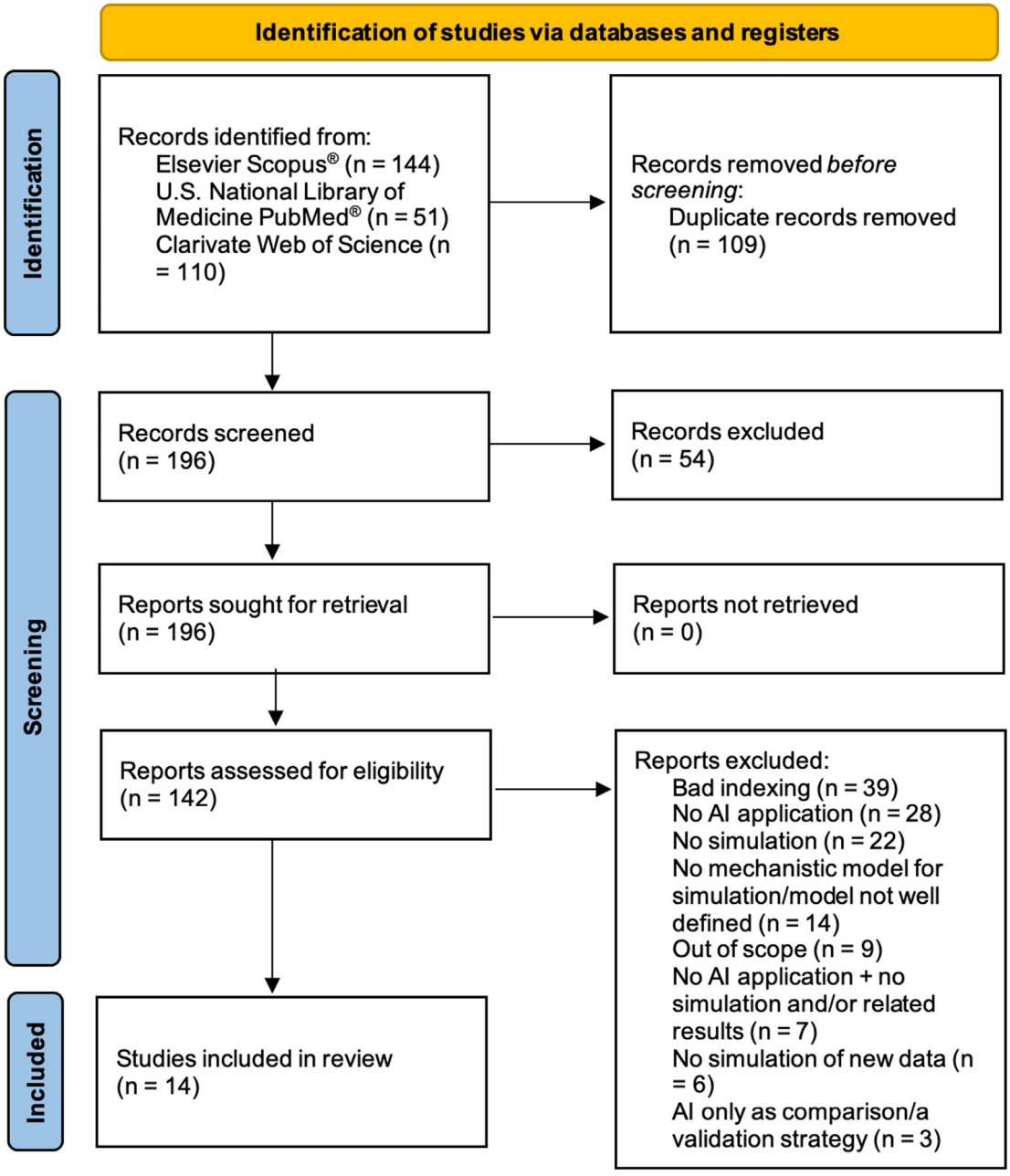


**Fig. 1.** PRISMA flowchart for the present systematic literature review. The three major search engines, Elsevier Scopus®, Clarivate Web of Science™and National Library of Medicine PubMed®, were enquired by using the queries reported in Table 1.

**Table 1**
Research queries for each of the three search engines enquired in the present systematic literature review. Before duplicates count, the queries were subsequently refined to exclude from the results the records labeled as published in the current year. * In Clarivate Web of Science™, the query was performed by checking on all the databases.

| Database | Query |
|---|---|
| Elsevier Scopus® | TITLE-ABS-KEY-AUTH (((“Machine Learning” OR “ReinforcementLearning”) AND (“Simulation” OR “Mechanistic Model”)) AND “Systems Biology”) |
| Clarivate Web of Science™ | ((((TS=(“Machine Learning”) OR TS= (“Reinforcement Learning”)) AND (TS=(“Simulation”) OR TS=(“Mechanistic Model”)))) AND TS=(“Systems Biology”)) |
| National Library of Medicine PubMed® | (((“Machine Learning”[Text Word]) OR (“Reinforcement Learning”[Text Word])) AND (“Simulation”[Text Word] OR (“Mechanistic Model”[Text Word]))) AND (“Systems Biology”[Text Word]) |

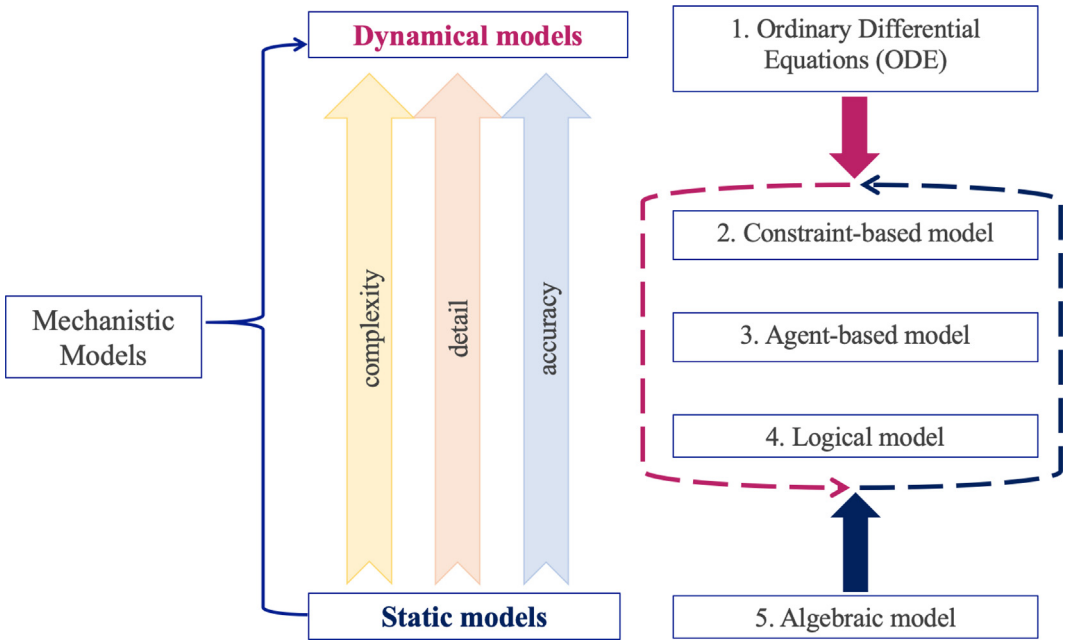


**Fig. 2.** Schematic representation of the main types of mechanistic models, ordered by complexity, degree of detail, and accuracy of the produced results.

sis that cellular systems can reach their steady-state very quickly, these models allow overcoming the problem due to the need to estimate the parameter values of complex ODE models, [8], as well as the complexity of the systems. Indeed, constraint-based models are based on the analysis of only metabolic fluxes involved in a specific process or cellular context. Their solution requires the definition of a series of constraints on flows, which determine not a single solution, but a set of possible ones. The benefits of these models include the ability to forecast growth rates, dynamical changes, and metabolite fluxes over the entire metabolic network, as well as the possibility to analyze and investigate the metabolic phenotypes induced by genetic changes. Constraint-based models are successfully applied to reconstruct and model the human microbiome, [33], to study the metabolic differences between several strains of bifidobacteria colonizing the gut of infants, [34], or to investigate the metabolic causes of treatment resistance in some urological cancers, [35].

#### 2.2.3. Agent-based models

Agent-based models (ABMs) are particular classes of models based on agents. In these models, each agent is designed with specific characteristics, i.e., physical or biochemical laws, that allow the simulation of the interaction with the other agents and with the environment and to respond by generating specific output signals, [36,37]. The agents are particles or entities characterized by i) autonomy, ii) modularity, and iii) ability to interact with surrounding agents, [38]. They are mainly employed to model processes such as, e.g., the tumor growth, [39–41], or the vascular adaptation, [42].

#### 2.2.4. Logical models

Logic-based models (LMs), which strike a reasonable balance between accuracy and complexity, enable the depiction of extensive biochemical networks without the need for in-depth knowledge of the mechanisms driving the interactions between the modeled species, [43]. Each specie has two states to represent it: ON and OFF. In particular, this state evolves dynamically until a stable condition of the global network is achieved. Logic-based models are successfully applied to model several cellular contexts: e.g., in [44], an ad-hoc logical model was employed to investigate the c-Met signal transduction network and its implication in tumor development. Flobak et al., in [45], proposed an approach based on LMs to represent a cell fate decision network in human gastric adenocarcinoma cells, predicting the synergistic inhibitory action of five combinations of anti-cancer drug treatments.

#### 2.2.5. Algebraic models

Several cellular processes, in a particular biological setting, evolve on varying temporal scales: especially, some processes progress slowly over time, while others quickly find a new equilibrium condition following a perturbation. Specifically, these last-mentioned evolutions may be modeled using simple and static algebraic models (AMs). The primary benefit of these models is their simplicity, making them particularly suitable in those cases where the system complexity requires modeling approaches entailing a limited computational effort. Several zero-orders models are exploited, e.g., to describe the constant dynamics underline the drug release from nanoparticles in the pharmacokinetics field, as in [46–48], or other biological processes such as ultrasensitivity [49,50].

### 2.3. Machine learning in systems biology

The recent success of ML techniques for large-scale biological data analysis has provided a complementary and, in some cases, competing alternative to more traditional model-based approaches, [51,52]. The use of ML in this area comes from the incompleteness of detailed knowledge concerning the effects of inhibitors or known biochemical reactions [53]. ML algorithms are well-suited for managing and processing large amounts of data, as well as identifying complex, multi-dimensional correlations that may not be easily discernible by humans. These algorithms can also handle noisy or incomplete data and are often able to generalize patterns they learn to new, unseen data, perfectly adapting to the main problems of systems biology. In addition, many ML algorithms are designed to handle ill-conditioned problems, in which small changes in the input data can lead to large changes in the output, [9]. Indeed, several optimization algorithms can be sensitive to the initial starting point, but ML algorithms are often able

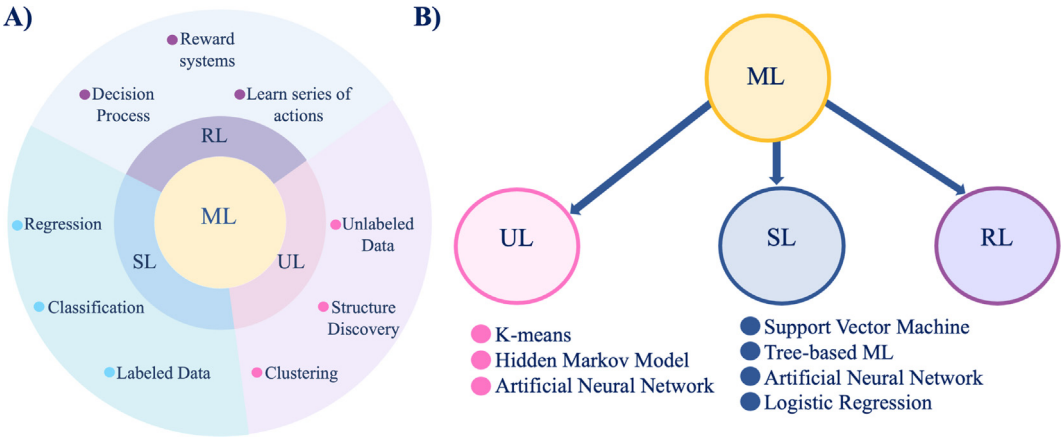


**Fig. 3.** A) Main characteristics of the different types of ML treated in this work; B) Schematic representation of the ML classes and the related algorithms covered in the selected paper. ML is the term chosen to indicate Machine Learning, RL represents Reinforcement Learning, while SL and UL mean Supervised and Unsupervised Learning, respectively.

to find good solutions even when starting from random initial conditions. Furthermore, recent trends involved in the integration of physics-based models with ML algorithms, to improve predictions and reduce errors in complex systems, as reported in [54]. In particular, Physic-informed Neural Networks (PINNs), [55], combines the strengths of physics-based models, which capture the underlying physical principles of a system, with the flexibility and adaptability of ML algorithms, which can learn from data and make predictions.

For all these reasons, ML and related techniques, such as Support Vector Machine (SVM), [56], Hidden Markov Model (HMM), [57], Decision Tree (DT), [58], and Neural Network (NN), [59], have been increasingly used to solve problems in several fields of SB. ML can be divided into three macro-areas, namely i) supervised learning (SL), ii) unsupervised learning (UL), depending on whether the class of an object to be classified is considered as a known value during the training of the ML algorithm, and iii) reinforcement learning (RL). While SL works with labeled data and is primarily employed for the classification of data and regression, UL works on unlabeled data, and it is mainly used for clustering and for investigating data and discovering hidden relationships between them, (Fig. 3 - A). Instead, concerning the RL, this kind of ML aims at fully automatic learning of decisions to be made through continuous interaction with the surrounding environment. Fig. 3 - B shows the main ML algorithms covered in this paper (grouped into the previously mentioned classes) which will be discussed below.

### 2.3.1. Supervised learning

*Support Vector Machines* Support Vector Machines (SVMs) are considered to be robust algorithms that exhibit a lower susceptibility to overfitting. Through the selection of a suitable kernel, SVMs can effectively handle non-linearly separable data in the feature space. Furthermore, the SVM algorithm aims to identify the hyperplane with the largest margin, which enhances its predictive capacity for accurately classifying novel instances, [60]. They are robust to high dimensional data since the complexity remains unaffected by the number of features [61]. Moreover, they have good generalization ability, even if training speed is low, and their performances depend strongly on the choice of model parameters [62]. SVMs are extensively used in the field of computational biology and SB [63–65]. For example, Liao et al. combined pairwise sequence similarity and SVMs for detecting remote protein evolutionary and structural relationships [66]. *Tree-based Machine Learning Algorithms* The simplest algorithm on which tree-based algorithms are based is the Decision Tree (DT). DT is easy to interpret and explain and can easily handle interactions between features. Moreover, being a non-parametric model, outliers do not affect the model in an important way and it can deal with not-linearly separable data. Among DTs, the most famous algorithms are ID3, [67], C4.5, [68], and CART, [69]; they differ in splitting criteria, namely Gini Coefficient, Gain Ratio, and Info Gain [70]. DT can handle a variety of data (nominal, numeric, textual), missing values, and redundant features. Moreover, they have good generalization ability, are robust to noise, and provide high performance for relatively small computational effort; on the contrary, DTs find it difficult to handle high dimensional data. DT uses the divide *et impera* approach which performs well if few highly relevant features are present but not very well if many complex interactions are present. Errors propagate through trees and become a serious problem as the number of classes increases [71,72]. Finally, DTs are susceptible to overfitting without an effective pruning strategy. To overcome this issue, Random Forest (RF) operates by training multiple DTs and returning the most recurring class among the results of all the trees in the ensemble, [73]. Tree-based ML algorithms are widely used in SB, [74], for instance, to integrate gene expression, demographic and clinical data to determine disease endotypes and to better understand complex diseases, [75], or to predict the effects of inhibiting contractility signaling on cell motility, [76], or for efficient identification of parameter relations leading to different signaling states, [77]. *Logistic Regression* Logistic regression (LR) is a statistical method in which a logistic curve is fitted to the dataset. This technique is applied when the dependent variable or target variable is dichotomous. Since it is a statistical learning algorithm, it has a probabilistic interpretation and the model can be updated to take new data easily by means of the gradient descent method [61]. Anyway, in order to have a reliable and robust model, it needs that the following assumptions has to be verified [78,79]: absence of multicollinearity, absence of outliers, the ratio between the sample size of the smallest class and the number of independent variables greater than 10 [80,81]. In SB, LR was used for biomarker identification, [82], in particular applications where LR was applied to identify blood-based multi-omics biomarkers for Alzheimer's disease, as in [83], or to predict lysine acetylation site, [84].

### 2.3.2. Unsupervised learning

*K-means* The K-means algorithm is generally the most known and used clustering method. Clustering is a very useful tool in data science when the labels of instances are not known a priori, and for this reason, it is an unsupervised machine-learning approach. It is a method for finding cluster structure in a data set characterized by the greatest similarity within the same cluster and the

greatest dissimilarity between different clusters. The k-means algorithm takes the input parameter, k, and partitions a set of n instances into k clusters so that the resulting intra-cluster similarity is high, while inter-cluster similarity is low (or, similarly, minimizing the squared-error function). Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster centroid or "center of gravity" [85]. However, K-means algorithms can be applied only when the clusters are known, therefore it cannot be applied when data with categorical attributes are involved. In the context of SB, K-means was used to improve the biological features of Weighted Gene Co-expression Network Analysis (WGCNA), [86]. *Hidden Markov Model* Hidden Markov modeling is a powerful statistical ML technique. Hidden Markov Models (HMMs) offer the advantage of having strong statistical foundations that are well-suited to several scenarios. Moreover, HMMs are computationally efficient to develop and evaluate thanks to the existence of established training algorithms [87]. HMMs are a formal foundation for making probabilistic models of linear sequence 'labeling' problems since they provide a conceptual toolkit for building complex models just by drawing an intuitive picture [88]. HMMs are recurring themes in computational biology since often biological sequence analysis is just a matter of putting the right label on each residue. For example, in genomics, it is possible to use HMMs to label nucleotides as exons, introns, or intergenic sequences, as reported in [89–91], while, in proteomics, HMMs applications are available, for example, in protein modeling [92].

### 2.3.3. Artificial neural network

Artificial Neural Networks (ANNs) are computational structures based on neural architectures similar to the human brain. An ANN consists of an input layer of neurons (or node, units), one or more hidden layers, and a final layer of output neurons, [93]. ANNs provide a powerful alternative to conventional techniques (i.e. statistical learning) which are often limited by strong assumptions of normality, linearity, variable dependence, absence of multicollinearity, absence of outliers, the precise ratio between the number of instances and features also in relation to the number of classes to predict. In the field of SB and bioinformatics, several studies appeared in the scientific literature using ANNs to develop models of the dynamics of gene expression, [94]. ANNs are widely used in metabolomics, [95], and in genomics where they are exploited, for example, to model omics data, [96].

### 2.3.4. Reinforcement learning

Reinforcement Learning (RL) is a distinct form of Machine Learning (ML) characterized by its capacity to acquire knowledge from the environment and to generate actions in a nearly autonomous way, without the use of prior knowledge, [97,98]. Several applications of RL are presented in SB, [99]: e.g., in [100], the authors proposed a new methodology based on RL, to improve the drug design. In [101], the authors introduced a novel framework based on both RL and anticancer drug sensitivity prediction model, named PaccMann$^{RL}$, able to generate molecules that are tailored for a given target. In proteomics, Zhu et al., [102], proposed an RL-based methodology to establish a protein interaction network.

## 2.4. Main evaluation metrics

In a typical data classification problem, several appropriate evaluation metrics are used to optimize the classifier, during the training step, and to measure the effectiveness of the trained classifier in the testing stage. Accuracy is the most used evaluation metric either for binary or multi-class classification problems, even if also other metrics are used to assess the effectiveness of ML models [103]. The most used evaluation metrics are reported in Table 2, where $t_p$, $t_n$, $f_p$, and $f_n$ denote the true positive, true negative, false positive, and false negative, respectively and they allow to create the confusion matrix as reported in Table 3.

**Table 2**
Evaluation metrics.

| Evaluation Metrics | *MathFormula* |
|---|---|
| Accuracy | $\frac{t_p+t_n}{t_p+t_n+f_p+f_n}$ |
| Error Rate | $\frac{f_p+f_n}{t_p+t_n+f_p+f_n}$ |
| Sensitivity | $\frac{t_p}{t_p+f_n}$ |
| Specificity | $\frac{t_n}{t_n+f_p}$ |
| Precision | $\frac{t_p}{t_p+f_p}$ |
| Recall | $\frac{t_p}{t_p+t_n}$ |
| F-measure | $\frac{2*Precision*Recall}{Precison+Recall}$ |
| Geometric-mean | $\sqrt{tp*tn}$ |

**Table 3**
Confusion matrix.

| | Actual Positive | Actual Negative Class |
|---|---|---|
| **Predicted Positive Class** | $t_p$ | $f_n$ |
| **Predicted Negative Class** | $f_p$ | $t_n$ |

The Area Under the ROC (Receiver Operating Characteristic) Curve (AUC) is one of the popular ranking-type metrics reflecting the overall ranking performance of a classifier. For two-class problems, the AUC value can be calculated as reported in (1).

$$AUC = \frac{S_p - n_p(n_n+1)/2}{n_p n_n} \tag{1}$$

Where $S_p$ is the sum of all positive examples ranked, while $n_p$ and $n_n$ denote the number of positive and negative examples respectively. The AUC was proven theoretically and empirically better than the accuracy metric, [104], for evaluating the classifier performance. The $R^2$ index is calculated starting from the regression analysis. The $R^2$ value of a regression reported in (2) is usually taken as the portion of the variance of the dependent variable accounted for by the explanatory variables.

$$R^2 = \frac{RSS}{TSS} \tag{2}$$

where RSS is the regression sum of squares (namely the deviations from the mean explained by the regression), while TSS is the total sum of squares, [104].

## 3. Hybrid approach based on MMs-ML

The hybrid approach arises from the fusion of both mechanistic-based model simulations (MMs) and ML techniques: this integration strategy allows to merge of the advantages of the two techniques, and, at the same time, overcoming their single application limits.

The investigate protocols (IPs) of the main hybrid models are shown in Fig. 4:

- IP-1 - starting from a well-defined mathematical model, this architecture is generally exploited to generate synthetic data capable of simulating the behavior of the system under different experimental conditions. This large amount of synthetic data, generated by the models in a fairly simple way, represents the input of ML algorithms, as in [105,106];
- IP-2 - in this architecture, structurally complementary to the IP-1, ML techniques are used to act as an input to classic simulation models, [107,108];

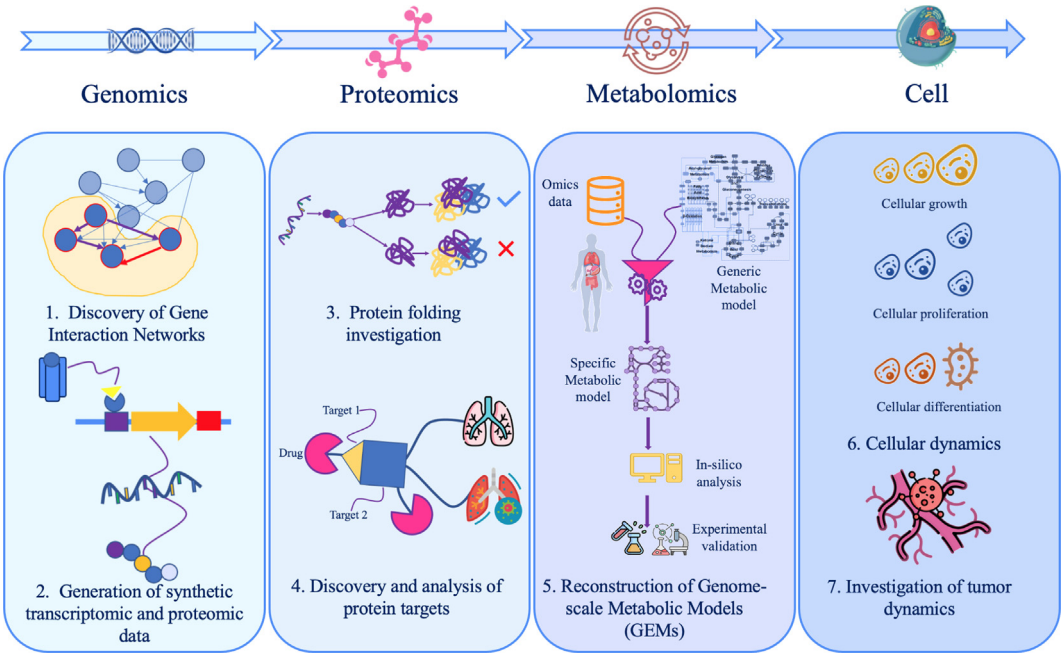


**Fig. 5.** Summary of the main applications of the hybrid approaches, implemented with the protocols shown in the Fig. 4, in the 4 biological fields found: Genomics, Proteomics, Metabolomics, and Cellular Dynamics.

**Table 4**
Selected papers grouped by biological application fields, with relative MMs, ML, and implementation protocol reported in Fig. 4. Genomics: ; Proteomics: ; Metabolomics: ; Cellular Dynamics: .

| Authors | MMs | ML | IP |
|---|---|---|---|
| Carrè et al. [111] | AMs | SVM | IP-1 |
| Miagoux et al. [111] | LMs | UL–not specified | IP-2 |
| Moore et al. [112] | LMs | DT/RF | IP-1 |
| Hua et al. [113] | ODEs | K-means/DT | IP-1 |
| Chua et al. [52] | ODEs | SVM | IP-4 |
| Oguz et al. [114] | ODEs | RF | IP-2 |
| Matsunaga & Sugita ([115] | AMs/ODEs | HMM | IP-1 |
| Derbalah et al. [116] | ODEs | ANNs | IP-4 |
| Biba et al. [117] | LMs | HMM | IP-1 |
| Szappanos et al. [118] | CBMs | RF/LR | IP-1 |
| Medlock & Papin [119] | CBMs | K-means/RF | IP-1 |
| Maeda et al. [120] | ODEs | RF | IP-2 |
| Sieburg et al. [121] | AMs | NN/RL | IP-3 |
| Zangooei et al. [122] | ABMs | RL | IP-3 |

IP-3 - similar to IP-2 but based on Reinforcement Learning (RL) techniques, which basically act as a controller, guiding the model towards the right representation of reality, [109];

IP-4 - represents a juxtaposition of the first two techniques, consisting of a first module determined by IP-1, and a second module coinciding with IP-2. It is mainly used to improve the models available with the information obtained from ML algorithms, [110].

From the systematic research carried out, 14 original articles and conference proceedings were found. Especially, Table 4 reports the 14 articles grouped by biological application fields, *Genomics, Proteomics, Metabolomics*, and *Cellular Dynamics*, with relative MMs, ML, and implementation protocol.

From the in-deep analysis of the 14 selected papers, we identified the most frequent applications of this hybrid approach in the four main branches of SB, as shown in Fig. 5. Specifically, the main applications of the hybrid approach range from the discovery of gene interaction networks and the generation of synthetic data, in genomics, the study of protein folding and the analysis of potential protein targets of specific drug therapies in proteomics, up to the reconstruction of the metabolic network, in metabolomics. Finally, hybrid approaches can be employed to describe cellular dynamics and the behavior of tumor cells.

### 3.1. Genomics

The regulation of genes is a cellular control mechanism that promotes the management and coordination of essential cellular functions. Understanding the intricate network of gene interactions underlying a biological process holds considerable implications in several fields of biology, pharmaceuticals, clinical research, and industry. SB model and reconstruct these interaction networks, thereby identifying all the connections between every single gene and any potential cross-talk between various pathways. DREAM is the most famous challenge in SB oriented to the inference of gene regulatory networks (GRNs). All participants propose new methodologies or strategies for network inference, exploiting and integrating all known knowledge in this field [123–126]. The first contribution presented by Carrè, [111], in which the authors proposed the Fasting Randomizing Algorithm for Network Knowledge (FRANK) to reconstruct large gene interaction networks, is aimed in this direction. GRNs thus produced contain all the characteristics supported by the literature, useful to generate synthetic gene expression data. The representation of the network via a direct graph is the starting point of FRANK. Specifically, each network encoded in a sparse matrix N consisting of two submatrixes *i)* the squared matrix A, with all the interactions of the type TF→TF, and *ii)* the non-squared matrix B, containing all the interactions of the type TF→TA. The interaction between two genes is represented in both the matrices through positive and negative values, i.e., representing the activation or inhibition action of a specific gene on another one. Given that the quantity of interactions is typically lower with respect to the genes, only a restricted set of positions in A exhibit non-zero values. These values are derived from $\mathcal{N}(\beta, 1)$, with $\beta$ a given parameter. Finally, the eigenvalues of matrix A are analyzed in order to assess the stability of the system, and matrix B is computed. After the definition of matrices A and B, the mathematical terms

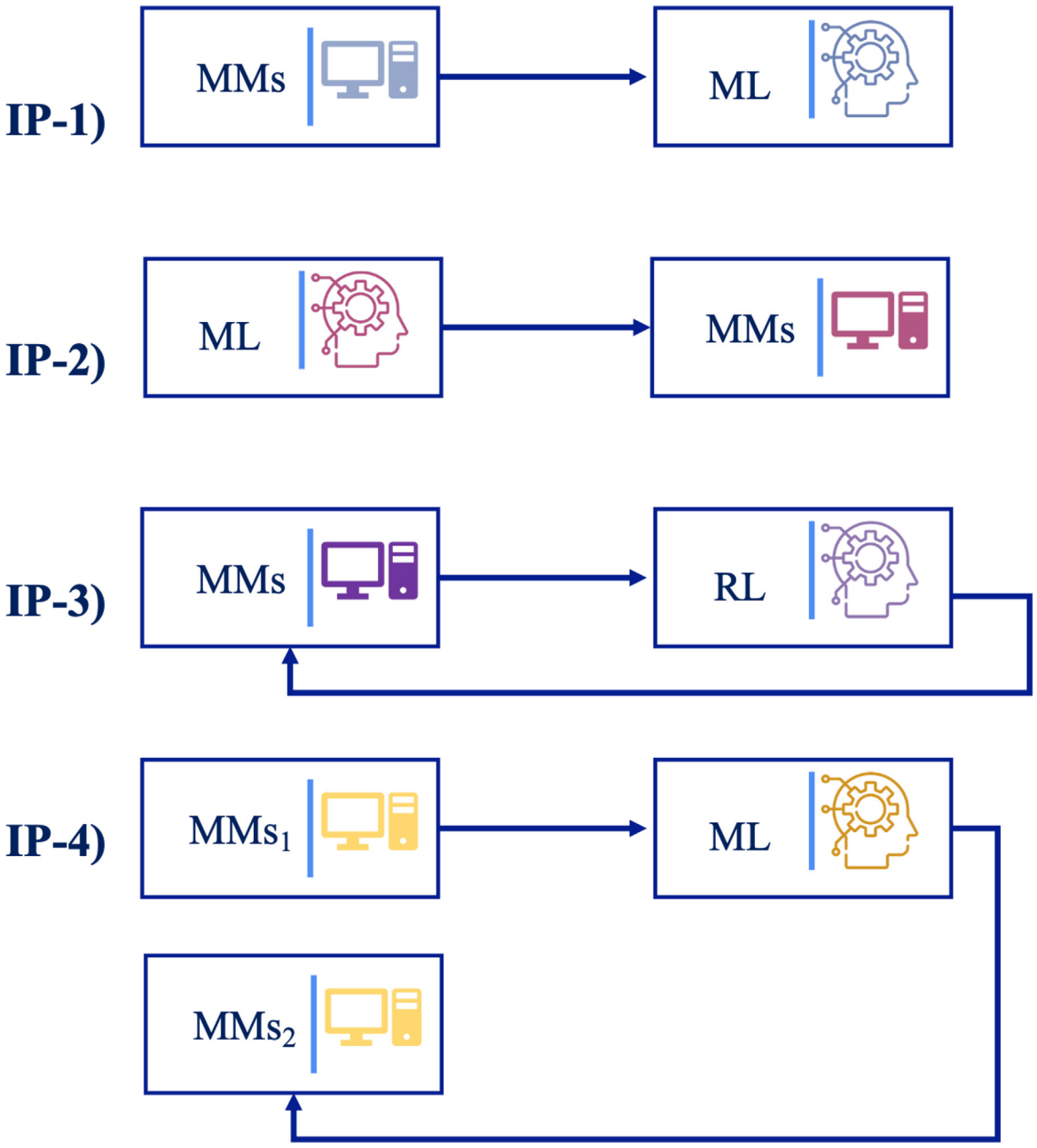


**Fig. 4.** Investigate Protocols (IPs) of the hybrid approaches present in the 14 selected papers. Especially, MMs are the acronym for Mechanistic Models (the subscript is indicative of two different types of MMs), while ML and RL mean Machine Learning and Reinforcement Learning, respectively.

describing the gene expression levels are generated. In particular, the vector X relative to the expression of genes related to TF and to TG target genes is determined as follows:

$$X_{TF}(t) = exp(V(t)) + \epsilon_{TF}(t)$$
$$X_{TG}(t) = exp(W(t)) + \epsilon_{TG}(t)$$

where both V(t) and W(t) represent the logarithmic expression of TF and TG respectively.

Based on the results produced by FRANK, SVM is applied to the previously simulated transcriptomic data and a set of connections in the network, known as “prior knowledge”, to benchmark the reconstructed GRN. Among the obtained results, the authors found that targets-oriented prior knowledge proved decisive for training SVM compared to TFs-oriented prior knowledge, and that SVM is resilient to typical errors that are often unavoidable in wet-lab experiments.

Still in the context of the modulation of gene expression by TFs, in [127], Miagoux et al. looked at the unidentified mechanism controlling the regulation of the primary TFs in rheumatoid arthritis (RA). Several genetic, epigenetic, and environmental factors are involved in this extremely complicated chronic inflammatory disease of the joints. The main goal was to offer a hybrid tool for the analysis of the success of therapy for each specific RA patient. The approach proposed by the authors consists of a first step in which an unsupervised ML technique was exploited to infer co-regulatory networks of TFs and target genes. Specifically, this step was performed by applying the CoRegNet R/Bioconductor package on the transcriptomics data and the TFs activity profile. This inferred co-regulatory network was enriched by information provided by a state-of-the-art disease map for RA. Then, genomic and transcriptomics data of treated RA patients were investigated to identify the key mutations associated with the response to anti-TNF treatment.

Finally, all this information was condensed into a mathematical system describing the dynamical behavior using Boolean formalism. As results, the authors found the implication of the IL6 and TGFB1 cascades as positive regulators of the expression of the TFs identified as master regulators.

On the other hand, Moore et al. present in their works [112,128] the Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) software solution, to generate synthetic data by emulating complex biological systems. This software combines a logical modeling approach with new and flexible ML techniques to achieve improved results. First of all, HIBACHI emulates all possible events, including post-translational changes and environmental factors, that may cause variations during the transcription and translation processes of a protein. In the second step, a logical model is used to describe how information is transferred from the genetic to the phenotypic level, taking into account all the variations defined in the first step. Then, HIBACHI allows for the definition of disease thresholds, based on the phenotype values generated in the previous step. Finally, the genetic programming (GP) module, mainly consists of binary trees (DT and RF), which provide a relatively simple way to generate variability in the solutions. As a proof-of-concept, the authors presented a study involving 2000 subjects split equally between samples and controls, as detailed in [128]. For each sample, 10 features, represented by an integer ranging from 0 to 2, to encode genetic variants, were collected. These data were subsequently exploited to detect and characterize high-order gene-gene interactions by analyzing the results of the multifactor dimensionality reduction technique, specifically looking at accuracy and information gain. Additionally, the authors ran several ML algorithms to demonstrate the utility of HIBACHI simulations for comparing 8 different models of epistatic interactions, achieving accuracies of about 85% for 2 out of 8 models. In [112], the work was extended to implement a three-objective fitness function aimed at optimizing the performance of one ML method with respect to another, while also reducing the complexity of the mathematical function that generates dataset labels. The simulations used in this work were different from those reported in the previous study. The results showed that it is possible to use HIBACHI to discover mathematical models that generate data for which LR, DTs, and RF perform differently.

In two of the three works, the hybrid approaches presented as applications in the genomics field result to adapt to the implementation scheme IP-1, while the other one fits the implementation scheme IP-2, as reported in Fig. 4.

### 3.2. Proteomics

In general, the analysis of a genome network alone is insufficient to explain pathophysiological phenotypes, since the knowledge of the transcription level of a certain gene is not sufficient to derive the concentration level of the associated protein, due to post-transcriptional regulation and alternative splicing mechanisms. Therefore, SB approaches are often applied to investigate the structure and dynamics of protein interaction networks using computational and statistical approaches. From the twelve works included in the study, we identified 4 contributions focusing on the application of hybrid techniques to protein interaction networks, highlighting the importance of integrating different computational and experimental methods.

For instance, Hua et al. proposed a data-driven modeling framework in [113] for the analysis of synthetic data generated from a mechanistic model, enabling a more comprehensive study of pathway behavior and identification of combinatorial targets for new therapies. In their work, the authors focused on the apoptotic pathway induced by Fas, which plays a critical role in various cellular processes and is implicated in the development of cancer or autoimmune diseases. Upon activation by the Fas ligand (FasL), this pathway triggers the apoptotic process by activating Casp3. However, the pathway is regulated by various proteins, such as XIAP or FLIP, which serve as inactivators of the apoptotic process by sequestering Casp3 or inhibiting its production. The proposed framework encompasses two main steps: i) the creation of a synthetic dataset by exploiting the mechanistic model describing the dynamics underpinning the activation of the apoptotic pathway, and ii) the application of ML techniques to investigate the relationship between the different concentrations of intermediate proteins and different phenotypes (healthy and sick). The mathematical model used in this framework comprises 22 biochemical reactions involving 11 different species. One species is kept constant at its initial value since it represents the input to the model (FasL), while one species represents the output (Casp3). The remaining nine species represent the intermediate proteins whose effects are to be investigated. The reactions are represented using simple mass-action (eq(3)) and transport (eq (4)) equations:

$$A + B \longleftrightarrow C \tag{3}$$

$$A \overset{k_f B}{\rightarrow} A^* \tag{4}$$

The mechanistic model, which consists of 22 biochemical reactions involving 11 different species, is described by ODEs that track the variations in the concentrations of species and complexes. To simplify the parameter estimation process, the original model was simplified and then used to generate synthetic data by running Monte Carlo simulations with different initial conditions for sets of species. The large amount of synthetic data obtained (about one million simulations) was clustered using k-means algorithms into three distinct classes based on the sensitivity of free-cleaved Casp3 production (insensitivity, medium sensitivity, and sensitivity). The authors then applied DT to the clustered dataset to identify potential links between the molecules inside the pathway that could explain the classifications. DT predicts XIAP and Fas as key components for the insensitive response on the production of Casp3 and was validated on a new simulated Monte Carlo dataset, achieving 71% prediction accuracy.

Chua et al. presented MASCOT (Machine Learning-based Prediction of Synergistic COmbinations of Targets) in [52], a technique that is capable of discovering any combination of targets from curated signaling networks and a desired therapeutic effect. The approach consists of two phases: first, MASCOT generates the candidate targets combination, and second, this combination is tested to simulate the related effects. The effectiveness of this approach was tested on the HRG-induced MAPK-PI3K signaling network, which is involved in several types of cancer, especially in proliferation and tumorigenesis. The signaling network is represented as an oriented hypergraph G = (V, E), which is then converted into a bipartite graph. Each species (node) is modeled through an ODE, and each reaction (edge) involving the species represents a contribution to the current ODE. This matrix is the input to MASCOT. Here, an ML-based step was designed to prioritize targets in signaling networks with respect to a disease node. To achieve this, a network-centric, ML-based approach called TAPESTRY, [129], was used. TAPESTRY deploys Tenet, [130], a recently proposed target characterization technique, which operates by using an SVM-based strategy aimed at (a) learning offline the optimal set of predictive topological features for characterizing known curated targets in a set of publicly-available signaling networks (such as the MAPK-PI3K signaling network), and (b) generating a set of characterization models based on these features. Once the characterization models and the disease-related signaling network with unknown targets of interest ("unseen network") have been obtained, TAPESTRY selects the "best" characterization model - which it should be adopted as its prioritization model - from the collection of characterization mod-

els of the candidate networks. TAPESTRY then calculates a prioritization score derived from the selected model and the dynamics of the unseen network, which is used to prioritize candidate targets. The simulation step utilizes the ODE model of the biological network, which is based on mass-action kinetics to model the production and consumption of the species. In this step, MASCOT modifies the matrix G to define a new derived-ODE system that simulates the effects of the target combination proposed by MASCOT. Later, given a signaling network G, a set of prioritized node rank W generated by TAPESTRY, a desired therapeutic effect $\zeta_{th}$ and the required combination size $S$, MASCOT identifies a set of synergistic target combinations $R$ which satisfies $\zeta_{th}$ and has minimal off-target effects $\zeta_{off}$. The inputs G and W are used to modify the drug targets and target activities. In addition, G is also used to simulate the target combination effects. Comparing the effects of heuristics with other implemented approaches, the Authors found that the triplet MASCOT-TAPESTRY-LOEWE (where the last one is a strategy/theory to seek for a combination index as a measure of the interaction effect between drugs in a combination) obtained the lowest off-target values and the best solutions set combinations, indicating the triplet could be useful as a guide for the discovery of potential targets and the evaluation of new target combinations.

In line with the previous studies, Oguz et al., [114], proposed a framework to investigate the implication of the network topology on the prediction of the protein abundance in the yeast. To this aim, a metaheuristic method based on differential evolution was used to explore a widening range of parameter vectors, finding two top-performing schemes. Specifically, one of these two models was more robust following parametric perturbations, and it was exploited to predict the phenotypes of 129 mutants, 86 of which were viable. The authors ranked the cell cycle proteins based on their contributions to the cumulative variability of relative protein abundance predictions: they found significant differences in contribution to the predictive variability of both proteins and modules of the cell cycle. Finally, they investigated the patterns generated from these predictions, and they assessed them through ML algorithms, such as RF, obtaining AucRoc values ranging from 0.82 to 0.88.

While in the previous three articles the authors investigated the protein interaction networks through hybrid approaches, Matsunaga & Sugita, in [115], studied the protein conformational dynamics to understand the incidence of the folding on the development of specific diseases. The main challenge in studying folding dynamics is represented by the gap between simulated and experimental data, due to the problematic repeatability of the energy balance between folded and unfolded states. To this aim, the authors proposed a new approach based on Markov State Model (MSM) to statistically approximate the long-time dynamics beyond the transitions between different discrete conformation states. The starting point is represented by the generation of a modular protein domain, WW domain, or WWP repeating motif, from nuclear magnetic resonance (NMR). These domains play a crucial role in folding since they mediate specific interactions with protein ligands. Using molecular dynamics (MD) simulations, they performed different simulations at different temporal intervals in order to generate the initial input of MSM as low-dimensional time-series data, by sampling regions between the two states. The transition probability between folded and unfolded states at $\tau$ is reported in the transition-probability matrix, $T(\tau)$, obtained by the results of MD simulations. Since this matrix could be affected by uncertainties or biases, two-step of ML is performed to refine the matrix $T(\tau)$, linking simulations and experimental data of photon trajectories. The two-step consists of: i) *Supervised-learning step*, where the $T(\tau)$ is estimated by counting transitions between the states; and ii) *Unsupervised-learning step*, where $T(\tau)$ is refined in order to reproduce the time-series data. The authors found that both steps were helpful in achieving novel insights into the folding mechanisms of the formin-binding protein WW domain. Specifically, the supervised learning step allowed them to tune the optimal number of states and the lag time of the transition-probability matrix for this specific problem, while the unsupervised learning step helped to improve $T(\tau)$ by incorporating previous knowledge acquired from single-molecule Forster resonance energy transfer data. Following the final step, the authors evaluated the data-assimilated MSM and obtained results consistent with independent experimental mutagenesis data. Additionally, they demonstrated that the strategy was robust to variations in the model of single-molecule experiments, particularly the Forster radius $R_o$. Based on the outcome of the investigation, the authors suggested that the data-assimilated MSM pathway could be used to improve force-field parameters and to understand the conformational transitions in proteins, nucleic acids, and other biomolecules.

Finally, Derbalah et al., in [116], propose a hybrid pipeline combining ODEs and ANNs to approximate and reduce the complexity of high-dimensional systems models. The authors tested their hybrid pipeline on the coagulation processes, describing both the in-vivo and in-vitro characteristics of this process. Concerning the type of mechanistic model employed in this work, the proposed pipeline is based on the so-called quantitative systems pharmacology (QSP) model describing a particular biological process through ODEs, with 62 state variables, and as many ODEs, and 184 and 188 parameters, respectively for in-vivo and in-vitro models. Given the relevant size of the model, ANNs have been employed in order to propose the right order reduction to be applied to the model. In particular, the original QSP was exploited to generate a conspicuous set of synthetic data, employed to train the tested ANNs. All the results, as well as the performances of several ANNs, were validated by computing the mean square error (MSE).

Concerning the implementation protocols, 2 of the 5 selected papers for proteomics follow the scheme IP-1, 2 follow the IP-4, while the last one follows the IP-2, as reported in Table 4 and shown in Fig. 4.

### 3.3. Metabolomics

Metabolism plays a crucial role in cellular life since it ensures the production of all the components required for the biosynthesis of vital building blocks such as amino acids, fatty acids, and nucleotides needed for cellular growth, as well as cellular-specific metabolism, [131]. Consequently, the study of metabolism is an additional application field for SB: MMs and ML are frequently used to simulate and investigate the metabolic behavior of complex biological systems. The combined effort between ML and MMs could be advantageous, for instance, in identifying latent phenomena in the metabolic network. In this regard, Biba et al., in [117], proposed a hybrid framework PRogramming In Statistical Modelling (PRISM), to model the aromatic amino acid metabolic pathway of the yeast. The authors chose to depict the analyzed metabolic network as a graph, where the nodes represent the enzyme-catalyzed reactions. In particular, each reaction, capable to convert two or more reacting metabolites into one or more products, was easily described through first-order logic representation. However, the above-mentioned representation is not enough to provide useful information on which of the reactions is the most likely, in a complex metabolic network. Furthermore, the probabilities of the reactions depend on several factors, such as e.g., environmental, chemical, and/or physical factors, and on the availability of metabolites, which in turn outline different metabolic scenarios. To address this limitation, the authors integrated ML methods into the PRISM framework to learn the probability distribution from observations, which were interpreted as model parameters. The accuracy of the

estimated probabilities was evaluated using the Root Mean Square Error (RMSE). The approach was tested on two different networks, a first Network which had no alternative branches in the pathways leading from node one, and a second one, which included an alternative path. In both cases, the results showed good accuracy and scalability.

Szappanos et al., in [118], proposed a new approach that systematically integrates measurements of genetic interactions between pairs of metabolic genes and simulated data to bridge the gap between theoretical models and experimental results. The proposed pipeline was tested and validated on the metabolic network of the yeast. Firstly, the genetic interaction map of the yeast metabolism was reconstructed from the large-scale synthetic genetic array (SGA). An interaction score, computed as $\epsilon = f_{12} - f_1 * f_2$ ($f_{12}$, and $f_1 * f_2$ represent double, and the single-mutant fitness, respectively), was assigned to each tested pair of genes, while a significant threshold was chosen to identify the type of interaction between pairs of genes (positive or negative). All the information previously obtained was employed to create a constraint-based model of the analyzed metabolic network, and a Flux Balance Analysis (FBA) was performed to simulate and make predictions on the interaction between pairs of genes. Finally, an ML method (based on RF) was developed to automatically generate hypotheses to explain the *in-vivo* compensation between genes. In this step, several changes were suggested to improve the fit of the model, including i) modifying the reversibility of reactions, ii) removing reactions, and iii) modifying the metabolic compounds involved in biomass production. As a result, they achieved an increase in Recall of 267% and an increase in Precision of 59%. Concerning the analyzed case study, an example of the proposed modification consist of the omission of glycogen from the compounds of biomass production, in accordance with the role of glycogen as an important energetic reserve in case of nutrient deficiency or cellular stress. Glycogen omission allowed the correction of two falsely predicted genetic interactions. Similarly, the suggestion of the removal of the two-step aspartate in the quinolinate pathway implicated in the NAD production allowed the correction of the prediction of four negative interactions between alternative NAD biosynthetic pathways in yeast.

Still, in the context of metabolic networks, Medlock & Papin, in [119], presented Automated Metabolic Model Ensemble-Driven Elimination of Uncertainty with Statistical learning (AMMEDEUS), to improve the quality of the reconstructed genome-scale metabolic models (GEMs). The proposed software drives the effective and targeted phase of curation of the GEMs, through different phases ranging from the creation of the model to its simulation, up to the application of supervised and unsupervised ML techniques. AMMEDEUS was tested on the reconstructed GEMs of 29 bacterial species. As in the previously presented works, the first step involves the creation of consistent GEMs from experimental data, generating an ensemble of models. These models were then used to simulate several biological scenarios and produce data. A following unsupervised learning step was performed on the simulated data to define the phenotypic clusters of the ensemble models, based on the similarity between simulated profiles. A subsequent supervised learning step was performed to predict cluster membership for a specific model, by using the variable parameters as input in the model. These two steps of ML (supervised and unsupervised) allow us to identify the structural variations that most influence the simulations; consequently, the authors propose these two steps of ML as crucial in the phase of curation of the GEMs. Based on their research, the authors emphasized the potential of the proposed framework, as well as the possibility of optimizing it to investigate how different supervised ML models affect the explainability of feature importance. This could lead to differences in the curation of GEMs.

Maeda et al., in [120] proposed a new method for estimating the Michaelis constant ($K_M$) in kinetic models. The devised method, called MLAGO (Machine Learning-Aided Global Optimization), combines ML and global optimization techniques to improve the accuracy and efficiency of $K_M$ estimation. In particular, ML was exploited to predict valid values for kinetic parameters from a curated dataset derived from the BRENDA database, [132]. Five different ML methods were tested and their performance was evaluated in terms of RMSE, getting the best result with the RF (RMSE=0.795), and testing it on the ODE models describing the metabolism of the carbon and nitrogen. The authors demonstrated the effectiveness of MLAGO by comparing its performance to other commonly used methods, showing that it outperforms them in terms of both accuracy and speed. The obtained results underline the efficiency of MLAGO as a potential tool to significantly improve the accuracy and reliability of kinetic modeling in biochemical research.

In the context of metabolomics applications, 3 papers, of the 4 selected, follow the implementation protocol IP-1, while the remaining one follows the IP-2, as reported in Table 4 and shown in Fig. 4.

### 3.4. Cellular dynamics

Models and simulations are also powerful tools for understanding the complex and dynamic functions of cells. Several modeling approaches have been exploited to investigate cellular biophysical characteristics and functions.

In this context, Sieburg et al. [121] addressed the challenge of estimating the life span of clones derived from hematopoietic stem cells (HSCs) and predicting their long-term performance using a model-based ML algorithm. HSCs generate a clone that maintains stemness and differentiates into the progeny, and the life span of the clone depends on the self-renewal capacity of the HSCs. To evaluate this capacity, the authors identified the percentage of donor-type cells (% DT) in the blood as the crucial predictor. Based on the shape of the experimental curve, the authors proposed the following Weibull model to predict the clonal life span:

$$D(t) = bt - at^{\alpha} \tag{5}$$

In eq. (5), $b$ and $a$ represent the average rate of cellular growth and loss, respectively, whereas $\alpha$ is the slope rate of the curve. Enforcing $D(T) = 0$, this equation can be exploited to directly evaluate the life span, T, as follows:

$$T = \frac{b}{a^{1/(\alpha-1)}} \tag{6}$$

A total of $10^7$ different combinations of parameters were tested and used to compute a specific life-span curve. Then, each simulated curve was compared with experimental data using the $\chi^2$ distance metric. Reinforcement learning (RL) was employed to train the search for high-quality configurations in a large Monte Carlo database of allowable combinations, resulting in an increase in the rate of correct predictors from 65% to 83%. Subsequently, the authors investigated the parameters that affect the life span of HSCs by exploiting a cellular model previously developed [133]: each simulated cell was characterized by the following set of parameters ($c$, $w$, $\tau$, $\theta$), where $c$ represents the type of cell (1 for HSC, and 2 for the differentiated ones, DIF), $w$ and $\tau$, its proliferative capacity and resistance to the differentiation, respectively. In contrast, $\theta$ considered the previous cellular divisions: these parameters were used to define *proliferation*, and *differentiation* rules. Several functions were tested to model and assess the decay motif related to the fate decision. The results allowed the authors to confirm the implication of self-renewal as a crucial parameter in determining the life span of the HSCs clones. Furthermore, the authors assessed

the intrinsic nonlinear synchronicity of the decay of the HSCs, excluding any random mechanism underlying this process.

The approach proposed by Sieburg et al. is just one example of the application of the hybrid approach to investigate particular aspects of cellular life. On the other hand, Zangooei et al., in [122], proposed a multiscale approach to investigate and predict the microvascular growth of tumor masses. Tumor growth consists of i) an initial avascular phase, and ii) a belated malignant vascular phase. The transition between these two phases was guided by the angiogenesis process, responsible for generating new blood vessels to meet the demand for nutrients necessary for tumor growth. The proposed computational model uses features extracted from contrast-enhanced micro-computed tomography images to simulate several phenomena related to breast cancer growth. The core of the approach consists of an ABM describing both cells and vessels as separate agents, which together form the tumor environment. The interaction between cells and vessels is modeled through a diffusion equation. Deep reinforcement learning (DRL) techniques allow the agents to learn in a specific environment during the training phase, generating data that are exploited in the testing phase, where the multiscale model predicts the best action using a neural network. For this study, the dataset generated during the training phase considered several attributes of cells and vessels, mainly related to the tumor microenvironment and their status (e.g., hypoxic, necrotic), which were used as the input layer (three nodes) of the neural network during the testing phase. Meanwhile, the output layer, consisting of five nodes, was set to the number of cell and vessel attributes. Pearson correlation analysis was used to compare simulated tumor growth data and results obtained using independent mathematical models.

Both the articles selected for the Cellular Dynamics section follow the implementation protocol IP-3, as in Fig. 4.

## 4. Discussions & conclusions

Simulations based on mechanistic mathematical models, as methods of artificial intelligence, represent two of the most promising and widely employed tools in SB. Individually, these two approaches present some pitfalls, which limit their applicability in specific contexts. An increasing number of scientific works suggest that such problems can be overcome by properly combining these two approaches. Hybrid approaches are highly promising since they can leverage, on the one hand, the explainability of mathematical models and the possibility to easily carry out numerical simulations (often named *in-silico* experiments in the biological context) and, on the other hand, the effectiveness of machine learning techniques in interpreting large amounts of experimental data.

In the present systematic review, we have analyzed the main investigation paradigms and fields of application of this hybrid approach in SB. In particular, from a detailed analysis of the literature on the subject, 14 research articles were selected, which present different applications of hybrid MMs-ML approaches. The analysis of these documents allowed us to identify genomics, proteomics, metabolomics, and cellular dynamics as the main application fields. Specifically, in the selected contexts, the hybrid MMs-ML techniques were mainly used to improve the reconstruction of gene, protein, and metabolic interaction networks, as well as for the development of robust tools to generate reliable *in-silico* data and to investigate the dynamics underlying the growth and differentiation of several types of cells.

Furthermore, according to the implementation protocols highlighted in [16], the selected works have been classified into 4 main classes shown in Fig. 4, featuring different combinations of the MMs and ML analysis phases. A particularly interesting result extracted from this study is related to the preferential hybrid schemes of the micro (genomics, proteomics, and metabolomics) and the macro (cellular dynamics) fields, for which most of the selected works present the implementation protocols IP-1 and IP-2, respectively, as summarized in Table 4. This demonstrates that, in the genomics, proteomics, and metabolomics fields, mechanistic models can be effectively exploited as means to generate large amounts of synthetic data that, in turn, can be leveraged to tune ML-based algorithms for the reconstruction of interaction networks. On the other hand, concerning the modeling of cellular dynamics, there is a tendency to prefer hybrid protocols based on reinforcement learning. These approaches, mostly relying on agent-based modeling, are inspired by biological evolution and functioning principles, where the dynamics of each cell are driven by the surrounding environment and the local interaction with other cells.

Finally, our hope is that this work may provide a valuable starting point to spark the interest of those researchers involved in both mechanistic modeling of biological systems and ML techniques applied to biological data and encourage the development of novel effective hybrid investigation frameworks in the field of SB. The present systematic review is part of a larger research project, [134,135], focused on the development of hybrid techniques that can exploit the robustness and reliability of MMs and the capability to learn from data typical of ML, to design robust pipelines for the reconstruction and the assessment of biological networks (e.g. GEMs) and improving clinical-decisions making.

## 5. Acronyms

ABMs: Agent-based Models; AMs: Algebraic Models; ANN: Artificial Neural Network; CBMs: Constraint-based Models; DT: Decision Tree; GEM: Genome-scale Metabolic Model; GRN: Gene Regulatory Network; HMM: Hidden Markov Model; IP: Investigate Protocol; LMs: Logical Models; ML: Machine Learning; MMs: Mechanistic Models; ODE: Ordinary Differential Equations; RF: Random Forest; RL: Reinforcement Learning; SB: Systems Biology; SL: Supervised Learning; SVM: Support Vector Machine; UL: Unsupervised Learning.

## 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] E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, Systems Biology: A Textbook, John Wiley & Sons, 2016.
[2] R.A. Meyers, Systems Biology, John Wiley & Sons, 2012.
[3] H.-Y. Chuang, M. Hofree, T. Ideker, A decade of systems biology, Annu. Rev. Cell Dev. Biol. 26 (2010) 721–744.
[4] H.C. Yeo, K. Selvarajoo, Machine learning alternative to systems biology should not solely depend on data, Brief. Bioinform. 23 (6) (2022) bbac436.
[5] T. Huang, J. Zhang, Z.-P. Xu, L.-L. Hu, L. Chen, J.-L. Shao, L. Zhang, X.-Y. Kong, Y.-D. Cai, K.-C. Chou, Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches, Biochimie 94 (4) (2012) 1017–1025.
[6] R.C. Eccleston, S. Wan, N. Dalchau, P.V. Coveney, The role of multiscale protein dynamics in antigen presentation and t lymphocyte recognition, Front. Immunol. 8 (2017) 797.
[7] O. Chang, F.A. Gonzales-Zubiate, L. Zhinin-Vera, R. Valencia-Ramos, I. Pineda, A. Diaz-Barrios, A protein folding robot driven by a self-taught agent, BioSystems 201 (2021) 104315.
[8] D. Machado, R.S. Costa, M. Rocha, E.C. Ferreira, B. Tidor, I. Rocha, Modeling formalisms in systems biology, AMB Express 1 (1) (2011) 1–14.
[9] M. Alber, A. Buganza Tepole, W.R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W.W. Lytton, P. Perdikaris, L. Petzold, et al., Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences, NPJ Digital Medicine 2 (1) (2019) 1–11.
[10] G.C.Y. Peng, M. Alber, A. Buganza Tepole, W.R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W.W. Lytton, P. Perdikaris, et al., Multiscale

modeling meets machine learning: What can we learn? Arch. Comput. Methods Eng. 28 (2021) 1017–1037.
[11] R.E. Baker, J.-M. Pena, J. Jayamohan, A. Jérusalem, Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol. Lett. 14 (5) (2018) 20170660.
[12] H. Subramanian, Combining scientific computing and machine learning techniques to model longitudinal outcomes in clinical trials, 2021, Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-176427.
[13] G. Yauney, P. Shah, Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection, in: Machine Learning for Healthcare Conference, PMLR, 2018, pp. 161–226.
[14] P. Desai, A. Sealy, S. Tedman, L. Wright, B. Biller, How simulation will impact the future of healthcare and life sciences, Proc. PharmaSUG (2021).
[15] L. von Rueden, S. Mayer, R. Sifa, C. Bauckhage, J. Garcke, Combining machine learning and simulation to a hybrid modelling approach: current and future directions, in: Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18, Springer, 2020, pp. 548–560.
[16] A. Greasley, Architectures for combining discrete-event simulation and machine learning, in: 10th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, SIMULTECH 2020, SciTePress, 2020, pp. 47–58.
[17] K. Sharafutdinov, S.J. Fritsch, M. Iravani, P.F. Ghalati, S. Saffaran, D.G. Bates, J.G. Hardman, R. Polzin, H. Mayer, G. Marx, et al., Computational simulation of virtual patients reduces dataset bias and improves machine learning-based detection of ARDS from noisy heterogeneous ICU datasets, IEEE Open J. Eng. Med. Biol. (2023).
[18] M.J. Page, D. Moher, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, et al., Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, Br. Med. J. 372 (2021).
[19] M.J. Grant, A. Booth, A typology of reviews: an analysis of 14 review types and associated methodologies, Health Inform. Libraries J. 26 (2) (2009) 91–108.
[20] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L.A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-p) 2015 statement, Syst. Rev. 4 (1) (2015) 1–9.
[21] K. Alden, J. Cosgrove, M. Coles, J. Timmis, Using emulation to engineer and understand simulations of biological systems, IEEE/ACM Trans. Comput. Biol. Bioinf. 17 (1) (2018) 302–315.
[22] L. Salerno, C. Cosentino, A. Merola, D.G. Bates, F. Amato, Validation of a model of the GAL regulatory system via robustness analysis of its bistability characteristics, BMC Syst. Biol. 7 (2013), doi:10.1186/1752-0509-7-39. Article n. 39.
[23] L. Salerno, C. Cosentino, G. Morrone, F. Amato, Computational modeling of a transcriptional switch underlying b-lymphocyte lineage commitment of hematopoietic multipotent cells, PLoS ONE 10 (7) (2015) e0132208.
[24] E.I. Parrotta, A. Procopio, S. Scalise, C. Esposito, G. Nicoletta, G. Santamaria, M.T. De Angelis, T. Dorn, A. Moretti, K.-L. Laugwitz, et al., Deciphering the role of wnt and rho signaling pathway in iPSC-derived ARVC cardiomyocytes by in silico mathematical modeling, Int. J. Mol. Sci. 22 (4) (2021) 2004.
[25] E. Kim, J.-Y. Kim, J.-Y. Lee, Mathematical modeling of p53 pathways, Int. J. Mol. Sci. 20 (20) (2019) 5179.
[26] C. Cosentino, R. Ambrosino, M. Ariola, M. Bilotta, A. Pironti, F. Amato, On the realization of an embedded subtractor module for the control of chemical reaction networks, IEEE Trans. Automat. Contr. 61 (11) (2016) 3639–3643.
[27] A. Procopio, S. De Rosa, M.R. García, C. Covello, A. Merola, J. Sabatino, A. De Luca, C. Indolfi, F. Amato, C. Cosentino, Experimental modeling and identification of cardiac biomarkers release in acute myocardial infarction, IEEE Trans. Control Syst. Technol. 28 (1) (2018) 183–195.
[28] A. Procopio, S. De Rosa, C. Covello, A. Merola, J. Sabatino, A. De Luca, C. Liebetrau, C.W. Hamm, C. Indolfi, F. Amato, et al., Estimation of the acute myocardial infarction onset time based on time-course acquisitions, Ann. Biomed. Eng. 49 (2021) 477–486.
[29] A. Procopio, M. Bilotta, A. Merola, F. Amato, C. Cosentino, S. De Rosa, C. Covello, J. Sabatino, A. De Luca, C. Indolfi, Predictive mathematical model of cardiac troponin release following acute myocardial infarction, in: 2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC), IEEE, 2017, pp. 643–648.
[30] S. Bakshi, V. Chelliah, C. Chen, P.H. van der Graaf, Mathematical biology models of parkinson's disease, CPT: Pharmacometric. Syst. Pharmacol. 8 (2) (2019) 77–86.
[31] F. Montefusco, A. Procopio, I.M. Bulai, F. Amato, M.G. Pedersen, C. Cosentino, Interacting with COVID-19: How population behavior, feedback and memory shaped recurrent waves of the epidemic, IEEE Control Syst. Lett. 7 (2022) 583–588.
[32] N.P. Stavros, P. Colombo, G. Colombo, M.R. Dimitrios, Design of experiments (doe) in pharmaceutical development, Drug. Dev. Ind. Pharm. 43 (6) (2017) 889–901.
[33] Y. Fan, O. Pedersen, Gut microbiota in human metabolic health and disease, Nat. Rev. Microbiol. 19 (1) (2021) 55–71.
[34] N.T. Devika, K. Raman, Deciphering the metabolic capabilities of bifidobacteria using genome-scale metabolic models, Sci. Rep. 9 (1) (2019) 1–9.
[35] J.H. Gunter, M. Kruithof-de Julio, E. Zoni, Personalized medicine for urological cancers: Targeting cancer metabolism, Front. Oncol. 12 (2022).
[36] G. An, Q. Mi, J. Dutta-Moscato, Y. Vodovotz, Agent-based models in translational systems biology, Wiley Interdiscip. Rev. Syst. Biol. Med. 1 (2) (2009) 159–171.
[37] E. Merelli, G. Armano, N. Cannata, F. Corradini, M. d'Inverno, A. Doms, P. Lord, A. Martin, L. Milanesi, S. Möller, et al., Agents in bioinformatics, computational and systems biology, Brief. Bioinformatics 8 (1) (2007) 45–59.
[38] C.M. Macal, M.J. North, Agent-based modeling and simulation, in: Proceedings of the 2009 winter simulation conference (WSC), IEEE, 2009, pp. 86–98.
[39] M. Ghadiri, M. Heidari, S.-A. Marashi, S.H. Mousavi, A multiscale agent-based framework integrated with a constraint-based metabolic network model of cancer for simulating avascular tumor growth, Mol. Biosyst. 13 (9) (2017) 1888–1897.
[40] Z. Wang, J.D. Butner, R. Kerketta, V. Cristini, T.S. Deisboeck, Simulating cancer growth with multiscale agent-based modeling, in: Seminars in cancer biology, volume 30, Elsevier, 2015, pp. 70–78.
[41] Z. Heidary, J. Ghaisari, S. Moein, S. Haghjooy Javanmard, The double-edged sword role of fibroblasts in the interaction with cancer cells; an agent-based modeling approach, PLoS ONE 15 (5) (2020) e0232965.
[42] A. Corti, M. Colombo, F. Migliavacca, J.F. Rodriguez Matas, S. Casarin, C. Chiastra, Multiscale computational modeling of vascular adaptation: a systems biology approach using agent-based models, Front. Bioeng. Biotechnol. (2021) 978.
[43] M.L. Wynn, N. Consul, S.D. Merajver, S. Schnell, Logic-based models in systems biology: a predictive and parameter-free network analysis method, Integr. Biol. 4 (11) (2012) 1323–1337.
[44] R. Franke, M. Müller, N. Wundrack, E.-D. Gilles, S. Klamt, T. Kähne, M. Naumann, Host-pathogen systems biology: logical modelling of hepatocyte growth factor and helicobacter pylori induced c-met signal transduction, BMC Syst. Biol. 2 (2008) 1–17.
[45] A. Flobak, A. Baudot, E. Remy, L. Thommesen, D. Thieffry, M. Kuiper, A. Lægreid, Discovery of drug synergies in gastric cancer cells predicted by logical modeling, PLoS Comput. Biol. 11 (8) (2015) e1004426.
[46] N. Malekjani, S.M. Jafari, Modeling the release of food bioactive ingredients from carriers/nanocarriers by the empirical, semiempirical, and mechanistic models, Compr. Rev. Food Sci. Food Saf. 20 (1) (2021) 3–47.
[47] A. Procopio, E. Lagreca, R. Jamaledin, S. La Manna, B. Corrado, C. Di Natale, V. Onesto, Recent fabrication methods to produce polymer-based drug delivery matrices (experimental and in silico approaches), Pharmaceutics 14 (4) (2022) 872.
[48] G.A. Hughes, Nanostructure-mediated drug delivery, Nanomed. Nanotechnol. Biol. Med. 1 (1) (2005) 22–30.
[49] G.J. Melen, S. Levy, N. Barkai, B.-Z. Shilo, Threshold responses to morphogen gradients by zero-order ultrasensitivity, Mol. Syst. Biol. 1 (1) (2005). 2005–0028
[50] J.E. Ferrell Jr., S.H. Ha, Ultrasensitivity part i: Michaelian responses and zero-order ultrasensitivity, Trends Biochem. Sci. 39 (10) (2014) 496–503.
[51] W. Gilpin, Y. Huang, D.B. Forger, Learning dynamics from large biological data sets: machine learning meets systems biology, Curr. Opin. Syst. Biol. 22 (2020) 1–7.
[52] H.E. Chua, S.S. Bhowmick, L. Tucker-Kellogg, Synergistic target combination prediction from curated signaling networks: Machine learning meets systems biology and pharmacology, Methods 129 (2017) 60–80.
[53] S.H. Muggleton, Machine learning for systems biology, in: International Conference on Inductive Logic Programming, Springer, 2005, pp. 416–423.
[54] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, Nature Rev. Phys. 3 (6) (2021) 422–440.
[55] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686–707.
[56] S. Suthaharan, S. Suthaharan, Support vector machine, Mach. Learn. Model. Algor. Big Data Classificat.: Think. Exampl. Effect. Learn. (2016) 207–235.
[57] S.R. Eddy, Hidden markov models, Curr. Opin. Struct. Biol. 6 (3) (1996) 361–365.
[58] Y. Freund, L. Mason, The alternating decision tree learning algorithm, in: icml, volume 99, 1999, pp. 124–133.
[59] W. Jin, Z.J. Li, L.S. Wei, H. Zhen, The improvements of BP neural network learning algorithm, in: WCC 2000-ICSP 2000. 2000 5th international conference on signal processing proceedings. 16th world computer congress 2000, volume 3, IEEE, 2000, pp. 1647–1649.
[60] W.S. Noble, What is a support vector machine? Nat. Biotechnol. 24 (12) (2006) 1565–1567.
[61] A. Singh, N. Thakur, A. Sharma, A review of supervised machine learning algorithms, in: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), Ieee, 2016, pp. 1310–1315.
[62] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 161–168.
[63] B. Naul, A review of support vector machines in computational biology, Protein Similarit. and (2009).
[64] W.S. Noble, et al., Support vector machine applications in computational biology, Kernel Methods Comput. Biol. 71 (2004) 92.
[65] A. Ben-Hur, C.S. Ong, S. Sonnenburg, B. Schölkopf, G. Rätsch, Support vector machines and kernels for computational biology, PLoS Comput. Biol. 4 (10) (2008) e1000173.

[66] L. Liao, W.S. Noble, Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships, J. Comput. Biol. 10 (6) (2003) 857–868.
[67] K.J. Cios, N. Liu, A machine learning method for generation of a neural network architecture: a continuous ID3 algorithm, IEEE Trans. Neural Networks 3 (2) (1992) 280–291.
[68] J.R. Quinlan, C4. 5: Programs for Machine Learning, Elsevier, 2014.
[69] B. Choubin, H. Darabi, O. Rahmati, F. Sajedi-Hosseini, B. Kløve, River suspended sediment modelling using the CART model: A comparative study of machine learning techniques, Sci. Total Environ. 615 (2018) 272–281.
[70] L. Rokach, O. Maimon, Top-down induction of decision trees classifiers-a survey, IEEE Trans. Syst. Man Cybern., Part C (Appl. Rev.) 35 (4) (2005) 476–487.
[71] X. Daniela, C. Hinde, R. Stone, Naive bayes vs. decision trees vs. neural networks in the classification of training web pages, Int. J. Comput. Sci. Iss. 4 (2009).
[72] N.B. Amor, S. Benferhat, Z. Elouedi, Naive bayes vs decision trees in intrusion detection systems, in: Proceedings of the 2004 ACM symposium on Applied computing, 2004, pp. 420–424.
[73] A.C. Lorena, L.F.O. Jacintho, M.F. Siqueira, R. De Giovanni, L.G. Lohmann, A.C. De Carvalho, M. Yamamoto, Comparing machine learning classifiers in potential distribution modelling, Expert Syst. Appl. 38 (5) (2011) 5268–5275.
[74] P. Geurts, A. Irrthum, L. Wehenkel, Supervised learning with decision tree-based methods in computational and systems biology, Mol. Biosyst. 5 (12) (2009) 1593–1605.
[75] C.R. Williams-DeVane, D.M. Reif, E. Cohen Hubal, P.R. Bushel, E.E. Hudgens, J.E. Gallagher, S.W. Edwards, Decision tree-based method for integrating gene expression, demographic, and clinical data to determine disease endotypes, BMC Syst. Biol. 7 (2013) 1–19.
[76] S. Kharait, S. Hautaniemi, S. Wu, A. Iwabu, D.A. Lauffenburger, A. Wells, Decision tree modeling predicts effects of inhibiting contractility signaling on cell motility, BMC Syst. Biol. 1 (1) (2007) 1–13.
[77] Y. Koch, T. Wolf, P.K. Sorger, R. Eils, B. Brors, Decision-tree based model analysis for efficient identification of parameter relations leading to different signaling states, PLoS ONE 8 (12) (2013) e82593.
[78] G. D'Addio, L. Donisi, G. Cesarelli, F. Amitrano, A. Coccia, M.T. La Rovere, C. Ricciardi, Extracting features from poincaré plots to distinguish congestive heart failure patients according to NYHA classes, Bioengineering 8 (10) (2021) 138.
[79] L. Donisi, C. Ricciardi, G. Cesarelli, A. Coccia, F. Amitrano, S. Adamo, G. D'addio, Bidimensional and Tridimensional Poincaré Maps in Cardiology: A Multiclass Machine Learning Study, Electronics (Basel) 11 (3) (2022) 448.
[80] M. van Smeden, K.G.M. Moons, J.A.H. de Groot, G.S. Collins, D.G. Altman, M.J.C. Eijkemans, J.B. Reitsma, Sample size for binary logistic prediction models: beyond events per variable criteria, Stat. Methods Med. Res. 28 (8) (2019) 2455–2474.
[81] L. Donisi, G. Cesarelli, E. Capodaglio, M. Panigazzi, G. D'Addio, M. Cesarelli, F. Amato, A logistic regression model for biomechanical risk classification in lifting tasks, Diagnostics 12 (11) (2022) 2624.
[82] K. Zhang, W. Geng, S. Zhang, Network-based logistic regression integration method for biomarker identification, BMC Syst. Biol. 12 (9) (2018) 113–122.
[83] M.N. Abdullah, Y.B. Wah, A.B.A. Majeed, Y. Zakaria, N. Shaadan, Identification of blood-based multi-omics biomarkers for alzheimer's disease using firth's logistic regression, vol 30 (2022) 1197–1218.
[84] T. Hou, G. Zheng, P. Zhang, J. Jia, J. Li, L. Xie, C. Wei, Y. Li, Lacep: lysine acetylation site prediction using logistic regression classifiers, PLoS ONE 9 (2) (2014) e89575.
[85] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd, 2006.
[86] J.A. Botía, J. Vandrovcova, P. Forabosco, S. Guelfi, K. D'Sa, U.K.B.E. Consortium, J. Hardy, C.M. Lewis, M. Ryten, M.E. Weale, An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks, BMC Syst. Biol. 11 (2017) 1–16.
[87] K. Seymore, A. McCallum, R. Rosenfeld, et al., Learning hidden markov model structure for information extraction, in: AAAI-99 workshop on machine learning for information extraction, 1999, pp. 37–42.
[88] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[89] S.R. Eddy, What is a hidden markov model? Nat. Biotechnol. 22 (10) (2004) 1315–1316.
[90] J. Henderson, S. Salzberg, K.H. Fasman, Finding genes in DNA with a hidden markov model, J. Comput. Biol. 4 (2) (1997) 127–141.
[91] E. Birney, Hidden markov models in biological sequence analysis, IBM J. Res. Dev. 45 (3.4) (2001) 449–454.
[92] A. Krogh, M. Brown, I.S. Mian, K. Sjölander, D. Haussler, Hidden Markov Models in Computational Biology: Applications to Protein Modeling, J. Mol. Biol. 235 (5) (1994) 1501–1531.
[93] S.-C. Wang, Artificial neural network, in: Interdisciplinary computing in java programming, Springer, 2003, pp. 81–100.
[94] J. Vohradsky, Neural network model of gene expression, FASEB J. 15 (3) (2001) 846–854.
[95] L.M. Hall, D.W. Hill, L.C. Menikarachchi, M.-H. Chen, L.H. Hall, D.F. Grant, Optimizing artificial neural network models for metabolomics and systems biology: an example using HPLC retention index data, Bioanalysis 7 (8) (2015) 939–955.
[96] T. Ching, X. Zhu, L.X. Garmire, Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data, PLoS Comput. Biol. 14 (4) (2018) e1006076.
[97] M. Botvinick, S. Ritter, J.X. Wang, Z. Kurth-Nelson, C. Blundell, D. Hassabis, Reinforcement learning, fast and slow, Trends Cogn. Sci. (Regul. Ed.) 23 (5) (2019) 408–422.
[98] P. Ladosz, L. Weng, M. Kim, H. Oh, Exploration in deep reinforcement learning: A survey, Inform. Fusion (2022).
[99] R.K. Tan, Y. Liu, L. Xie, Reinforcement learning for systems pharmacology-oriented and personalized drug design, Expert Opin. Drug Discov. 17 (8) (2022) 849–863.
[100] S.K. Gottipati, B. Sattarov, S. Niu, Y. Pathak, H. Wei, S. Liu, S. Blackburn, K. Thomas, C. Coley, J. Tang, et al., Learning to navigate the synthetically accessible chemical space using reinforcement learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 3668–3679.
[101] J. Born, M. Manica, A. Oskooei, J. Cadow, G. Markert, M.R. Martínez, PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning, Iscience 24 (4) (2021) 102269.
[102] F. Zhu, Q. Liu, X. Zhang, B. Shen, Protein interaction network constructing based on text mining and reinforcement learning with application to prostate cancer, IET Syst. Biol. 9 (4) (2015) 106–112.
[103] M. Hossin, M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, Int. J. Data Min. Knowl. Manag. Process 5 (2) (2015) 1.
[104] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (3) (2005) 299–310.
[105] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, V. Koltun, Carla: An open urban driving simulator, in: Conference on robot learning, PMLR, 2017, pp. 1–16.
[106] T.M. Deist, A. Patti, Z. Wang, D. Krane, T. Sorenson, D. Craft, Simulation-assisted machine learning, Bioinformatics 35 (20) (2019) 4072–4080.
[107] K. Cranmer, J. Brehmer, G. Louppe, The frontier of simulation-based inference, Proc. Natl. Acad. Sci. 117 (48) (2020) 30055–30062.
[108] A. Sobester, A. Forrester, A. Keane, Engineering Design via Surrogate Modelling: A Practical Guide, John Wiley & Sons, 2008.
[109] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, et al., Solving rubik's cube with a robot hand, arXiv preprint arXiv:1910.07113 (2019).
[110] S. Bergmann, N. Feldkamp, S. Strassburger, Emulation of control strategies through machine learning in manufacturing simulations, J. Simul. 11 (1) (2017) 38–50.
[111] C. Carré, A. Mas, G. Krouk, Reverse engineering highlights potential principles of large gene regulatory network design and learning, npj Syst. Biol. Appl. 3 (1) (2017) 1–15.
[112] J.H. Moore, M. Shestov, P. Schmitt, R.S. Olson, A heuristic method for simulating open-data of arbitrary complexity that can be used to compare and evaluate machine learning methods, in: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium, World Scientific, 2018, pp. 259–267.
[113] F. Hua, S. Hautaniemi, R. Yokoo, D.A. Lauffenburger, Integrated mechanistic and data-driven modelling for multivariate analysis of signalling pathways, J. R. Soc. Interface 3 (9) (2006) 515–526.
[114] C. Oguz, L.T. Watson, W.T. Baumann, J.J. Tyson, Predicting network modules of cell cycle regulators using relative protein abundance statistics, BMC Syst. Biol. 11 (1) (2017) 1–24.
[115] Y. Matsunaga, Y. Sugita, Linking time-series of single-molecule experiments with molecular dynamics simulations by machine learning, Elife 7 (2018) e32668.
[116] A. Derbalah, H.S. Al-Sallami, S.B. Duffull, Reduction of quantitative systems pharmacology models using artificial neural networks, J. Pharmacokinet. Pharmacodyn. 48 (2021) 509–523.
[117] M. Biba, S. Ferilli, N.D. Mauro, T. Basile, A hybrid symbolic-statistical approach to modeling metabolic networks, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2007, pp. 132–139.
[118] B. Szappanos, K. Kovács, B. Szamecz, F. Honti, M. Costanzo, A. Baryshnikova, G. Gelius-Dietrich, M.J. Lercher, M. Jelasity, C.L. Myers, et al., An integrated approach to characterize genetic interaction networks in yeast metabolism, Nat. Genet. 43 (7) (2011) 656–662.
[119] G.L. Medlock, J.A. Papin, Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning, Cell Syst. 10 (1) (2020) 109–119.
[120] K. Maeda, A. Hatae, Y. Sakai, F.C. Boogerd, H. Kurata, Mlago: machine learning-aided global optimization for michaelis constant estimation of kinetic modeling, BMC Bioinform. 23 (1) (2022) 455.
[121] H.B. Sieburg, B.D. Rezner, C.E. Muller-Sieburg, Predicting clonal self-renewal and extinction of hematopoietic stem cells, Proc. Natl. Acad. Sci. 108 (11) (2011) 4370–4375.
[122] M.H. Zangooei, R. Margolis, K. Hoyt, Multiscale computational modeling of cancer growth using features derived from microCT images, Sci. Rep. 11 (1) (2021) 1–17.
[123] F. Liu, S.-W. Zhang, W.-F. Guo, Z.-G. Wei, L. Chen, Inference of gene regulatory network based on local bayesian networks, PLoS Comput. Biol. 12 (8) (2016) e1005024.
[124] M. Bansal, G.D. Gatta, D. Di Bernardo, Inference of gene regulatory networks and compound mode of action from time course gene expression profiles, Bioinformatics 22 (7) (2006) 815–822.
[125] F. Montefusco, A. Procopio, D.G. Bates, F. Amato, C. Cosentino, Scalable reverse-engineering of gene regulatory networks from time-course measurements, Int. J. Robust Nonlinear Control (2022) e1005024.

[126] Y. Xu, J. Chen, A. Lyu, W.K. Cheung, L. Zhang, DyndeepDRIM: A dynamic deep learning model to infer direct regulatory interactions using time-course single-cell gene expression data, Brief. Bioinformatics 23 (6) (2022) bbac424.
[127] Q. Miagoux, V. Singh, D. de Mézquita, V. Chaudru, M. Elati, E. Petit-Teixeira, A. Niarakis, Inference of an integrative, executable network for rheumatoid arthritis combining data-driven machine learning approaches and a state-of-the-art mechanistic disease map, J. Pers. Med. 11 (8) (2021) 785.
[128] J.H. Moore, R. Amos, J. Kiralis, P.C. Andrews, Heuristic identification of biological architectures for simulating complex hierarchical genetic interactions, Genet. Epidemiol. 39 (1) (2015) 25–34.
[129] H.E. Chua, S.S. Bhowmick, J. Zheng, L. Tucker-Kellogg, Tapestry: Network-centric target prioritization in disease-related signaling networks, in: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2016, pp. 108–117.
[130] H.E. Chua, S.S. Bhowmick, L. Tucker-Kellogg, C.F. Dewey Jr, Tenet: topological feature-based target characterization in signalling networks, Bioinformatics 31 (20) (2015) 3306–3314.
[131] J. Nielsen, Systems biology of metabolism: a driver for developing personalized and precision medicine, Cell Metab. 25 (3) (2017) 572–579.
[132] L. Jeske, S. Placzek, I. Schomburg, A. Chang, D. Schomburg, Brenda in 2019: a european elixir core data resource, Nucleic Acids Res. 47 (D1) (2019) D542–D549.
[133] H.B. Sieburg, O.K. Clay, The cellular device machine development system for modeling biology on the computer, Complex Syst. 5 (6) (1991) 575–602.
[134] C. Ricciardi, G. Cesarelli, A.M. Ponsiglione, G. De Tommasi, M. Cesarelli, M. Romano, G. Improta, F. Amato, Combining simulation and machine learning for the management of healthcare systems, in: 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), IEEE, 2022, pp. 335–339.
[135] A. Procopio, G. Cesarelli, S. De Rosa, L. Donisi, C. Critelli, A. Merola, C. Indolfi, C. Cosentino, F. Amato, A combined simulation and machine learning approach to classify severity of infarction patients, in: 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), IEEE, 2022, pp. 283–288.