Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Large-scale knowledge graph representations of disease processes

Matti Hoch[1], Shailendra Gupta[1] and Olaf Wolkenhauer[1,2]

Abstract

Today, a wide range of technologies and data types are available when studying disease-relevant processes. Therefore, a major challenge is integrating data from different technologies covering different levels of functional cellular organization. This motivates approaches that start with a bird's-eye perspective, initially considering as many molecules, cell types, and cellular functions as possible. Knowledge graphs (KGs) provide such a perspective through graphically structured representations of the functional connections between biological entities. However, linking KGs of disease processes with experimental or clinical data requires their curation in a large-scale, multi-level layout. The resulting heterogeneity leads to new challenges in KG curation, data integration, and analysis. Existing approaches for small-scale applications must be adapted or combined into multi-scale tools to analyze multi-omics data in KGs. This short review reflects upon the large-scale KG approach to studying disease processes. We do not review all modeling approaches but focus on a personal perspective on.

Addresses

[1] Institute of Computer Science, Department of Systems Biology & Bioinformatics, University of Rostock, Germany

[2] Leibniz Institute of Food Systems Biology at Technical University Munich, Germany

Corresponding author: Wolkenhauer, Olaf (olaf.wolkenhauer@uni-rostock.de)

Current Opinion in Systems Biology 2024, **38**:100517

This review comes from a themed issue on **Mathematical Modelling (2023)**

Edited by **Jana Wolf** and **Kevin Thurley**

For complete overview of the section, please refer the article collection - Mathematical Modelling (2023)

Available online 30 April 2024

https://doi.org/10.1016/j.coisb.2024.100517

Introduction

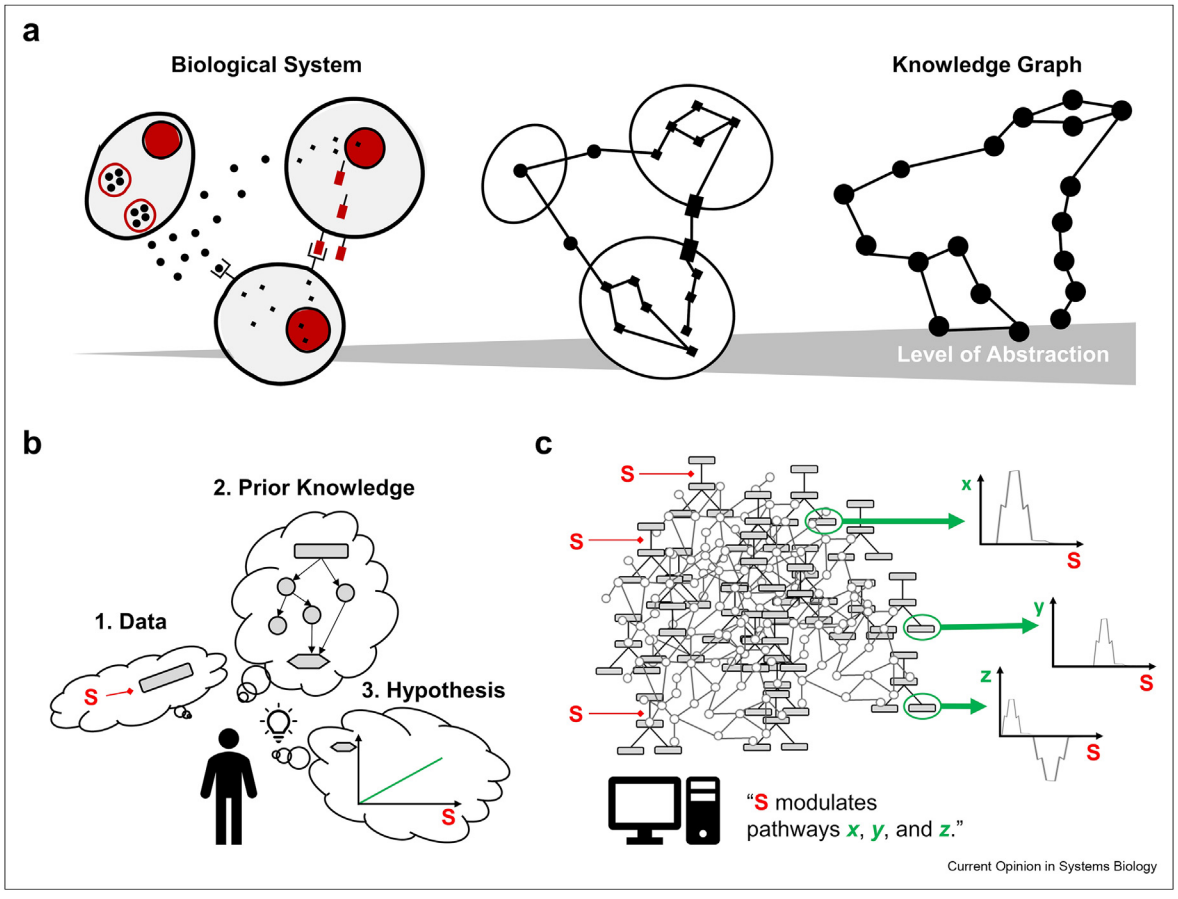
Systems biology approaches can be looked at as a 50-year-long history. Metabolic Control Analysis developed in the 1970s and put the network into the spotlight: to understand the role or function of molecules, we need to study them through their interactions in a network. The availability of technologies to generate transcriptomic data spurred gene regulatory and metabolic network modeling in the 1990s, and quantitative proteomics motivated the mathematical modeling of signal transduction pathways for over 20 years.

By studying molecular networks and their role or function in disease-related processes, we are no longer limited to one technology. Disease projects often generate multi-omics datasets, providing information about multiple molecule types. For models, this implies many variables and various levels of functional organization that need to be accounted for.

Knowledge graphs (KG) are computational representations that encode entity relationships in a graph-like structure (Figure 1(**a**)). Entities are called nodes, representing molecules, cell types, tissues, or abstractions of higher-level processes, depending on the model granularity. The relations linking them are referred to as edges, covering conceptual, functional, or physicochemical interactions across levels of structural organization (e.g., from molecules to tissues and organs) and levels of functional organization (e.g., from molecular reactions to physiological processes). Linking experimental or clinical disease data with information in a KG can identify meaningful patterns in raw data that would otherwise remain hidden [1].

Small-scale KG models focus on functionally confined metabolic or signaling pathways where the general causalities are usually well understood, and modeling aims to accurately quantify molecule concentrations. Large-scale KGs, conversely, encompass broader interactions, often entire cellular systems and interactions across multiple levels of structural and functional organization. In 2014, Weinberg reflected upon these issues by outlining the history of moving from small-scale to large-scale systems biology approach in cancer research [2]. The complexity of various genetic, epigenetic, and environmental factors manifesting throughout cellular signaling described by Weinberg also applies to other diseases. Modeling disease mechanisms requires characterizing the underlying biological processes at multiple scales [3]. On the spatial scale, diseases do not act as enclosed systems but in dependence on multiple tissues and organs that can cause systemic symptoms. On a

Figure 1



Concepts of knowledge graph (KG) approaches in systems biology. (**a**) KGs are abstractions of biological systems, representing underlying relationships in a graph structure. (**b**) Researchers perform biomolecular experiments with a hypothesis in mind. They connect data, either experimental conditions, like a stimulus (S), or previous results, with a basic understanding of the system under observation. (**c**) Similarly, computational KG approaches combine prior knowledge with new data but enable the analysis of complex, non-linear processes on a much larger scale.

temporal scale, cellular processes like gene expression or metabolic reactions occur for minutes to hours, and single molecular interactions occur on the nanosecond scale, while diseases manifest over days, weeks, or even years. Modeling these processes with small-scale approaches becomes insufficient to represent the inherent heterogeneity. Applying quantitative approaches also becomes challenging due to the large number of parameters and uncertainties. Thus, large-scale KGs usually utilize logic modeling to infer causal links between complex systems of disease processes with a high degree of non-linearity and the data (Figure 1(**b,c**)) [4].

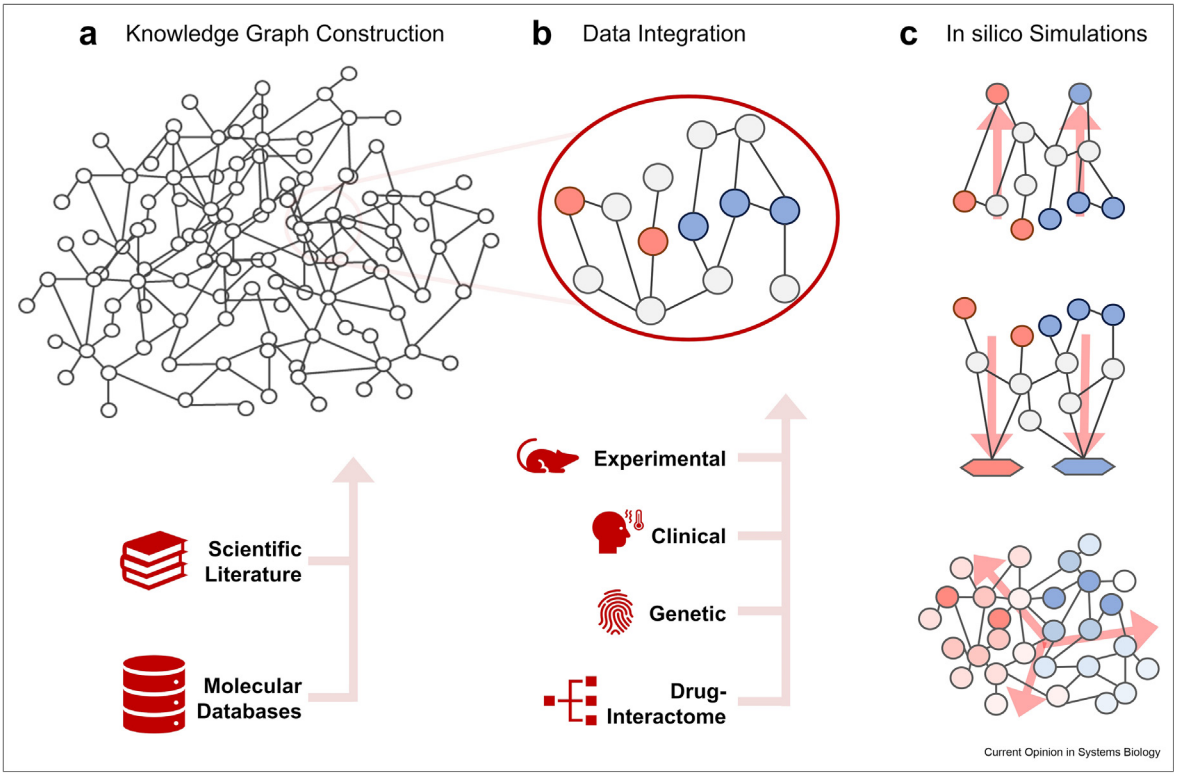
Over the last decade, independent research efforts have utilized interactive large-scale KG representations of disease mechanisms as publicly available and community-driven platforms, collectively called Disease Maps [5]. They are crucial in understanding disease progression and identifying potential therapeutic targets. Examples of established Disease Maps are the Atlas of Inflammation Resolution (AIR) [6], the Sarcopenia Map [7], and the COVID-19 Disease Map [8].

This short review gives an overview of methodologies for creating and analyzing large-scale KGs to investigate human diseases. Mentioning the vast number of research projects that successfully utilized KG in disease research would go far beyond the scope of this work. For detailed reviews of KG methods and applications, we refer to the extensive works of other colleagues [1,9–11]. Instead, we will focus on specific challenges we and other groups have encountered in KG modeling and the scientific efforts with the potential to overcome them. These can be divided into three parts, namely KG construction, data integration, and *in silico* simulations, as shown in Figure 2.

Knowledge graph construction

KGs can be constructed either knowledge-driven, i.e., from literature in which experimentally validated

Figure 2



Key components of biomolecular data analysis using knowledge graphs.

knowledge about the relationships is encoded, or data-driven, when cause and consequence in the graph are derived directly from experimental data. The data-driven approach is often called "network inference," a predominant example being the inference of gene regulatory networks (GRNs) from transcriptomics data. Correlation does not automatically infer causation, and a significant challenge is having approaches for multiple data types. Marku and Vera recently extensively reviewed this topic [12]. The main challenge lies in ensuring experimental data's integrity and discerning between causality and mere correlation. Biological feedback mechanisms cause cycles in graphs, which reduce the pool of mathematical approaches suitable for the analysis of large-scale networks.

Data-driven and large-scale knowledge-driven KGs, like protein–protein interaction (PPI) or gene-regulatory network (GRN) databases, are curated in a simple interaction format (SIF). The SIF describes the direct interactions between two entities, usually as an assumed causality between a source and a target. More complex representations, such as metabolic reactions, transport processes, or complex formations, require special formats and, inherently, manual curation. Aiming for standardization of such, **SBGN** (Systems Biology Graphical Notation) provides a standardized visual language, ensuring that diagrams detailing complex interactions remain universally interpretable. **SBML** (Systems Biology Markup Language) offers a structured format for encoding these models, ensuring seamless sharing and integration across computational tools. The BioModels database contains numerous computational models from different approaches, including SBML models of small and medium-sized KGs with their parameterization if available [13]. For the definition of KGs as large-scale multilevel models, the BioModels database is a source of information for subnetworks but rarely for the entire disease maps. KG resources and file formats have been extensively reviewed in Ref. [14].

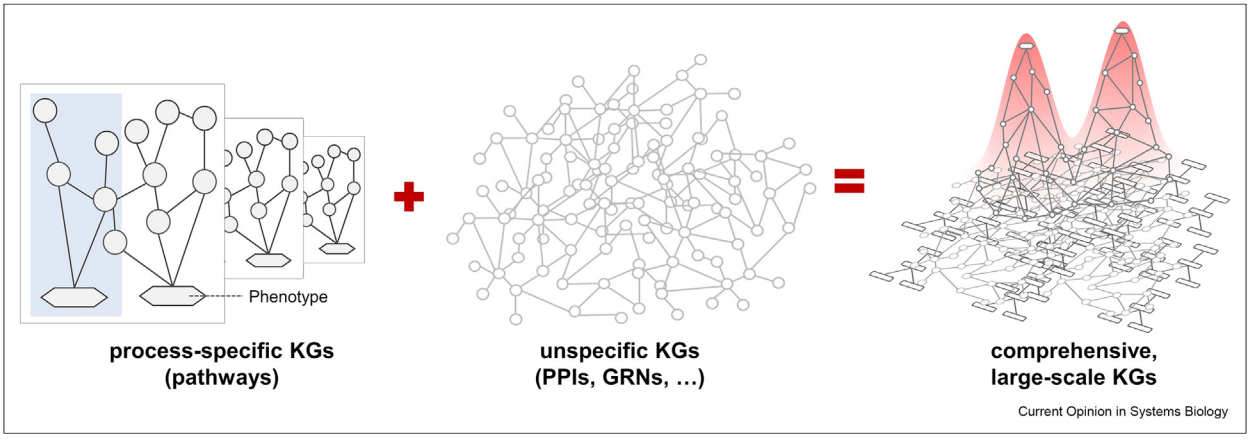
Manual, knowledge-driven curation requires that the KGs are designed to offer the curator and the user an intuitive presentation. This is achieved by modularization into smaller KGs with functional or spatial specificity. Since KGs of different biological systems usually overlap in some biological pathways, these modules can be reused and adapted, significantly reducing the

modeling effort, provided that standardization is ensured. These modularized KGs are then called subgraphs or, especially in the context of disease maps, submaps. Pathway databases such as **KEGG** [15]**, REACTOME** [16]**, and WikiPathways** [17] provide such modularized KGs of biological processes, curating expansive information on molecular pathways for various organisms, which is invaluable for researchers designing their models. However, interoperability between resources is still compromised due to different standards for KG representation. Tools for file conversion have been developed, but usually only between two formats, and many of them are not supported anymore. The **MINERVA** platform, employed by many publicly available disease map projects, provides interactive visualizations, automated annotation of KGs, and a REST API for converting some KG formats [18,19].

Given the diversity of diseases and the variability of associated data, it is difficult to create KGs that are both concise and comprehensive [10]. Converting disease heterogeneity into a single, homogeneous KG can overlook cell type specifications and lose information on spatial or temporal dynamics. Combining KGs from multiple sources with different curation details in a multi-level layout can help sustain granularity. For example, combining small KGs of process-specific signaling with large-scale KGs of unspecific PPIs results in a comprehensive KG that describes connections between pathways (Figure 3). Automated workflows for crafting context-specific KGs have emerged as a solution to streamline modeling. For example, OmniPath offers robust API documentation coupled with R and Python packages, enabling users to formulate comprehensive KGs tailored to their needs [20]. The MINERVA team recently published "Automap", an automated pipeline to create disease maps from Orphanet or HPO IDs by scanning existing KG databases and disease maps. While these tools significantly enhance data accessibility, the challenge remains to ensure interaction confidence, especially given the diversity of evidence sources. It is necessary to compare the reliability of interactions from the respective databases, as the same interaction may be misinterpreted or differently represented in different sources. With KG size, the risk of incorporating erroneous connections or missing pivotal interactions increases, potentially leading to dense models that may lack precision and accuracy. Villaveces et al. introduced MIscore, a scoring mechanism that evaluates the consistency of overlapping interactions [21].

With the availability of large-scale data collections and the rapid advances of artificial intelligence over the past decade, machine learning is now also being explored in several directions. By predicting protein structures, alpha fold has been applied to rank the confidence of physical interactions of proteins in human interactomes [22]. Not only are these approaches already showing great value in the study of molecular structures, but KG curation and network inference are now also being addressed. The Integrated Network and Dynamical Reasoning Assembler (INDRA) generates mechanistic KGs by harnessing text-mining approaches and information from pathway databases [23]. A recently developed automated causation inference tool, CausalPath, compares prior knowledge with proteomics data to create causal interaction models of high confidence [24]. Segarra-Queralt et al. employed machine learning on experimental data to optimize KGs curated from literature databases [25].

Figure 3



The curation of large-scale Knowledge Graphs (KGs) by combining manually curated KGs of specific processes with unspecific KGs generated from large databases. PPIs – Protein-Protein-Interactions; GRNs – Gene Regulatory Networks.

The rise of single-cell technologies has enhanced the resolution of biological data, paving the way for more refined computational models. Such high-resolution data provides more details of disease processes, ranging from tissue dynamics to individual cellular behavior. Given that diseases manifest across multiple cell types, developing models that consider cellular specificity has become necessary. While this granularity could create high-resolution models, it also presents challenges regarding accuracy and robustness. Qualitative approaches are generally more feasible for single-cell mapping and do not require cell-specific parameterization. Another challenge is the high degree of inherent and apparent randomness characterizing biological systems [26]. While deterministic models can adeptly capture average behaviors over multiple samples, they fall short when applied to individual samples, such as single cells. The noise becomes more disruptive with increasing resolution, necessitating a shift to stochastic approaches, including Bayesian and Markov models, for single-cell sampling [27]. Modeling approaches incorporating stochastic elements into systems biology models can yield more profound insights [28].

When diving into the high resolution of individual cells, the spatial dynamics must also be considered. Processes that might seem straightforward in isolation can exhibit profound structural complexity through interactions with neighboring cells. Diseases like cancer, where microenvironments play a pivotal role, necessitate models that can capture these spatial nuances. On a larger scale, when considering individual agents like cells and their interactions in an environment, ABM (Agent-Based Modeling) becomes valuable. In this approach, each agent operates based on a set of rules. Their collective actions can lead to emergent phenotypic changes, often observable at tissue or organ levels. Dutta-Moscato et al. developed an ABM of liver fibrosis progression through liver cell proliferation, reconstructing histological data *in silico* [29]. The PhysiBoSS platform was developed from the 3D ABM software PhysiCell and the MatLab Boolean modeling tool MaBoss [30]. It provides a multi-scale platform that combines agent-based simulations at the cell/tissue level with underlying molecular signaling. By alternating simulations at different scales and using the results of one scale as inputs to another, PhysiBoss efficiently handles the challenge of granularity. Pushing the computational limits of ABMs, so-called "Giga Scale Models" aim to bring cellular models up to macroscopic scales [31].

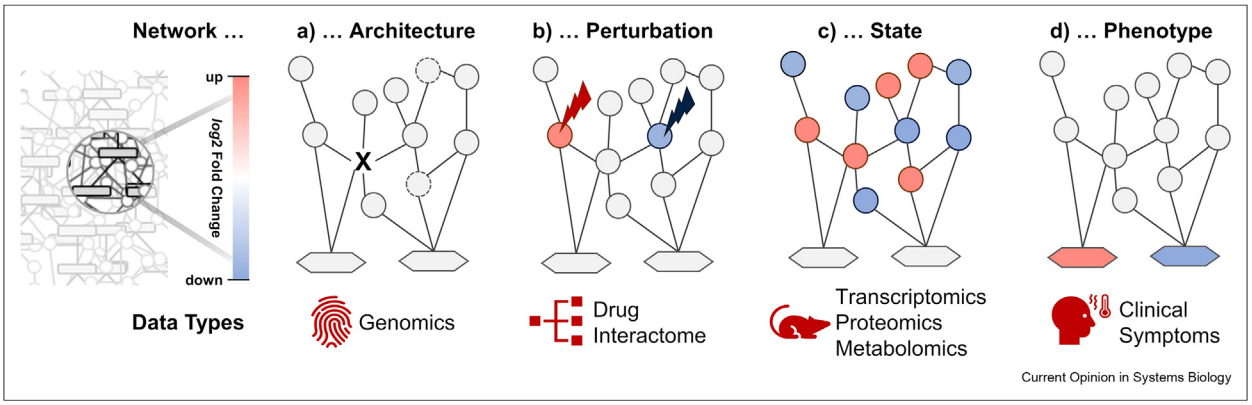
Data integration

The rapid growth of omics technologies brings both opportunities and challenges. While researchers now have access to unparalleled amounts of biological data, the field of modeling is challenged with translating measurements into numerical parameters for model components and, subsequently, aligning diverse data types of different scales. Each dataset, often from different sources and platforms, has unique formats, underlying biases, and specific challenges (Figure 4). Combining different data types could lead to losing details essential for the disease, such as cellular context, temporal dynamics, or critical environmental factors. In multi-omics data, it is crucial to align omics modalities with their corresponding prior knowledge representations, a task that, while seemingly straightforward, introduces new challenges. With solely transcriptomics data, the gene expression values are often mapped directly to their protein products. However, when additionally integrating proteomics data, a distinction is essential between mapping to mRNA or protein components. Clinical measurements are often measured using different approaches and individuals, and even the same parameters can be assessed using different methods or scoring systems. Consequently, their mapping to KGs requires manual curation, ideally in consultation with clinical experts.

Multi-omics data integration thus necessitates the expansion of KGs with new layers. The "Mergeomics 2.0" web server provides a streamlined workflow for integrating multi-omics disease association studies [32]. By focusing on disease-specific data, the platform employs multiple data filtering and enrichment steps before mapping the data onto KGs to identify key disease drivers using network analyses. Bodein et al. proposed a comprehensive workflow for integrating longitudinal, untargeted, multi-omics data by combining data- and knowledge-driven curation of multi-layer networks and the performance of analyses within and between the layers [33]. While these methods for data integration differ due to the data origin and analytical context, they emphasize that modularization and iterations of network-based data processing steps become essential for integrating complex datasets. Chen et al. recently reviewed multi-omics data integration methods in disease research [34].

With the increasing quality and quantity of biological data, the challenge of effective parameterization, especially for complex multicellular or multimolecular systems, is proving difficult. Therefore, many efforts are aimed at finding solutions for computationally efficient parameterization of models from large data sets. Bayesian approaches drastically improve parameterization from experimental data and uncertainty quantification [35,36]. Neural-mechanistic hybrid approaches become increasingly popular, harnessing the power of machine learning and mechanistic modeling to optimize parameterization tasks [37]. Machine learning can not only be used for identifying parameters in KGs but also to reconstruct missing reactions. Chen et al. developed CHESHIRE, a method for data-independent prediction of missing reactions in genome-scale metabolic models [38]. Despite these computational advances,

Figure 4



Biomolecular data types and their representation in knowledge graphs (KGs). (**a**) Population-wide references of molecular interactions form the foundational architecture of knowledge graphs. However, individual mutations can introduce alterations, necessitating adjustments to the graph's topology. (**b**) External stimuli, physiological or pathological, serve as perturbations in the biological system, inducing specific changes to graph elements. (**c**) Data from biological experiments, notably from omics technologies, capture snapshots of a system's state at specific moments. Dynamic behavior predictions can be formulated by analyzing differences between successive system states. (**d**) Clinical traits, or observable characteristics, are integrated as higher-level elements in the graph, linked to the molecular interactions they arise from.

reproducibility remains an ongoing challenge, hampered by missing parameters in published materials, omitted codes, or gaps in documentation [39].

In silico simulations

Quantitative approaches can be challenging, demand extensive datasets for parameterization, and may offer an incomplete picture due to diseases not being confined to isolated pathways [40]. Qualitative, usually non-parametric, solutions, such as topological analysis, Boolean logic, or Petri nets, can be more accessible [1]. Generally speaking, deriving information from KGs requires understanding their topological structure. The connectivity of nodes and edges in the graph, locally or globally, can be expressed from different perspectives as numerical features called centrality measures. These features can be used for further analysis, e.g., to enhance the performance of statistical enrichment [41,42] or machine learning approaches [43,44]. Without any modifications, the signal transduction in a KG can be simulated from its topology alone [45]. This enables, for example, the consideration of feedback or the aggregation of signals in downstream circuits [46]. Boolean modeling has allowed for assessing steady state and mechanistic simulations of diseases by evaluating the ON/OFF states of nodes through simple logic rules [7,47]. Petri nets offer a more granular approach by defining explicit rules for signal transduction pathways, thus enabling the modeling of complex biological signaling mechanisms [48,49].

When we recognize diseases as multi-layer and multi-scale processes, it becomes evident that no single modeling strategy can capture their full complexity. Qualitative models that focus on tissue—level interactions lack the detail of underlying molecular interactions, whereas pathway models may be too narrowly designed and might only answer concrete questions at the molecular level. Consequently, several efforts in the last decade aimed at finding solutions to combine types of modeling to overcome their limitations and expand the analytical scope [50,51]. Maldonado et al. developed a multi-scale model of non-alcoholic fatty liver disease by employing Quasi-Steady State Petri Nets (QSSPN), which combine genome-scale metabolic networks with ordinary differential equation (ODE) models of protein signaling and compartmental transport kinetics [52]. Considering KGs as multi-level models, instead of directly combining approaches, less complex methods can be used to extract subgraphs to which more detailed approaches can then be applied. Khan et al. extracted so-called core regulatory networks (CRNs) from large-scale KGs for perturbation simulation using Boolean modeling [53]. Liu et al. provide a detailed review of hybrid modeling approaches combining quantitative/qualitative and deterministic/stochastic models [54].

During the COVID-19 pandemic, systems biology approaches experienced unprecedented clinical interest, e.g., for drug repurposing [55,56]. The field of systems pharmacology seeks to utilize disease models to inform drug development and therapeutic interventions. However, the translatability of any model depends on its accuracy and validity. Models must also account for the uncertainties inherent in biological systems. Despite advances in tools and methods for quantifying uncertainties, they are not universally applied. This gap is particularly notable given biological systems' inherent

variability and unpredictability, ranging from molecular fluctuations to patient-to-patient variability. The abstract nature of biological KGs leads to many biases in disease models. "Knowledge bias" refers to the fact that molecules are better studied than others and, consequently, are more interconnected in databases [57]. These biases are reflected in the analysis and affect model validation, as established disease genes are more likely to be prioritized in test data sets than others in a self-fulfilling prophecy. Molotkov et al. developed a method to detect such "validation bias" in gene prioritization approaches, suggesting that it might impede the translation of models into practice [58].

The path from a computational model to a clinical application is fraught with challenges, from validation issues to regulatory considerations. An effort to bring systems biology towards the clinical is the creation of so-called "digital twins" referring to models that can be adapted using patient data to allow personalized predictions of treatment outcomes [59]. Cardiovascular research is pioneering the field of digital twins, possibly due to the more feasible translatability of computational models in electrophysiology [60]. The systems immunology community recently published a roadmap toward an "immune digital twin" [61]. A starting point for translating *in silico* models to clinics is their application to retrospective studies to test performance on clinical data. For example, the Universal Immune System Simulator, an agent-based model of the immune system, was applied to predict drug effects and mirror the results of their clinical trials [62].

Summary

Mathematical modeling of molecular networks has a tradition of about 50 years. What we refer to as Metabolic Control Analysis (MCA) was developed in the 1970s [63]. The central idea of MCA is that we can understand a metabolic network through sensitivity analysis, that is, systematic perturbations and the observation of relative changes. These efforts put the network into the spotlight: to understand the role or function of molecules, we need to study them through their interactions in a network. The availability of technologies to generate transcriptomic data spurred the modeling of gene regulatory networks in the 1990s, and quantitative proteomics motivated the mathematical modeling of signal transduction pathways for over 20 years.

The increasing complexity of data in systems biology requires advanced modeling techniques. Metabolic pathways are well-researched, and large amounts of data are available for parameterization. They can be modeled at a large scale on the cellular scale and for numerous examples with reasonably well-understood kinetics. Traditional single-pathway models of gene regulation and cell signaling often need to be more comprehensive to capture the heterogeneity of disease-associated data. Cellular communication networks are more complex, partially due to the heterogeneity of interaction types, which makes modeling approaches more challenging. The notion of "crosstalk" between pathways reflects this struggle between what signaling networks 'are' and our assumptions about them. Numerous cell types and molecules are involved in many disease-relevant processes, motivating different approaches. Systems biology approaches are evolving in response to this, and large-scale, multi-level KGs are one direction in which we are moving.

Curation overhead, dependencies, and limited interoperability complicate the application of *in silico* tools in wet lab research. Ready-to-use software packages with automated KG curation and streamlined parameterization and analysis pipelines could be better translated into clinical and experimental practice. The availability of multi-omics data has also created the potential for advanced modeling approaches. Hybrid models, which combine the structured representation of KGs and the adaptability of machine learning, promise improved accuracy and depth in understanding disease mechanisms. Their ability to integrate different data types and scales suggests that multi-scale models will become increasingly essential in systems biology research.

Systems biology approaches have evolved from focusing on single technologies and methodologies to workflows that curate and analyze a multitude of data using a range of methodologies. The availability of evidence through databases and our ability to extract information from publications expand our toolset. With many systems biology approaches targeting disease-relevant phenomena, it is wonderful to see how many efforts have come closer to clinical practice. There is still a gap between basic research of molecular mechanisms and therapeutic decisions, but for diagnostic and prognostic purposes, the analysis of classical clinical patient data with molecular data remains a promising direction.

Since around 2016, any large dataset that can be represented in sequences or images allows the application of modern machine learning approaches. Many types of data are generated through imaging, but other data, like, for example, molecular structures, can also be interpreted as sequences. Generative machine learning algorithms may, in the future, also be able to help with a problem that any generation of modelers has faced so far. Despite an increasing volume of data and data types for understanding cellular processes, there is too often a lack of sufficiently rich, quantitative time course datasets. While technological advances to generate experimental data have, for a long time, driven research in molecular and cell biology, it is great to see methodological advances to analyze, model, and interpret data, opening up new questions and directions.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors have been supported by Heel GmbH to create the Atlas of Inflammation Resolution, which is referred to in this review, alongside a reference to the publication of the AIR. This "disease map" is publicly available but since it is an example for the topic of the mini review we feel obliged to declare this relationship, even we do not see it as "competing interest".

Data availability

No data were used for the research described in the article.

Acknowledgements

The authors acknowledge support through funding of the German Federal Ministry for Education and Research (BMBF), as part of their eMed initiative (Project: Melautim, grant number: 01ZX1905B). The authors were also supported by grants from Biologische Heilmittel Heel GmbH, Baden-Baden, Germany. The funding supported the creation of the Atlas of Inflammation Resolution (AIR).

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. * Garrido-Rodriguez M, Zirngibl K, Ivanova O, Lobentanzer S, Saez-Rodriguez J: **Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks**. *Mol Syst Biol* 2022, **18**:1–15, https://doi.org/10.15252/msb.202211036.

Garrido-Rodriguez et al. provided a very comprehensive overview of existing resources, tools, and techniques for knowledge graph curation and analysis. The article covers a wide range of analysis techniques, from network theory to Bayesian and machine learning approaches, and is therefore highly relevant to the broader systems biology community.

2. Weinberg RA: **Coming full circle – from endless complexity to simplicity and back again**. *Cell* 2014, **157**:267–271, https://doi.org/10.1016/j.cell.2014.03.004.

3. Berlin R, Gruen R, Best J: **Systems medicine disease: disease classification and scalability beyond networks and boundary conditions**. *Front Bioeng Biotechnol* 2018, **6**, 389105, https://doi.org/10.3389/FBIOE.2018.00112/BIBTEX.

4. Barsi S, Szalai B: **Modeling in systems biology: causal understanding before prediction?** *Patterns* 2021, **2**, 100280, https://doi.org/10.1016/j.patter.2021.100280.

5. Mazein A, Ostaszewski M, Kuperstein I, Watterson S, Le Novère N, Lefaudeux D, *et al.*: **Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms**. *NPJ Syst Biol Appl* 2018, **4**, https://doi.org/10.1038/S41540-018-0059-Y.

6. Serhan CN, Gupta SK, Perretti M, Godson C, Brennan E, Li Y, *et al.*: **The Atlas of Inflammation Resolution (AIR)**. *Mol Aspects Med* 2020, **74**, https://doi.org/10.1016/J.MAM.2020.100894.

7. Hoch M, Ehlers L, Bannert K, Stanke C, Brauer D, Caton V, *et al.*: **In silico investigation of molecular networks linking gastrointestinal diseases, malnutrition, and sarcopenia**. *Front Nutr* 2022, **9**, https://doi.org/10.3389/fnut.2022.989453.

8. ** Ostaszewski M, Niarakis A, Mazein A, Kuperstein I, Phair R, Orta-Resendiz A, *et al.*: **COVID19 Disease Map, a computational knowledge repository of virus-host interaction mechanisms**. *Mol Syst Biol* 2021, **17**, https://doi.org/10.15252/MSB.202110387.

The COVID-19 disease map is an exceptional example of collaboration in the systems biology community. Numerous modeling research groups were able to adapt their existing modeling efforts to the SARS-CoV-2 and publish a comprehensively curated resource of COVID-19 disease mechanisms in a matter of months.

9. Galindez G, Sadegh S, Baumbach J, Kacprowski T, List M: **Network-based approaches for modeling disease regulation and progression**. *Comput Struct Biotechnol J* 2023, **21**:780–795, https://doi.org/10.1016/J.CSBJ.2022.12.022.

10. Yue R, Dutta A: **Computational systems biology in disease modeling and control, review and perspectives**. *NPJ Syst Biol Appl* 2022, **8**, https://doi.org/10.1038/s41540-022-00247-4.

11. Yang Y, Lu Y, Yan W: **A comprehensive review on knowledge graphs for complex diseases**. *Brief Bioinform* 2023, **24**, https://doi.org/10.1093/BIB/BBAC543.

12. Marku M, Pancaldi V: **From time-series transcriptomics to gene regulatory networks: a review on inference methods**. *PLoS Comput Biol* 2023, **19**, https://doi.org/10.1371/JOURNAL.PCBI.1011254.

13. Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, Xavier A, *et al.*: **BioModels – 15 years of sharing computational models in life science**. *Nucleic Acids Res* 2020, **48**:D407–D415, https://doi.org/10.1093/NAR/GKZ1055.

14. Touré V, Flobak Å, Niarakis A, Vercruysse S, Kuiper M: **The status of causality in biological databases: data resources and data retrieval possibilities to support logical modeling**. *Brief Bioinform* 2020, **2020**:1–15, https://doi.org/10.1093/bib/bbaa390.

15. Kanehisa M: **Toward understanding the origin and evolution of cellular organisms**. *Protein Sci* 2019, **28**:1947–1951, https://doi.org/10.1002/PRO.3715.

16. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, *et al.*: **The reactome pathway knowledgebase 2022**. *Nucleic Acids Res* 2022, **50**:D687–D692, https://doi.org/10.1093/NAR/GKAB1028.

17. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, *et al.*: **WikiPathways: connecting communities**. *Nucleic Acids Res* 2021, **49**:D613–D621, https://doi.org/10.1093/NAR/GKAA1024.

18. Gawron P, Ostaszewski M, Satagopam V, Gebel S, Mazein A, Kuzma M, *et al.*: **MINERVA-a platform for visualization and curation of molecular interaction networks**. *NPJ Syst Biol Appl* 2016, **2**, https://doi.org/10.1038/NPJSBA.2016.20.

19. Hoksza D, Gawron P, Ostaszewski M, Hasenauer J, Schneider R: **Closing the gap between formats for storing layout information in systems biology**. *Brief Bioinform* 2020, **21**:1249–1260, https://doi.org/10.1093/BIB/BBZ067.

20. Türei D, Korcsmáros T, Saez-Rodriguez J: **OmniPath: guidelines and gateway for literature-curated signaling pathway resources**. *Nat Methods* 2016, **13**:966–967, https://doi.org/10.1038/NMETH.4077.

21. Villaveces JM, Jiménez RC, Porras P, Del-Toro N, Duesbury M, Dumousseau M, *et al.*: **Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study**. *Database* 2015, **2015**, https://doi.org/10.1093/DATABASE/BAU131.

22. ** Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, *et al.*: **Towards a structurally resolved human protein interaction network**. *Nat Struct Mol Biol* 2023, **30**:216–225, https://doi.org/10.1038/S41594-022-00910-8.

The authors applied the predictive power of AlphaFold to evaluateprotein–protein interaction networks. Their work shows that the future of modeling could be independent of experimental data. Instead, machine learning could be used to predict causal interactions and reaction kinetics from the physicochemical structure of molecules alone.

23. Bachman JA, Gyori BM, Sorger PK: **Automated assembly of molecular mechanisms at scale from text mining and curated databases**. *Mol Syst Biol* 2023, **19**, https://doi.org/10.15252/MSB.202211325.

24. Babur Ö, Luna A, Korkut A, Durupinar F, Siper MC, Dogrusoz U, *et al.*: **Causal interactions from proteomic profiles: molecular data meet pathway knowledge**. *Patterns* 2021, **2**, https://doi.org/10.1016/J.PATTER.2021.100257.

25. Segarra-Queralt M, Neidlin M, Tio L, Monfort J, Monllau JC, González Ballester M, *et al.*: **Regulatory network-based model to simulate the biochemical regulation of chondrocytes in healthy and osteoarthritic environments**. *Sci Rep* 2022, **12**, https://doi.org/10.1038/S41598-022-07776-2.

26. Ilan Y: **Order through disorder: the characteristic variability of systems**. *Front Cell Dev Biol* 2020, **8**, https://doi.org/10.3389/FCELL.2020.00186.

27. Tiberi S, Walsh M, Cavallaro M, Hebenstreit D, Finkenstädt B: **Bayesian inference on stochastic gene transcription from flow cytometry data**. *Bioinformatics* 2018, **34**:i647–i655, https://doi.org/10.1093/BIOINFORMATICS/BTY568.

28. Simoni G, Vo HT, Priami C, Marchetti L: **A comparison of deterministic and stochastic approaches for sensitivity analysis in computational systems biology**. *Brief Bioinform* 2020, **21**:527–540, https://doi.org/10.1093/BIB/BBZ014.

29. Dutta-Moscato J, Solovyev A, Mi Q, Nishikawa T, Soto-Gutierrez A, Fox IJ, *et al.*: **A multiscale agent-based in silico model of liver fibrosis progression**. *Front Bioeng Biotechnol* 2014, **2**, https://doi.org/10.3389/FBIOE.2014.00018.

30. Letort G, Montagud A, Stoll G, Heiland R, Barillot E, MacKlin P, *et al.*: **PhysiBoSS: a multi-scale agent-based modelling framework integrating physical dimension and cell signalling**. *Bioinformatics* 2019, **35**:1188–1196, https://doi.org/10.1093/BIOINFORMATICS/BTY766.

31. Montagud A, Ponce-de-Leon M, Valencia A: **Systems biology at the giga-scale: large multiscale models of complex, heterogeneous multicellular systems**. *Curr Opin Syst Biol* 2021, **28**, 100385, https://doi.org/10.1016/J.COISB.2021.100385.

32. Ding J, Blencowe M, Nghiem T, Ha SM, Chen YW, Li G, *et al.*: **Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics**. *Nucleic Acids Res* 2021, **49**:W375–W387, https://doi.org/10.1093/NAR/GKAB405.

33. Bodein A, Scott-Boyer MP, Perin O, Lê Cao KA, Droit A: **Interpretation of network-based integration from multi-omics longitudinal data**. *Nucleic Acids Res* 2022, **50**:E27, https://doi.org/10.1093/NAR/GKAB1200.

34. Chen C, Wang J, Pan D, Wang X, Xu Y, Yan J, *et al.*: **Applications of multi-omics analysis in human diseases**. *MedComm* 2023, **4**:e315, https://doi.org/10.1002/MCO2.315.

35. Linden NJ, Kramer B, Rangamani P: **Bayesian parameter estimation for dynamical models in systems biology**. *PLoS Comput Biol* 2022, **18**, https://doi.org/10.1371/JOURNAL.PCBI.1010651.

36. Zhang X, Su Y, Lane AN, Stromberg AJ, Fan TWM, Wang C: **Bayesian kinetic modeling for tracer-based metabolomic data**. *BMC Bioinf* 2023, **24**, https://doi.org/10.1186/S12859-023-05211-5.

37. Faure L, Mollet B, Liebermeister W, Faulon JL: **A neural-mechanistic hybrid approach improving the predictive power of genome-scale metabolic models**. *Nat Commun* 2023, **14**, https://doi.org/10.1038/S41467-023-40380-0.

38. Chen C, Liao C, Liu YY: **Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning**. *Nat Commun* 2023, **14**, https://doi.org/10.1038/S41467-023-38110-7.

39. * Tiwari K, Kananathan S, Roberts MG, Meyer JP, Sharif Shohan MU, Xavier A, *et al.*: **Reproducibility in systems biology modelling**. *Mol Syst Biol* 2021, **17**, https://doi.org/10.15252/MSB.20209982.
The authors have evaluated numerous published models over the past year and found that a high percentage of the results are not reproducible. They identify the reasons for this shortcoming and make suggestions for improvement. This article is intended to encourage researchers to be more careful about the reproducibility of their work.

40. Vidal M, Cusick ME, Barabási AL: **Interactome networks and human disease**. *Cell* 2011, **144**:986–998, https://doi.org/10.1016/J.CELL.2011.02.016.

41. Zito A, Lualdi M, Granata P, Cocciadiferro D, Novelli A, Alberio T, *et al.*: **Gene set enrichment analysis of interaction networks weighted by node centrality**. *Front Genet* 2021, **12**, 577623, https://doi.org/10.3389/fgene.2021.577623.

42. Hoch M, Smita S, Cesnulevicius K, Lescheid D, Schultz M, Wolkenhauer O, *et al.*: **Network- and enrichment-based inference of phenotypes and targets from large-scale disease maps**. *NPJ Syst Biol Appl* 2022, **8**:13, https://doi.org/10.1038/s41540-022-00222-z.

43. Liu H, Zhang W, Nie L, Ding X, Luo J, Zou L: **Predicting effective drug combinations using gradient tree boosting based on features extracted from drug-protein heterogeneous network**. *BMC Bioinf* 2019, **20**:645, https://doi.org/10.1186/s12859-019-3288-1.

44. Liu H, Zhang W, Nie L, Ding X, Luo J, Zou L: **Predicting effective drug combinations using gradient tree boosting based on features extracted from drug-protein heterogeneous network**. *BMC Bioinf* 2019, **20**:645, https://doi.org/10.1186/s12859-019-3288-1.

45. Lee D, Cho KH: **Topological estimation of signal flow in complex signaling networks**. *Sci Rep* 2018, **8**:1–11, https://doi.org/10.1038/s41598-018-23643-5.

46. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J, *et al.*: **High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes**. *Oncotarget* 2016, **8**:5160–5178, https://doi.org/10.18632/ONCOTARGET.14107.

47. Singh V, Naldi A, Soliman S, Niarakis A: **A large-scale Boolean model of the rheumatoid arthritis fibroblast-like synoviocytes predicts drug synergies in the arthritic joint**. *NPJ Syst Biol Appl* 2023, **9**, https://doi.org/10.1038/S41540-023-00294-5.

48. Koch I, Büttner B: **Computational modeling of signal transduction networks without kinetic parameters: Petri net approaches**. *Am J Physiol Cell Physiol* 2023, **324**:C1126–C1140, https://doi.org/10.1152/AJPCELL.00487.2022/ASSET/IMAGES/MEDIUM/C-00487-2022R01.PNG.

49. Grunwald S, Speer A, Ackermann J, Koch I: **Petri net modelling of gene regulation of the Duchenne muscular dystrophy**. *Biosystems* 2008, **92**:189–205, https://doi.org/10.1016/J.BIOSYSTEMS.2008.02.005.

50. D'Alessandro LA, Samaga R, Maiwald T, Rho SH, Bonefas S, Raue A, *et al.*: **Disentangling the complexity of HGF signaling by combining qualitative and quantitative modeling**. *PLoS Comput Biol* 2015, **11**, https://doi.org/10.1371/JOURNAL.PCBI.1004192.

51. Sego TJ, Aponte-Serrano JO, Gianlupi JF, Glazier JA: **Generation of multicellular spatiotemporal models of population dynamics from ordinary differential equations, with applications in viral infection**. *BMC Biol* 2021, **19**, https://doi.org/10.1186/S12915-021-01115-Z.

52. Maldonado EM, Fisher CP, Mazzatti DJ, Barber AL, Tindall MJ, Plant NJ, *et al.*: **Multi-scale, whole-system models of liver metabolic adaptation to fat and sugar in non-alcoholic fatty liver disease**. *NPJ Syst Biol Appl* 2018, **4**, https://doi.org/10.1038/S41540-018-0070-3.

53. Khan FM, Marquardt S, Gupta SK, Knoll S, Schmitz U, Spitschak A, *et al.*: **Unraveling a tumor type-specific regulatory core underlying E2F1-mediated epithelial-mesenchymal transition to predict receptor protein signatures**. *Nat Commun* 2017, **8**:1–15, https://doi.org/10.1038/s41467-017-00268-2.

54. * Liu F, Heiner M, Gilbert D: **Hybrid modelling of biological systems: current progress and future prospects**. *Brief Bioinform* 2022, **23**, https://doi.org/10.1093/BIB/BBAC081.
As discussed in our work, hybrid modeling approaches are becoming increasingly popular due to the heterogeneity of biological systems. The review paper by Liu et al. summarizes quantitative or qualitative and deterministic or stochastic methods and discusses the advances and challenges in hybrid approaches.

55. Zhou Y, Liu Y, Gupta S, Paramo MI, Hou Y, Mao C, *et al.*: **A comprehensive SARS-CoV-2-human protein-protein interactome reveals COVID-19 pathobiology and potential host therapeutic targets**. *Nat Biotechnol* 2023, **41**:128–139, https://doi.org/10.1038/S41587-022-01474-0.

56. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, *et al.*: **A SARS-CoV-2 protein interaction map reveals targets for drug repurposing**. *Nature* 2020, **583**:459–468, https://doi.org/10.1038/S41586-020-2286-9.

57. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH: **Too many roads not taken**. *Nature* 2011, **470**:163–165, https://doi.org/10.1038/470163A.

58. Molotkov I, Artomov M: **Detecting biased validation of predictive models in the positive-unlabeled setting: disease gene prioritization case study**. *Bioinform Adv* 2023, **3**, https://doi.org/10.1093/BIOADV/VBAD128.

59. Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, Gustafsson M, *et al.*: **Digital twins to personalize medicine**. *Genome Med* 2019, **12**, https://doi.org/10.1186/S13073-019-0701-3.

60. Corral-Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, *et al.*: **The "Digital Twin" to enable the vision of precision cardiology**. *Eur Heart J* 2020, **41**:4556–4564B, https://doi.org/10.1093/EURHEARTJ/EHAA159.

61. Laubenbacher R, Niarakis A, Helikar T, An G, Shapiro B, Malik-Sheriff RS, *et al.*: **Building digital twins of the human immune system: toward a roadmap**. *NPJ Digit Med* 2022, **5**, https://doi.org/10.1038/S41746-022-00610-Z.

62. Maleki A, Crispino E, Italia SA, Di Salvatore V, Chiacchio MA, Sips F, *et al.*: **Moving forward through the in silico modeling of multiple sclerosis: treatment layer implementation and validation**. *Comput Struct Biotechnol J* 2023, **21**:3081–3090, https://doi.org/10.1016/J.CSBJ.2023.05.020.

63. Sauro HM: **50 Years of metabolic control analysis**. *Interface Focus* 2024, **14**, https://doi.org/10.1098/RSFS.2023.0080.