Papers for pre-read for *Mechanistic, statistical, and causal modeling in drug discovery*

The three review papers introduce modeling techniques that are commonly used in drug discovery.

- Zhang et al. (2020) introduces a hierarchy of models with regard to the scales at which we study drug-body interactions.
- Badillo et al. (2020) introduces statistical and machine learning techniques to model data.
- Michoel and Zhang (2023) introduces the concept and the practice of causal modeling.

Please read them before the course. In case of questions, please contact Jitao David Zhang.

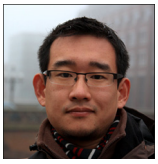
ELSEVIER

Teaser It is high time to integrate computational models at multiple levels to understand drug mechanism and safety.

Multiscale modelling of drug mechanism and safety

Jitao David Zhang[1,2], Lisa Sach-Peltason[1], Christian Kramer[1], Ken Wang[1] and Martin Ebeling[1]

[1] Pharma Early Research and Development, Roche Innovation Center Basel, F. Hoffmann-La Roche, Grenzacherstrasse 124, 4070 Basel, Switzerland
[2] University of Basel, Department of Mathematics and Computer Science, Spiegelgasse 1, 4051 Basel, Switzerland



Jitao David Zhang was a Marie-Curie Fellow at the European Institute of Bioinformatics, undertaking PhD research in computational biology at the German Cancer Research Center, and then joining Roche in 2011. His research, often embedded in collaborations with industrial and academic partners, focuses on *in silico* aspects of integrated approaches to understanding drug mechanism and safety. David co-develops new techniques supporting discovery projects, such as molecular phenotyping, which deciphers how drug candidates modulate biological pathways. David enjoys publishing open-source software and open-access publications, working with students and postdoctoral researchers, and teaching applied mathematics and informatics in drug discovery at the University of Basel.

Here, we introduce models at three levels—molecular level, cellular and omics level, and organ and system level—that study drug mechanism and safety in preclinical drug discovery. The models differ in both their scope of study and technical details, but are all rooted in mathematical descriptions of complex biological systems, and all require informatics tools that handle large-volume, heterogeneous, and noisy data. We present principles and recent developments with examples at each level and highlight the synergy by a case study. We proffer a multiscale modelling view of drug discovery, call for a seamless flow of information in the form of models, and examine potential impacts.

``Prédire n'est pas expliquer' (To predict is not to explain) – René Thom"

Introduction

A drug discovery program ideally maximises the knowledge about a molecule before progressing to test it in humans. Understanding the mechanism, namely knowing how and why a molecule works in *in vitro* and *in vivo* disease models (or not), is desired in both target-based and phenotypic discovery programs. It not only satisfies curiosity, but also enriches our knowledge in biology, guides optimisation of the molecule, enables targeted therapies and drug combinations, and reveals unexpected disease targets [1].

Understanding drug mechanism and safety is more than target identification and describing its mechanism of action. Target identification, reviewed in [2,3], is an important step towards mechanistic understanding. Identification of the mechanism of action (MoA), the specific biochemical interaction between a drug and its target through which the substance exerts its pharmacological effect, is another step forward. However, target and MoA are not all there is to know about the mechanism of a drug for several reasons. First, proteins, RNAs, and metabolites form complex intracellular and intercellular biological networks. The drug–target interaction is propagated and amplified through the networks. Therefore, a drug can regulate genes and

Corresponding author. Zhang, J.D. (jitao_david.zhang@roche.com)

https://doi.org/10.1016/j.drudis.2019.12.009

pathways in a cell identity- and context-specific manner [4]. Some effects are only secondary to MoA, for instance, negative feedback regulation of the target. Nevertheless, they are part of the mechanism for better or worse. In addition, many modalities, including small molecules and nucleotide-based agents, display polypharmacology, interacting with multiple targets, and with varying affinities [5–7]. Therefore, MoA is a simplified discrete model of a continuous spectrum of binding affinities at best. Moreover, drugs of distinct MoAs can have similar pharmacological effects when regulation of distinct upstream targets converges on the same downstream pathway [8]. Last but not least, even when the target and molecular MoA are unknown, drug mechanism can still be elucidated experimentally and quantitatively. Therefore, we call for a systematic, multiscale study of drug mechanism, expanding the scope of target and MoA identification.

A favourable safety profile of a module is at least as important as a favourable efficacy profile. Given that both mechanism and safety share the same physicochemical and biological foundations (drug-target interaction, gene/pathway perturbation, etc.), they are treated similarly, and sometimes simultaneously, from a modelling perspective. Mechanistic understanding of safety findings informs about the safety margin of a molecule, dissects target-related from molecule-specific or class-specific toxicity, and enables the identification of molecules with better safety profiles. By serendipity, mechanistic understanding of safety findings can repurpose a molecule and reveal unexpected insight into human biology, probably best highlighted by the history of sildenafil (Viagra®). More generally, experts consider that both refinement of *in silico* tools and greater mechanistic understanding might provide future opportunities to better identify drug safety liabilities [9].

Drug mechanism and safety are studied traditionally in animal models *in vivo* and in cellular systems *in vitro*. Since the 1970s, computational approaches have emerged, engaging *in vitro* and *in vivo* assays on the one hand, and mathematical and computational modelling *in silico* on the other hand. By the scale of the study, mathematical and computational models in preclinical drug discovery are classified into three levels: (i) *molecular-level modelling*, which models molecular structure and interactions between molecules, using physicochemical principles as well as molecular modelling and simulation techniques; (ii) *cellular- and omics-level modelling*, which probes all molecules of a particular kind in cells, their spatial organisation, as well as cellular morphology, using high-throughput techniques as well as bioinformatics and statistical analysis; and (iii) *organ- and system-level modelling*, which examines how drug and body interact and affect each other over time.

The three modelling approaches correspond to microscopic, mesoscopic, and macroscopic descriptions of complex systems, respectively [10]. On the molecular level, a large number of interacting entities (RNA species, proteins, metabolites, etc.) exist and only a small subset of relevant interactions can be modelled. On the cellular and omics level, well-designed experiments can reveal gene-, pathway-, and network-level changes following perturbation. The information is particularly valuable if the cell system in use is relevant for the disease or for the safety-relevant question [11]. On the organ and system level, pharmacokinetic (PK) and pharmacodynamic (PD) properties of selected drug candidates are modelled and scrutinised to ensure appropriate exposure and target engagement [12].

Population modelling, which examines distributions of PK and PD parameters in a human population, represents the fourth level of modelling, the highest in the hierarchy. Although it is beyond the scope of this review, population modelling is an important instrument that characterises individualised dose–exposure–response relationships [13,14].

The models form a hierarchy in which models at lower levels are the basis of models at higher levels. Models at each level have their own virtues and limitations, of which scientists and decision-makers must be aware to avoid being misinformed or misguided. Furthermore, to understand the mechanism and safety profile of a molecule, no individual model at any single level would suffice. Instead, it calls for a multiscale modelling approach.

We can implement such an approach in two steps. First, we establish mathematical, ideally mechanistic, models of drug–body interactions on the molecular level, the cellular and omics level, and the organ and system level. Next, we integrate modelling results at different levels, in the hope of gaining a holistic view of how and why the drug exerts both pharmacological and toxicological effects.

Results, both on individual levels and integrated across different levels, of multiscale modelling are impactful. On one side, they can bring deeper and sometimes unexpected insight into biology, pharmacology, and toxicology. On the other side, they allow scientists and decision-makers to quantitatively assess the probability of success of discovery projects. Such assessments can be crucial, given the high attrition rate of the industry [15], to prioritise and accelerate projects with a higher probability of success. Early but encouraging results were reported by colleagues from AstraZeneca [12] and from Merck & Co. Inc. [16], who both adopted assessment systems based on the integration of evidence from multiple levels and saw productivity increasing.

Multiscale modelling needs both knowledge and data to thrive. On the one hand, it requires human expertise that translates biological knowledge into mathematical descriptions and models that can be handled by computational tools, and the human mind to interpret the results. On the other hand, it requires high-quality and reproducible data that fulfil the Findable, Accessible, Interoperable, and Reusable (FAIR) standard [17]. To enable multiscale modelling of drug mechanism and safety, experimentalists and modelling scientists should work together to create a model-driven process that starts with hypotheses, which are rejected, accepted or refined, and retested iteratively, to gradually improve the models (Fig. 1).

Modelling preclinical drug mechanism and safety at three levels

Molecular-level modelling

At the most fundamental level, drugs exert their effects by interacting with molecular targets, including DNA, RNA, proteins, and so on. We illustrate the principles of molecular-level modelling by choosing drug–protein interactions as examples because proteins constitute most efficacy targets [18]. Although the therapeutic potential of many proteins remains to be explored [19], DNA and RNA (including epigenetic modifications), the other two components of the central-dogma model of molecular biology, are increasingly pursued as targets [20–23]. Molecular-level

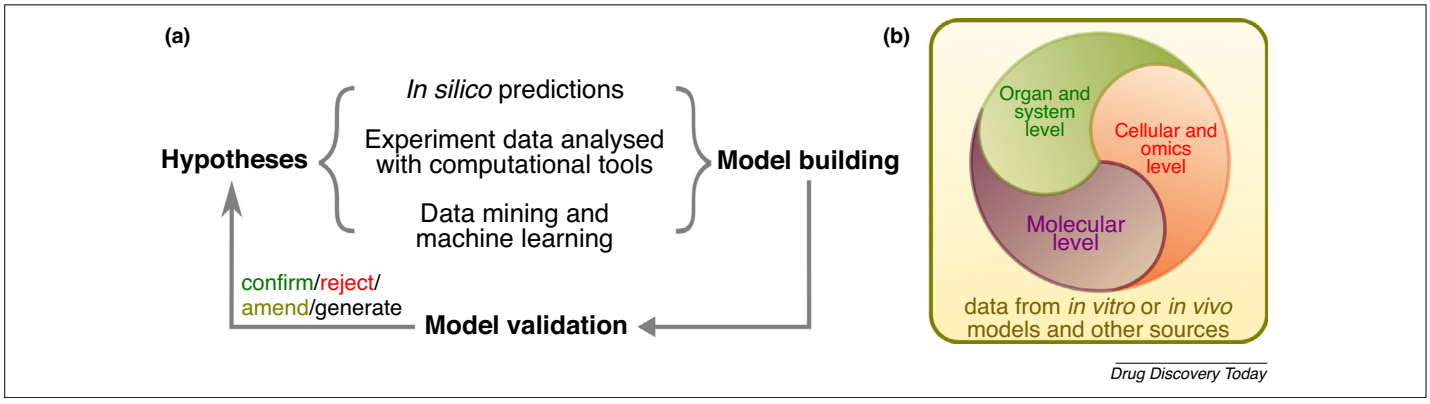


FIGURE 1

Multiscale modelling empowers preclinical drug discovery. **(a)** A drug discovery program can be modelled as an iterative process of hypothesis generation, model building, model verification (not shown for simplicity), model validation, and actions that confirm, reject, or amend the working hypotheses, and last but not least, generate new hypotheses for further testing. Computational approaches can make predictions alone, analyse experiment data, and mine existing data and identify patterns to build models. **(b)** Models are built at molecular, cellular and omics, and organ and system levels. Models of different scales inform and benefit from each other. Data generation, management, interpretation, and integration are equally important to characterise drug mechanism and safety by multiscale modelling.

modelling is also applied in these cases, with sometimes different techniques being used.

When we design novel drugs or improve existing ones, we need to tailor the molecules to have strong interactions with the desired target protein(s) and maximally weak interactions with any undesired off-target proteins. The strength of the interactions largely depends on the shape and electrostatic complementarity of the drugs with their protein binding site, famously coined the lock-and-key principle by E. Fischer [24]. Figure 2 provides an overview of molecular modelling techniques.

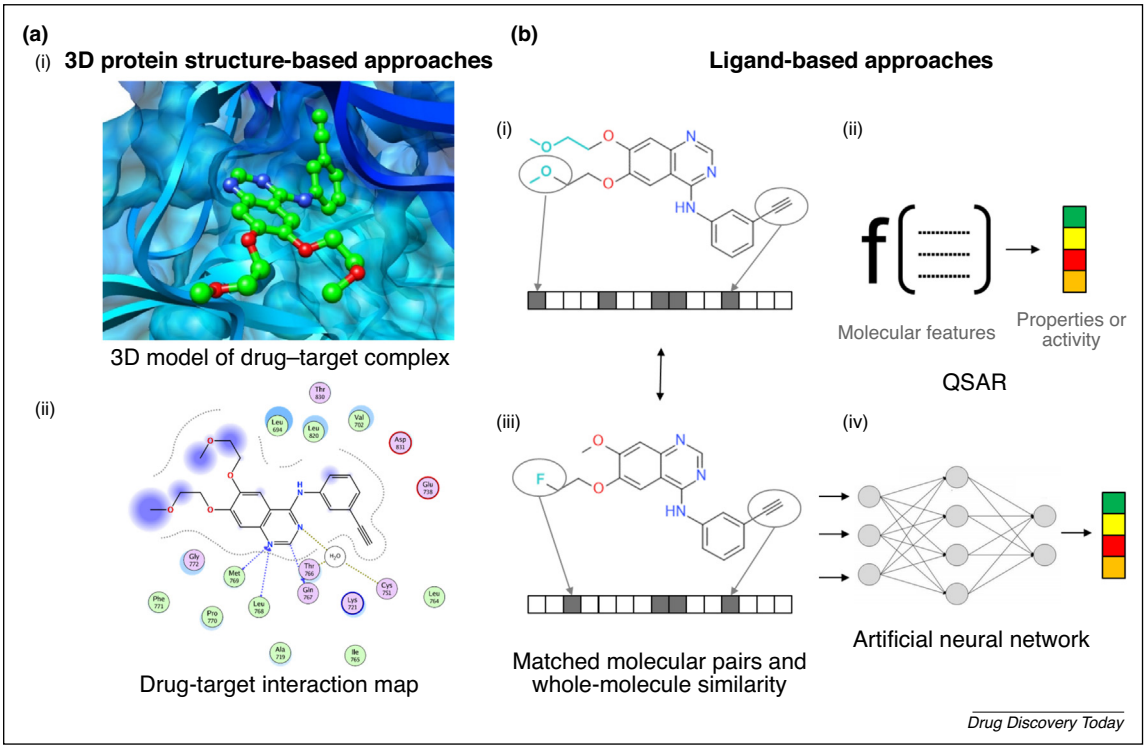


FIGURE 2

Overview of molecular-level modelling techniques. **(a)** 3D protein structure-based approaches. Molecular interactions between a drug and its target protein are key to the biological activity of the drug. Modelling these interactions requires a 3D model of the protein structure and the binding of the drug [(i) 3D model of erlotinib in complex with its target Epidermal Growth Factor Receptor, EGFR]. Different types of interaction between the drug and its target can be modelled at the atomistic or molecular level [(ii) interaction map of erlotinib with its target EGFR and solvent; created with Molecular Operating Environment (MOE) 2019.01, Chemical Computing Group]. **(b)** Ligand-based approaches. Comparison of chemical structures can be used to infer physicochemical and absorption, distribution, metabolism, and excretion (ADME) properties or biological activities. Matched molecular pair analysis focuses on local structural modifications and their effect on properties or activity, whereas molecular similarity analysis takes into account features of entire molecule structures, often encoded in a binary fingerprint [2D structure of (i) erlotinib and (iii) a closely related analogue CHEMBL2068878 [201]; modified parts are highlighted in cyan). Quantitative predictions can be derived from molecular features, using either classical quantitative structure–activity relationship (QSAR) regression methods (ii) or neural networks (iv).

Structure-based molecular modelling

To build models for molecule–target interactions, some structural information is required as a starting point. Currently, X-ray crystallography is the major experimental source for 3D protein-structure data, with currently more than 150 000 protein structures publicly available in the Protein Data Bank (PDB) [25]. Over the past few years, cryo-electron microscopy has emerged as an alternative technology that can deliver atomic-resolution structures of proteins, especially larger protein aggregates and membrane proteins [26]. Many structures of G-protein-coupled receptors (GPCRs) and ion channels have been solved, which now offers the possibility to apply structure-based drug design even for these transmembrane protein targets [27].

Although there are still several drug targets for which no 3D protein structure is known, that gap is becoming smaller [28]. In cases where no 3D structure is available, 3D homology models can be developed if structures of related proteins are known [29], for instance with tools such as SWISS-MODEL [30]. If no 3D structure of any related amino acid sequence is known, *de novo* protein structure prediction methods can be used. They are not yet highly reliable, but the field is developing steadily, as measured by the regular Critical Assessment of protein Structure Prediction (CASP) assessments, and has made promising advances, for instance with neural-network based methods [31,32].

With a 3D model available for the protein target, models of the drug–protein complex can be generated. In the case of small-molecule drugs, this can be done using either automated docking programs or manual modelling, or a combination thereof. The quality of the model can be assessed by docking scores that are based on atomistic force fields, interaction frequency statistics, or other empirical formulae [33]. Recently, machine-learning-based scoring functions have been introduced, which might one day outperform classic scoring functions [34].

Model building is usually a process where the protein target is kept rigid, and only the ligand is relaxed and fit to the binding site to save computational resources. In a further refinement step, candidate models can be energy minimised by molecular dynamics or other methods to allow for a better fit [35]. In laboratory experiments, protein–ligand binding is typically measured at room temperature, where both protein and ligand are flexible. This can be simulated in molecular dynamics simulations. Molecular dynamics simulations can also be used to test whether the modelled protein–ligand complex is stable and, therefore, likely to be true [36,37]. Simulations have greatly improved over the past few years thanks to sophisticated atomic force fields that make fewer simplifications of the underlying physics. Further improvements in modelling the underlying physics, including longer simulated time frames, could lead to even broader application in drug design [38].

3D protein–ligand interaction models are used to understand the atomistic interactions that stabilise the complex and to estimate whether a compound will show measurable binding energy. They can further be used to develop strategies for how molecules have to be modified to optimise the interaction or get rid of it in the case of off-targets [39]. 3D models can be even more impactful if they are validated with crystal structures, especially because compound binding can induce conformational changes that are not seen in the model. Important criteria for validation are: (i) whether a 3D model can explain previously observed experimental structure–activity relationships (SARs); and (ii) whether they are at least able to enrich compounds known to bind to the target among a set of decoy compounds.

Ligand-based molecular modelling

Another important class of model used in drug design are purely statistical ligand-based models. These can be rather simple matched-molecular pair models, where molecular transformations are associated with changes in molecular properties [40]. On a whole-molecule level, entire chemical structures can be represented by sets of features or descriptors in two or three dimensions [41], and the resulting representations are analysed to infer favourable molecular properties based on the 'similarity-property' principle [42]. The traditional way of modelling relationships between chemical compounds and correlating their properties to chemical structures is now being complemented, and partly even superseded, by recent advances in machine-learning algorithms [43–45]. Complex quantitative SAR (QSAR) or pharmacophore models have been generated, where a large set of computed molecular properties is used to train machine-learning models, such as random forests, support vector machines, deep neural networks, or an ensemble of models, that correlate descriptors with experimental measurements. In contrast to 3D protein–ligand models, extrapolation with statistical models needs caution because molecules of interest and their properties might not be well represented in the training set [46]. However, they come with several advantages compared with 3D models, because they are fast to calculate, can be systematically improved with more experimental training data, and typically come with a statistical estimate of the reliability of their predictions [47].

Applications of molecular-level models in drug design and discovery

In a practical drug design setting, both types of model are used for different applications. 3D models are used to optimise compound series for on- and off-target binding where a 3D structure of the target is available. They are mainly applied to generate hypotheses for beneficial chemical modifications of known structural classes in lead identification and lead optimisation campaigns [39]. Statistical models are typically used to predict physicochemical properties and absorption, distribution, metabolism, and excretion (ADME) properties of (virtual) candidate compounds. The quality of many physicochemical property predictions is often sufficient such that experiments for these properties are only occasionally necessary to spot check and improve the model for specific compound classes [48]. ADME predictions are usually followed up with experiments, but the models can also be used stand-alone to prioritise among otherwise promising compounds for synthesis [49]. During the early phases of drug discovery, when it comes to purchasing external compounds and selecting compound classes for further development, models are even more important because experimental data are either not available or usually too sparse to rank compound classes.

A special case is off-target prediction. Finding off-targets that lead to specific toxicities and ultimately adverse effects can be a daunting task. Several cases have been reported where statistical QSAR-type models have helped to identify off-targets of a given compound class [50,51]. Although the prediction of target activity is only coarse, those predictions might be sufficient to prioritise a

few out of many potential off-targets that can then be experimentally followed up. The knowledge of predicted off-targets can also be exploited for drug repurposing, where currently unknown interactions between ligands and predicted targets are exploited for novel therapeutic uses of known compounds [52]. Purely ligand similarity-based methods, as well as combinations of ligand and 3D protein modelling, have been applied successfully to predict unforeseen targets and to identify repurposing opportunities for existing drugs [53,54]. In general, however, it remains challenging to predict on- and off-targets *in silico* with high precision and recall [55].

An emerging application of known and predicted drug–target interactions is the selection and application of chemical probes to understand the molecular mechanisms in phenotypic disease models [56]. Compounds that are well characterised in terms of their molecular MoA, off-target profile, and safety can be used to perturb disease models. Molecules that evoke desired changes of the observed phenotype are analysed regarding their annotated targets and profiled with omics experiments. Systematic computational network analysis techniques are applied to infer molecular targets that are likely causal for the observed phenotype [57]. These target hypotheses can be used to focus laborious biochemical target validation experiments and ultimately lead to a better understanding of the molecular mechanisms underlying the phenotypic model. This approach can help turn a phenotypic drug discovery project into a target-based one, with the benefit of making molecular-level modelling approaches applicable and, thus, accelerating the drug discovery program [58]. Further consideration from the cellular and omics modelling perspective is provided in the subsection '*Chemogenomic and functional genomics tools*'.

Although earlier discussions focus on small molecules, molecular modelling is also indispensable to other modalities. The success of antibodies would not be imaginable without a detailed understanding of the SAR or without creative designs that lead to new MoAs [59]. Antisense oligonucleotides (ASOs) are another beneficiary of molecular modelling. Given that ASOs exert their functions by DNA/RNA interaction mediated by complementary sequence matching [60], it is feasible to model target-binding affinity and off-target potential by integrating biological sequence analysis, RNA structure, and physicochemical models [6].

In summary, molecular modelling has an important role, particularly during lead identification and optimisation. This process can be time-consuming and cumbersome, with success taking many years until a good candidate molecule is found. Computational models can be used to reduce the number of compounds that have to be synthesised, making the entire process more time- and cost-effective.

Cellular- and omics-level modelling

A cell is the smallest unit of life where the mechanism of a drug manifests. Omics technologies quantify and characterise all molecules of one kind (e.g., DNA, RNA, proteins, or metabolites) in one or more cells. Imaging technologies, such as high-content microscopy, capture morphological changes of cells upon perturbation, as well as expression changes of selected genes with either tagged molecules or surrogate markers, such as substrates or products of biochemical reactions [61]. Complementary to these technologies, functional genomics tools, among others chemical mutagenesis, RNA interference (RNAi), overexpression, and CRISPR-Cas9 loss-of-function and gain-of-function screens, offer ways to manipulate thr activity and function of individual genes [62]. Novel cellular models are used for disease and safety modelling, including induced pluripotent stem cells [63] and microphysiological systems, such as organ-on-a-chip [64] and organoids [65]. Taken together, technological and biological tools deliver huge volumes of data that allow computational tools to characterise drug mechanism and safety at the cellular and omics levels.

Properly designed, conducted, and analysed comparative experiments are key to building cellular- and omics-level models. The experiments are comparative in the sense that an absolute value, such as the speed of light, is not of interest; rather, a molecule is compared with controls or with other molecules of known mechanism [66]. Given the high dimensionality, large volume, and noisy nature of omics and imaging data, it is necessary to use appropriate statistical models to infer differences and to incorporate prior knowledge of biological systems to infer the mechanism and safety profiles of the molecule.

Modelling with omics data

Omics approaches are indispensable to modern drug discovery. They combine high-throughput biophysical and biochemical assays with statistical modelling, data mining, network analysis, and machine learning to offer a comprehensive view of gene-, pathway-, and network-level regulation by drug candidates [67–69]. Models trained with data from nontoxic and toxic compounds identify signatures applicable in preclinical safety screening [70,71].

Remarkably, omics data and models build an essential link between molecular-level modelling and organ-system level and population modelling. They can characterise molecular variations in a human population, such as single-nucleotide polymorphisms (SNPs) and other structural variants, expression levels, and splicing isoforms, which inform the status of the drug target and other genes that potentially affect PK/PD profiles in healthy or diseased individuals. For instance, a pharmacogenomics study of GPCRs demonstrated the power of omics linking models at other levels, where integrative analysis of the genetic architecture of a large population recovered individual differences in response to pharmacological modulation [72].

RNA-sequencing-based gene expression profiling is currently the most widely used omics technique to understand drug mechanisms, as judged by statistics from public data repositories. Complementary to standard full-transcriptome sequencing, several platforms exist that lower cost and reduce information redundancy. Examples include L1000, where ~1000 genes were selected to maximise the retained variance of gene expression data [67], DRUG-Seq, which is empowered by miniaturisation and shallow sequencing [73], and molecular phenotyping, where ~1200 pathway reporter genes were selected to infer pathway activity [68]. Independent of the variant, a typical data analysis workflow includes mapping reads to a reference genome or transcriptome, technical and biological quality control (QC) [74], dimension reduction and clustering, and differential gene expression analysis [75]. The differential expression profile is then compared with those of compounds with known MoA to infer the mechanism, following the 'guilt-by-association' principle [67]. In parallel, the profile can be interpreted with prior knowledge of gene regulatory

networks to infer latent genes and pathways that are likely regulated by the compound, for instance using gene-set enrichment analysis [76,77].

Our knowledge of gene regulatory networks, particularly how they vary by cell identities and by context, is scarce. One of the few things that we are certain about is that these networks are intrinsically complex. It is challenging to recover direct binding targets unambiguously from gene-expression data. Nevertheless, gene expression can offer valuable insight into the mechanism and safety profile of compounds, and comparative studies with more than one compound can reveal common and distinct mechanisms underlying the same phenotype. In a previous study, for instance, we identified two compounds in a phenotypic screening that convert stem cell-derived white adipocytes to display the brown-adipocyte phenotype. RNA-sequencing data analysis revealed that both molecules inhibit the Janus kinase/signal transducers and activators of transcription (JAK/STAT) signalling pathway. Whereas one molecule, tofacitinib, is a specific JAK3 inhibitor, the other, R406, shows polypharmacology, inhibiting multiple pathways simultaneously [78]. In another study, specific inhibition of epidermal growth factor (EGF) target genes was identified with molecular phenotyping. Together with other results, the finding helps to establish inhibition of EGF uptake as a trait of nephrotoxic antisense drugs [71].

RNA sequencing (RNA-seq) provides more information than gene expression. Splicing isoform quantification, for instance, led to the finding that the *CD44* isoform status predicts response to treatment with an anti-CD44 antibody both in cancer cell lines and in patients [79]. Another example, differential splicing of the *SMN2* gene, is discussed in detail later.

Over the past few years, RNA-seq technology has evolved from profiling many cells ('bulk-mode') to being able to sequence single cells ('single-cell mode') [80]. This progress, together with other single-cell technologies such as flow cytometry [81], mass cytometry time-of-flight (CyTOF) [82], and nuclear magnetic resonance (NMR) [83], motivated computational approaches to study biology at the resolution of single cells [84,85]. Although still in their early days, single-cell techniques have greatly expanded our ability to understand how drug perturbation affects individual cells and the communication between them.

Beyond the transcriptional level, proteomics characterises either expression or post-translational modifications of proteins *in vitro* or *in vivo* [86]. For small molecules permissive to chemical modifications, chemoproteomics identifies binding targets of modified drug candidates using the 'bait-and-prey' principle [87]. Modification-free methods, such as antibody-based cellular thermal shift assay (CETSA) and mass-spectrometry-based variants, are also available [88,89]. Compared with RNA-seq, where both devices and software solutions are more standardised, the data generation and analysis landscape of proteomics used to be fragmented. Recently, several community initiatives have benchmarked statistical analysis strategies and established best practices [90–92].

Imaging

Imaging is a sister technique of omics to study the cellular effects of drug candidates. Traditionally, it uses optical systems and computer-vision techniques to analyse cell morphology [93,94]. Emerging optics-free molecular imaging systems, such as NMR, positron emission tomography (PET) [95], and DNA microscopy [96], hold the promise of revealing the spatial organisation of biological molecules of interest.

Computational approaches are indispensable to imaging, because the sheer amount of data prohibits manual analysis, and the set-up of image assays and the analysis pipeline engage and influence each other [97]. Relevant computational concepts and non-commercial software are reviewed in [98].

Whole-slide imaging in pathology with automated image analysis is an important application that links cellular and organ-level information [99]. Cell imaging can also be used to cluster drug candidates based on their MoA [93,94]. This is of particular relevance for antimicrobial discovery because morphological changes can be concordant with the MoA of bacteria killing [100,101]. When profiling compound libraries that are annotated with target information, high-content imaging was reported to be able to predict biological activity, namely back-translating from phenotype to target [102], which, if generalisable, can be useful to identify target profiles of poorly characterised compounds.

Chemogenomics and functional genomics tools

Complementary to quantification methods, perturbations of different kinds can be used in combination with a drug candidate to interrogate its target and MoA. This orthogonal perturbation can be induced by chemical probes [33] or functional genomics tools [42]. The underlying assumption is that the MoA of the drug candidate is mediated by one type of biological molecule, and knockdown, knockout, or inhibition of that type of molecule abolishes the pharmacological effect of the drug. Similarly, overexpression or activation augments the effect. A combinatorial screening can reveal the biological molecule of interest.

One pitfall of these approaches is that, given the complex structure of gene networks, modulation of many genes can partially or fully antagonise the effect of the molecule, leading to a long list of candidates for validation. When the chemogenomics or functional genomics tool displays off-target effects, target deconvolution is even more challenging. Nevertheless, with the rapid development of *in vitro* and *in vivo* use of gene-editing systems, functional genomics approaches will likely have a huge impact on target identification and validation [103]. In such campaigns, bioinformatics tools are used to design specific agents to interfere with the expression and biological function of the putative target, and statistical models, equipped with prior knowledge, are engaged to identify the most likely targets that deserve further experimental validation.

A call for model-driven omics data interpretation and integration

We argue that different cellular and omics disciplines need to be combined and to be supplemented by models of various kinds, including molecular-level and organ- and system-level models, to enable scientific understanding of a biological process. Figure 3, a simplified diagram inspired by Thomas *et al.* [104], illustrates our point.

The centre diagram of Fig. 3 describes schematically the homeostasis of arginine biosynthesis in *Escherichia coli*. Arginine, the final biosynthesis step of which is catalysed by ornithine transcarbamylase (OTCase), binds to a transcriptional repressor and allosterically triggers its binding to promoter elements in both the OTCase and the repressor genes. When arginine is lacking,

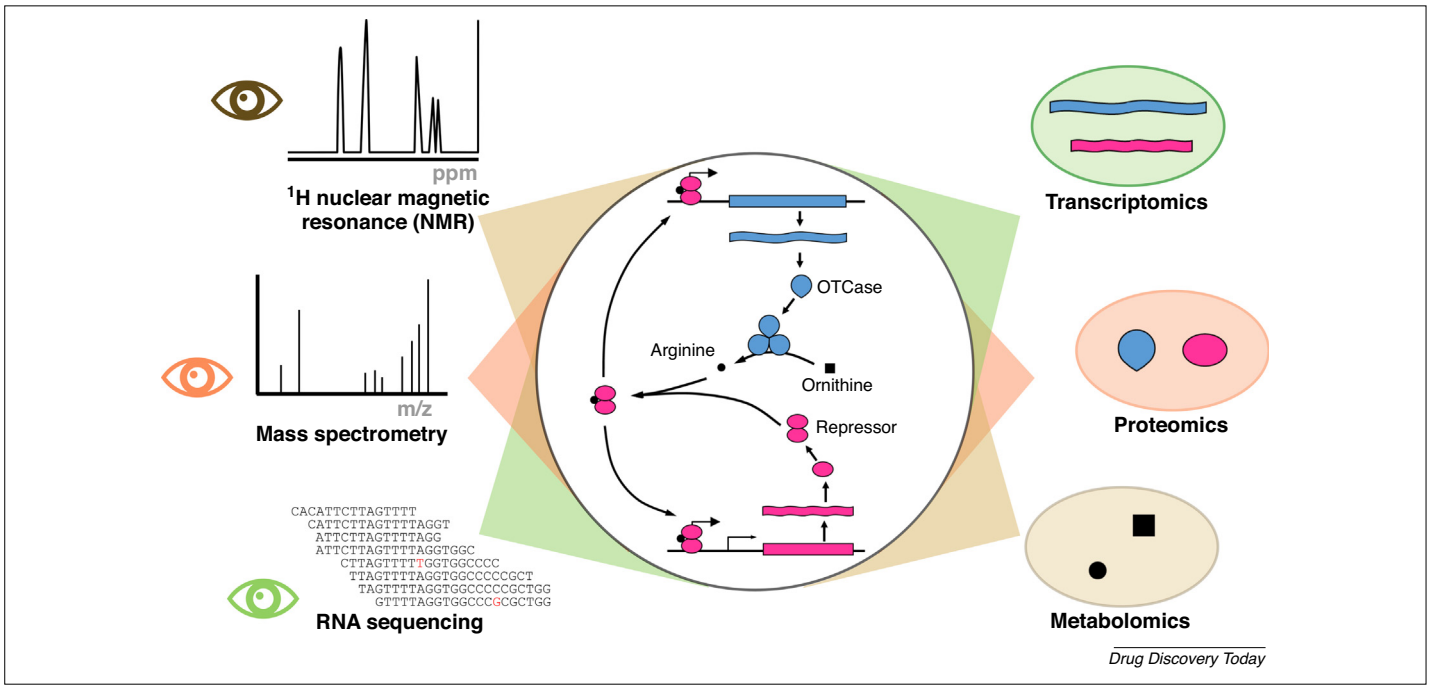


FIGURE 3

Omics data are projections of high-dimensional biological space. The centre diagram illustrates the homeostasis of arginine biosynthesis in *Escherichia coli*. Detailed descriptions can be found in the main text. Eyes indicate different perspectives of omics studies. High-throughput technologies, with ^{1}H nuclear magnetic resonance (NMR), mass spectrometry, and RNA sequencing as examples illustrated on the left side, examine all molecules of a particular kind in the cell system. They generate data of the metabolome (all metabolites), proteome (all proteins), and transcriptome (all RNAs), respectively, shown on the right side. These data, despite their high dimensionality, are low-dimensional projections of high-dimensional biological space. The same principle also applies to cellular imaging data.

repression is relieved, allowing new arginine to be produced by OTCase. Low-level, constitutive expression of the repressor gene ensures that the repressor will be available to bind arginine again once its levels have risen above a certain threshold.

The three projections surrounding the central diagram in Fig. 3 represent metabonomic, proteomic, and transcriptomic views of the system, respectively. They are generated by high-throughput technologies, namely NMR, mass spectrometry, and RNA-seq, as depicted. Based on repeated measurements of the system under various conditions, it will be possible to establish positive (ornithine and arginine) and negative (repressor and OTCase) correlations in the amounts of the analytes for any one of the three omics disciplines. However, a comprehensive understanding of even this simple network would not be possible based on any one of them alone. Thus, based on this example, albeit arguably theoretical, and other observations, we call for caution in the use of model-free approaches to drug mechanism and safety (Box 1).

A classical approach towards studying such a relatively small network would be to derive a schematic network diagram with (abstract) nodes and edges, either from literature data, dedicated molecular interaction databases, or based on correlation analyses of omics data generated under a range of conditions. Such a schema, coloured using quantitative data from an omics experiment, is often used in publications to cover the network aspect of a biological system of interest. With more complex data available, collected under different conditions, time points, or perturbations, statistical analyses enable the construction of correlation networks or Bayesian networks, suggesting functional or regulatory links between the measured entities. From an algorithmic point of view, correlation networks are liable to high false-positive and false-negative rates, even using single-cell expression data [105]. One structural drawback of Bayesian networks is that they produce directed acyclic graphs (DAGs), which cannot model feedback loops, a feature considered to be central to biological processes. Despite these limitations, it is possible to improve the results of network inference by using single-cell time-series data [106], by following the consensus of diverse algorithms [107], and by combining information from several omics disciplines, such as genetics, proteomics, and genomics [108,109].

It is important to distinguish between modelling the data and modelling the underlying biological processes. The latter types of model are mechanistic in nature. Ideally, inferring correlation or a Bayesian network model from an omics experiment will ultimately inspire a mechanistic model of the underlying biological process, but they are not guaranteed to be mechanistic *per se*. As the simple example in Fig. 3 demonstrates, there are several key aspects for understanding the mechanism and, ultimately, the dynamics of a biological network, that are nonobvious even from a combined omics point of view. These include: (i) the nature of the functional protein complexes, including the repressor dimer and the OTCase trimer; (ii) the concrete physical interactions, including those of the small molecule with repressor proteins, those of the repressor protein with DNA-binding sites, and those of the enzyme catalysis that leads to the production of arginine; and (iii) the presence and functional relevance of regulatory elements in the gene promoters.

BOX 1

Beware of model-free omics'Omics' approaches, by definition of the neologism, cover the entirety of detectable (and, typically, quantifiable) species of a certain kind, such as all mRNAs, all proteins, all miRNAs, and so on. Thus, their appeal stems from a combination of comprehensiveness and lack of bias: a properly done transcriptomics study will not miss any present transcript, and it does not make any prior assumptions about the relevance or role of any one of the measured entities. All the possibly attainable information about the biological system under study, it appears, must be contained.

Based on these premises, it is clear that omics data lend themselves readily for modern data analysis approaches, and can even be considered examples of 'big data'. Given a sufficiently large collection of, say, genetics or proteomics data sets, most of the biologically relevant patterns (of heritability, or of protein interactions) should be detectable in an almost 'model-free' approach, based on either statistical measures of correlation and significance, or causal models where the assumptions of causal inference are met. The only aspects of such studies that depend on scientific model assumptions are the identification of the relevant '-ome', and the knowledge of how to detect and measure its instances.

However, these aspects turn out to be very important and often underestimated in their consequences.

First, omics approaches never imply the measurement of all relevant players in a biological system. For example, transcriptomics, proteomics, and metabonomics represent three projections onto partially correlated subspaces of biological entities, none of which will provide more than the most superficial understanding by itself. Merging the three in 'systems biology' or 'integrated omics' efforts requires much stronger use of state-of-the-art biological models. Foregoing such integrated approaches leads in essence to a 'targeted approach' based at least on the (far-reaching) assumption that certain species of biological entities will turn out to be more relevant for a given question. From there, it is just one more step to truly targeted approaches (e.g., to monitor only the 'kinome' of protein kinases, or the 'pathway reporter genes' of the molecular phenotyping platform), methods that depend heavily on assumptions of scientific models.

Second, the theory and practice of generating omics data has a substantial influence on our interpretation of data and derived results. A well-known example is the era of hybridisation-based 'array' methods in transcriptomics. While opening the door for feasible, meaningful, and affordable omics data generation, they were, at least in their early years, mostly agnostic of splice variants. This must clearly have had an influence on the generation of hypotheses and interpretation of data, because what we cannot assess experimentally is the subject of speculation only and, if no improvement of methodology is provided, might even tend to be forgotten over time: it essentially 'ceases to exist'. Modern omics approaches studying the entirety of genetic enhancer elements, CCCTC-binding factor (CTCF) binding sites, histone modifications, or 5-methylcytidine provide evidence that the discovery of new, biologically relevant 'omes' has certainly not yet come to an end.

Finally, scientific models become vital when it comes to establishing crosslinks with other domains discussed in this review, the molecular and the organism level, and to transfer knowledge between them.

These concepts ultimately need to be incorporated into a mathematical model of this biological process, especially when our aim is to build mechanistic models predicting and explaining biological phenomena.

As discussed by Thomas *et al*. [104], the next step towards a truly integrated understanding of the network is therefore to build a mechanistic model, such as the simple one in Fig. 3, and to analyse its feedback loops, steady states, and parameter space using 'logical network analysis'. Such an analysis not only allows the identification of stable states or homeostatic loops, but also predicts effects of mutants, such as a repressor mutant with reduced affinity to arginine, or a promoter mutant that lacks the repressor binding site, at least in a qualitative way.

The ultimate, and often the admittedly distant, goal is to learn the kinetic parameters of the biological processes being studied, to describe the system using differential equations, and to simulate fully integrated dynamical models of all the involved components [110]. The availability of such models would then open the possibility to run truly *in silico* experiments and to computationally predict effects of perturbations that were never observed in an experiment.

The increased complexity in human and eukaryotes compared with prokaryotes makes network inference with omics data an intimidating task [107]. Nevertheless, examples abound where insight into human disease biology is gained by data integration [109] and modelling [111]. Readers interested in the mathematical and methodological aspects of these approaches are welcome to consult two well-written reviews [112,113]. We believe that predictive models of disease biology and drug safety integrating molecular, cellular and omics data will substantially increase our ability to prioritise targets and to understand drug mechanism and safety. Such models can be integrated into larger organ- and system-level models, as described later.

Organ- and system-level modelling

Pharmacokinetics (PK), the fate of drug substances administered to a living system, and pharmacodynamics (PD), the biochemical and physiological effects of drug substances, jointly determine the efficacy and safety profile of a drug. Mechanistic or semimechanistic mathematical models are important tools to quantify both the PK and PD profiles of drug candidates. Such models integrate information at the organ and whole-body level to establish the relationship between dose, exposure (measured in plasma, at the target site, and at off-target sites), and response, including efficacy and adverse effects. The applications of PK and PK/PD modelling in drug discovery include: (i) quantitative evaluation of the therapeutic window and, hence, informing go-no-go decisions [114,115]; (ii) integration and translation of preclinical data to inform the first-in-human dose [116–118]; (iii) informing clinical study design and drug labelling [119–121]; (iv) characterisation of patient-to-patient variability and informing optimal dose for subpopulations [122,123]; and (iv) extrapolation to special subpopulations, such as the paediatric population [119]. The adoption of PK/PD models has been strongly encouraged by health authorities, for instance through the Model-Informed Drug-Development Pilot Program (MIDD) from the US Food and Drug Administration (FDA).

Although empirical models have been widely and frequently used in PK and PK/PD modelling, here we focus on physiologically

based or mechanistically detailed models, where human physiology and pathophysiology is described mathematically as an underlying system with which the compound interacts. These models are intrinsically multiscale. On the molecular level, they can describe subcellular processes, such as enzyme reactions [124], ion channel kinetics [125,126], transporter activities [127], and subcellular signalling processes [128]. On the cellular level, they can model processes such as proliferation and apoptosis [129] as well as cell–cell interactions [130]. One level above, they can model organ- and system-level processes, for instance, whole-heart electrophysiological activities [131] and processes involved in ADME [132]. Last but not least, physiologically based or mechanistically detailed models can characterise individual variability in a population [133].

The modelling of physiological and/or pathophysiological systems is based on our understanding of the underlying biology. The models are parameterised by mechanistic data, for instance, cell proliferation rate, receptor expression level, or ion channel electrophysiology. In areas where biology is not well understood and data are scarce, an empirical or semimechanistic approach is required. Otherwise, if the biology is well studied and data are available from both human and preclinical species, these models can quantitatively capture interspecies differences and serve as a powerful tool to translate between preclinical discovery and clinical development [134–136]. If observational data confirm model predictions, we gain confidence in our understanding of the biological system and accept our hypothesis of the drug. Otherwise, if the data contradict model predictions, and if we are convinced that the data are free from bias and error, we shall revise the model and, thus, gain new insight.

To model the drug effect on the system, the models integrate compound-specific mechanistic data, such as ADME properties and target or off-target site interaction data, to predict PK, safety endpoint, or efficacy biomarkers under different dosing regimens. Identical or similar model structures can be parameterised to study the same processes or mechanisms in different physiological systems (e.g., preclinical species, healthy adults, and paediatric or patient subpopulations). Mathematically, these models are often systems of ordinary differential equations (ODEs), describing variables changing over time, for example, drug concentration in time course or the number of tumour cells at each time point. Partial differential equations (PDEs) [137], stochastic differential equations [138,139], and agent-based models [140,141] are also occasionally used. A schematic illustration of the physiological based systems model is shown in Fig. 4.

To highlight how these models assist drug mechanism and safety understanding, we focus here on two aspects: physiological-based PK (PBPK) models, and systems models for safety, using cardiac safety models as an example.

PBPK models

PK profiles of drugs are classically modelled by compartment models with first-order kinetics (see [142] for a concise introduction and [143] for a comprehensive treatment). Compartments in such models do not necessarily have a physiological meaning. By contrast, a PBPK model comprises compartments that represent organs or tissues (e.g., liver, heart, brain, etc.), and these compartments are connected by blood or lymph flow. The model is typically a system of coupled ODEs and contains physiological parameters such as intestinal fluid volume, intestinal pH, blood flow rate, organ/tissue volumes, and transporter and metabolic enzyme expression levels. These system parameters can be adjusted to reflect different physiology conditions. For instance, the demographic parameter and the metabolic enzyme abundance parameter can be adjusted to build a virtual patient cohort with a particular ethnic and genetic background [144]. The model uses drug-specific parameters, such as solubility, permeability, plasma protein binding, tissue-to-plasma partition coefficient, enzyme kinetics and intrinsic clearance, most of which are determined

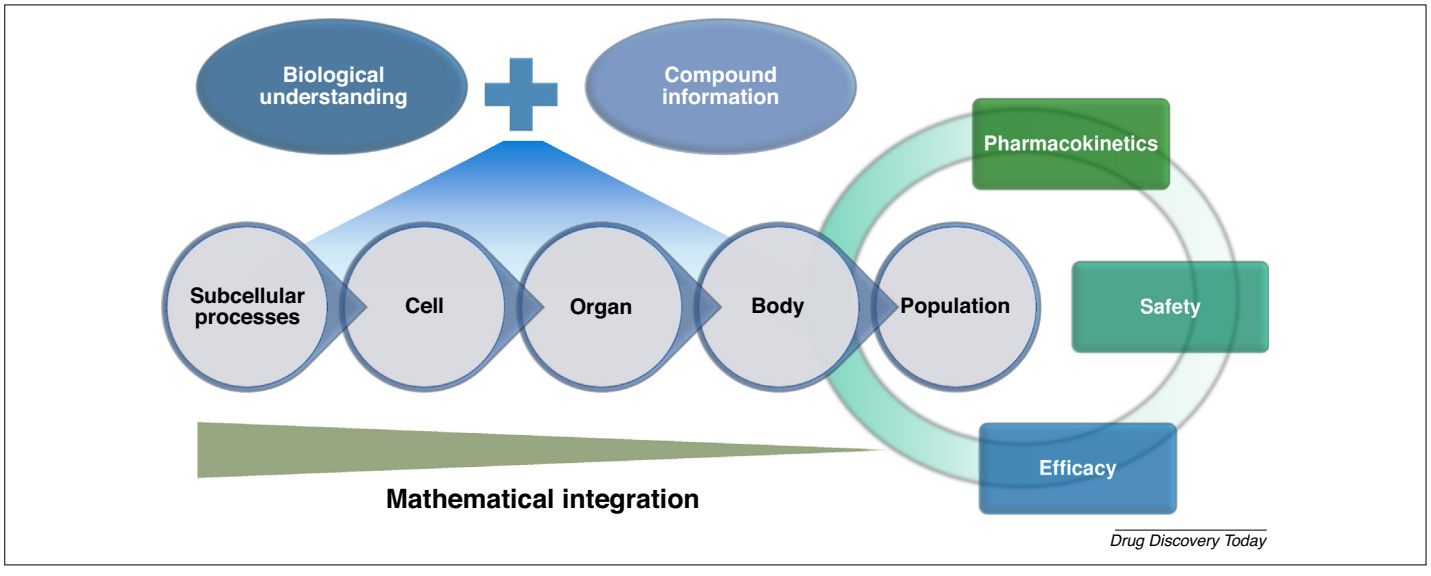


FIGURE 4

Schematic illustration of system- and organ-level models. Such models often take the form of a system of coupled ordinary differential equations (ODEs), with biological understanding and compound information encoded as parameters. Information of different scales is integrated to inform pharmacokinetics, safety, and efficacy profiles of drugs over time.

by preclinical mechanistic studies. The compound distribution into each organ can be typically modelled as either perfusion limited, in which the distribution is limited by the blood flow rate, or permeability limited, in which the distribution is limited by the ability of the compound crossing the membrane. The basic concept of the PBPK model, in particular for small molecules, was nicely reviewed by Jones and Rowland-Yeo [134]. Reviews and examples of PBPK models for biologics can be found elsewhere [145–147].

PBPK, especially small-molecule PBPK, is one of the most established systems models and has had a significant regulatory-included impact in the pharma industry. Between 2008 and 2017, the FDA's Office of Clinical Pharmacology received 130 investigational new drug (IND) applications and 94 new drug applications (NDA) that contain PBPK modelling, excluding *de novo* PBPK analysis that informed the regulatory decision for 30 submissions [119]. By the end of 2015, the European Medicines Agency (EMA) had received a total of 67 submissions of PBPK models, 20 of which were suggested by the regulator [148]. Between 2014 and 2016, the Japanese Pharmaceuticals and Medical Devices Agency (PMDA) received 17 submissions of PBPK analysis in NDAs [149]. The main areas of PBPK regulatory applications include, but are not limited, to prediction of drug–drug interactions (DDI) mediated by P450 enzymes [150,151] or transporters [127,152], prediction of paediatric exposure to inform the dosing regimen [153], prediction of exposure for patients with organ impairment [154–156], and prediction of food effects [156].

Systems models for safety

Systems models offer an integrative approach to combine mechanistic preclinical safety data with PK and patient-specific characteristics. They can assess safety liabilities under different dosing regimens in different populations quantitatively and identify mechanisms underlying liabilities, which may offer insight for compound optimisation. Similar to PBPK models, these safety models serve as a platform and are not restricted to particular disease areas.

Cardiac safety models are some of the most advanced systems safety models and have gained regulatory attention and encouragement [157]. These models describe cardiac electrophysiology at multiple levels mathematically. At the cellular level, the opening, closing, and inactivation of cardiac ion channels, the activities of ion pumps and exchangers, the intracellular calcium ion handling, and the cyclic transmembrane potential change are modelled. At the tissue and/or organ level, both the electrical coupling between cardiac myocytes and the electrical signal propagation through the heart are modelled. At the whole-body level, the electrical signal propagation to the torso surface and, hence, the electrocardiogram signal, are modelled [158]. The cellular electrophysiology is modelled as coupled ODEs and the organ and/or torso models are governed by the bidomain model, a PDE model, coupled with the cellular ODE models.

Similar to PBPK models, cardiac safety models also require a set of systems parameters, such as ion channel kinetics (e.g., transition rates between different states), ion channel conductance, which is associated with their abundance, as well as heart and torso anatomy. The models integrate information on drug interaction with different ion channels, which for instance can be inferred from *in vitro* ion channel inhibition data from the patch-clamp experiment, to predict the impact on cellular action potential, whole-heart electrical signals, or electrocardiogram (ECG). One typical aim of such predictions is to assess the proarrhythmic potential of drug candidates. The development and use of cardiac models for safety assessment have been reviewed elsewhere [159]. Recently, the FDA-led Comprehensive *in vitro* Proarrhythmia (CiPA) initiative published a series of papers on training [160,161] and validation [162] of a model to predict the risk of torsade de pointes, a rare but potentially fatal arrhythmia, which demonstrated the feasibility of establishing a predictive cardiac safety model.

Apart from guiding compound and dose selection to minimise proarrhythmic risk, these models also inform the intensity of clinical ECG monitoring and drug labelling, including prohibited concomitant medications [121,163]. In November 2018, a concept paper from the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) was released [163]. It indicated that the cardiac safety model, following the path of the PBPK model, will soon become another systems model in the regulatory context. Furthermore, other systems safety models are emerging, such as those that predict drug-induced liver injury [164].

Although we have focussed on applying systems models for safety so far, systems modelling has also been widely used to quantitatively describe disease biology. For example, a model constructed by Clausznitzer *et al.* described lipid dysregulation in the brain of patients with Alzheimer's disease (AD). The model predicted that activation of sphingosine-1-phosphate receptor 5 (S1PR5) can reverse the lipid dysregulation in AD and, therefore, that S1PR5 may serve as a promising target [165]. In addition to building confidence in a target and revealing potential new targets, these models are also used to optimise dosing regimens [166,167] and to stratify patients into responders and nonresponders [168].

Towards multiscale modelling: a case study

We illustrated earlier the principles of the three levels of modelling. The different modelling approaches share mathematics as the common language to describe the complex system of human biology and how drugs assert their effects. They also share informatics as the common tool to handle data, perform computations, make inferences, and transform models and data into human interpretable results. Moreover, they complement each other to shed light on human disease biology and the mechanism and safety profile of drugs in a holistic view.

In the parable 'Blind men and an elephant', each blind man perceives an elephant, depending on where he touches it, as a snake (the trunk), a fan (the ear), a tree trunk (the leg), a wall (the side), a rope (the tail), or a spear (the tusk) [169]. Each model only reveals part of the truth at best. To understand the elephant, it is necessary to combine different views, synthesise the evidence and models into a unifying model, and put that model to test. In our case, drug mechanism and safety is the elephant, and clinical trials are the test. We foresee that a multiscale modelling approach will improve our understanding of drug mechanism and safety profiles over isolated modelling approaches.

To illustrate the potential, let us examine a case elephant. Small-molecule splicing modifiers of the human survival of motor neuron 2, centromeric gene (*SMN2*) in the context of spinal muscular

atrophy (SMA) provide a recent, concrete example of how multiscale modelling can help elucidate mechanism and safety profiles of a drug candidate. Discussions below are based on published data [170–173].

The transcriptomes of patient cells in the presence or absence of test compounds were profiled with RNA sequencing. Although differential expression profiles were readily available using standard algorithms, these effects were deemed secondary to the primary effect of the drug, namely corrected splicing of the *SMN2* gene and potentially other, off-target, genes. In support of this notion, a chemoproteomics analysis with protein extracts from patient cells provided evidence for an interaction of a derivative of the candidate compound with the cellular splicing machinery.

Therefore, a second analysis was based on only those sequencing reads that uniquely represented one of the two possible alternatives in any of the about ~300 000 alternative, local splicing events described in the RefSeq human transcriptome database [174]. Based on this, every local splicing event (as opposed to a full-length splice variant, which can depend on the combination of any number of local alternative splicing events) was monitored and characterised by changes in a 'percent spliced in' score (ΔPSI) [171]. Although all of the tested compounds exerted comparable effects on the target splicing event in *SMN2*, they showed different, although largely overlapping, profiles in terms of additional, off-target events. A comparable analysis by Palacino *et al.* [170] had already revealed a preferred splice site consensus pattern of GA|GUAAGU in targeted splice sites, as opposed to the canonical donor site AG|GURAGU, where the vertical bars indicate the exon–intron boundary. Direct binding of the drug candidates to the complex formed from such 5′ donor splice sites and U1 snRNA was then demonstrated by solution-state ^{1}H NMR, and a structural model was derived (Fig. 2 in [171]). Through this integrated approach, patterns in an omics data set were ultimately explained in terms of ever more-refined mechanistic models, which finally elucidated the MoA at an atomic resolution [173]. The findings strengthened confidence in the original models and the compounds under investigation.

To explain the differences in off-target potential between different candidate compounds, an additional search for shared sequence patterns was performed that identified a 12-base pair-long stretch of purine residues ~25 base pairs upstream of the crucial donor splice sites in both the primary target and the most persistent off-target splice site across all the experiments performed. This sequence stretch had already been described as an exonic splicing enhancer (ESE) in the case of the target *SMN2* gene [175]. Subsequent analysis using surface plasmon resonance confirmed that the most selective compounds gained their increased selectivity through the joint interaction with the splice site and this purine-rich ESE [171,173].

As summarised in [172], substantial additional efforts in the areas of pharmacology, drug metabolism and PK (DMPK), and nonclinical safety were required to further understand the *in vivo* effects of the studied compounds and to ultimately allow a few of them to progress to clinical trials. A particularly interesting example is the characterisation of 'compound 2' (Fig. 1 in [172]). After oral administration in rodents and cynomolgus monkeys, it was found to be dealkylated *in vivo* (Fig. 5 in [172]), yielding a related 'compound 6', which reached plasma concentrations of up to 9% of those of the originally administered 'compound 2', but being tenfold more potent than the parent compound. This metabolite was also peripherally restricted (i.e., did not reach the brain) and, thus, showed a different overall profile compared with the parent 'compound 2'. These findings led to the design and characterisation of more advanced and advantageous compounds as detailed in [172].

In summary, transcriptomics data and their context-specific analysis have inspired chemoproteomics, structural and bioinformatics pattern searches, and DMPK analyses that, in combination, provided deep insight into the mechanism and safety profile of risdiplam, a drug candidate for SMA, enabling its progression to clinical trials. Thus, multiscale modelling integrates learnings across disciplines to accelerate discovery and development.

Concluding remarks and outlook

Here, we introduced molecular-level, cellular- and omics-level, and organ- and system-level models, examined their applications, and illustrated the potential impact of a multiscale modelling approach in drug discovery with the case study of risdiplam.

The concept of multiscale modelling has been proposed in many different areas. In physics and material science, it summarises mathematical concepts and tools that exploit the disparity of different scales and provides solutions to multiphysics problems [176,177]. In biochemistry, multiscale modelling that integrates classical molecular modelling and quantum models had a fundamental role in elucidating complex biological and chemical systems, for which Martin Karplus, Michael Levitt, and Arieh Warshel were awarded the Nobel Prize in Chemistry 2013. In biology, it has been realised that a multiscale modelling approach is indispensable to understand the hierarchical nature and the complexity of diverse biological systems [178–180]. In drug discovery, we propose multiscale modelling as a process of building mathematical, preferably mechanistic, models at individual levels and integrating them. We believe that multiscale modelling will not only lift productivity and reduce the attrition rate, but also lead to deeper and novel insights into disease biology and, ultimately, to new therapeutics.

We have limited our discussions to preclinical research. We note that other aspects of drug discovery and development, including biomarker and clinical development, have benefited from model-driven approaches, such as pharmacometrics and quantitative systems pharmacology, as comprehensively reviewed in [181] and [182], respectively. The trend of using models to inform decisions is also observed beyond the pharma industry. For instance, European Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) legislation already accepts predictions made by QSAR models in special cases [183].

Although the short-term goal of multiscale modelling is to enrich the knowledge of a particular molecule, either to gain confidence or to stop and fail early, one key long-term goal is to accumulate knowledge in disease biology, pharmacology, and toxicology in the form of testable, predictive, and ideally personalised, models. Reverse translation, the practice of extracting patterns and building models from clinical trials as well as other real-world evidence to inform preclinical discovery projects, is a viable option to feed population models and patient-relevant data into the multiscale modelling schema [184]. Multiscale modelling

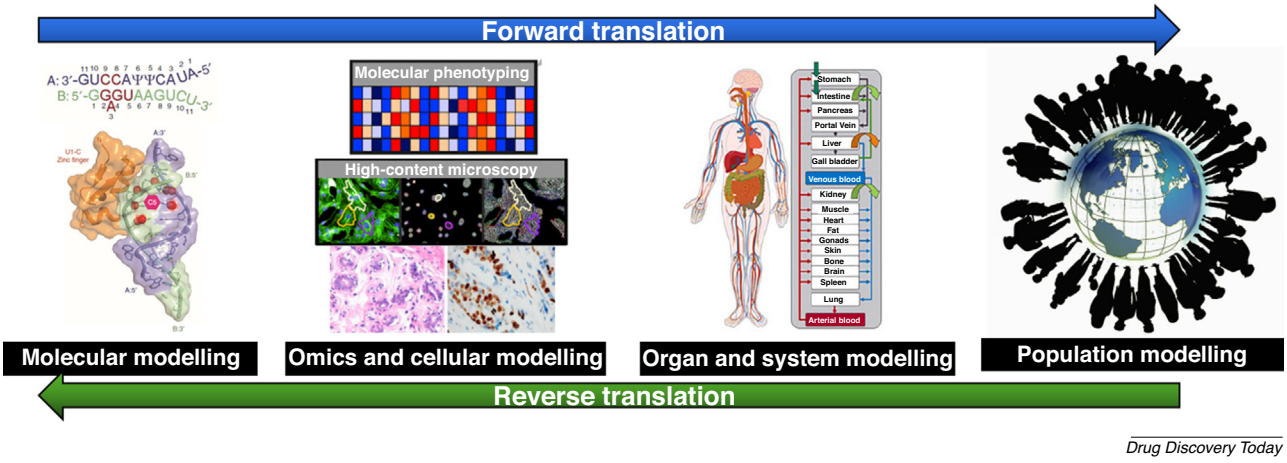


FIGURE 5

Multiscale modelling as a computational engine serves forward and reverse translation in drug discovery. Modern drug discovery depends on both forward translation, the process of turning molecular findings into drugs, as well as reverse translation, the process of turning real-world evidence and data into knowledge that catalyses future therapeutics. Multiscale modelling supports both processes by modelling biological systems at different levels and integrating models of different scales. Adapted from [171] (molecular modelling), [68] (cellular and omics modelling), [132] (organ and system modelling), and obtained from pixabay.com (user: geralt) under the Pixabay License (population modelling).

as a computational engine has the potential to serve both forward and reverse translation, as illustrated in Fig. 5.

Understanding drug mechanism and safety takes hard, creative work and a long time. Take acetylsalicylic acid (Aspirin®) as an example. It was first synthesised and manufactured between 1897 and 1899. Not until 2019 was a new mechanism reported, in which acetylsalicylic acid directly acetylates cyclic GMP–AMP synthase (cGAS) to prevent DNA-induced autoimmunity [185]. Although it might appear anecdotal, historical reviews suggest that the time span between drug discovery and mechanism and safety understanding can take years to decades, well beyond the usual timescale of a preclinical drug discovery program [186,187].

The discrepancy between the time and effort needed for a mechanistic understanding of drug mechanism and safety at multiple levels and business life-cycles of the pharma industry implies that new ways of research are needed. Collaborations between industry, academia, and research institutes to build multiscale models jointly might be beneficial. Indeed, technical solutions that allow private and practical pharmacological collaboration already exist [188]. Community-wide efforts, such as CASP assessments and DREAM challenges, participated in by both academic and industrial research groups, have demonstrated their value to identify pros and cons of individual models and to generate consensus models that outperform individual models [107,189]. We foresee that more interdisciplinary research across boundaries will transform the process of understanding drug mechanism and safety.

In contrast to the disciplines mentioned earlier, where multiscale modelling has been established and applied, multiscale modelling in drug discovery and development is still in its infancy. Its scope needs to be refined, its methodology to be established, and its impact to be scrutinised. Key open questions that are beyond the scope of this review beg answers. For instance, how can data and models, in line with the FAIR principles, be shared and managed so that they can be reused for other purposes than the original quest? Are there mathematical methods and models available, for example, Bayesian optimisation, as some experts have suggested, that have the promise to take the human out of the loop and to jointly optimise parameters of heterogeneous models [190]? If the answer is 'yes', what prevents us from building and using them? If the answer is 'no', at least for a given period of time, how can we make sure that the quantitative outputs of multiscale models are meaningfully processed by humans and translated into appropriate decisions?

Neither answers to these questions nor any success recipe are available to us. Nevertheless, we are inspired by the success of multiscale modelling in many scientific fields and are convinced of its potential in drug discovery as well. We invite all interested researchers to join forces with us in implementing and practising it. In Box 2, we share ten open questions, in the hope of initiating a dialogue with the community and learning from best practices.

Despite the many open questions and uncertainties, we envision that the implementation of multiscale modelling in drug discovery, in which modelling activities at different scales synergise and inform each other, will lead to insight in disease biology, which in turn will translate into effective and safe medicines. We are witnessing more powerful and accurate tools for molecular modelling and dynamics [191], accumulating single cell-level knowledge of human biology at temporal and spatial resolutions [192–194], physiology-emulating biological modelling systems [64], development of precise genome-editing tools [22,195], higher degree of automation [196], and last but not least, increased computational power and scalability [197]. It is hoped that these progresses will catalyse future models that operate in the 'what-if' mode. By combining *in silico* modelling and simulation with (semi-)automatic *in vitro* biological assays intelligently, the 'what-if' models shall assist drug-discovery scientists in maximising the knowledge of their drug candidate before human clinical testing.

BOX 2

Ten open questions on multiscale modelling of drug mechanism and safety We identified a set of open questions on how our three levels of modelling can benefit from each other. We appreciate open discussions and experience sharing within the community.

1 How can we assess off-target relevance in a cell-identity-, dose-, and MoA-specific manner, integrating evidence from different levels?
2 How good do compound property predictions need be so that they are useful in PK/PD modelling?
3 How can we integrate molecular modelling results as prior information for omics studies?
4 How can we effectively translate omics readouts, which are often relative because of the comparative nature of experiments, into parameters of organ- and system-level models?
5 How can we use the output of basic models (e.g., QSAR predicted ADME properties or ion channel IC_{50} values) as input for physiologically based PK or PK/PD models, particularly for early-stage compounds when experimental data are not available?
6 How can we use cellular and omics information to identify relevant cellular and/or subcellular processes to be included in the PK/PD model and, hence, determine the model structure?
7 How can we use molecular, cellular, and omics information, such as the abundance of particular proteins and pathway activity, to inform parameter values of system models?
8 How can we connect models in preclinical research with clinical, population models, for both forward and reverse translation?
9 What methodologies and tools can we use to facilitate and assist multiscale modelling?
10 How can we scientists working in drug discovery implement and achieve multiscale modelling as a community?

Acknowledgements

The authors thank the many colleagues who shared insight and provided inputs during years of collaboration. The names, if all listed, shall consume a significant proportion of the expected length of the article. Nevertheless, we would like to particularly thank Manfred Kansy and Holger Fischer for their knowledge of, and insight into, quantitative aspects of drug discovery, and for their suggestions to improve the manuscript. Corinne Solier and Michael Prummer offered valuable feedback. We thank Fabian Birzele, Jérôme Hert, Barbara Endler-Jobst, Juergen Hammer, Benjamin Ribba, and Thomas Singer for their support.

We are in debt to Roald Hoffmann and Jean-Paul Malrieu for their captivating tripartite essay on 'Simulation versus Understanding' [198–200], which we discovered among others from the blog *In The Pipeline* by Derek Lowe during the revision process. We thank four anonymous reviewers for their invaluable criticisms and insightful suggestions. J.D.Z. thanks students of the 'Applied Mathematics and Informatics in Drug Discovery' course and colleagues supporting the lecture series at the Department of Mathematics and Informatics, University of Basel, for inspiration and motivation.

J.D.Z. and M.E. wish to dedicate their work to Clemens Broger (2017†), a pioneer of bioinformatics in drug discovery, and a man true to himself.

This work was funded by F. Hoffmann-La Roche Ltd.

References

1 Schenone, M. *et al.* (2013) Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* 9, 232–240
2 Ziegler, S. *et al.* (2013) Target identification for small bioactive molecules: finding the needle in the haystack. *Angew. Chem. Int. Ed.* 52, 2744–2792
3 Comess, K.M. *et al.* (2018) Emerging approaches for the identification of protein targets of small molecules - a practitioners' perspective. *J. Med. Chem.* 61, 8504–8535
4 Miller-Jensen, K. *et al.* (2007) Common effector processing mediates cell-specific responses to stimuli. *Nature* 448, 604–608
5 Wang, Y. *et al.* (2016) Evidence-based and quantitative prioritization of tool compounds in phenotypic drug discovery. *Cell Chem. Biol.* 23, 862–874
6 Hagedorn, P.H. *et al.* (2017) Managing the sequence-specificity of antisense oligonucleotides in drug discovery. *Nucleic Acids Res.* 45, 2262–2282
7 Gao, J. *et al.* (2018) Small molecule interactome mapping by photoaffinity labeling reveals binding site hotspots for the NSAIDs. *J. Am. Chem. Soc.* 140, 4259–4268
8 Malone, C.F. *et al.* (2017) mTOR and HDAC inhibitors converge on the TXNIP/thioredoxin pathway to Cause catastrophic oxidative stress and regression of RAS-driven tumors. *Cancer Discov.* 7, 1450–1463
9 Weaver, R.J. and Valentin, J.-P. (2019) Today's challenges to de-risk and predict drug safety in human 'mind-the-gap'. *Toxicol. Sci.* 167, 307–321
10 Lachowicz, M. (2011) Microscopic, mesoscopic and macroscopic descriptions of complex systems. *Probab. Eng. Mech.* 26, 54–60
11 Horvath, P. *et al.* (2016) Screening out irrelevant cell-based models of disease. *Nat. Rev. Drug Discov.* 15, 751–769
12 Morgan, P. *et al.* (2018) Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat. Rev. Drug Discov.* 17, 167–181
13 Sheiner, L. and Wakefield, J. (1999) Population modelling in drug development. *Stat. Methods Med. Res.* 8, 183–193
14 Mould, D.R. and Upton, R.N. (2012) Basic concepts in population modeling, simulation, and model-based drug development. *CPT Pharmacomet. Syst. Pharmacol.* 1, 1–14
15 Waring, M.J. *et al.* (2015) An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.* 14, 475–486
16 Brown, F.K. *et al.* (2017) Data to decisions: creating a culture of model-driven drug discovery. *AAPS J.* 19, 1255–1263
17 Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1–9
18 Santos, R. *et al.* (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* 16, 19–34
19 Oprea, T.I. *et al.* (2018) Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* 17, 317–332
20 Matsui, M. and Corey, D.R. (2017) Non-coding RNAs as drug targets. *Nat. Rev. Drug Discov.* 16, 167–179
21 Setten, R.L. *et al.* (2019) The current state and future directions of RNAi-based therapeutics. *Nat. Rev. Drug Discov.* 18, 421–446
22 Fellmann, C. *et al.* (2017) Cornerstones of CRISPR-Cas in drug discovery and therapy. *Nat. Rev. Drug Discov.* 16, 89–100
23 Jones, P.A. *et al.* (2016) Targeting the cancer epigenome for therapy. *Nat. Rev. Genet.* 17, 630–641
24 Fischer, E. (1894) Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte Dtsch Chem. Ges.* 27, 2985–2993
25 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242
26 Fernandez-Leiro, R. and Scheres, S.H.W. (2016) Unravelling biological macromolecules with cryo-electron microscopy. *Nature* 537, 339–346
27 Hilger, D. *et al.* (2018) Structure and dynamics of GPCR signaling complexes. *Nat. Struct. Mol. Biol.* 25, 4–12

28 Hauser, A.S. *et al.* (2017) Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discov.* 16, 829–842
29 Hillisch, A. *et al.* (2004) Utility of homology models in the drug discovery process. *Drug Discov. Today* 9, 659–669
30 Waterhouse, A. *et al.* (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303
31 Evans, R. *et al.* (2018) De novo structure prediction with deep-learning based scoring. *Annu. Rev. Biochem.* 77, 363–382
32 AlQuraishi, M. (2019) End-to-end differentiable learning of protein structure. *Cell Syst.* 8, 292–301
33 Li, J. *et al.* (2019) An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdiscip. Sci. Comput. Life Sci.* 11, 320–328
34 Wallach, I. *et al.* (2015) AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv* 2015, 1510.02855v1
35 Lee, G.R. and Seok, C. (2016) Galaxy7TM: flexible GPCR–ligand docking by structure refinement. *Nucleic Acids Res.* 44, W502–W506
36 Chen, Y.-C. (2015) Beware of docking! *Trends Pharmacol. Sci.* 36, 78–95
37 Yuan, H. and Merck D3R Team, (2017) *MD Simulation in Pose Refinement and Scoring Using AMBER Workflows*. Merck
38 Klepeis, J.L. *et al.* (2009) Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* 19, 120–127
39 Bissantz, C. *et al.* (2010) Medicinal chemist's guide to molecular interactions. *J. Med. Chem.* 53, 5061–5084
40 Kramer, C. *et al.* (2018) Learning medicinal chemistry absorption, distribution, metabolism, excretion, and toxicity (ADMET) rules from cross-company matched molecular pairs analysis (MMPA). *J. Med. Chem.* 61, 3277–3292
41 Nicholls, A. *et al.* (2010) Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.* 53, 3862–3886
42 Maggiora, G. *et al.* (2014) Molecular similarity in medicinal chemistry. *J. Med. Chem.* 57, 3186–31204
43 Sellwood, M.A. *et al.* (2018) Artificial intelligence in drug discovery. *Future Med. Chem.* 10, 2025–2028
44 Chen, H. *et al.* (2018) The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250
45 Lo, Y.-C. *et al.* (2018) Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* 23, 1538–1546
46 Sahigara, F. *et al.* (2012) Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 17, 4791–4810
47 Sheridan, R.P. (2013) Using random forest to model the domain applicability of another random forest model. *J. Chem. Inf. Model.* 53, 2837–2850
48 Meier, R.J. (2019) A way towards reliable predictive methods for the prediction of physicochemical properties of chemicals using the group contribution and other methods. *Appl. Sci.* 9, 1700
49 Lobell, M. *et al.* (2006) In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem* 1, 1229–1236
50 Keiser, M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* 25, 197
51 Awale, M. and Reymond, J.-L. (2017) The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. *J. Cheminform.* 9, 11
52 Cha, Y. *et al.* (2018) Drug repurposing from the perspective of pharmaceutical companies. *Br. J. Pharmacol.* 175, 168–180
53 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181
54 Zhou, H. *et al.* (2015) Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Sci. Rep.* 5, 11090
55 Mathai, N. *et al.* (2019) Validation strategies for target prediction methods. *Brief Bioinform.* . http://dx.doi.org/10.1093/bib/bbz026
56 Bunnage, M.E. *et al.* (2013) Target validation using chemical probes. *Nat. Chem. Biol.* 9, 195–199
57 Hill, S.M. *et al.* (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* 13, 310–318
58 Jones, L.H. and Bunnage, M.E. (2017) Applications of chemogenomic library screening in drug discovery. *Nat. Rev. Drug Discov.* 16, 285–296
59 Beck, A. *et al.* (2010) Strategies and challenges for the next generation of therapeutic antibodies. *Nat. Rev. Immunol.* 10, 345–352
60 Bennett, C.F. and Swayse, E.E. (2010) RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. *Annu. Rev. Pharmacol. Toxicol.* 50, 259–293
61 Perlman, Z.E. *et al.* (2004) Multidimensional drug profiling by automated microscopy. *Science* 306, 1194–1198
62 Nijman, S.M.B. (2015) Functional genomics to uncover drug mechanism of action. *Nat. Chem. Biol.* 11, 942–948
63 Lei, C.L. *et al.* (2017) Tailoring mathematical models to stem-cell derived cardiomyocyte lines can improve predictions of drug-induced changes to their electrophysiology. *Front. Physiol.* 8, 986
64 McAleer, C.W. *et al.* (2019) Multi-organ system for the evaluation of efficacy and off-target toxicity of anticancer therapeutics. *Sci. Transl. Med.* 11, eaav1386
65 Takahashi, T. (2019) Organoids for drug discovery and personalized medicine. *Annu. Rev. Pharmacol. Toxicol.* 59, 447–462
66 Bailey, R. (2008) *Design of Comparative Experiments*. Cambridge University Press
67 Subramanian, A. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437–1452
68 Drawnel, F.M. *et al.* (2017) Molecular phenotyping combines molecular information, biological relevance, and patient data to improve productivity of early drug discovery. *Cell Chem. Biol.* 18, 624–634
69 Moffat, J.G. *et al.* (2017) Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat. Rev. Drug Discov.* 16, 531–543
70 Zhang, J.D. *et al.* (2014) Data mining reveals a network of early-response genes as a consensus signature of drug-induced *in vitro* and *in vivo* toxicity. *Pharmacogenomics J.* 14, 208–216
71 Moisan, A. *et al.* (2017) Inhibition of EGF uptake by nephrotoxic antisense drugs *in vitro* and implications for preclinical safety profiling. *Mol. Ther. Nucleic Acids* 6, 89–105
72 Hauser, A.S. *et al.* (2018) Pharmacogenomics of GPCR drug targets. *Cell* 172, 41–54
73 Ye, C. *et al.* (2018) DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat. Commun.* 9, 4307
74 Zhang, J.D. *et al.* (2017) Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genomics* 18, 277
75 Love, M.I. *et al.* (2015) RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research* 4, 1070
76 Goeman, J.J. and Bühlmann, P. (2007) Analysing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23, 980–987
77 Geistlinger, L. *et al.* (2019) Towards a gold standard for benchmarking gene set enrichment analysis. *BioRxiv* 2019, 674267
78 Moisan, A. *et al.* (2015) White-to-brown metabolic conversion of human adipocytes by JAK inhibition. *Nat. Cell Biol.* 17, 57–67
79 Birzele, F. *et al.* (2015) CD44 isoform status predicts response to treatment with anti-CD44 antibody in cancer patients. *Clin. Cancer Res.* 21, 2753–2762
80 Luecken, M.D. and Theis, F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15, e8746
81 Edwards, B.S. and Sklar, L.A. (2015) Flow cytometry: impact on early drug discovery. *J. Biomol. Screen.* 20, 689–707
82 Anchang, B. *et al.* (2018) DRUG-NEM: Optimizing drug combinations using single-cell perturbation response to account for intratumoral heterogeneity. *Proc. Natl. Acad. Sci. U. S. A.* 115, E4294–E4303
83 Nikolaev, Y. *et al.* (2019) Systems NMR: single-sample quantification of RNA, proteins and metabolites for biomolecular network analysis. *Nat. Methods* 16, 743
84 Tanay, A. and Regev, A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541, 331–338
85 Lotfollahi, M. *et al.* (2019) scGen predicts single-cell perturbation responses. *Nat. Methods* 16, 715–721
86 Liu, J.J. *et al.* (2018) *In vivo* brain GPCR signaling elucidated by phosphoproteomics. *Science* 360, eaao4927
87 Moellering, R.E. and Cravatt, B.F. (2012) How chemoproteomics can enable drug discovery and development. *Chem. Biol.* 19, 11–22
88 Molina, D.M. *et al.* (2013) Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science* 341, 84–87
89 Savitski, M.M. *et al.* (2014) Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* 346, 1255784
90 Choi, M. *et al.* (2017) ABRF Proteome Informatics Research Group (iPRG) 2015 Study: detection of differentially abundant proteins in label-free quantitative LC–MS/MS experiments. *J. Proteome Res.* 16, 945–957
91 Hogrebe, A. *et al.* (2018) Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat. Commun.* 9, 1–13
92 Donnelly, D.P. *et al.* (2019) Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat. Methods* 16, 587–594
93 Fetz, V. *et al.* (2016) Target identification by image analysis. *Nat. Prod. Rep.* 33, 655–667
94 Scheeder, C. *et al.* (2018) Machine learning and image-based profiling in drug discovery. *Curr. Opin. Syst. Biol.* 10, 43–52
95 Rudin, M. and Weissleder, R. (2003) Molecular imaging in drug discovery and development. *Nat. Rev. Drug Discov.* 2, 123–131

96 Weinstein, J.A. *et al.* (2019) DNA microscopy: optics-free spatio-genetic imaging by a stand-alone chemical reaction. *Cell* 178 (1), 229–241.e16
97 Boutros, M. *et al.* (2015) Microscopy-based high-content screening. *Cell* 163, 1314–1325
98 Smith, K. *et al.* (2018) Phenotypic image analysis software tools for exploring and understanding big image data from cell-based assays. *Cell Syst.* 6, 636–653
99 Webster, J.D. and Dunstan, R.W. (2014) Whole-slide imaging and automated image analysis: considerations and opportunities in the practice of pathology. *Vet. Pathol.* 51, 211–223
100 Zoffmann, S. *et al.* (2019) Machine learning-powered antibiotics phenotypic drug discovery. *Sci. Rep.* 9, 1–14
101 Bray, M.-A. *et al.* (2016) Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* 11, 1757–1774
102 Simm, J. *et al.* (2018) Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chem. Biol.* 25, 611–618
103 Ahmad, G. and Amiji, M. (2018) Use of CRISPR/Cas9 gene-editing tools for developing models in drug discovery. *Drug Discov. Today* 23, 519–533
104 Thomas, R. *et al.* (1995) Dynamical behaviour of biological regulatory networks-I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.* 57, 247–276
105 Chen, S. and Mar, J.C. (2018) Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* 19, 232
106 Munsky, B. *et al.* (2012) Using gene expression noise to understand gene regulation. *Science* 336, 183–187
107 Marbach, D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804
108 Zhu, J. *et al.* (2012) Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol.* 10, e1001301
109 Argelaguet, R. *et al.* (2018) Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* 14, e8124
110 Karr, J.R. *et al.* (2012) A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401
111 Jensen, K.J. *et al.* (2016) Network architecture predisposes an enzyme to either pharmacologic or genetic targeting. *Cell Syst.* 2, 112–121
112 Bersanelli, M. *et al.* (2016) Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 17, S15
113 Le Novère, N. (2015) Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.* 16, 146–158
114 Lavé, T. *et al.* (2007) Challenges and opportunities with modelling and simulation in drug discovery and drug development. *Xenobiotica* 37, 1295–1310
115 Dockendorf, M.F. *et al.* (2018) Leveraging model-informed approaches for drug discovery and development in the cardiovascular space. *J. Pharmacokinet. Pharmacodyn.* 45, 355–364
116 Agoram, B.M. (2009) Use of pharmacokinetic/pharmacodynamic modelling for starting dose selection in first-in-human trials of high-risk biologics. *Br. J. Clin. Pharmacol.* 67, 153–160
117 Zou, P. *et al.* (2012) Applications of human pharmacokinetic prediction in first-in-human dose estimation. *AAPS J.* 14, 262–281
118 Graaf, vander P.H. and Benson, N. (2018) The role of quantitative systems pharmacology in the design of first-in-human trials. *Clin. Pharmacol. Ther.* 104 797–797
119 Grimstein, M. *et al.* (2019) Physiologically based pharmacokinetic modeling in regulatory science: an update from the U.S. Food and Drug Administration's Office of Clinical Pharmacology. *J. Pharm. Sci.* 108, 21–25
120 Reddy, V.P. *et al.* (2018) Development, verification, and prediction of osimertinib drug–drug interactions using PBPK modeling approach to inform drug label. *CPT Pharmacomet. Syst. Pharmacol.* 7, 321–330
121 Vicente, J. *et al.* (2018) Mechanistic model-informed proarrhythmic risk assessment of drugs: review of the 'CiPA' initiative and design of a prospective clinical validation study. *Clin. Pharmacol. Ther.* 103, 54–66
122 Knibbe, C.A.J. and Danhof, M. (2011) Individualized dosing regimens in children based on population PKPD modelling: are we ready for it? *Int. J. Pharm.* 415, 9–14
123 Zhou, Q.-T. *et al.* (2017) Meropenem dosing based on a population pharmacokinetic–pharmacodynamic model in elderly patients with infection of the lower respiratory tract. *Drugs Aging* 34, 115–121
124 Hu, Z.-Y. *et al.* (2014) Physiologically based pharmacokinetic modeling of impaired carboxylesterase-1 activity: effects on oseltamivir disposition. *Clin. Pharmacokinet.* 53, 825–836
125 Hodgkin, A.L. and Huxley, A.F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544
126 O'Hara, T. *et al.* (2011) Simulation of the undiseased human cardiac ventricular action potential: model formulation and experimental validation. *PLoS Comput. Biol.* 7 e1002061
127 Zhao, Y. and Hu, Z.-Y. (2014) Physiologically based pharmacokinetic modelling and *in vivo* I/Ki accurately predict P-glycoprotein-mediated drug-drug interactions with dabigatran etexilate. *Br. J. Pharmacol.* 171, 1043–1053
128 Wangorsch, G. *et al.* (2011) Time-resolved in silico modeling of fine-tuned cAMP signaling in platelets: feedback loops, titrated phosphorylations and pharmacological modulation. *BMC Syst. Biol.* 5, 178
129 Deveaux, W. *et al.* (2019) Defining rules for cancer cell proliferation in TRAIL stimulation. *NPJ Syst. Biol. Appl.* 5, 1–8
130 de Pillis, L.G. *et al.* (2005) A validated mathematical model of cell-mediated immune response to tumor growth. *Cancer Res.* 65, 7950–7958
131 Trayanova, N.A. and Winslow, R. (2011) Whole-heart modeling. *Circ. Res.* 108, 113–128
132 Kuepfer, L. *et al.* (2016) Concepts in PBPK modeling: how to build a PBPK/PD model. *CPT Pharmacomet. Syst. Pharmacol.* 5, 516–531
133 Emoto, C. *et al.* (2019) A theoretical physiologically-based pharmacokinetic approach to ascertain covariates explaining the large interpatient variability in tacrolimus disposition. *CPT Pharmacomet. Syst. Pharmacol.* 8, 273–284
134 Jones, H.M. and Rowland-Yeo, K. (2013) Basic concepts in physiologically based pharmacokinetic modeling in drug discovery and development. *CPT Pharmacomet. Syst. Pharmacol.* 2, 63
135 Howell, B.A. *et al.* (2012) *In vitro* to *in vivo* extrapolation and species response comparisons for drug-induced liver injury (DILI) using DILIsym™: a mechanistic, mathematical model of DILI. *J. Pharmacokinet. Pharmacodyn.* 39, 527–541
136 Delporte, M.-L. *et al.* (2017) *Dose Selection in Entry into Human (EIH) Studies: Learnings from a Retrospective Survey on EiH Studies with Small Molecules Conducted between 2004 and 2016 at Hoffman-La Roche.* Roche
137 Okada, J. *et al.* (2015) Screening system for drug-induced arrhythmogenic risk combining a patch clamp and heart simulator. *Sci. Adv.* 1, e1400142
138 Jha, S.K. and Langmead, C.J. (2012) Exploring behaviors of stochastic differential equation models of biological systems using change of measures. *BMC Bioinformatics* 13, S8
139 Donnet, S. and Samson, A. (2013) A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. *Adv. Drug Deliv. Rev.* 65, 929–939
140 Wang, Z. *et al.* (2015) Integrated PK-PD and agent-based modeling in oncology. *J. Pharmacokinet. Pharmacodyn.* 42, 179–189
141 Kather, J.N. *et al.* (2017) Silico modeling of immunotherapy and stroma-targeting therapies in human colorectal cancer. *Cancer Res.* 77, 6442–6452
142 Mortensen, S.B. *et al.* (2008) *Introduction to PK/PD Modelling - With Focus on PK and Stochastic Differential Equations.* Technical University of Denmark
143 Gabrielsson, J. (2016) *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications* (5th edn), Apotekarsocieteten
144 Wang, H. *et al.* (2016) Evaluating a physiologically based pharmacokinetic model for predicting the pharmacokinetics of midazolam in Chinese after oral administration. *Acta Pharmacol. Sin.* 37, 276–284
145 Wong, H. and Chow, T.W. (2017) Physiologically based pharmacokinetic modeling of therapeutic proteins. *J. Pharm. Sci.* 106, 2270–2275
146 Niederalt, C. *et al.* (2018) A generic whole body physiologically based pharmacokinetic model for therapeutic proteins in PK-Sim. *J. Pharmacokinet. Pharmacodyn.* 45, 235–257
147 Glassman, P.M. and Balthasar, J.P. (2019) Physiologically-based modeling of monoclonal antibody pharmacokinetics in drug discovery and development. *Drug Metab. Pharmacokinet.* 34, 3–13
148 Luzon, E. *et al.* (2017) Physiologically based pharmacokinetic modeling in regulatory decision-making at the European Medicines Agency. *Clin. Pharmacol. Ther.* 102, 98–105
149 Sato, M. *et al.* (2017) Quantitative modeling and simulation in PMDA: a Japanese regulatory perspective. *CPT Pharmacomet. Syst. Pharmacol.* 6, 413–415
150 Baneyx, G. *et al.* (2014) Physiologically based pharmacokinetic modeling of CYP3A4 induction by rifampicin in human: Influence of time between substrate and inducer administration. *Eur. J. Pharm. Sci.* 56, 1–15
151 Wagner, C. *et al.* (2016) Predicting the effect of CYP3A inducers on the pharmacokinetics of substrate drugs using physiologically based pharmacokinetic (PBPK) modeling: an analysis of PBPK submissions to the US FDA. *Clin. Pharmacokinet.* 55, 475–483

152 Hanke, N. *et al.* (2018) PBPK models for CYP3A4 and P-gp DDI prediction: a modeling network of rifampicin, itraconazole, clarithromycin, midazolam, alfentanil, and digoxin. *CPT Pharmacomet. Syst. Pharmacol.* 7, 647–659
153 Johnson, T.N. *et al.* (2019) Development of a physiologically based pharmacokinetic model for mefloquine and its application alongside a clinical effectiveness model to select an optimal dose for prevention of malaria in young Caucasian children. *Br. J. Clin. Pharmacol.* 85, 100–113
154 Morcos, P.N. *et al.* (2018) Effect of hepatic impairment on the pharmacokinetics of alectinib. *J. Clin. Pharmacol.* 58, 1618–1628
155 Rhee, S. *et al.* (2017) Physiologically based pharmacokinetic modelling and prediction of metformin pharmacokinetics in renal/hepatic-impaired young adults and elderly populations. *Eur. J. Drug Metab. Pharmacokinet.* 42, 973–980
156 Tistaert, C. *et al.* (2019) Food effect projections via physiologically based pharmacokinetic modeling: predictive case studies. *J. Pharm. Sci.* 108, 592–602
157 Strauss, D.G. *et al.* (2019) Comprehensive *In vitro* Proarrhythmia Assay (CiPA) Update from a Cardiac Safety Research Consortium/Health and Environmental Sciences Institute/FDA Meeting. *Ther. Innov. Regul. Sci.* 53, 519–525
158 Pathmanathan, P. and Gray, R.A. (2018) Validation and trustworthiness of multiscale models of cardiac electrophysiology. *Front. Physiol.* 9, 106
159 Davies, M.R. *et al.* (2016) Recent developments in using mechanistic cardiac modelling for drug safety evaluation. *Drug. Discov. Today* 21, 924–938
160 Dutta, S. *et al.* (2017) Optimization of an in silico cardiac cell model for proarrhythmia risk assessment. *Front. Physiol.* 8, 106
161 Li, Z. *et al.* (2017) Improving the in silico assessment of proarrhythmia risk by combining hERG (Human Ether-à-go-go-Related Gene) channel–drug binding kinetics and multichannel pharmacology. *Circ. Arrhythm. Electrophysiol.* 10 (2), 1–12
162 Li, Z. *et al.* (2019) Assessment of an in silico mechanistic model for proarrhythmia risk prediction under the CiPA initiative. *Clin. Pharmacol. Ther.* 105, 466–475
163 ICH (2018) *Final Concept Paper ICH S7B and E14 Q&A, Endorsed by the MC with Support of the Assembly on 15.* ICH
164 Watkins, P.B. (2019) The DILI-sim initiative: insights into hepatotoxicity mechanisms and biomarker interpretation. *Clin. Transl. Sci.* 12, 122–129
165 Clausznitzer, D. *et al.* (2018) Quantitative systems pharmacology model for Alzheimer disease indicates targeting sphingolipid dysregulation as potential treatment option. *CPT Pharmacomet. Syst. Pharmacol.* 7, 759–770
166 Moore, H. (2018) How to mathematically optimize drug regimens using optimal control. *J. Pharmacokinet. Pharmacodyn.* 45, 127–137
167 Ribba, B. *et al.* (2018) Prediction of the optimal dosing regimen using a mathematical model of tumor uptake for immunocytokine-based cancer immunotherapy. *Clin. Cancer Res.* 24, 3325–3333
168 Milberg, O. *et al.* (2019) A QSP model for predicting clinical responses to monotherapy, combination and sequential therapy following CTLA-4, PD-1, and PD-L1 checkpoint blockade. *Sci. Rep.* 9, 1–17
169 Haefner, J.W. (2005) *Modeling Biological Systems: Principles and Applications.* Springer
170 Palacino, J. *et al.* (2015) SMN2 splice modulators enhance U1–pre-mRNA association and rescue SMA mice. *Nat. Chem. Biol.* 11, 511–517
171 Sivaramakrishnan, M. *et al.* (2017) Binding to SMN2 pre-mRNA-protein complex elicits specificity for small molecule splicing modifiers. *Nat. Commun.* 8, 1476
172 Ratni, H. *et al.* (2018) Discovery of risdiplam, a selective survival of Motor Neuron-2 (SMN2) gene splicing modifier for the treatment of spinal muscular atrophy (SMA). *J. Med. Chem.* 61, 6501–6517
173 Campagne, S. *et al.* (2019) Structural basis of a small molecule targeting RNA for a specific splicing correction. *Nat. Chem. Biol.* 15, 1191–1198
174 O'Leary, N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745
175 Cléry, A. *et al.* (2011) Molecular basis of purine-rich RNA recognition by the human SR-like protein Tra2-β1. *Nat. Struct. Mol. Biol.* 18, 443–450
176 Weinan, E. and Lu, J. (2011) Multiscale modeling. *Scholarpedia* 6, 11527
177 Weinan, E. (2011) *Principles of Multiscale Modeling.* Cambridge University Press
178 Marée, A.F.M. *et al.* (2006) Polarization and movement of keratocytes: a multiscale modelling approach. *Bull. Math. Biol.* 68, 1169–1211
179 Coveney, P.V. and Fowler, P.W. (2005) Modelling biological complexity: a physical scientist's perspective. *J. R. Soc. Interface* 2, 267–280
180 Walpole, J. *et al.* (2013) Multiscale computational models of complex biological systems. *Annu. Rev. Biomed. Eng.* 15, 137–154
181 Visser, S.A.G. *et al.* (2013) Model-based drug discovery: implementation and impact. *Drug Discov. Today* 18, 764–775
182 Milligan, P.A. *et al.* (2013) Model-based drug development: a rational approach to efficiently accelerate drug development. *Clin. Pharmacol. Ther.* 93, 502–514
183 Benfenati, E. *et al.* (2011) The acceptance of in silico models for REACH: requirements, barriers, and perspectives. *Chem. Cent. J.* 5, 58
184 Mariathasan, S. *et al.* (2018) TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* 554, 544–548
185 Dai, J. *et al.* (2019) Acetylation blocks cGAS activity and inhibits Self-DNA-induced autoimmunity. *Cell* 176 (6), 1447–1460.E14
186 Drews, J. (2000) Drug discovery: a historical perspective. *Science* 287, 1960–1964
187 Gerald, M.C. (2013) *The Drug Book: From Arsenic to Xanax, 250 Milestones in the History of Drugs.* Sterling
188 Hie, B. *et al.* (2018) Realizing private and practical pharmacological collaboration. *Science* 362, 347–350
189 Choobdar, S. *et al.* (2019) Assessment of network module identification across complex diseases. *Nat. Methods* 16, 843–852
190 Shahriari, B. *et al.* (2016) Taking the human out of the loop: a review of Bayesian optimization. *Proc. IEEE* 104, 148–175
191 Chmiela, S. *et al.* (2018) Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* 9, 1–10
192 Regev, A. *et al.* (2017) Science forum: the human cell atlas. *eLife* 6, e27041
193 Burgess, D.J. (2019) Spatial transcriptomics coming of age. *Nat. Rev. Genet.* 20, 317
194 Satpathy, A.T. *et al.* (2019) Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* 37, 925–936
195 Smits, A.H. *et al.* (2019) Biological plasticity rescues target activity in CRISPR knock outs. *Nat. Methods* 16, 1087–1093
196 Schneider, G. (2018) Automating drug discovery. *Nat. Rev. Drug Discov.* 17, 97–113
197 Lyu, J. *et al.* (2019) Ultra-large library docking for discovering new chemotypes. *Nature* 566, 224
198 Hoffmann, R. and Malrieu, J-P. (XXXX) Simulation vs understanding a tension, in quantum chemistry and beyond. Part A: stage setting. *Angew. Chem. Int. Ed.* https://doi.org/10.1002/anie.201902527.
199 Hoffmann, R. and Malrieu, J-P. (XXXX) Simulation vs understanding a tension, in quantum chemistry and beyond. Part B: the march of simulation, for better or worse. *Angew. Chem. Int. Ed.* https://doi.org/10.1002/anie.201910283.
200 Hoffmann, R. and Malrieu, J-P. (XXXX) Simulation vs understanding a tension, in quantum chemistry and beyond. Part C: toward consilience. *Angew. Chem. Int. Ed.* https://doi.org/10.1002/anie.201910285.
201 Gaulton, A. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.* 45, D945–D954

An Introduction to Machine Learning

Solveig Badillo[1,*,†], Balazs Banfai[1], Fabian Birzele[1], Iakov I. Davydov[1], Lucy Hutchinson[1], Tony Kam-Thong[1], Juliane Siebourg-Polster[1], Bernhard Steiert[1] and Jitao David Zhang[1]

In the last few years, machine learning (ML) and artificial intelligence have seen a new wave of publicity fueled by the huge and ever-increasing amount of data and computational power as well as the discovery of improved learning algorithms. However, the idea of a computer learning some abstract concept from data and applying them to yet unseen situations is not new and has been around at least since the 1950s. Many of these basic principles are very familiar to the pharmacometrics and clinical pharmacology community. In this paper, we want to introduce the foundational ideas of ML to this community such that readers obtain the essential tools they need to understand publications on the topic. Although we will not go into the very details and theoretical background, we aim to point readers to relevant literature and put applications of ML in molecular biology as well as the fields of pharmacometrics and clinical pharmacology into perspective.

The advent of data availability and growth of computational power, combined with the arrival of novel learning methods, has led to a number of breakthroughs in many scientific areas. This includes biological and clinical research, where applications range from molecular biology[1] to image data analysis[2] and clinical practice.[3] However, the idea of a computer learning some abstract concepts—like humans do constantly—has been around at least since the 1950s when the first neural networks[4] were developed. Even before that, other methods like Bayesian statistics and Markov chains were used with a similar idea in mind. Many of these methods are known to the pharmacometrics and clinical pharmacology community by different naming conventions. On the left, we indicate the machine learning terminology and, on the right, the usual statistics naming (based on Tibshirani https://statweb.stanford.edu/~tibs/stat315a/glossary.pdf):

- network, graphs ⇔ model
- weights ⇔ parameters
- learning ⇔ fitting
- generalization ⇔ test set performance
- supervised learning ⇔ regression or classification
- unsupervised learning ⇔ density estimation, clustering
- features ⇔ covariates or explanatory variables

The main difference to more traditional approaches lies very much in the two distinct cultures of statistical modeling. This has been eluded to nearly 2 decades ago by Breiman.[5] Here, we extend his definition by incorporating physiological models in one of the cultures. In particular, culture 1 involves specifying a model to describe the observed data, and culture 2 aims to solve the problem by taking an algorithmic modeling approach, thus inherently leading to models with a higher number of free parameters and complex interactions. This complexity can pose challenges to the interpretation of the model (so called "black box" problem). The approaches typically used in pharmacometric applications fall into culture 1, where an underlying model is assumed based on pharmacological principles and understanding of drug properties. Such models are usually physiologically interpretable. Most machine learning (ML) approaches fall into culture 2, where no explicit model is specified, and a computer is responsible for identifying associations in the observed data. These models tend to be difficult to interpret physiologically, however, significant progress was made over the years in the interpretability of ML models.[6,7] Today, many aspects of a black box model can be interpreted using proper tools.[8]

In this paper, we aim to support readers to develop the intuition needed to understand how computers can learn or help humans to identify patterns in data. The foundational ideas of ML are highlighted, but we do not describe the details and theoretical background of available ML methods. We point the interested readers to other articles or books, such as "The Elements of Statistical Learning"[9] (referred as ESL), and we refer to examples of their application in molecular biology, drug discovery, drug development, and clinical pharmacology.

We first introduce the concepts of data points, features, feature spaces, and similarity measures and then dive deeper into the two main domains of machine learning, namely unsupervised and supervised learning, touching key aspects and examples. In the case of unsupervised learning, computers are tasked to identify yet unknown patterns in data without pre-existing knowledge like groups or classes, whereas in the case of supervised learning, computers are tasked to learn how to predict the class or the value of yet unobserved data points based on a concept (often also called a "model") that has been derived from a training dataset. **Figure 1** shows a taxonomy of the different methods described in this paper and can be

[1]Pharmaceutical Sciences, Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, Basel, Switzerland.
*Correspondence: Solveig Badillo (solveig.badillo@roche.com)
Authors in alphabetical order. All authors contributed equally.
[Correction added on 6th March, 2020, after first online publication: Author contribution text was added].
[†]S.B. was employed by Soladis Group during the time when the manuscript was written.
Received October 8, 2019; accepted January 15, 2020. doi:10.1002/cpt.1796

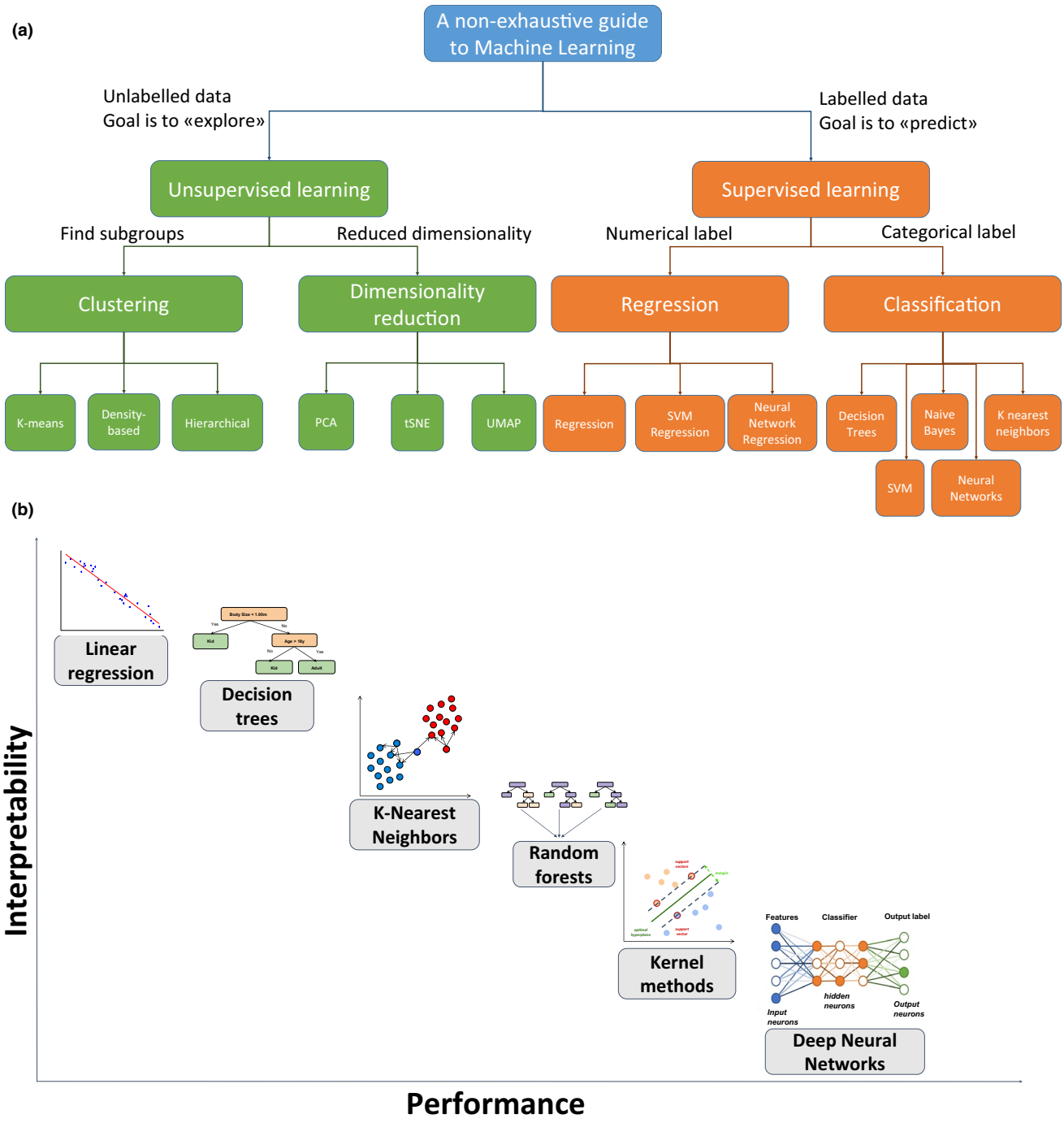


Figure 1 Taxonomy and overview of main machine learning (ML) algorithms. (**a**) Taxonomy of the different methods presented. (**b**) Overview of ML methods. The spectrum of available methods ranges from simpler and more interpretable to more advanced algorithms with potentially higher performance at the expense of less interpretability. Position of methods on the figure is qualitative and in practice depends on the number of free parameters, model complexity, data type, and the exact definition of interpretability used.[8] PCA, principal component analysis; SVM, support vector machine; tSNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection.

used as a reference, albeit nonexhaustive, on what scenario is suitable to apply which ML tool. Please note that all the unsupervised methods are also applicable in the case when labels are available.

DATA AND FEATURES

In ML, we deal with data and datasets. A dataset is composed of multiple data points (sometimes also called samples), where each data point represents an entity we want to analyze. Therefore, a data point can represent anything like a patient or a sample taken from a cancer tissue. Many of the issues related to data are universal and affect not only ML approaches but any quantitative discipline, including pharmacometrics.

To compile the dataset, one has measured and collected a number of features (i.e., data that describe properties of the data

points). Those features can be categorical (predefined values of no particular order like male and female), ordinal (predefined values that have an intrinsic order to them like a disease stage), or numerical (e.g., real values). For a patient in a clinical setting, these could be (combinations of) the patient's demographics, disease history, results of blood tests, or more complex and high dimensional measures, like gene expression profiles in a particular tissue or all single nucleotide polymorphisms that represent the patient's unique genome.

Each feature represents one dimension of the feature space and the concrete value of a feature for a particular data point places the point in a defined place in this dimension of the space. Taken together, all the values of all features of a data point is called a feature vector. The more features we have collected for the dataset, the higher the dimensionality of the resulting feature vector and the feature space. Obviously, as the dimensionality increases, visualization of all dimensions of the feature space becomes intractable and we have to rely on the computer to identify the relevant patterns or have to apply dimensionality reduction methods, as explained later in the section "Dimensionality Reduction."

Clinical pharmacologists are usually familiar with longitudinal data, such as pharmacokinetic (PK) and pharmacodynamic (PD) profiles, where the time-dependency plays a central role. In fact, models used in pharmacometrics are based on equations that can be justified based on physiology and pharmacology, which yield insights into the time-evolution of the system. This is similar to, for example, physical problems, such as weather forecasts, where air flow and temperature lead to a certain temporal behavior of the system. In ML, including time as a distinguished continuous variable into respective algorithms, remains challenging and is an area of active research. As of now, several options exist to include time-dependent data in ML datasets: Either directly where each time point represents a feature, or via transformations, such as Fourier transform or B-splines, resulting in coefficients of basic functions that can be considered as features. Alternatively, Recurrent Neural Networks (RNNs) can be used to handle longitudinal data, as outlined in the section "Recurrent Neural Network." However, all these approaches have the limitation of—directly or indirectly—discretizing the time-dimension.

Most ML algorithms are designed to handle high-dimensional datasets. Hence, derived features from the existing data are often included, such as log-transformed data, products, and ratios of features, or more advanced combinations. Such data transformation is an important preprocessing step that can have a profound effect on the model performance. Therefore, it is always a good idea to use available domain knowledge and expertise to come up with relevant features, a process sometimes referred to as feature engineering.

Data quality plays a crucial role in ML. Carefully chosen ML methods and visual inspection defend against extreme values or outliers. Missing data, however, can be challenging. Not all the methods support data missingness, and again data transformation could be required as a preprocessing step in such cases. There are various ways to impute missing data, the performance of which depends on the dataset and the method used.[10] The most trivial approach to the imputation is to replace a missing value with the feature mean across all the samples where it is defined. This, however, sometimes can cause overfitting[11] (also see the section "Performance Measures and the Issue of Overfitting").

It is also essential to scrutinize any bias in the data (e.g., selection bias). Preferably, samples for the ML should be an unbiased random subset of the population. In practice, this is rarely the case, and there are some biases in the data. These biases can affect the ability of the model to generalize beyond the training dataset (and even the test dataset if both share a similar bias). An example of such a generalization problem is a model that is supposed to learn how to distinguish a wolf from a husky by animal characteristics, but eventually turns out to simply identify patches of snow on the photograph.[6] There are various approaches to mitigate bias (e.g., one could down-weight or completely exclude biased samples or features).[12] In particular, propensity scores are useful when estimating the effect of a therapeutic intervention.[13] Inspection of the feature importance provides valuable information about the magnitude and the effect of the bias,[6,7] which is recommend to be used for checking the trustworthiness of ML models.

Many clinical classification datasets are unbalanced, meaning that one or more classes are underrepresented. This could pose difficulties for many ML algorithms, including artificial neural networks and gradient boosting methods. One way to mitigate this problem is undersampling/oversampling the majority/minority class, respectively, or tweaking the misclassification cost in the objective function.[14]

Finally, for many applications, it is important to define a similarity or distance measure between two data points in the feature space. The simplest distance measure would be the Euclidean distance:

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \quad (1)$$

between the numerical feature vectors of two data points A and B, for features $i = 1 \ldots n$, but depending on the type of data we are dealing with there can be many other and sometimes much more complex distance or similarity measures, such as cosine similarity[15] or similarity scores of two biological sequences.[16]

Main takeaways

- Transforming input data and feature engineering may improve the model.
- Missing data requires imputation.
- Biases in the data should be scrutinized.
- Unbalanced datasets require amendment of the model.
- Meaningful measures of similarity between the samples should be defined.

UNSUPERVISED LEARNING

In exploratory data analysis, we often do not know the true "labels," or we might want to examine the naturally emerging patterns in the data. For this purpose, we can use unsupervised learning methods, like clustering, frequent pattern detection, and dimensionality reduction. Here, we will focus particularly

on clustering and dimensionality reduction as they have many applications in molecular biology and clinical practice.

Clustering

The goal of applying clustering methods is to identify relevant subgroups in a given dataset without having a predefined hypothesis on the properties subgroups might have. For example, in a cohort of patients with a particular disease, we might want to identify subtypes that represent distinct biological mechanisms driving the disease based on molecular measures taken.[17]

A cluster is a subset of the data which are "similar" to each other, whereas points belonging to different clusters are more "different." There are multiple approaches to clustering that use different underlying algorithms to group data points by their "similarity." All of them have advantages and disadvantages and needed to be selected carefully depending on the application and properties of the data.

One simple approach to clustering is *k*-means clustering.[18] Here, the number of clusters to be identified is predefined by a user-selected parameter *k*. Each cluster is represented by a cluster center, which is an artificial data point that represents the mean (or median) value of all points assigned to this cluster. In the beginning, *k* cluster centers, known as "seeds," are randomly placed in the feature space. The algorithm then iterates through two steps. In step one ("assignment"), data points are assigned to the cluster represented by the closest center. In step two ("center shift"), the position of each cluster center is updated based on the composition of the clusters after step one. After a number of iterations, this will usually converge to a local optimum where cluster assignments do not or only marginally change. The result of such a process is visualized in **Figure 2b**. Although the procedure is intuitive, its major drawback is that usually the clustering is strongly influenced by the value of *k*, and more often than not the true number of clusters in the data is unknown *a priori*. Because there is rarely a clear cut right or wrong answer in clustering, further cluster investigation is required to identify meaningful clusters, which can be challenging particularly in the light of a high-dimensional feature space.

Another group of methods for clustering is density-based clustering.[19] In density-based methods, a cluster represents a part of the feature space where data points are dense. Data points belonging to the regions of the feature space with low density are considered to be noise. One of the well-known density-based clustering algorithms is Density-Based Spatial Clustering of Applications with Noise.[20] Density-based clustering does not require a predefined value setting the number of clusters and provides a reproducible result. Further, it is able to also identify complex cluster shapes, like the one shown in **Figure 2c**.

In hierarchical clustering analysis, the goal is to build a hierarchy of clusters (ESL, chapter 14).[9] One simple approach to hierarchical clustering is neighbor joining. First, all pairwise distances between all data points in the dataset are computed. Later, in every step of an iterative process, the two data points with smallest distance are grouped together. This results in a tree-like cluster structure, as displayed in **Figure 2a** on the left side and top of the heatmap where the branch lengths of the tree represent the distances of samples. To arrive at a discrete set of clusters like with *k*-means a distance threshold has to be chosen at which the tree is cut horizontally. Again, there is no optimal way of selecting such a threshold and many reasonable solutions may exist. Hierarchical clustering can be used alone, or used in combinations with heatmaps (e.g., **Figure 2a**) to visualize selected or all features, for instance, gene expression data.

Dimensionality reduction

The number of features and, therefore, the dimensionality of the feature space can be very high with tens of thousands of measures per sample. Not only does this make data visualization challenging but also the analysis is challenging. In particular, analysis of high-dimensional datasets can be associated with a phenomenon known as the "curse of dimensionality,"[21] which refers to data sparsity and counterintuitive geometrical properties in high-dimensional

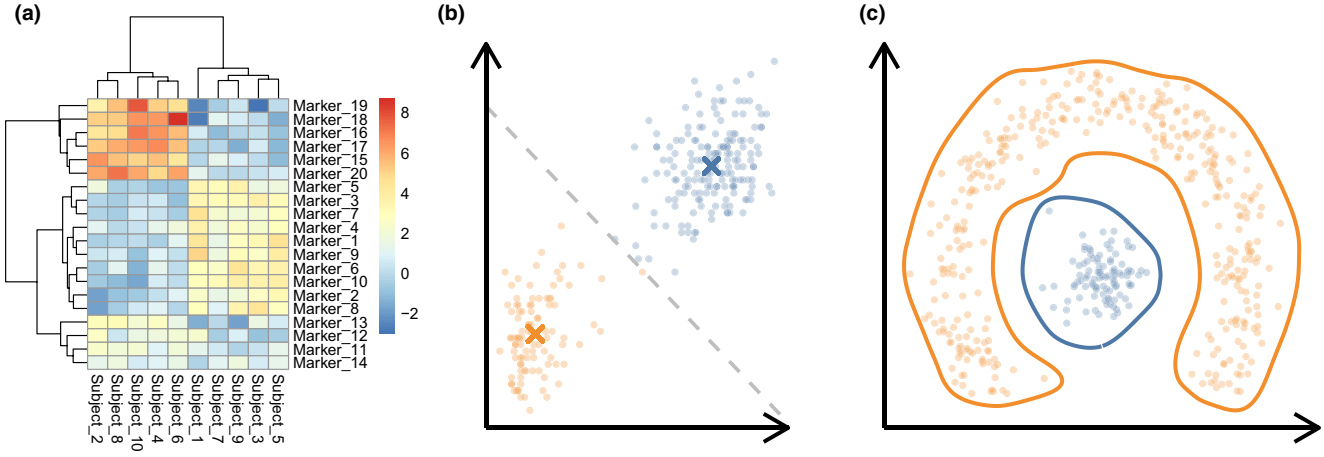


Figure 2 Overview of the results of different clustering approaches. (**a**) Shows the results of a two-dimensional hierarchical clustering. The two dendrograms visualize the similarity across samples and also across the markers measured. Such visualization is frequently used in biology for gene expression or other -omics technology readouts. (**b**) Shows the outcome of a classical clustering using *k*-means with a selected value of $k = 2$. Resulting clusters are usually convex and every point is assigned to one cluster, namely the one which is represented by the closest center point (marked by X). (**c**) Shows the result of a density-based clustering. Please note that the approach can identify nonconvex cluster forms, such as the orange cluster.

spaces. The "curse of dimensionality" poses challenges on most data analysis approaches, including but not limited to ML.

To mitigate such problems dimensionality reduction methods might be applied. Dimensionality reduction can aid data visualization by transforming each high-dimensional data point into two or more dimensions while keeping the majority of the variability and relative distances. Furthermore, dropping uninformative features could improve the model performance and convergence time. Although some of these methods, like principal component analysis, have even been developed long before the term ML has been coined,[22] others, like t-Distributed Stochastic Neighbor Embedding[23] or Uniform Manifold Approximation and Projection,[24] were developed recently and address complex challenges arising in data analysis. There is also a powerful neural network-based dimensionality reduction approach called autoencoder. For details on how to apply dimensionality reduction in biomedical data, we would like to refer the reader to a recent review.[25]

Examples of unsupervised ML applications

Clustering is widely used when analyzing high-dimensional data, such as transcriptomic, metabolomic, and proteomic experiments. Typically, hierarchical clustering would be used to identify main factors affecting the readouts as well as for identification of modules with high degree of coregulation. In single-cell sequencing, nonhierarchical clustering is used to understand which cell types are present in the sample. Clustering is also used to identify relationships among patients, tissues, diseases, or even disease symptoms.[26–29] Drug compounds themselves may also be clustered based on gene expression, sensitivity, and target protein properties[30–32] with the goal of guiding drug discovery.

Dimensionality reduction is routinely used in transcriptomic and other -omics experiments, usually to identify outliers and potential batch effects. In single-cell sequencing, Uniform Manifold Approximation and Projection or t-Distributed Stochastic Neighbor Embedding are used both for data visualization and for subsequent clustering.[24] Dimensionality reduction is also used to visualize the high-dimensional chemical space[33] or as a preprocessing step to improve performance of an ML model.[34]

Main takeaways

- Clustering can be used to understand structure in data by grouping similar observations together.
- *k*-means clustering is a simple yet powerful tool, however, the number of clusters must be specified in advance.
- Density-based methods do not require a prespecified number of clusters and allows identification of complex patterns in the data.
- Hierarchical clustering provides an overview of the relationship on multiple levels.
- Dimensionality reduction is used not only for data visualization but also to drop uninformative features.

SUPERVISED LEARNING

In a supervised learning problem, the computer is fed training data with observations and the corresponding known output values. The goal is to learn general rules (also often called a "model") that map inputs to outputs, so that it will be possible to predict the output for new unseen data, where we have observed input values but not their associated output.

There are two main categories of supervised learning: (i) classification where the output values are categorical, and (ii) regression where the output values are numeric.

In subsequent sections, the context of model fitting in supervised learning and the common issue of overfitting are introduced. Then, we explain how the performance is evaluated for classification and regression tools (i.e., how to assess the quality of mapping from inputs to outputs by the algorithm). This aspect is essential, as the merit of adopting ML methods often centers around the prospect of obtaining higher performance with the trade-off of interpretability. Understanding the different performance metrics enables better evaluation of the merits of a proposed model, as opposed to an assumption that an ML solution could always outperform a traditional approach.

We then dive into some of the existing classification and regression methods, starting off at the shallow end, where interpretation of the models is still straightforward, and progressing toward more ML-centric approaches where performance triumphs, often at the expense of interpretability. **Figure 1** summarizes the available spectrum of methods with respect to performance and interpretability. This section concludes with a nonexhaustive review of the applications of supervised learning methods in biology and, particularly, clinical pharmacology.

Performance measures and the issue of overfitting

The goal of a learning algorithm is to learn a concept or function (= a model) that describes the observed training data and is able to generalize on new independent data by avoiding both underfitting and overfitting.

The performance of a model is evaluated by methods that allow model assessment (i.e., estimating how well a given model performs in general and model selection; and the estimation of the performance of different models to choose the most adequate model). Some of these methods are highlighted in the next sections.

Model fitting. The model parameters are estimated based on observed data in the training set. To derive the optimal parameter values (e.g., for coefficients and weights), a distance measure between model and data is defined and minimized numerically. Independently of the metric chosen, the goal of model fitting is always to estimate the parameters by minimizing the distance, also called *loss function* or *cost function*, with two requirements:

- The model should provide predicted values that are close to observed ones on the training set, otherwise we say that it *underfits* and has a high *bias*.
- The model should generalize beyond the training set. A model that *overfits* predicts well on the training set but poorly on an independent test set, often because it is too complex for the data. In this case, we also talk about *high variance*.

In the following, we will call *objective function* any function that is optimized to estimate the model parameters.

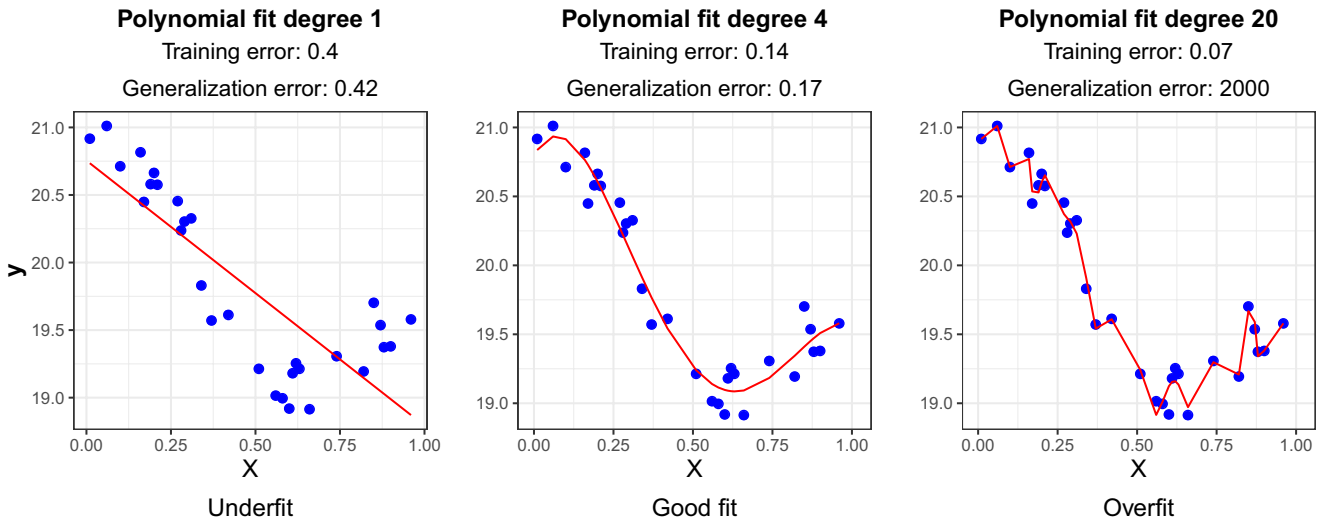


Figure 3 Illustration of the underfitting/overfitting issue on a simple regression case. Data points are shown as blue dots and model fits as red lines. Underfitting occurs with a linear model (left panel), a good fit with a polynomial of degree 4 (center panel), and overfitting with polynomial of degree 20 (right panel). Root mean squared error is chosen as objective function for evaluating the training error and the generalization error, assessed by using 10-fold cross-validation.

In the regression case, **Figure 3** illustrates the issue of underfitting and overfitting in the context of regression. Underfitting can occur when the model is too simple or when the features extracted from the data are not informative enough (**Figure 3**, left panel). Overfitting often occurs when the model is too complex or there are too many features over a small set of training examples (**Figure 3**, right panel).

This underfitting/overfitting issue is also often referred to as the bias/variance trade-off, which comes from the expression of the expected prediction error, including both bias and variance terms. The bias is an indication of the average error of the model for different training sets: It is the discrepancy between average of predicted values and the true mean we are trying to predict. The variance reflects the sensitivity of the model to the training set: For a given point, it corresponds to the spread of predicted values around their mean.

To minimize the predicted error, there is a trade-off between minimizing bias and variance: Increasing model complexity decreases bias but increases variance. To build less complex models, different techniques exist summarized under the term regularization. The principle consists in modifying the objective function by adding penalization terms that will influence parameter estimation. L1 and L2 regularization are the most common ones (ESL, sections 3.4.1 and 3.4.2).[9]

Different categories of loss functions. Different objective functions can be chosen to measure the distance between observed data and values predicted by the model. Some of the distance metrics used in practice can be associated to a *likelihood*. The likelihood indicates how probable it is to observe our data according to the selected model. The most common use of a likelihood is to find the parameters that make the model fit optimally to the data (i.e., the maximum likelihood parameter estimates). Usually, the negative logarithm of the likelihood is minimized and considered as objective function because it has favorable numerical properties. Similarly, in ML metrics, such as mean squared error, logistic objective, or cross-entropy, are used to find optimal parameters or assess the fitness of the model.

In practice, analytical calculation of maximum likelihood or minimal loss may not be feasible, and it is often necessary to use a numerical optimization algorithm to solve for the best parameter values. *Gradient descent* is such an algorithm, where we first define an objective function for which we want to minimize and then *iteratively* update the values of the parameters in the direction with the steepest decrease (first-order derivative) of the objective function until a convergence to a minimum distance is deemed reached. In the scenario of a nonconvex objective function, the success of finding a global minimum, as opposed to landing in some local minima, will depend on the choice of the initial set of parameter values, the learning rate (i.e., step size of each iteration) and the criterion for convergence. The reader can refer to ref. [35] for details on convex and nonconvex optimization processes. Stochastic gradient descent is an additional trick that can further speed up the optimization by randomly sampling a training dataset and summing the distances across this subset of training data points for approximating the objective function.

General principle of model selection and assessment. The problem of overfitting shows that the model performance on the training set is not a good indicator of its performance on a new dataset. We will detail below the principles of model performance evaluation in a supervised learning setting.

The general principle of model selection is as follows: When there are enough data, we separate them into three subsets—training, validation, and test sets. The training set is used to build different models, whereas the validation set is subsequently used to choose the algorithm and select the hyperparameters, if needed. Then, the model with the best performance on the validation set is selected. Finally, the test set enables to assess the generalization error, also called *test error*, which is the prediction error over a test dataset

that was not used during the training.[9] It is important to note here that the generalization error could be higher than expected when the original dataset is biased (see the section "Data and features"). Validating the model against a fully independent test dataset is the gold-standard method of assessing the generalizability of the model.

When the dataset is too small to extract a decent validation set, it is, for example, possible to use cross-validation techniques to select model hyperparameters. After putting aside a subset of the data for testing, k-fold validation consists of dividing the training set into k subsets, $k-1$ subsets being used for training and the last one to assess the performance. This process is repeated k times, each k subset being used once for validation, and the performance scores from each subset are then averaged for each set of hyperparameters to test. The k-fold cross-validation procedure is summarized in **Figure 4**. To choose between different learning algorithms[36] nested cross-validation can be used.

Indicators of model complexity vs. goodness of fit. In pharmacometrics, model selection is usually based on quantitative measures that summarize how well the model fits the data, often with penalties for overfitting. The most commonly used are the Akaike information criterion and Bayesian information criterion. They penalize the number of model parameters and reward goodness of fit, measured through likelihood. The Akaike information criterion is formalized as:

$$\mathrm{AIC} = 2M - 2\ln(\hat{\mathcal{L}}), \tag{2}$$

with the number of parameters M and the maximum likelihood $\hat{\mathcal{L}}$.

In contrast, the Bayesian information criterion:

$$\mathrm{BIC} = \ln(n)\cdot M - 2\ln(\hat{\mathcal{L}}), \tag{3}$$

takes into account the number of data points n.

These model selection approaches are rarely used in ML, partly due to the complexity of datasets and the associated violation of distributional assumptions. Instead, approaches like cross-validation are more commonly used (Clustering).

Performance measures for model assessment. For regression models, we typically use the mean squared error, or other types of average objective functions, to compare model performance on training and test set. For two-class classification problems, common performances measures are often derived from the "confusion matrix" shown in **Figure 5** and briefly described below.

- Precision, corresponding to the ratio of correctly predicted positive values to the total number of predicted positive values.
- Recall, also called true positive rate (TPR) corresponding to the ratio of correctly predicted positive values to the total number of positive values in the dataset.
- False Positive Rate (FPR), corresponds to the proportion of negative values predicted incorrectly.
- Accuracy, corresponding to the number of correctly predicted values divided by the total number of predicted values.
- Area under the ROC curve (AUC): Receiver operating characteristic (ROC) curves show the TPR (recall) and FPR dependence. In binary classification, each point on the ROC curve is located by choosing different thresholds for classification of y_i in positive or negative class. The top left corner of an ROC curve

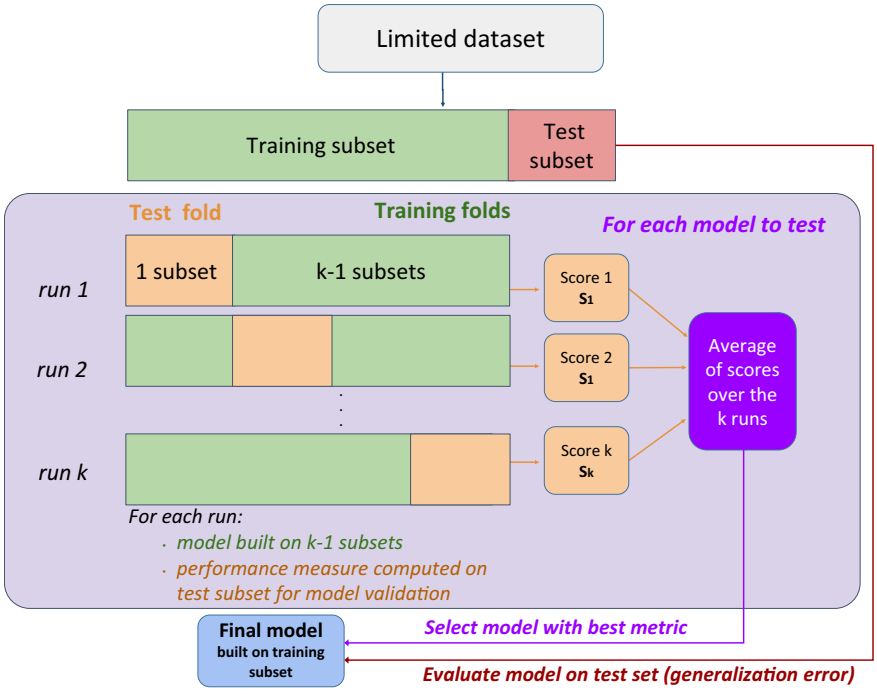


Figure 4 Illustration of the general principles of supervised learning in the case of a limited dataset. To assess the generalization ability of a supervised learning algorithm, data are separated into a training subset used for building the model and a test subset used to assess he generalization error.

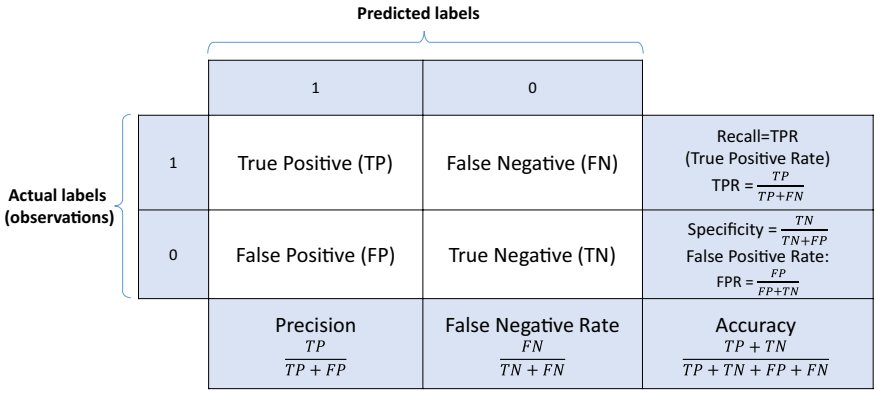


Figure 5 Confusion matrix for two-class problems. The confusion matrix indicates how successful the algorithm was at predicting labels in a binary classification problem where labels take values 0 (called "negative") or 1 (called "positive") by evaluating the predicted vs. the real labels. Every data point in the test set belongs to one of the four categories and different measures can be derived from these numbers.

is the ideal case with 100% of positive values correctly classified (TPR = 1) and 0% of positive values incorrectly predicted at 0 (FPR = 0). As it is ideal to maximize the TPR while minimizing the FPR, a larger area under the ROC curve (AUC) is better.

Some of these metrics could be generalized for multiclass problems, where there are more than two different labels in the dataset. However, the metrics mentioned above are noncontinuous with respect to model parameters, hence, parameter optimization may be challenging when they are used as objective function. A continuous alternative and widely used metric previously mentioned in the section "Model fitting" is cross-entropy (ESL, chapter 9),[9] which not only accounts for the most likely prediction but also for the prediction score (prediction confidence).

k-Nearest neighbors

We start our overview on existing learning methods with a method that skips the learning step completely and, therefore, does not lead to an explicit model that is being learned from the training data. As we will discuss later, this is also one of its biggest shortcomings. This type of learning is also often referred to as "instance-based learning" and, in our particular example, "*k*-nearest neighbor learning" (kNN).[37]

In these approaches, learning simply consists of storing all the existing, labeled data points (i.e., the training data) in a database. When a new, yet unclassified example is observed, the algorithm will place it in the *n*-dimensional feature space based on its feature values. For each data point in the database, we now compute the distance (e.g., a Euclidean distance or other, more complex ones) to this new data point in order to identify its *k* closest neighbors. In a second step, we examine the known labels of these *k*NNs in our database. Say we have chosen *k* to be nine and we observe seven of the nearest neighbors to be labeled as class X whereas two of them are labeled as class Y. In this case, we would assign our new data point to the class X as the majority of its neighbors are of this class. An extension of this simple approach would be to weight the importance of the neighbors to the classification by their distance to the new data point. Despite being very straightforward and simple, it proves to be a very effective classification method in practice. It is very efficient when it comes to training (i.e., storing the data in the database) and efficient implementations for computing the *k*NNs exist.

So, what are the challenges to this approach? The most obvious one is that because there is no "learning step," the *k*NNs approach does not identify the features that are really relevant to predict the class of a new case. Therefore, even though in a 20-dimensional feature space, where only 2 might be really relevant for the classification, the distance will be computed taking all 20 dimensions into account. Thus, the *k* nearest data points returned by the query will be highly influenced by irrelevant features or noise (see also "Dimensionality reduction" on how to remove some of those features). As a consequence, the resulting classification will be driven by noise rather than the real underlying pattern in the data. In this aspect, the approach suffers from the same challenge that also clustering approaches (see the section "Clustering") are facing, which are often summarized as the "curse of dimensionality."[21]

Naive Bayes

The second and very intuitive learning approach we would like to introduce is naive Bayes. It is based on computing simple statistics from a given training dataset as the learning step following a straightforward (but naive) application of the Bayesian formula for conditional probability in order to obtain a classification. Due to its simplicity it is also often used to obtain a baseline classification performance that other, more involved methods have to improve upon. It can best be explained by a simple example.

Let us assume we have training dataset with patients suffering either from a harmless cold or an influenza (flu) infection. We have measured two features for each patient, namely fever (high, low, or no) and pain (strong, low, or no). For each patient, we know through a laboratory test if the patient had an influenza infection or not. We now want to learn from these data and apply it to diagnose a new patient (where we have no laboratory test available) using the naive Bayes approach.

As a learning step, we count for each feature value how often it occurs in the influenza and in the cold patient group (e.g., to obtain the probability for high fever under the condition of the patient having a flu and so on). The result of this learning step might be seen in **Table 1**.

Table 1 Illustration of naive Bayes: Example of learning step results on flu dataset, showing the probabilities of features values given the patient category

Features	Fever			Pain		
Classes	High	Low	No	Strong	Low	No
Influenza (Flu)						
P (Flu) = 0.1	P(Fever = High|Flu) = 0.95	P(Fever = Low|Flu) = 0.05	P(Fever = No|Flu) = 0	P(Pain = Strong|Flu) = 0.75	P(Pain = Low|Flu) = 0.20	P(Pain = No|Flu) = 0.05
Cold						
P(Cold) = 0.9	P(Fever = High|Cold) = 0.1	P(Fever = Low|Cold) = 0.4	P(Fever = No|Cold) = 0.5	P(Pain = Strong|Cold) = 0.3	P(Pain = Low|Cold) = 0.3	P(Pain = No|Cold) = 0.4

Table 1 summarizes probability of each feature given the category of patient and shows that in the whole patient population the probability for a patient having an influenza infection is 0.1, whereas the probability for a normal cold is 0.9.

Once we have generated these values and, therefore, completed the "learning step" by analyzing our dataset, naive Bayes makes a now naive assumption, which is that all these features are conditionally independent of one another. In reality, this is rarely true and there are more advanced Bayesian learning methods that do not make this assumption. However, the assumption allows for a straightforward application of the Bayesian theorem. For details (i.e., formulas) on how to derive this classifier, we would like to refer the reader to further reading material (ESL, chapter 6).[9] In brief, the probability of a certain label (flu or cold) for a new test item can be computed as the product of the single conditional feature probabilities (fever and pain) that are observed for the data point times the probability for the class (flu or cold). The class with the maximal posterior likelihood is selected as the predicted class for the test item. Assuming we have a test person with an unknown diagnosis for influenza or cold, and we know that this person shows up with high fever and a high level of pain, we would compute the likelihood for influenza as:

$$P(\text{Fever} = \text{High}|\text{Flu}) \cdot P(\text{Pain} = \text{Strong}|\text{Flu}) \cdot P(\text{Flu}) = 0.95 \cdot 0.75 \cdot 0.1 = 0.07125. \quad (4)$$

In the same way we would compute the likelihood for a cold as:

$$P(\text{Fever} = \text{High}|\text{Cold}) \cdot P(\text{Pain} = \text{Strong}|\text{Cold}) \cdot P(\text{Cold}) = 0.1 \cdot 0.3 \cdot 0.9 = 0.027. \quad (5)$$

For a patient that presents to the doctor with high fever and strong muscular pain or headache, this results in a (nonnormalized) posterior probability for an influenza infection of 7.125% and in a probability of 2.7% for a normal cold. Therefore, the patient suffers more likely from a flu than from a cold.

In many aspects, naive Bayes, therefore, formalizes how humans might learn from experience.

Decision trees, random forests, and gradient boosting

Decision trees are an essential building block for many ML algorithms. They have been used for at least 50 years.[38,39] The idea behind decision trees is very intuitive and best represented in a visual form (e.g., **Figure 1**). Depending on the problem, decision tree leaf nodes have classes, probabilities, or continuous values in case of regression. In the early days of ML, decision trees have been used to solve pharmacological problems, such as dosing, toxicology, and diagnostics.[40–42] Although usage of decision trees is intuitive, the question is how to construct such trees from the available data. A few famous approaches worth mentioning are CART[43] and ID3.[44]

Currently, decision trees are almost never used in ML in their original form. One of the reasons being is the fact that decision trees are prone to overfitting. Nevertheless, decision trees became the building block for two widely used approaches: Random decision forests and gradient boosting frameworks.

Both random decision forests and tree-based gradient boosting use a set (ensemble) of trained decision trees to predict the outcome variable. The crucial difference between tree-based gradient boosting and random decision forests is on how trees are created.

In case of random forests, the algorithm constructs hundreds or thousands of deep decision trees ("strong predictors"). Each of those trees is likely overfitted, however, by combining the outputs of multiple trees we can solve the overtraining problem. On the contrary, in a gradient boosting algorithm, such as XGBoost or CatBoost, each of the trees is a shallow decision tree ("weak predictor"), and the algorithm iteratively decreases the classification error over time by adding more and more trees.

Today, gradient boosting methods show a great performance both in publications and ML competitions. Even without hyperparameter tuning, they usually provide excellent performance with a relatively low computational cost.[11] On the other hand, random forests are usually less prone to overfitting[45] and require less parameter tuning.[46] This makes random decision forests attractive for smaller datasets or as a baseline method for benchmarking.

Tree ensemble methods can be used for classification tasks, as well as for regression. In both cases, tree outputs are averaged, which can create a smooth output function.

Kernel methods: Support vector machines and regression

Kernel methods and, more specifically, support vector machine (SVM) for classification and support vector regression (SVR) for continuous output have found applications in computational biology for their ability to be robust against noise and to work with high-dimensional datasets found in genetics, transcriptomics, and proteomics.[47] Concretely in a more recent example, SVR was used for delineating cell compositions from bulk transcriptomics data.[48]

This section first offers a brief overview of the key concepts highlighting the notions of kernel transformations, an objective function with a lossless region, and a regularization term.[49,50] The emphasis will be placed on providing the reasoning behind why this is a more versatile method in dealing with multiple inputs where their effects on the output are unknown and can be postulated to span into nonlinear functions.

Background. Similar to all regression methods, the objective of SVR is to postulate a function on the input(s) that can help estimate for the observed output. Likewise, for SVM, the goal is to find the optimal decision boundary that separates the classes. As the name suggests, the core concept behind SVM/regression is the ability to objectively choose a subset of training data called *support vectors*. These support vectors define the model, which is usually a hyperplane in some feature space. To achieve this, several notions need to be introduced.

- An ε-*insensitive loss* function allows for residual less than ε, to be considered lossless and, thus, not part of the support vectors factored in to estimate the output-input function.
- A *regularization term* is added to the objective function with the aim of searching for a model to describe the relationship between the input and output variables such that the hyperplane is kept as flat as possible.
- *Slack variables* can be introduced to allow for training errors, termed soft margin, when the output is found outside the ε-insensitive region. By introducing slack variables, tolerance for the residual term to be greater than ε is made.
- A *kernel function* allows us to work in a higher dimension space, feature space. A kernel function applied in the input space corresponds to a dot product in the feature space where similarity measures are computed. This is achieved without having to explicitly map the input data from the input space to some feature space by some mapping function ϕ.

With all these concepts at hand, we are now capable of fitting a model with some thickness, known as a tube introduced by the ε-insensitive loss function, whereas the regularization term controls for the flatness of this hyperplane in some feature space defined by the kernel function. **Figure 6** illustrates these basic concepts of SVM.

Kernel trick and choice. SVR can capture nonlinear target functions, which map the multivariate inputs to the output. More precisely, the kernel trick means that a kernel:

$$k(\mathbf{x}_i,\mathbf{x}_j) := \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \tag{6}$$

applied to a set of inputs in the input space is equivalent to computing the dot product as a similarity measure in some feature space. This is achieved without having to explicitly perform a pre-mapping of the inputs, $\mathbf{x}_i$, with a mapping function Φ. A kernel function calculated in the input space corresponds to a dot product in some feature space if and only if it is a symmetric positive definite function.[51,52]

The choice of the radial basis function kernel,

$$\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle := k(\mathbf{x}_i,\mathbf{x}_j)_{\mathrm{RBF}} = \exp^{-\gamma(\|x_i - x_j\|)^2} \tag{7}$$

is often made as it can be expanded to a feature space of infinite dimensions. Although radial basis function covers a wide range

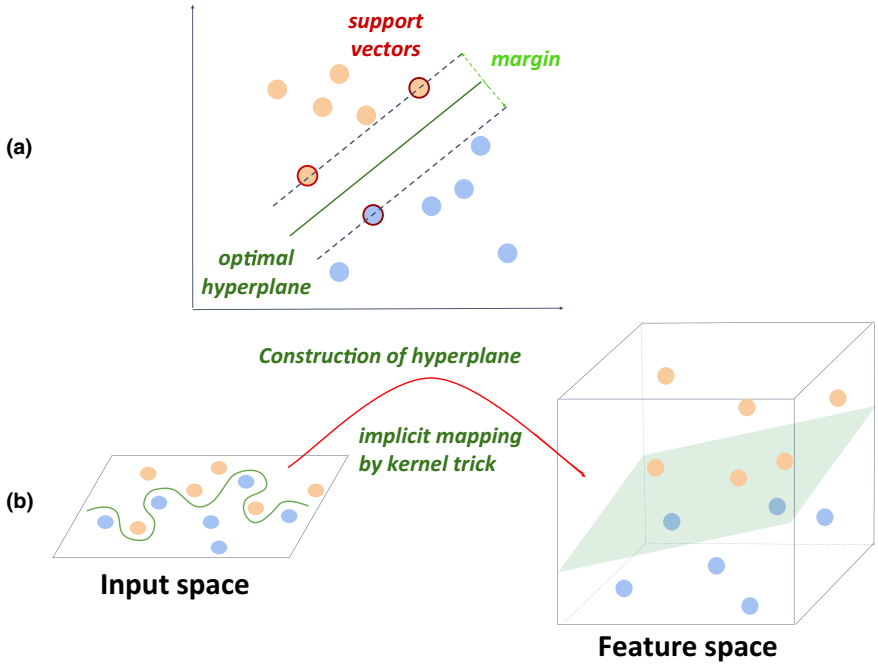


Figure 6 Illustration of support vector machine (SVM) principles. (**a**) Illustration of a simple case where hyperplane separate two groups directly in inputs space. (**b**) Illustration of performing nonlinear classification by implicitly mapping inputs into high-dimensional feature spaces where data points can be separated by a hyperplane.

of possible effects, it leads to harder interpretation of the eventual model. In practice, the selection of the kernel function is based on computational efficiency. Other popular kernels include linear and polynomial kernels.[47]

Neural networks

Background. Neural networks constitute a collection of neurons and edges, drawing its origins from circuit analysis. Different weights can be applied to each edge connecting the neurons. At each neuron, an activation function is applied to a weighed input signal to generate an output signal. A sigmoidal function is often used, consisting of a first order lowpass filter of a unit step function. Such sigmoidal function has the advantages of yielding bounded output and of being continuously differentiable, which is needed in the backward propagation step to tune the weights (parameters of the model), see steps defined in the section "Recurrent Neural Network."

Neurons are further subdivided into an input layer, hidden layer(s), and output layer, as shown in **Figure 7a**. The hidden layers perform the layer of abstraction needed to go from the input layer to the output layer. The number of hidden layers define whether the system is a shallow learning system (with one or a few hidden layer) or deep learning (with many hidden layers). There is an inherent trade-off between the number of hidden layers and time required to train the model. For this reason, although the core concept embedded in the neural network is not a novel one, it has found a resurgence of applications due to recent advances in computational power.

The most basic type is known as *feedforward* neural network, as information is just propagated from the input layer to the hidden layer(s) and finally to the output layer. The current state of the system is not defined by any past state; hence, it represents a memoryless system.

In the following, illustrative examples of neural networks are described: recurrent neural networks, long short-term memory networks, and gated recurrent networks. Further notable neural networks that are out of scope for this article but we recommend further reading on are convolutional neural networks,[53] encoder-decoder networks,[54,55] and generative models.[56]

Recurrent neural network. Recurrent neural networks are a class of neural networks dedicated to time series datasets as they factor in the inherent sequential relationship observed in the data of one time point to another. It has found success in what is known in the field as *sequential data*, where the order or time sequence of the signal plays a role, namely in natural language processing and time series forecasting. More closely related to our field of research, it has found application in predicting outcomes from electronic health records, where the richness comes inherently from the sequence correlation structure of the data to recommend swift and even anticipatory actions to be taken by the medical staff.[57] Rephrasing the question to solve a modeling conundrum in the pharmacometrics field is only starting to emerge at the time when this paper was drafted. Tang *et al*. present one of the rare attempts on how to use ML (here: RNNs) to characterize the PK of remifentanil and compared the results to the pharmacometrics gold-standard method NONMEM.[58] Although nonstandard PK models were used for the comparison and the generalizability of the results can be challenged, Tang *et al*. make a valuable contribution in exemplifying where RNNs could be used in pharmacometrics.

The basic form of an RNN is shown in **Figure 7b**, where each current state (at time t) is defined by a combination of the previous state of the system and the current input, which is similar to the concept of classical dynamic systems. The weights for each edge can be determined as to how far back to look into, similar to a time constant. Contrary to feedforward neural network, an identical weight is shared across in the individual neuron unit block across all the earlier discrete time steps.

At the core of the RNN, it consists of an input sequence defined by $x(t)$, an output sequence as defined by $o(t)$, a hidden or system state sequence as defined by $h(t)$, as well as a chained submodules of repeated units.

The steps needed to train an RNN model are as follows:

1. Define a network architecture and initialize the model with random weights and biases.
2. Perform a forward propagation to compute the estimated output.
3. Calculate the error at the output layer.

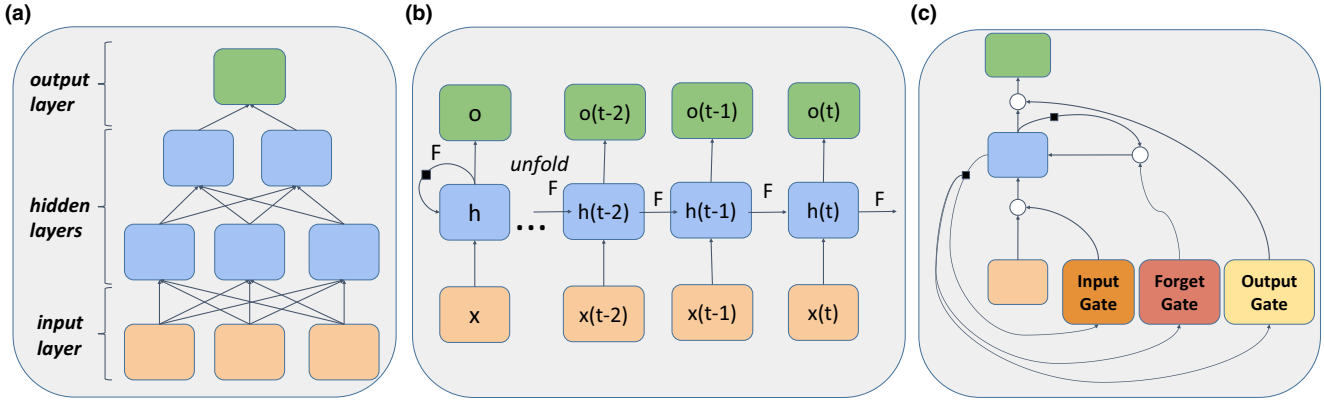


Figure 7 Neural networks. (**a**) Basics of feedforward neural networks. (**b**) Unfolding of recurrent neural networks. (**c**) Extensions of recurrent neural networks with gating units. Black square represents a delay of one discrete time step.

4. Perform a backward propagation to update the weights using an optimization approach.
5. Repeat steps 2–4 for the number of epochs (or iterations) until the loss function value is deemed minimized.

Extensions from this *Vanilla* RNN were developed to address the problems of unstable gradient problem (e.g., the vanishing gradient problem and the more serious counterpart of instability caused by an exploding gradient). These problems at their core are due to multiplications (under the influence of numerical errors) introduced in the backward propagation in relation of the error estimates to the parameters along each layer of the neural network. In other words, the vanishing gradient causes information that needs to be captured from a time point further away from the current time and, thus, renders the model weak to capture valuable stored memory with longer time lag. In the less common event that at least one partial derivative violates the requirement for stability, translating to the state matrix of having at least one eigenvalue > 1, this will lead to an exploding gradient problem, a known problem in traditional dynamic system for discrete time. The remit used to address this fundamental problem will be described more in two well-known extensions of RNN (long short-term memory (LSTM) and gated recurrent network (GRU)).

There has been many different variants and development in RNN research, each novel method serves to address a different problem ultimately leading to the development of more robust models. For example, to circumvent the unstable gradient problem, gradient clipping of forcing the gradient to a threshold has been proposed in ref. [59,60], but by far the most widely accepted method is the inclusion of *gating units*.

Long short-term memory and gated recurrent network. LSTM is part of a larger family of gated RNNs that retain and forget information with the introduction of gating units. More specifically, three gating units can be included in the system, as shown in **Figure 7c**. First, a direct copying or clearing of the state altogether can be controlled by the forget gate. A similar approach is also handled by the input gate to decide whether to include the current input signal as part of the update of the state. The amount of information to retain from the previous state signal and from the perturbation input signal is learned at each time step.[61] The system needs to learn long-term time dependencies by retaining information but it must also occasionally learn to clear information from its current state.[62] Consequently, solving the vanishing and exploding gradient problems. Finally, an output gate can be introduced, although less common, as a gating mechanism to decide which output signal gets fed back to the system.

A simpler rendition and, thus, faster training implementation can be found in GRU. GRUs address the same problem of unstable gradients and represent a new addition to this family of RNN extensions. The core difference between LSTM and GRU is that the latter omits the output gate and uses simpler reset and update gates.[63] In theory, however, LSTM should perform better as it can up-weigh or down-weigh information from longer time-distance/lag.

Examples of supervised ML applications in clinical pharmacology

Models in clinical pharmacology have typically been established by translating physiological and pharmacological principles to systems of differential equations and using expectation-maximization algorithms to estimate the model parameters. This mechanistically motivated approach has proven useful in many applications and is a well-established component of drug development programs. Potentially due to the success of these established approaches, only few examples of applying ML methods to clinical pharmacology problems exist up to now. Ryu *et al*. trained a deep neural network on a large curated database covering 192,284 drug-drug interactions in order to predict drug-drug and drug-food interactions for prescriptions, dietary recommendations, and new molecules.[64] Combining datasets from multiple studies to create large databases increases the potential to use ML to tackle broad clinical pharmacology questions.

ML has also been used to bridge drug discovery and clinical development. For example, Hammann *et al*. were able to predict incidence of adverse events from a molecule's chemical structure using a decision tree method.[65] Similarly, Lancaster and Sobie implemented SVMs to predict risk of Torsades de Pointes from *in vitro* data.[66]

In the area of personalized safety, ML has been used by Daunhawer *et al*. to personalize safety in the context of hyperbilirubinemia in neonates.[67] The authors used lasso and random forests to make predictions from clinical datasets. Furthermore, reinforcement learning was used by Gaweda *et al*. to personalize pharmacological anemia management.[68] A similar approach was used to develop a "closed loop" system for glucose control by combining a mathematical model, a glucose sensor, and a reinforcement learning model.[69] Chavada *et al*. and Hennig *et al*. investigated the feasibility of Bayesian feedback for dose adjustment of antibiotics.[70,71] The area of personalized healthcare could greatly benefit from using ML models that recommend dose adjustments in real time. In a recent study, an ML-type control algorithm was integrated with existing structural PK/PD models that are familiar to pharmacometricians and the resulting closed-loop control system was found to outperform a sensor-assisted pump.[69]

Main takeaways

- Supervised learning methods infer models based on labeled output-input pairs of the training dataset.
- Performance metrics are used to assess the classification and regression models to avoid overfitting of the training dataset.
- Many supervised learning methods exist with different trade-off between interpretability and performance.
- RNN is a special form of neural network that represents a dynamic system in discrete time.
- Examples of the applications of these supervised learning methods in computational biology and particularly clinical pharmacology are beginning to emerge.

DISCUSSION

In this tutorial, we have introduced some fundamental methods of ML that are likely to be of interest to the clinical pharmacology

and pharmacometrics community. Our brief introduction is supplemented with a range of relevant references. We have provided context by mentioning examples relevant to drug development. We conclude by summarizing how the fields of ML and clinical pharmacology are currently situated and by providing an outlook on how we expect to see further integration of the fields in the future. Advanced statistical methods are not new to pharmacometricians; in fact, such methods have been used to describe PK and PD phenomena for some time. For example, Bayesian methods are a well-established component of pharmacometric approaches.[72,73] It seems, therefore, likely that as statistical and ML approaches become more established and more prominent in the pharmaceutical industry, pharmacometricians will be among those who take advantage of these methods. Furthermore, new opportunities to investigate other clinical questions, such as patient stratification from high-dimensional baseline characteristics, may become possible in clinical pharmacology using ML approaches.

Several of the examples where ML approaches have been applied to clinical pharmacology questions include the integration of "classical" modeling techniques, such as specifying a structural model based on mechanistic understanding, and ML approaches.[69–71] Classical pharmacometric approaches are based on pharmacological principles that reflect hypotheses generated from the understanding of physiology and drug properties. It is unlikely that these models will be completely replaced by ML approaches in the near future. However, when the datasets and problems are more complex, many unknown influences and relationships exist and the focus is on interpolation and fast evaluation, pharmacometrics might benefit from applying ML-type methods. Going forward, we expect that fusing this understanding with ML models could lead to very effective models in the future. A recent perspective article provides more detail on applications of ML in clinical pharmacology.[74]

In the age of big data, there are many new opportunities for ML in clinical pharmacology. For example, data generated from wearable devices pose new challenges on how they can be linked to PK data in the future. In addition, access to real-world data could provide strong evidence for covariates, supplement control datasets, and bolster models that have been trained on small datasets.

In pharmacometric approaches, a predictive model is typically established by integrating a structural model and relevant data. The structural model substantially constrains the solution space and, therefore, relatively little data are required to fit the model. On the contrary, in neural networks, model structure is not prespecified and, thus, comparatively much more data are required for building a predictive model. It is also important to note that we are still very much at the infancy stage of understanding at which point the merger of larger data with these novel ML methods can be beneficial for performance as compared with more traditional methods. The following challenge[75] on time series forecasting shows that combinations of classic statistical and ML methods produce the most accurate forecasting and, thus, suggest it as a way forward. One of the main drivers of success of the pharmacometric approaches is that the models include a thorough understanding of the processes of drug absorption, distribution, metabolism, and elimination. The established models are highly predictive and, thus, find wide use in supporting drug development. Due to this success, despite the arrival of ML, classical pharmacometrics approaches are not expected to decrease in importance and activity. In contrast, they can be enhanced and improved by knowledge and insight distilled by ML methods and models.

An ongoing challenge for members of the clinical pharmacology community who wish to use ML methods is the inherent prevalence of longitudinal data. So far, there are many ML methods that rely on baseline features to make predictions, but relatively few examples where longitudinal data are used.

Overall, we expect that there will never be a universal, one-size-fits-all approach to which modelers from different fields converge. We note that there are many areas of potential synergy where modeling fields overlap in the remit of drug development. The clinical pharmacology community will continue to base their analyses on pharmacological principles and will gradually build in new ML elements to their workflow, strengthening their models further. In addition, the clinical pharmacology community will be able to enhance the range of questions they are able to address by using ML approaches.

FUNDING

I.I.D. is a recipient of the Roche Postdoctoral Fellowship. This study is funded by F. Hoffmann-La Roche Ltd.

CONFLICT OF INTEREST

All authors declared no competing interests for this work.

1. Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C. & Collins, J.J. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).
2. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
3. Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
4. Kleene, S.C. Representation of Events in Nerve Nets and Finite Automata (RAND Project Air Force, Santa Monica, CA, 1951) <https://apps.dtic.mil/docs/citations/ADA596138>.
5. Breiman, L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, 199–231 (2001).
6. Ribeiro, M.T., Singh, S. & Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA*, August 13–17, 2016.
7. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2014).
8. Lipton, Z.C. The mythos of model interpretability. *ACM Queue* **16**, 31–57 (2018).
9. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer New York Inc., New York, NY, 2008).
10. Jerez, J.M. *et al*. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif. Intell. Med.* **50**, 105–115 (2010).

11. Dorogush, A.V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).
12. Cortes, C., Mohri, M., Riley, M., Rostamizadeh, A., Freund, Y., Györfi, L., Turán, G. & Zeugmann, T. (eds.) Sample selection bias correction theory. In *Algorithmic Learning Theory*. 38–53 (Springer, Berlin, Heidelberg, 2008).
13. Lee, B.K., Lessler, J. & Stuart, E.A. Improving propensity score weighting using machine learning. *Stat. Med.* **29**, 337–346 (2010).
14. Newby, D., Freitas, A.A. & Ghafourian, T. Coping with unbalanced class data sets in oral absorption models. *J. Chem. Inf. Model.* **53**, 461–474 (2013).
15. Hu, L.-H., Huang, M.-W., Ke, S.-W. & Tsai, C.-F. The distance function effect on k-nearest neighbor classification for medical datasets. *Springerplus* **5**, 1304 (2016).
16. Borozan, I., Watt, S. & Ferretti, V. Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. *Bioinformatics* **31**, 1396–1404 (2015).
17. Collisson, E.A., Bailey, P., Chang, D.K. & Biankin, A.V. Molecular subtypes of pancreatic cancer. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 207–220 (2019).
18. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
19. Kriegel, H.-P., Kröger, P., Sander, J., Zimek, A. Density-based clustering. *Wiley Interdisc. Rev Data Min. Knowl. Discov.* **1**, 231–240 (2011).
20. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X.A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 226–231 (AAAI Press, Portland, OR, 1996).
21. Zimek, A., Schubert, E. & Kriegel, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min.* **5**, 363–387 (2012).
22. Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901).
23. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learning Res.* **9**, 2579–2605 (2008).
24. Becht, E. *et al*. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
25. Nguyen, L.H. & Holmes, S. Ten quick tips for effective dimensionality reduction. *PLoS Comput. Biol.* **15**, e1006907 (2019).
26. Wang, C., Machiraju, R. & Huang, K. Breast cancer patient stratification using a molecular regularized consensus clustering method. *Methods* **67**, 304–312 (2014).
27. Cooper, G.S., Bynum, M.L. & Somers, E.C. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *J. Autoimmun.* **33**, 197–207 (2009).
28. Walsh, D. & Rybicki, L. Symptom clustering in advanced cancer. *Support. Care Cancer* **14**, 831–836 (2006).
29. Genotype-Tissue Expression (GTEx) Consortium *et al.* The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
30. Zhang, J.D., Berntenis, N., Roth, A. & Ebeling, M. Data mining reveals a network of early-response genes as a consensus signature of drug-induced in vitro and in vivo toxicity. *Pharmacogenomics J.* **14**, 208–216 (2014).
31. Gemma, A. *et al*. Anticancer drug clustering in lung cancer based on gene expression profiles and sensitivity database. *BMC Cancer* **6**, 174 (2006).
32. Koch, M.A. & Waldmann, H. Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. *Drug Discov. Today* **10**, 471–483 (2005).
33. Reutlinger, M. & Schneider, G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph. Model.* **34**, 108–117 (2012).
34. Ezzat, A., Wu, M., Li, X.-L. & Kwoh, C.-K. Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods* **129**, 81–88 (2017).
35. Nesterov, Y. *Lectures on Convex Optimization*. Vol. **137** (Springer, Berlin, Germany, 2018).
36. Cawley, G.C. & Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
37. Mitchell, T.M. *Machine Learning* (McGraw-Hill Inc., New York, NY, 1997).
38. Belson, W.A. Matching and prediction on the principle of biological classification. *J. R. Stat. Soc. Ser. C Appl. Stat.* **8**, 65–75 (1959).
39. Krischer, J.P. An annotated bibliography of decision analytic applications to health care. *Oper. Res.* **28**, 97–113 (1980).
40. Shortliffe, E.H., Buchanan, B.G. & Feigenbaum, E.A. Knowledge engineering for medical decision making: A review of computer-based clinical decision aids. *Proc. IEEE* **67**, 1207–1224 (1979).
41. Bach, P.H. & Bridges, J.W. A decision tree approach for the application of drug metabolism and kinetic studies to in vivo and in vitro toxicological and pharmacological testing. *Arch. Toxicol. Suppl.* **8**, 173–188 (1985).
42. Jordan, T.J. & Reichman, L.B. Once-daily versus twice-daily dosing of theophylline: a decision analysis approach to evaluating theophylline blood levels and compliance. *Am. Rev. Respir. Dis.* **140**, 1573–1577 (1989).
43. Breiman, L., Friedman, J., Olshen, R. & Stone, C. Classification and regression trees. *Wadsworth Int. Group* **37**, 237–251 (1984).
44. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).
45. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
46. Segal, M.R. Machine Learning Benchmarks and Random Forest Regression. Technical Report, (Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco, CA, 2003).
47. Ben-Hur, A., Ong, C.S., Sonnenburg, S., Schölkopf, B. & Rätsch, G. Support vector machines and kernels for computational biology. *PLoS Comput. Biol.* **4**, e1000173 (2008).
48. Newman, A. *et al*. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
49. Vapnik, V. & Lerner, A. Pattern recognition using generalized portrait method. *Automat. Rem. Contr.* **24**, 774–780 (1963).
50. Smola, A.J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
51. Mercer, J. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. Lond. B Bio. Sci.* **415–446**, (1909).
52. Schölkopf, B. & Smola, A.J. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (The MIT Press, Cambridge, MA, 2002).
53. Krizhevsky, A., Sutskever, I. & Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25**, 1097–1105 (2012).
54. Cho, K.*et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
55. Cho, K., Van Merriënboer, B., Bahdanau, D. & Bengio, Y.On the properties of neural machine translation: encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).
56. Kingma, D.P., Mohamed, S., Rezende, D.J. & Welling, M. Semi-supervised learning with deep generative models. NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, **2**, 3581–3589 (2014).
57. Choi, E.*et al*. Using recurrent neural network models for early detection of heart failure onset. *OUP Academic* (2016) <https://academic.oup.com/jamia/article/24/2/361/2631499>
58. Tang, J.-T., Cao, Y., Xiao, J.-Y. & Guo, Q.-L. Predication of plasma concentration of remifentanil based on Elman neural network. *J. Cent. South Univ.* **20**, 3187–3192 (2013).
59. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. International conference on machine learning, Edinburgh, June 26–July 1, 1310–1318 (2013).

60. Bengio, Y., Boulanger-Lewandowski, N. & Pascanu, R. Advances in optimizing recurrent networks. IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, March 25–30, 8624–8628 (2012).
61. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
62. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge, MA, 2016) <http://www.deeplearningbook.org>
63. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. Workshop on Deep Learning in NIPS, Montreal, December 8–13, (2014).
64. Ryu, J.Y., Kim, H.U. & Lee, S.Y. Deep learning improves prediction of drugdrug and drugfood interactions. *Proc. Natl. Acad. Sci.* **115**, E4304–E4311 (2018).
65. Hammann, F., Gutmann, H., Vogt, N., Helma, C. & Drewe, J. Prediction of adverse drug reactions using decision tree modeling. *Clin. Pharmacol. Ther.* **88**, 52–59 (2010).
66. Lancaster, M.C. & Sobie, E. Improved prediction of drug-induced torsades de pointes through simulations of dynamics and machine learning algorithms. *Clin. Pharmacol. Ther.* **100**, 371–379 (2016).
67. Daunhawer, I. *et al*. Enhanced early prediction of clinically relevant neonatal hyperbilirubinemia with machine learning. *Pediatr. Res.* **86**, 122 (2019).
68. Gaweda, A.E. *et al*. Individualization of pharmacological anemia management using reinforcement learning. *Neural Networks* **18**, 826–834 (2005).
69. Benhamou, P.-Y. *et al*. Closed-loop insulin delivery in adults with type 1 diabetes in real-life conditions: a 12-week multicentre, open-label randomised controlled crossover trial. *Lancet Dig. Health* **1**, e17–e25 (2019).
70. Chavada, R., Ghosh, N., Sandaradura, I., Maley, M. & Van Hal, S.J. Establishment of an $AUC_{0\text{-}24}$ threshold for nephrotoxicity is a step towards individualized vancomycin dosing for methicillin-resistant staphylococcus aureus bacteremia. *Antimicrob. Agents Chemother.* **61**, e02535–16 (2017).
71. Hennig, S., Holthouse, F. & Staatz, C.E. Comparing dosage adjustment methods for once-daily tobramycin in paediatric and adolescent patients with cystic fibrosis. *Clin. Pharmacokinet.* **54**, 409–421 (2015).
72. Dansirikul, C., Morris, R.G., Tett, S.E. & Duffull, S.B. A Bayesian approach for population pharmacokinetic modelling of sirolimus. *Br. J. Clin. Pharmacol.* **62**, 420–434 (2006).
73. Lunn, D.J., Best, N., Thomas, A., Wakefield, J. & Spiegelhalter, D. Bayesian analysis of population PK/PD models: general concepts and software. *J. Pharmacokinet. Pharmacodyn.* **29**, 271–307 (2002).
74. Hutchinson, L. *et al*. Models and machines: how deep learning will take clinical pharmacology to the next level. *CPT Pharmacomet. Syst. Pharmacol.* **8**, 131–134 (2019).
75. Makridakis, S., Spiliotis, E. & Assimakopoulos, V. The M4 competition: results, findings, conclusion and way forward. *Int. J. Forecast.* **34**, 802–808 (2018).

ELSEVIER

KEYNOTE (GREEN)

Causal inference in drug discovery and development

Tom Michoel [1], **Jitao David Zhang** [2,3,*]

[1] Computational Biology Unit, Department of Informatics, University of Bergen, Postboks 7803, 5020 Bergen, Norway
[2] Pharma Early Research and Development, Roche Innovation Centre Basel, F. Hoffmann-La Roche, Grenzacherstrasse 124, 4070 Basel, Switzerland
[3] Department of Mathematics and Computer Science, University of Basel, Spiegelgasse 1, 4051 Basel, Switzerland

Tom Michoel is professor in bioinformatics at the University of Bergen. His research focuses on developing methods for causal inference from multi-omics data.

Jitao David Zhang is a computational biologist working in drug discovery. His research focuses on multiscale-scale causal modelling for risk-benefit assessment of drug candidates.

To discover new drugs is to seek and to prove causality. As an emerging approach leveraging human knowledge and creativity, data, and machine intelligence, causal inference holds the promise of reducing cognitive bias and improving decision-making in drug discovery. Although it has been applied across the value chain, the concepts and practice of causal inference remain obscure to many practitioners. This article offers a nontechnical introduction to causal inference, reviews its recent applications, and discusses opportunities and challenges of adopting the causal language in drug discovery and development.

Keywords: causality; causal inference; DAG (standing for directed acyclic graph); drug discovery; drug development; reverse translation

Introduction

Causal inference is the process of identifying causal effects based on prior knowledge, hypothesis, and correlations observed in data. This article aims at equipping practitioners of drug discovery and development, especially those without formal training in mathematics and statistics, with necessary knowledge to start working with causal inference at three levels: (i) to recognize situations where causal inference is advantageous and to understand why a causal model is needed (Section *What is causal inference and why do we need it?*), (ii) to perform causal inference with software (Section *A step-by-step guide of causal inference*), and (iii) to learn and get inspiration from recent applications (Section *Literature review and case studies*). Finally, we discuss opportunities and challenges of adopting the causal language in drug discovery and development. A glossary is offered in the Supplementary File 1 in the Supplementary material online for readers to look up definitions of important terms.

Causal inference identifies causations from correlations. In statistics, *correlation* means the relationship between two random variables, whether causal or not. For instance, high temperatures lead to higher ice cream sales; therefore, the variables *temperature* and *ice cream sales* are correlated. Similarly, high temperatures make wildfire more likely to happen; therefore, the variables *temperature* and *wildfire frequency* are correlated. Apparently, *ice cream sales* and *wildfire frequency* are correlated but do not cause each other. The task of causal inference in this context is to infer

* *Corresponding author. Zhang, J.D. (jitao_david.zhang@roche.com)*

the causal relationship between *temperature, ice cream sales,* and *wildfire frequency* based on observations.

There is neither a consensus definition of causality nor a single way to identify it (Box 1). It is partially because the nature of causation is still under debate,[1] and partially because even simple causality such as *higher temperature causes higher ice cream sales* can be broken down into an infinite chain of causal relationships that can involve entities across both physical (e.g., molecules, receptors, cells, organs and systems, society) and time scales.

In this article, we introduce the statistical causal inference approach and define *causality* as a probabilistic relationship that satisfies four conditions: *regular probabilistic update*, *manipulation*, *counterfactual condition*, and *mechanism of action*.

Take the example of testing whether a drug is causal for halting deadly cancer progression. *Regular probabilistic update* means that taking the drug modifies the conditional probability of dying of the disease within a defined time window, regardless of where and when the trial happens. *Manipulation* means that drug treatment shows additional benefit even if we consider all other factors affecting patient survival, such as age and comorbidities. *Counterfactual condition* means that the death of a patient would not have been postponed had the drug not been taken. Finally, *mechanism of action* means that we understand why the drug prolongs patient survival, such as by activating tumour-infiltrating immune cells. Taken together, the four conditions ensure both statistical correlation and mechanistic understanding. They embody the causality criteria proposed by Austin Bradford Hill[2] and implement a practical test for the philosophical reasoning of establishing causality in healthcare.[3]

What is causal inference and why do we need it?

Why causality matters in drug discovery and development

Causality is indispensable for predicting outcomes of intervention, for answering counterfactual *what if* questions, and for human understanding. Although it takes time and effort to identify causes from correlations, the investment is rewarding. The history of drug discovery abounds with cases where causality inspires repurposing of a drug, developing new classes of chemical matters, or refuting the use of a drug.

One prominent example is the history of thalidomide.[4] Chemically, thalidomide is a mixture of two enantiomers: the sedative (*R*)-enantiomer and the teratogenic (*S*)-enantiomer. Although its use was first correlated with an unusually high incidence of birth defects,[5] it took decades to identify the cause of the teratogenicity, that is, the degradation of SALL4 protein via the E3 ubiquitin ligase complex.[6] In the meantime, thalidomide was found to be potentially effective against drug-resistant multiple myeloma. Follow-up study based on the first causal finding offered a causal explanation for this correlation, namely that thalidomide induces protein degradation of key transcription factors *Ikaros* (IKZF1) and *Aiolos* (IKZF3) with the same ubiquitin apparatus. Furthermore, the same causal link between thalidomide and protein degradation has inspired a flourishing search for new modalities, including proteolysis-targeting chimera (PROTAC)[7] and other multispecific drugs.[8] The ups and downs of thalidomide, in which causality plays a central role in translating serendipitous findings into new discoveries, is therefore described as Shakespearean.[4]

Repurposing drugs to treat COVID-19 offers another recent example demonstrating the importance of causality. Much literature reported a correlation between treatment with drugs targeting one or both human sigma receptors (σ_1 and σ_2), such as hydroxychloroquine, and negative modulation of SARS-CoV-2 infection. Researchers, however, noticed the intriguing discrepancy that although most compounds have comparable affinity against sigma receptors, the cellular antiviral activities showed large variation. It raised the question whether pharmacological modulation of sigma receptors is causal for SARS-CoV-2 inhibition. The question led Tummino *et al.*[9] to find that phospholipidosis, instead of target-mediated mechanisms, underlies the antiviral activity of many drugs. Phospholipidosis, determined by physicochemical properties of the drugs, results in dose-limiting toxicity and can be reliably predicted experimentally and computationally.[10] The lack of causal relationship between pharmacological modulation of sigma receptors and antiviral activities highlights the importance of dissecting causal mechanisms from correlations, from the opposite angle of the example of thalidomide.

Although correlations suffice for predictions and (wrong) explanations, causality enables us to understand why a drug works or causes harm and to act rationally. What is the relationship between correlation and causation? How can we systematically identify causations from correlations?

Distinguishing causation from correlation

Causality is often hidden and manifests itself by correlations, though correlations are not always due to direct causality. There are six reasons why we observe a correlation between two events: *causation*, *confounding*, *coincidence*, *conspiracy*, *collider,* and *chronology*.

We assume that two variables, x and y, depict expression levels of two proteins, X and Y, in a population of cells. If x and y are correlated, six scenarios are possible:

1. Causation: expression of X causes expression of Y, or expression of Y causes expression of X. We may use additional knowledge to favour one direction over the other. For instance, if X is a known transcription factor but Y is not, we may favour the model X → Y (→reads causes) over Y → X.
2. Confounding: a third, potentially unobserved, protein U causes expression of both X and Y, i.e. X ← U → Y.
3. Coincidence: the correlation is solely by chance. If so, we shall observe diminishing correlations as we collect more data. We can perform statistical inferences, such as permutation test and bootstrapping, to test how likely we observe the correlation by chance.
4. Conspiracy, or deliberate selection of data: the correlation is due to deliberate manipulation of data or the sampling process. We may, for instance, create a good correlation by removing data from all cells where two proteins are not correlated.
5. Collider: besides conspiracy, nondeliberate selection of observations from the general population can also cause correlation. Even if x and y are not correlated in all cells (imagine

random points confined in a circle), by considering only cells in which at least one of two proteins is expressed over a certain threshold (imagine removing points between six and nine o'clock), we observe a negative correlation between x and y. This is known as the Berkson's paradox[11] or the admission rate bias.[12]

6. Chronology: If both X and Y change their expression during the life cycle of a cell, they are correlated in time series even if neither protein causes change of expression of the other. Although time can be considered a confounder, it is not a cause per se of either protein's expression. Such independent trends with regard to time are systematically explored by Hans Reichenbach in The Direction of Time.[13] This is a subject of intensive debate in contemporary philosophy.[14]

Collectively, we call the six mechanisms the *6C model of correlation*. The model helps us distil real causes from empirical correlations by excluding the possibility that alternative mechanisms are at work. In particular, data and statistical models alone cannot tell causation apart from confounding. Figure 1 illustrates a 'toy' example in the same line of the protein expression example above. It demonstrates that in order to identify causal relationships from correlations, we need a model of either knowledge or hypotheses about how data is generated and to operate the model with statistical modelling.

For some applications, correlation is all we need. In the toy example illustrated, if our goal is to predict y given x in similar yet unobserved cells of the same population, correlation suffices, and there is little need to distinguish between causation and confounding. Off-the-shelf statistical and machine-learning models are good enough, because they exploit correlations between input variables and the target variable.

Delineating causation from correlation is essential for many other applications. If our task is to predict the expression of Y in a cell where we inhibit the expression of X (Figure 1g), such as by gene silencing, it is essential to know which of the models below is the correct one: $X \rightarrow Y$, or $Y \rightarrow X$, or $X \leftarrow U \rightarrow Y$. Because if $X \rightarrow Y$ is true, expression of Y will become residual; given the other two models, expression of Y will not be affected. If we do not have a model, collecting more data does not help us with the task.

The simple example illustrates three critical points: (i) causality is required to predict the outcome of intervention; (ii) data alone are not able to tell causality, no matter how much data we have and however complex or fancy the algorithm is; and (iii) we need both scientific models, that is, prior knowledge and hypothesis, and data to infer causality.

The takeaway message is that although statistical and machine-learning models are useful if we ask for correlation, we need additional tools to recover causality from correlations if our goal is to intervene in the system. While correlation helps us to predict or even to explain,[15,16] causality helps us to understand and to act.

Causal modelling with directed acyclic graphs (DAGs)

Directed acyclic graphs, or DAGs, are used as computational models of causal inference. DAGs contain nodes, which represent variables, and edges, which represent causal relationships. Like the two-variable cases described above, DAGs represent either knowledge or hypotheses about causal relationships between the variables. The edge, that is, the causal relationship, can have any discrete, linear, or nonlinear functional form. A lack of edge between any pair of nodes means that we exclude the possibility that they have direct causality on each other. An indirect causal relationship may still exist in such cases due to the propagation of causality through the DAG, as we shall see below.

Three three-node structures are prevalent in DAGs. Understanding these common structures, known as *motifs*, allows us to interpret more complex causal models:

1. The pipe (Figure 2a) describes the simplest chain of causality: $X \rightarrow Z \rightarrow Y$.
2. The fork (Figure 2b) describes a variable Z causing both X and Y: $X \leftarrow Z \rightarrow Y$.
3. The collider (Figure 2c) describes the situation where a variable Z is caused by both X and Y: $X \rightarrow Z \leftarrow Y$.

A pipe transduces causality in a chain, such as *Vemurafenib binding to V600E-mutated BRAF → reduced BRAF signalling → reduced tumour size*. If the intermediate cause is manipulated, such as by mutations enhancing BRAF signalling, the causal effect from drug on tumour size will diminish or even reverse. Statistically speaking, X and Y are marginally correlated, and they are independent conditional on Z.

A fork, like the temperature example raised before, generates correlations between variables that do not cause each other, for example: *PI3K phosphorylation ← RAS signalling → BRAF proliferation*. If the common cause is present, both its effects are correlated; if the common cause is absent, or if we focus on the subset of the data where the common cause takes the same value, the effects become independent. Similar to the situation of pipes, X and Y are marginally correlated, and they become independent conditional on Z.

Colliders behave very differently from pipes and forks. For instance, cells that lose the expression of the BOP1 (block of proliferation 1) gene become resistant against vemurafenib.[17] The mechanism can be described by a collider structure: *BRAF V600E → increased phosphorylation of ERK1/2 ← loss of BOP1*, namely both BRAF V600E mutation (V600E for short) and loss of BOP1 (BOP1↓) increases ERK1/2 phosphorylation (↑pERK1/2). As long as either condition is satisfied, we can observe ↑pERK1/2. Assuming that they are the only two factors influencing pERK1/2, the collider creates a dependency between V600E and BOP1↓. When ↑pERK1/2 is true, either V600E or BOP1↓ or both must exist; otherwise, neither V600E nor BOP1↓ is possible. Even if V600E and BOP1↓ are statistically independent from each other, that is, knowing V600E does not give us any information about whether BOP1 is lost or not, the two events become correlated once we observe ↑pERK1. In contrast to pipes and forks, X and Y are marginally independent but become correlated conditional on Z.

Knowledge of causal structure allows us to better recognize cognitive bias. We demonstrate this point with the example of colliders. With simulated data generated by the three three-node structures in Figure 1, we show that we can remove or

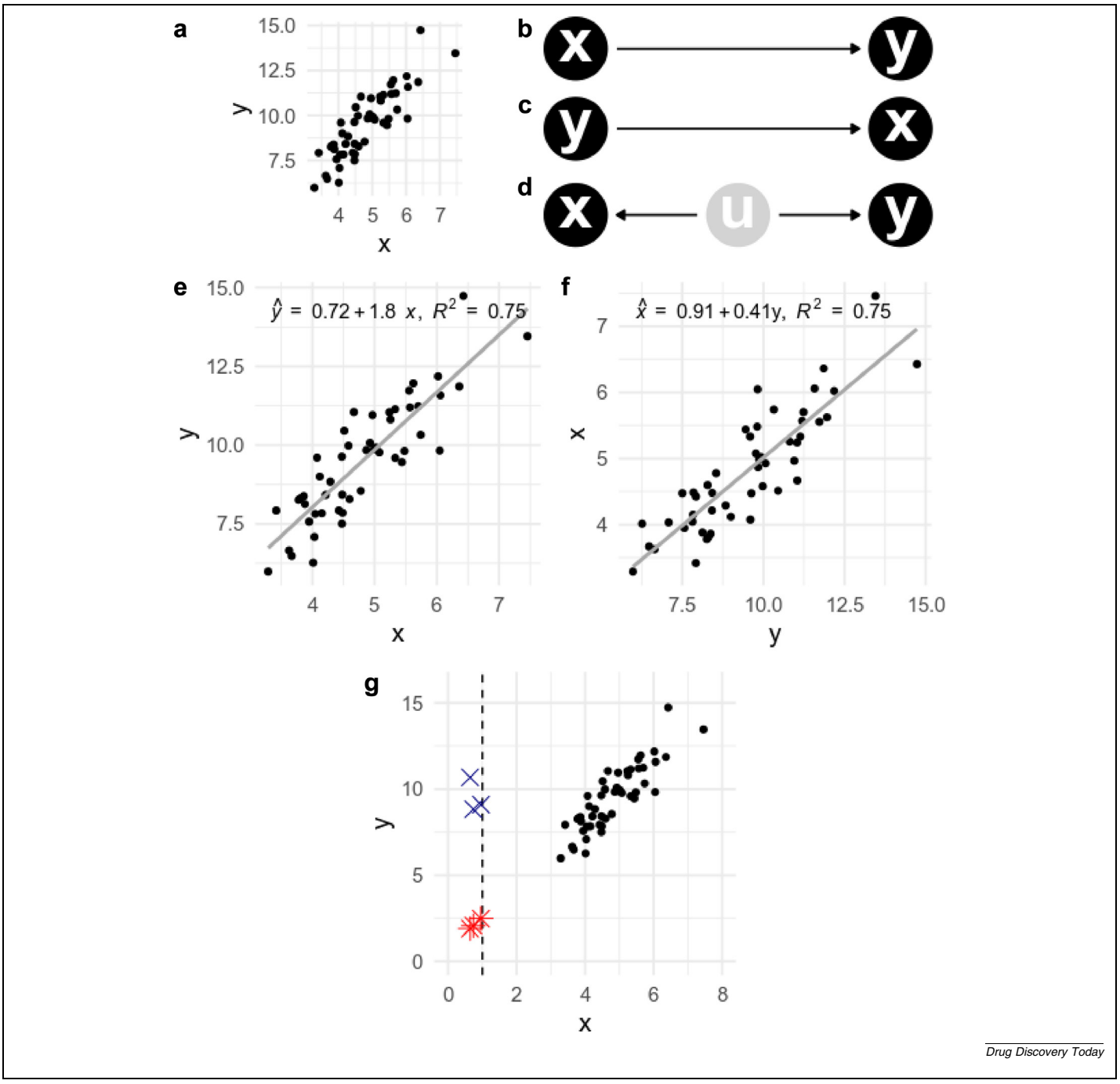


FIGURE 1

Distinguishing between correlation and causation. (a) Correlation between two variables *x* and *y*, which represents the expression levels of two proteins X and Y in a population of cells, respectively. Each dot represents one cell. The data were simulated by a linear relationship. (b-d) Three causal models that can generate the correlation observed in (a). two direct causal relationships (panel b and c) and one with a confounding variable (panel d). Coincidence, conspiracy, collider, and chronology as reasons generating correlations are not shown for simplicity. (e-f) Data and statistical models alone cannot tell the direction of causation, or whether confounding variables exist. Panel e shows the linear regression with *x* as independent variable and *y* as dependent variable using data shown in panel A. Panel f shows the regression with the same dataset, with the role of *x* and *y* swapped. The correlation coefficients (R^2) of both regression models are identical. (g) Predicting the outcome of interventions requires causal models. Recall that we use *x* and *y* to denote expression of two proteins X and Y in a population of cells. And now we reduce the expression of X artificially to 1.0 (dashed line in black). Depending on the causal structure, the outcome of Y may be near to the red stars (model in panel b) or blue crosses (models in panel c or d).

create correlation between *x* and *y* if we stratify by the third variable *z*, depending on the graph structure. The conclusion holds if *z* is a continuous variable and if we regress it out in a statistical model. Imagine now we face a dataset with two variables, *x* and *y*; we observe a good linear correlation; and our goal is to predict the value of *y*, given a new set of values *x*. Does a linear model suffice? Not necessarily, because if there is a variable *z* forming a collider with *x and* y, and its value changes for the

new set of values x, then our linear model is likely to fail. Even the simplest causal graphs show us it is important to know which factors affect the generation of data.

Besides the pipes, forks, and colliders, we highlight two four-node motifs:

1. The descendant describes a variant of the collider, where the variable caused by two variables X and Y is not directly observed, but its descendant, namely another variable that is caused by the unobserved variable, is observed. Because we often use the variable U to indicate an unobserved variable, the DAG can be represented as $X \rightarrow U \leftarrow Y$ and $U \rightarrow Z$. The structure, for instance, is useful to model the effect of known risk factors (X) and treatment (Y) on disease status (U). Though U may not be directly observed, a biomarker (Z), which is partially affected by U, can be measured.
2. The paw or 3-pan describes a class of motifs, including the one shown in Figure 5a. The tail variable is known as the instrumental variable. Models with instrumental variables are applicable in a wide variety of contexts, including Mendelian randomization, a technique for target identification to be described in Section 4, and the analysis of clinical trial data with noncompliance, that is, when some patients in the treatment arm do not take the medicine as prescribed.[18]

DAGs consisting of these and other motifs[19] can represent causal relationships between any number of variables. We refer readers interested in reading and understanding DAGs better to several outstanding tutorials.[20–22]

Because DAG structure specifies marginal and conditional correlations, it is possible to rank models by their plausibility, given data, and to learn the strength of causal relationships. The two tasks are known as *causal identification*, or *causal discovery*, in which the truth value of a claim of the form 'C causes E' is determined, and quantitative *causal estimation*, in which a numerical value s (strength) is estimated for a claim of the form 'C has an effect on E.' Causal discovery is more challenging than causal estimation, and discovered causal effects usually require causal estimation before they can be used for high-stakes applications.[23,24] Nevertheless, it plays an important role in reconstructing causal networks for disease understanding and target identification, especially because we lack much biological knowledge on how the observed data are generated. See Section 4 for details.

Relationship between causal inference and established modelling techniques

Causal inference completes statistical modelling and mechanistic modelling, two commonly used techniques in drug discovery and development (Table 1). On one hand, statistical modelling, including pattern recognition–based machine-learning models, sometimes termed *artificial intelligence* (AI), aims at identifying prediction-relevant features and functions that transform the features to approximate the target variable and thereby exploit correlations. The models may or may not have a causal interpretation. On the other hand, mechanistic modelling uses mathematical models of biological processes to describe and predict how components, information, or energy of a system evolve with time. Examples include pharmacokinetics and pharmacodynamics models and quantitative system pharmacology models based on ordinary or partial differential equations, agent-based models, and hidden Markov models. Although mechanistic models usually have a causal interpretation and are powerful for predicting the outcome of an intervention, we often miss information of important factors that affect the system's behaviour.

The three modelling approaches complement and benefit from each other. Reciprocal benefits between statistical and causal models are apparent. By discovering and quantifying causal relationships between observed variables, we can render statistical models' causal interpretations. By using advanced machine-learning approaches, we may learn highly complex functional forms describing the causal effect[25] or discover causality from high-dimensional data.[26] Practically, because we are often limited by the volume of high-quality data, causal inference may likely profit from parsimonious statistical models and Bayesian approaches.[27]

Results of causal inference help refine and improve existing mechanistic models by including causal factors and removing confounding factors. In turn, outputs from mechanistic models complement the current causal inference regime by predicting time-dependent outcomes of intervention, in particular those of components that regulate each other.

Causal inference for experimental data and observational data

The power of causal inference to connect and enhance existing modelling approaches is particularly prominent in the analysis of observational studies. Classically, we distinguish observational studies from controlled experimental studies. In a controlled experiment, we assign test objects, such as animals in preclinical experiments, to groups. One group receives the treatment and the other group does not. If we go one step further to require that the grouping is randomised with regard to any relevant attributes of the test objects (passage, sex, body weight, etc.), then we have a randomised controlled experiment, a gold standard to establish causality. In an observational study, in contrast, we measure or survey members of a sample without trying to affect them. Such data can come from epidemiological studies, electronic health records (EHRs), insurance claims, and other data that come from natural (instead of controlled) experiments, such as omics and behavioural data of healthy individuals and patients.

Several reasons make it imperative to use causal models to analyse data generated by observational studies. First, it allows us to integrate knowledge and hypotheses about inevitable biases in the data generation process. Second, we can investigate the causal effect of independent variables of interest while considering other variables that affect the outcome, known as *covariates*. Third, we can potentially resolve the effects of variables that influence both the independent variable and the outcome, known as *confounding variables*. If we analyse observational data as if we were handling randomised experimental data, without considering the causal structure underlying the data generation and collection, we may derive false, sometimes ridiculous, conclusions.[28]

Even if a study is set up as a classical randomised experiment, various reasons may break the randomization and demand causal

TABLE 1

A comparison of data modelling techniques.

	Examples	Prediction for independent and identically distributed samples	Prediction for outcome of intervention	Answering *what if*, counterfactual questions	Data-driven discovery
Mechanistic models	Pharmacokinetics and physically based pharmacokinetic models	Yes	Yes	Yes	Maybe
Causal models	See discussion and examples	Yes	Yes	Yes	Maybe
Statistical models	Cox regression model for survival, statistical tests, (generalised) linear models for *omic* data, support vector machines, random forest, neural networks	Yes	No	No	Yes

Adapted from Table 1.1 in Peters *et al.*[137].

inference to identify the treatment effect. For instance, in clinical trials, noncompliance is a common issue, where some patients do not take the treatment as prescribed. To identify the real treatment effect while acknowledging the noncompliance, it is necessary to employ a causal model, with the *instrumental variable* model illustrated above as the simplest example. Other issues that have been successfully addressed by causal inference include missing data,[29,30] such as when patients drop out of the trial, and intercurrent events, that is, events occurring after randomization that can either preclude observation of the outcome of interest or affect its interpretation, as defined by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use in its Guideline E9(E1), such as the development of antidrug antibodies.[31–34] The book by Guido Imbens and Donald Rubin provides an outstanding introduction to the analysis of these and other aberrations from classical randomised experiments with causal inference.[18]

Causal inference can leverage both experimental and observational data to generate insight. Eichler *et al.*[35] proposed a new framework, known as *threshold crossing*, that generates evidence by synthesising historical randomised clinical trials and real-world data resources. A recent landscape assessment and elicited comments have also shown that causal inference penetrates and contributes significantly to clinical trial study design and analysis by judging the suitability and integrating real-world data from observational studies.[36,37]

Finally, we emphasise the importance of study design for both experimental and observational studies. It is widely acknowledged that well-designed randomised controlled experiments, both in preclinical drug discovery[38] and in clinical drug development,[39] are the gold standard of verifying causal relationships. The quality and strength of evidence provided by an observational study and causal inference is also determined largely by the study design.[40] Study design should be scrutinised prior to causal inference in order to gauge whether we can answer the question of interest.

A step-by-step guide to causal inference

We introduce causal inference as an iterative process of six steps: *modelling* → *identification* → *estimation* → *refutation* → *refinement* → *application*. They are illustrated in Figure 3 and detailed below. To assist practitioners acquiring hands-on experience with causal inference, we offer complementary interactive tutorials Supplementary material. The tutorials, implemented in the programming languages Python (with *Jupyter Notebooks*) and R (with *Rmarkdown)*, are available at https://github.com/Accio/causal_-drug_discovery.

Step 1: Modelling

To start, we construct a DAG to model causal mechanisms that generated the observed data by synthesising common sense, scientific knowledge, or explicit assumptions. If multiple hypotheses exist, we may build several DAGs and subject them to analysis and comparison.

In case there is little knowledge and hypothesis available, we may apply causal discovery techniques to raise hypotheses about the mechanism of data generation. Commonly used methods include Bayesian networks,[41] factor analysis,[42] and encoder-decoder and other deep generative models.[43] Limitations of such methods include (i) they are often technically challenging (*NP-hard* problems), (ii) the results often include many alternative hypotheses that explain the data equally well, and (iii) the proposed models may not be causal, that is, they fail to be validated by interventional studies. Nevertheless, the discovered causal graphs, combined with expert review and curation, may serve as reasonable starting points.

Step 2: Identification

Once a causal model is set up, we test whether we can answer the question of interest quantitatively. The quantity that addresses the causal question, such as *how strong is the effect of the drug on disease progression*, is known as the *estimand* (see more details in Supplementary File 1, in the Supplementary material online). In this step, our goal is to assess whether we can estimate the estimand from the data at all, because some graph structures prohibit us from doing so.

In a causal model, the estimand is usually the causal effect pointing from a treatment (cause) to its target (effect) variable, which can be a regression coefficient or other numeric values that quantify the strength of the causal relationship. The causal graph model structure determines which estimators are available, and the numerical values of estimates are gauged by the users to interpret the results and challenged by refutation analysis.

An interesting and important result from theoretical studies is that we can assert whether an estimand can be identified using

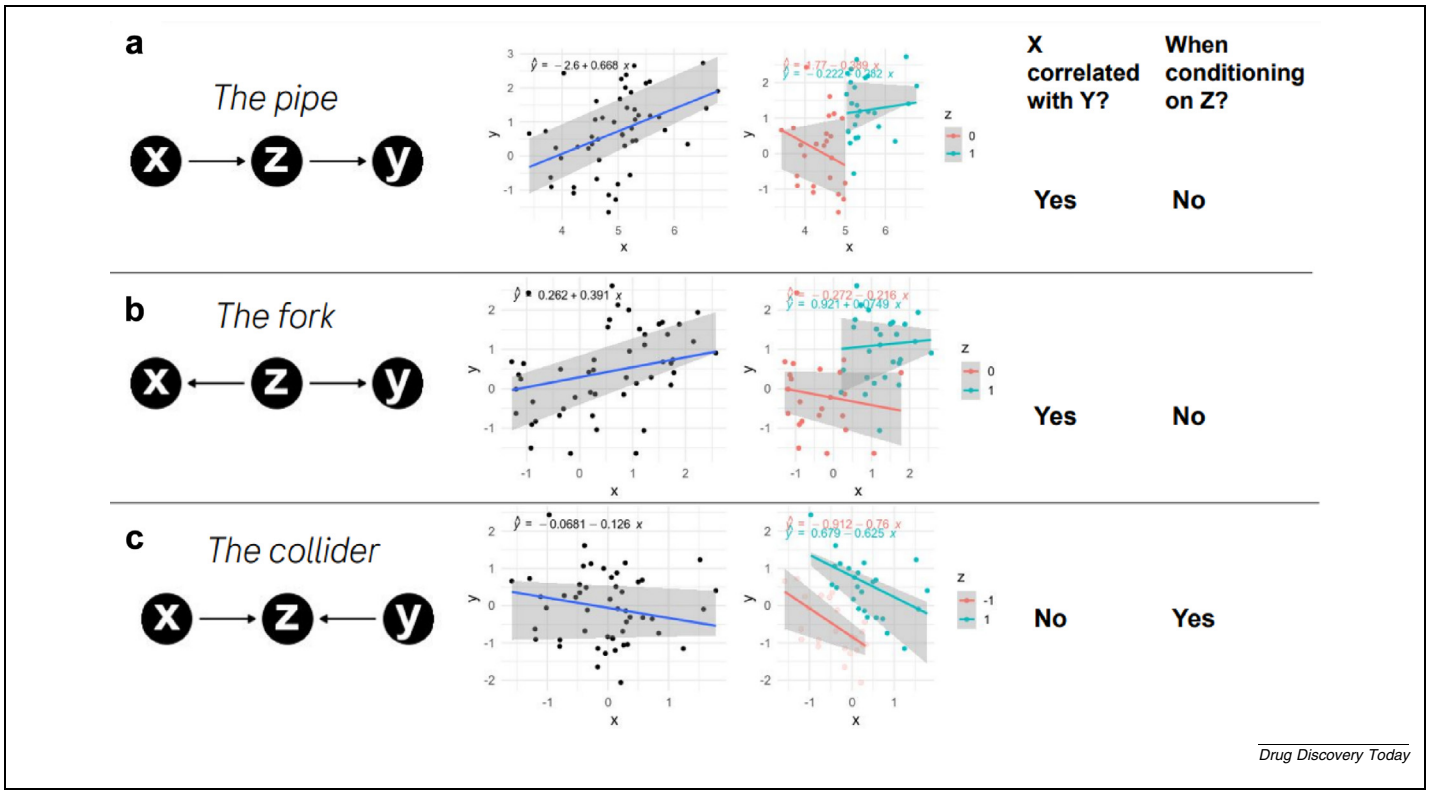


FIGURE 2

Common 3-node structures in DAGs. (a) Left. the pipe structure consists of a chain of three variables. The model was used to generate a simulated dataset of 50 data points. We specified that *x* follows a Gaussian distribution with mean of 5 and standard deviation, *z* takes the value of 0 if $x < 5$ and 1 otherwise, and *y* follows another Gaussian distribution with mean defined by 2*z and standard deviation. Middle. simulated data visualised with scatter plot. Each dot represents one data point. Both plots show *x* on the X-axis and *y* on the Y-axis. The positions of the points are the same in both plots. In the left plot, the regression line (blue) and its confidence interval (grey) are shown for $y \sim x$; in the right plot, the regression lines and confidence intervals are shown dependent on the value of Z. A simple visual assistance is that if *x* and *y* are marginally (mid-left plot) or conditionally (mid-right plot) correlated with each other, then the confidence interval should *not* contain any horizontal grid line. In this case, it is clear that while *x* is correlated with *y*, the correlation is broken if we condition on *z*. This summary is shown on the right panel of the plot. (b) Similar to panel a, with the fork structure on the left. Simulation rules (data points N = 50). *z* follows a Bernoulli distribution with a probability of success of 0.5. Both *x* and *y* follow normal distribution with mean defined by *z* and standard deviation. The interpretation of middle and right panels is comparable to panel a. (c) Similar to panel a and b, with the collider structure on the left. Simulation rules (data points N = 50). Both *x* and *y* follow normal distribution with 0 mean and standard deviation. The value of *z* is 1 if x + y > 0 and −1 otherwise. The interpretation of middle and right panels is comparable to panel a and b. The code to generate both the simulated data and the visualisations is available at https://github.com/Accio/causal_drug_discovery/blob/main/2021-12-CausalSalad.Rmd.

graphical models alone, independent of the choice of the estimator and without access to the data. This means that we can tell whether the question can be answered at all by causal inference *before* we collect or generate any data, which is usually the most time- and labour-consuming step in knowledge acquisition.

In this step, we need to identify the target estimand by specifying the quantity of interest in the DAG model. Usually, we use software to identify whether the estimand can be estimated based on the graph structure. If the estimand is not identifiable, we have little choice other than modifying our model, redefining our question, or admitting that the problem cannot be solved with causal inference. Otherwise, the model is said to be *identifiable*, and we can continue with the estimation step.

A causal model is intrinsically a generative model, namely it can be used to generate simulated data. Once the identifiability is established, a common practice is to simulate data with the model with specified parameters, to run the inference with simulated data, and to examine both whether the simulated data reassemble observed data and whether the estimate is consistent with the input parameters. Inference with simulated data offers us experience with the model, opportunities to identify bugs and flaws of the model before the real run, and confidence in the model and the choice of estimator.

Step 3: Estimation

Next, we collect existing data or generate new data according to the DAG. The importance of quality control of data cannot be overestimated: *garbage in, garbage out*. Data may be filtered so that only the subset that satisfies the DAG structure is used for the estimation.

Once we have decided on the estimand and the data, we can estimate the causal effect by identifying an appropriate estimator and deriving the estimate from the data, which is usually done by software (see Box 2 for popular open-source software packages for causal inference).

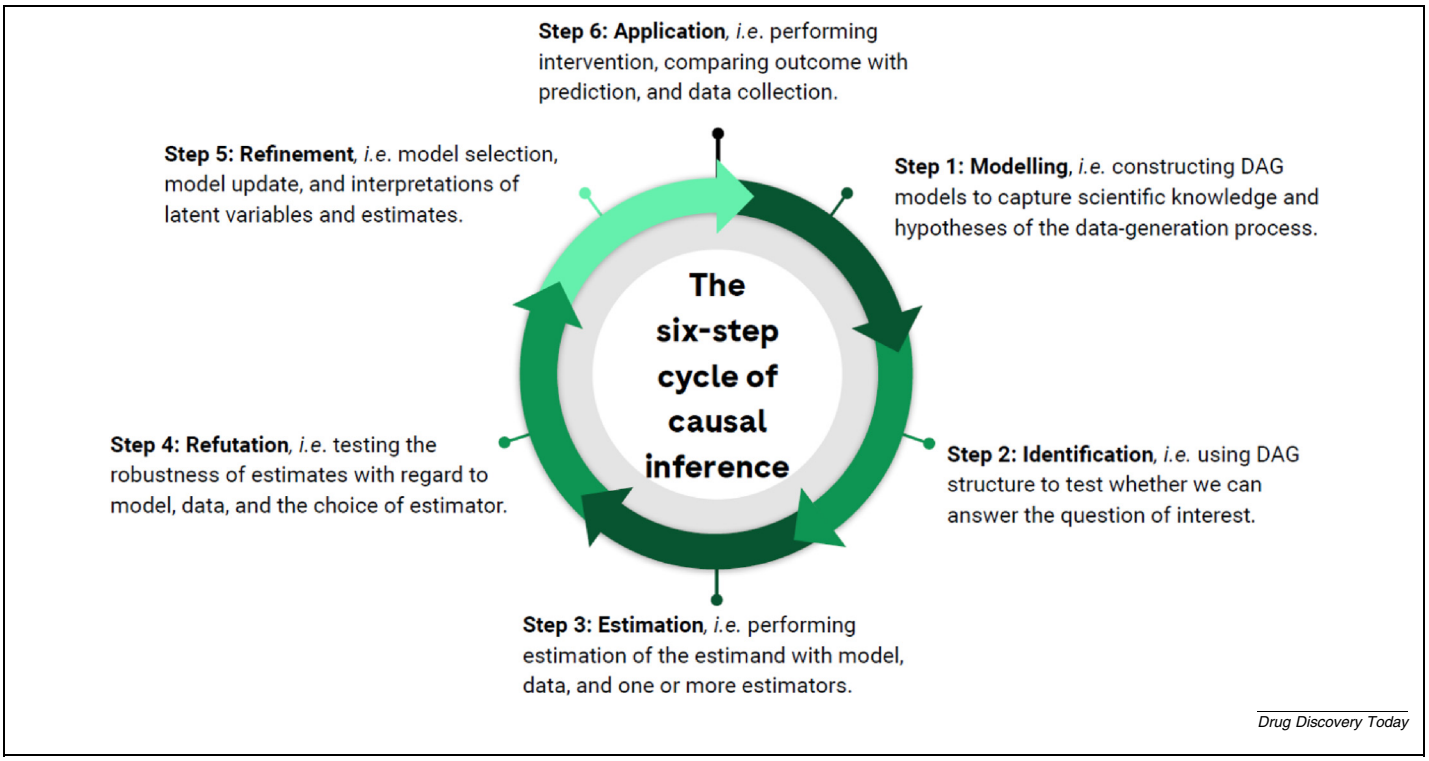


FIGURE 3
A six-step model of causal inference, detailed in Section *A step-by-step guide of causal inference*.

BOX 2 Emerging software tools for causal inference.

- https://github.com/microsoft/dowhy[44]
- Dagitty[150] in R (GNU), https://www.dagitty.net/
- https://github.com/babylonhealth/counterfactual-diagnosis
- https://github.com/lingfeiwang/findr
- https://github.com/causal-machine-learning
- https://github.com/uber/causalml
- https://github.com/microsoft/econml
- https://github.com/google/CausalImpact
- https://github.com/IBM/causallib
- https://github.com/uhlerlab/causaldag
- https://github.com/tlverse/tlverse, with tutorial at https://tlverse.org/tlverse-handbook/.

The choice of estimator depends on the DAG structure and the nature of the data. The popular *DoWhy* package, for instance, offers diverse estimators depending on the graph structure, including (i) (generalised) linear regression; (ii) stratification, matching, or weighting by propensity score, namely the probability of each unit being assigned to a treatment group given a set of observed covariates; (iii) using instrumental variables; and (iv) machine-learning–based estimators implemented in the *EconML* package. Although the choice of appropriate estimators is context- and question-specific, several measures can help us make better choices, especially (i) using simulated data, (ii) learning from case studies and past experience, and (iii) applying refutation techniques, discussed in the next step.

Step 4: Refutation

Although estimation is commonly deemed the most important step in causal inference, in high-stakes applications such as drug discovery and development, we need to test the robustness of our estimates. This is achieved by *refutation*, a collective name of many modelling techniques to test the strength, validity, and our confidence of the estimated causal effect. Commonly used techniques are listed in Supplementary File 1, in the Supplementary material online.

In addition, one can perform data partition or bootstrapping to estimate the variability of the estimates. Furthermore, bespoke refutation techniques can be used to address application-specific questions, such as noncompliance in randomised clinical trials.[44] We refer interested readers to a recent review about these and other techniques known as *sensitivity analysis*.[45]

Statistical causal inference is not the only way to identify causality (Box 1). Alternative evidence from other models, such as results from mechanistic models and results from randomised controlled experiments, should be used to strengthen or challenge our belief in the output of the causal model.

BOX 1 An anarchy of approaches to causality.

Despite a long history of causality research in multiple disciplines, especially statistics, computer science, epidemiology, healthcare, and philosophy of science, there is no consensus way of defining and inferring causality.[28] Impactful approaches for causal inference include the counterfactual outcome framework, the Campbell's framework, Bayesian decision theory, and the directed acyclic graph (DAG) approach.[142] We note that they are not mutually exclusive. In this review, we particularly focus on the counterfactual framework and the DAG approach.

A variety of philosophical and scientific approaches have been developed to investigate causality. Below is a list of approaches that mostly impacted the authors:

- The constant conjunction by David Hume.
- Path analysis pioneered by Sewall Wright in 1920s.[143]
- Potential outcomes proposed by Jerzy Neyman in 1923, which revealed the possibility of achieving causal inference with randomised trials.[144]
- Hill's causality criteria in 1960s.[2(p19)]
- Ignorability proposed by Donald Rubin in 1970s, known as unconfoundedness to epidemiologists and selection on observables by economists, which lead to propensity scores.[18,145,146]
- Bayesian network, DAGs, and do operators proposed between 1980s and 2000s by Judea Pearl.[83,147]
- Representation learning for causal inference.[132]
- Causal machine learning, aiming at individualised causal inference using high-dimensional features to subdivide population, with emerging work in medicine and healthcare.[47,148]
- Learning causal relationships, representing a more substantial challenge than causal effect learning.[24]
- Challenges of estimation from finite sample, see Shipley.[149]

We close this box with a quote from Richard McElreath: 'There is no method for making causal models other than science. There is no method to science other than honest anarchy.'

Step 5: Refinement

Even when the estimated causal effects withstand refutation analysis, they are seldom the sole goal of our investigation. Given that we lack a comprehensive understanding for virtually all biological problems, we often need to *refine* the model, which includes selection from multiple models, addition of new variables, interpretation of latent variables and of estimates, and answering *what if* questions. The refined model can be used, on one hand, to guide design of new experiments to further refute or update the model and, on the other hand, to guide interventions.

Step 6: Application

One of the ultimate goals of causal inference is to perform interventions in the real world. The outcomes can be further analysed in the causal inference framework, therefore closing the loop.

In contrast to a 'hypothesis-free' paradigm, where one expects to learn mechanisms generating the data from data alone, the cycle of *modelling, identification, estimation, refutation, refinement,* and *application* integrates knowledge, hypothesis, and data to address scientific questions. Interdisciplinary teamwork is required to construct explicit models of causality, to collect estimand-specific data, and to identify causal relationships.

Literature review and case studies

Literature review

To gain an overview of applications of causal inference in drug discovery and development, we compiled a list of publications by querying the MEDLINE/PubMed database. We found more than 800 scientific publications (Table S1 in the Supplementary material online) up to September 2022. As a comparison, we also queried publications on machine learning and artificial intelligence and stratified the number of publications by years (Figure 4a). We note two interesting patterns: (i) publications on causal inference in drug discovery have increased almost steadily since 1990, and (ii) the topic receives much less coverage than machine learning and artificial intelligence. Despite the scope of the latter two concepts being admittedly much broader, the patterns underscore the importance of populating the causal mindset and language among practitioners.

We classified the literature by applications and observed that causal inference has been applied throughout the value chain of drug discovery and development (Figure 4b,c). The majority of publications are about analysis of observational studies, especially drug safety, pharmacovigilance, and real-world data. Methodological papers and publications reporting applications for target identification and assessment, clinical trial design and analysis, and biomarker studies follow with distance. Applications in other areas, in particular preclinical and translational research, are still scarce.

Given the broad application of causal inference, and given that applications for clinical data, real-world data, and observational studies have been extensively reviewed elsewhere,[36,37,46,47] we focus below on case studies of causal inference for disease understanding and target identification in translational research.

Translational research includes activities that aim at establishing the causal relationship between drug candidates and disease progression. It has two interlinked components: forward and reverse translation. Forward translation predicts drug dosing and risk/benefit ratio in patients with *in silico* models and data from *in vitro*, *ex vivo*, and *in vivo* animal models and microdosing human studies. Reverse translation informs and improves decision-making in forward translation by analysing data collected in clinical trials, real-world data, as well as data generated with clinical stage or marketed molecules in preclinical models. Forward and reverse translation form a closed loop and complement each other to validate or refute the proposed causal relationship between treatment and patient health status.

Although translational research has many facets, we choose disease understanding and target identification to highlight the impact of causal inference for two main reasons. First, leveraging the causal engine in the target identification and selection phase may have the strongest potential in reducing costs to society of disease and of drug discovery.[48–50] Second, applications in this area involve heterogeneous and high-dimensional data, such as

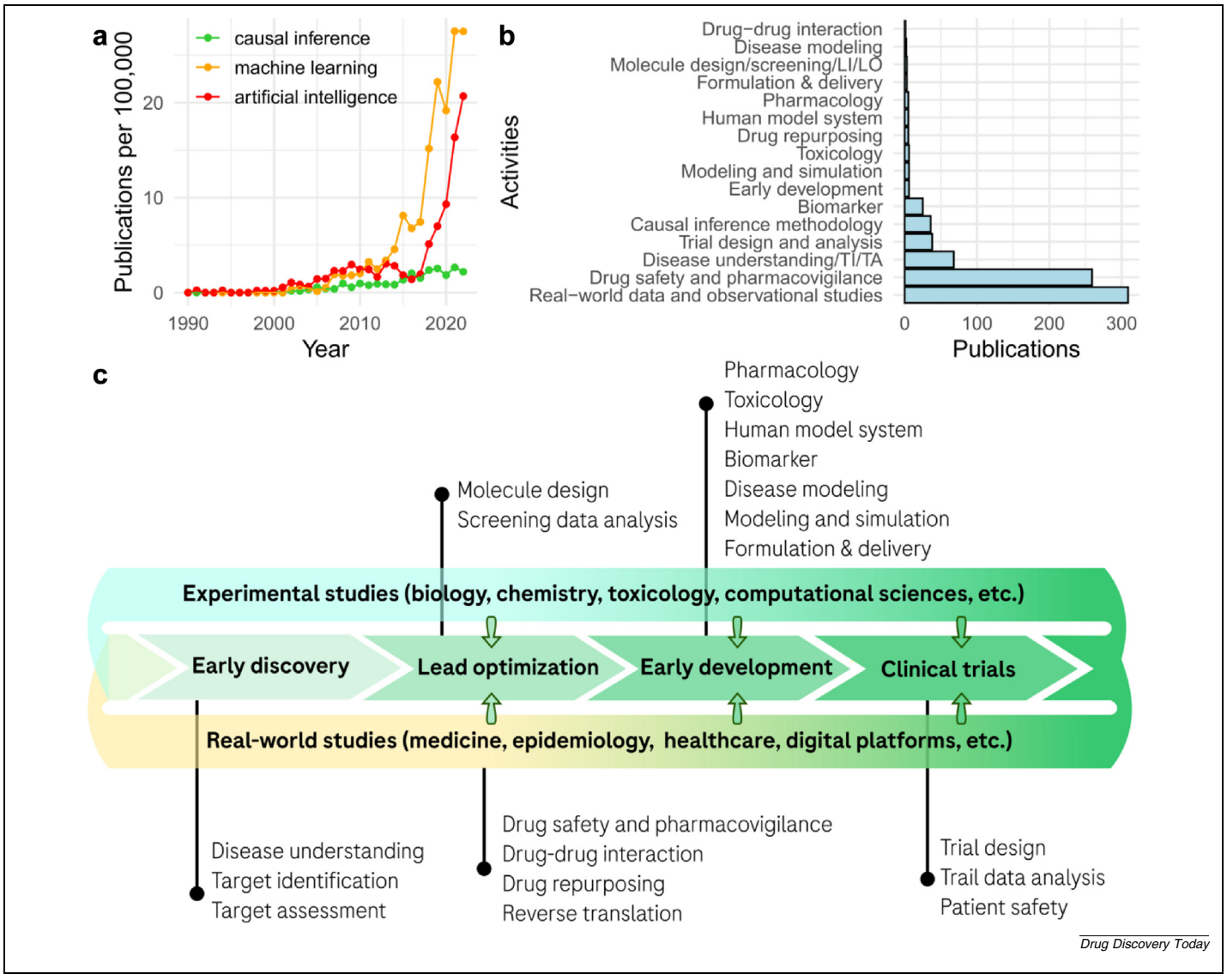


FIGURE 4

Established and emerging applications of causal inference along the value chain of drug discovery and development. (a) Publications indexed by MEDLINE/PubMed with keywords *causal inference* (green), *machine learning* (yellow), and *artificial intelligence* (red). Only peer-reviewed papers relevant for drug discovery and development are included. Numbers of publications are divided by the total number of publications and multiplied by 100,000. (b) Classification of publications on causal inference in drug discovery and development into activity categories. We read the abstract, and if available, the full text, of identified causal-inference publications and manually categorised them. The activities are reversely ordered by the number of total publications. (c) We positioned activities found in the literature review in the context of both forward and reverse translation. In the middle we show a simplified diagram of forward translation consisting of four steps. early discovery, lead optimization, early development, and clinical trials. Above and below we show reverse translation, i.e. analysis of experimental and observational real-world data, and its feedback to forward translation. Applications of causal inference are highlighted besides the lollipops.

EHRs and genomic and other omic data. Although the data type varies in other tasks of translational research, such as risk/benefit assessment, biomarker selection, and patient stratification and enrichment, the concepts, software tools, and practice can be transferred.

Learning causal associations from natural experiments and observational studies

Although target identification was driven by druggability and *in vitro* or animal disease models for a long time, reverse translation is upending this pattern. Causal inference contributes evidence to support or disaffirm drug targets by analysing data from natural experiments and observational studies. The need for causal inference is particularly strong in emerging fields such as microbiome medicine, where high-quality data are relatively scarce and associative analyses are still prevalent.[51] Similarly, causal inference is needed when the study of disease aetiology is challenging, such as in neuroscience. Path analysis, a precursor to and variant of causal inference, has been applied to prioritise or refute targets in neurogenesis, Parkinson's disease, and autism spectrum disorder.[52,53,54,55] See Supplementary File 1 (in the Supplementary material online) for a detailed case study with Alzheimer's disease.

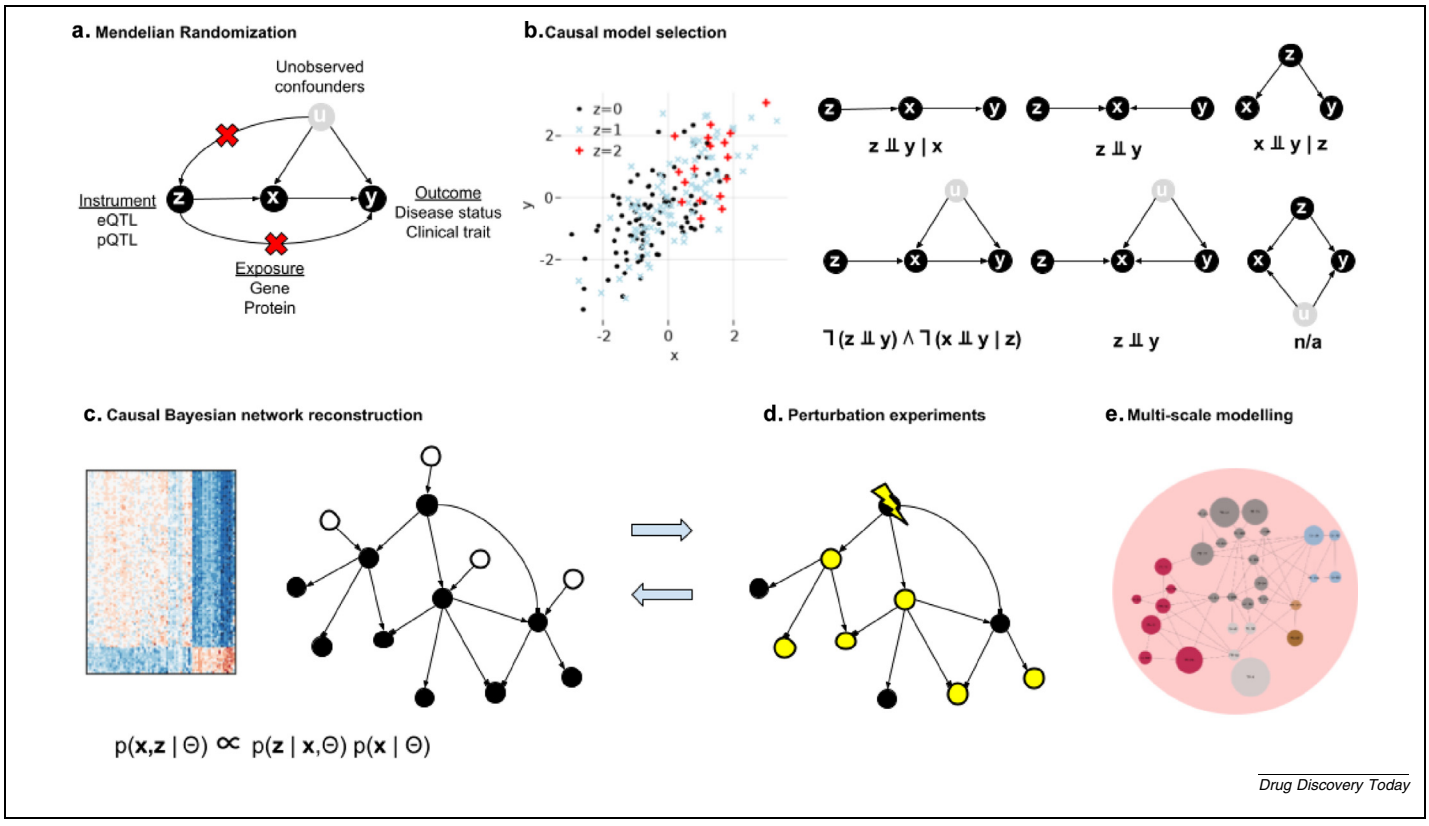


FIGURE 5

Causal inference for disease understanding and target identification. (a) Mendelian Randomization (MR) estimates the causal effect of an exposure ("x", e.g. a gene or protein) on an outcome ("y", e.g. disease status or clinical trait) using a genetic instrument for the exposure ("z", an eQTL or pQTL for a gene or protein exposure). MR is applicable if the instrument is independent of unobserved confounders ("u") and if the outcome is associated with the instrument only through the exposure (no alternative paths) (indicated by red crosses). (b) Simulated scatter plot of two coexpressed genes X and Y, with samples coloured according to the genotype of a genetic marker Z for X. Model selection determines the causal model underlying the data by testing the conditional independencies implied by each causal DAG. Representative causal models without (top) and with (bottom) unobserved confounders are shown. Model selection assumes that the edge Z → X must be included in the model and that Z is independent of any unobserved confounders. (c) When a larger set of coexpressed genes is considered (left), the data is modelled by a causal Bayesian network (right). Black nodes represent genes (x) and white nodes genetic instruments (z) that are used to orient the causal directions of gene-gene edges, and the Bayesian network represents the probability distribution of jointly observing *x* and z given a set of model parameters Θ learned from the training data. (d) Controlled perturbation experiments give direct information about the variables causally downstream of the perturbed node (yellow nodes) and can be used to refute or refine networks reconstructed from observational data. (e) Models inferred at one scale (e.g. cell-type or tissue-specific gene or protein networks) can be integrated into higher-level models. In the figure, nodes represent tissue-specific causal Bayesian networks (node colour, tissue; node size, network size) and edges represent "eigengene" similarities. Panel E obtained from Talukdar, Husain A., et al. Cell systems 2.3 (2016). 196–208 under CC-BY-NC-ND licence.

Application of causal inference on multiomic data (genomics, transcriptomics, proteomics, metabolomics) leads us to identify causal genes, pathways, and gene regulatory networks that have a direct effect on disease states.[56–59] Genome-wide association studies (GWASs) have mapped the genetic architecture of common diseases in humans[60] and have identified genetic loci that affect drug response and susceptibility to adverse drug reactions.[61] Drugs targeting genetically supported targets were reported to be twice as likely to be successful as other drugs.[62,63] Exploiting GWAS data in drug discovery is challenging, however, because there are generally hundreds to thousands of genetic risk variants, mostly lying in noncoding genomic regions and each typically contributing only a small amount of risk.

Mendelian randomization (MR)[64,65] is a statistical approach that uses genetic variants as randomised instruments ('natural experiments') to identify causal associations between heritable traits. It is based on the fact that genotypes are independently assorted and randomly distributed in a population by Mendel's laws, or in a less strict manner, independent of each other outside haplotype blocks,[66] and not affected by environmental or genetic confounders that affect both the exposure (target manipulation) and the outcome (disease state). If it can be assumed that a genetic locus affects the outcome only through the exposure, then the causal effect of the exposure on the outcome can be derived from their relative associations to the genetic locus, which acts as an instrumental variable[67] (Figure 5a).

By integrating GWAS with studies that map genetic effects on the transcriptome or proteome (expression or protein quantitative trait loci; eQTLs or pQTLs, respectively), MR can identify causal associations between molecular traits and disease.[68] When applied to proteomic data, inferred causal associations between proteins and disease can suggest drug repurposing opportunities[69] or new candidate drug targets.[70,71] An advantage of MR is that it can be performed using summary statistics alone.[72] A lim-

itation is that MR tests effects of molecular traits on diseases one by one and that no molecular pathways are reconstructed. This is important because the candidate causal factors may act indirectly or redundantly with other factors, or the candidate factors may not be druggable themselves but affect intermediate druggable targets.

When genetics and multiomic data from individuals in a segregating population are merged, comparison of independent, causative, and reactive DAGs can orient the direction of causality among QTLs and pairs of correlated genes or proteins[73,74] (Figure 5b). Statistical model selection compares the conditional independence relationships implied by each DAG and uses maximum likelihood to select the model that best fits the available data. Statistical significance is estimated by expressing the conditional independence implications of the tested DAGs as combinations of likelihood ratio tests. Either their maximum *P*-value is treated as an omnibus hypothesis test[75–77], or false discovery estimation is used to express the results of the individual tests as probabilities of their null or alternative hypothesis being true.[78,79] Accounting for hidden confounders requires additional assumptions on the genetic instrument, usually satisfied if the instrument is a *cis*-acting eQTL or pQTL for the 'exposure' variable, and involves a trade-off between a high false-negative rate or an increased false-positive rate.[79,80]

Bayesian networks combine the results of pairwise causal inferences into a causal network model of the underlying biological system. Bayesian networks are probabilistic machine-learning models for expressing both prior knowledge and inferred conditional independence and causal relationships among variables.[81–83] A Bayesian gene network consists of a DAG which connects regulatory genes to their targets[84,85] (Figure 5c). The structure of a Bayesian network is inferred from the data using score-based or constraint-based methods,[82] where the results of pairwise causal inference tests are used to constrain the DAG search space.[84–90] This is done by expressing the joint likelihood of observing both data types as a standard likelihood term for observing gene expression data, given the DAG, and an additional likelihood term for observing the genotype data, given the expression data and causal interactions implied by the DAG[84,85,88–90] (Figure 5c).

To bridge from the molecular scale to disease states, clinical phenotypes are not usually modelled as nodes in the causal network. Instead, genome-wide omic data are first partitioned into coexpression modules associated with clinical traits,[91–94] and causal networks are learned for each module separately and connected in a higher-level Bayesian or coexpression network where each node is a module 'eigengene,' a representation of the module's expression profile by its first principal component[95] (Figure 5e). This higher-level network can model tissue-specific and intertissue communication processes that are the result of the 'collective' states of molecular networks and that are more proximal to physiological disease processes.[58,93,94,96]

As in almost all other areas of genomics, single-cell technologies are creating new opportunities for learning causal disease mechanisms from population-based data.[97] Cost-effective strategies for generating single-cell RNA-sequencing data across individuals have been developed,[98,99] and the first large-scale population genetics studies have been published recently.[100–104] A particularly attractive application will be to exploit the fact that allele-specific expression can be mapped in single cells.[105] This will allow us to quantify eQTL effect sizes[106] at the level of an individual instead of a population, which will lead to a better understanding of the variation of causal effects across individuals.

Learning causal associations from controlled perturbation experiments

Having considered the impact of causal inference on reverse translation, now we turn to its impact on forward translation by analysing controlled perturbation experiments (Figure 5d).

For target-based discovery programs, canonical gene knockout or perturbation experiments in model systems establish causality in a manner similar to randomised controlled trials. When these experiments are combined with genome-wide readouts, they provide insights into the direct and indirect causal targets of the perturbed gene. Systematic small-molecule and perturbation screens using RNAi technology followed by transcriptome sequencing have been performed in cancer cell lines.[107] CRISPR-based technology combined with single-cell RNA sequencing is now rapidly expanding the scope of conducting genome-wide perturbation screens in human cells.[108–112]

Causal inference from perturbation experiments allows us to disentangle direct from indirect effects and predict the outcome of a new intervention or perturbation. Nested effect models are probabilistic graphical models to infer a genetic hierarchy from the nested structure of observed perturbation effects.[113,114] A more formal approach uses causal DAGs and *do*-calculus to model causal effects from perturbation screens.[115,116]

Pioneers and experts in phenotypic drug discovery have long realised the importance of establishing causal relationships. Moffat *et al.*[117] proposed a model of *chain of translatability*, namely from the chemical matter to the assay phenotype, to the preclinical disease model, and finally to human disease. We note the multiscale nature of this chain of translatability and the resemblance of challenges faced by target-based programs and those faced by phenotypic programs, except for the target identification and assessment question. A key open question is whether causal inference, and in particular causal discovery techniques, can help phenotypic programs identify molecular targets and mode of action.

A fundamental issue with forward translation is that cellular or animal models differ from human physiology. Therefore, a critical additional request is that the identified link must be persistent in humans in the form of *in vitro–in vivo* translation, governed by physiological constraints, and between-species translation, governed by evolutionary forces.

In order to decide whether experimental findings can be extrapolated across domains that differ both in their distributions and in their inherent causal characteristics, we need prior scientific knowledge about the invariance of certain mechanisms, represented in 'selection diagrams,' DAGs in which the causal mechanisms are explicitly encoded and in which differences in populations are represented as local modifications of those mechanisms.[118,119] Switching between populations corresponds to conditioning on different values of the variables that locate the mechanisms where structural discrepancies between

the two populations are suspected to take place. The transportability problem can then be formulated and decided using *do*-calculus and graphical criteria on the selection diagrams.

Translational study is undergoing a profound change. Although animal models are traditionally used as a bridge between *in vitro* experiments and studies with humans, the field is witnessing strong interest and investment in human model systems such as stem cell–derived cell lines, primary cells, organoids, organ-on-a-chip, or other microphysiological systems. They are currently being tested for many purposes, especially efficacy and safety assessment, prediction and modelling of pharmacokinetic and pharmacodynamic parameters, biomarker identification, and disease modelling.[120]

From the point of view of causal inference, the key question is how much the causal model relevant for the disease and the drug candidate in human systems is conserved in model systems, let it be *in vitro*, *in vivo*, or human model systems. To answer this question, it is necessary to establish causal models with each model system and to compare the models between the systems. So far, the translational value of most human model systems is judged by *looking alike*, that is, to assess similarity in morphology and/or omic profiles with regard to primary human material. In order to mimic the causal relationship, more work is required to assess these systems by *functioning alike*, that is, to assess their response upon perturbation and conditions that mimic disease aetiology and development.[121,122]

Discussion

This review addresses three questions: (i) When do we need a causal model? (ii) How can causal inference be performed? (iii) Which activities in drug discovery have been or will be empowered by causal inference? With the 6C model of correlation and causation in mind, we argued that causal models are essential for predictive modelling of the outcome of intervention, for answering *what if* questions, and for understanding disease biology and why drugs work (or do not). We introduced causal inference as a six-step iterative process with hands-on examples in the complementary tutorial. Finally, we performed an extensive literature review and discussed recent application of causal inference in translational research.

As discussed in the introduction, causality is multiscale. This renders causal inference a suitable tool to investigate interactions between drugs and human biological systems, which are *per se* multiscale.[123] Leveraging advances in statistical modelling and mechanism-based multiscale modelling has the potential to transform heterogeneous data into unified knowledge, which helps us understand the impact of genetic, epigenetic, and environmental factors on drug action.[124]

An emerging consensus is that the quality of *in silico* model predictions outweighs the speed of prediction, and qualitative statements that enable scientists to focus on promising targets and regions of chemical space are often more impactful than quantitative predictions for decision-making.[125,126] In line with this, we foresee three key opportunities for causal inference: (i) high-quality prediction for predictive modelling of interventional outcomes, including out-of-distribution predictions[125,126]; (ii) individualised causal inference and estimation of heterogeneous treatment effects[127]; and (iii) causally and counterfactually based decision-making.[20,128]

Besides the opportunities, we foresee three major challenges to be overcome by researchers and practitioners of causal inference in the coming decade: (i) the methodology of causal inference needs to be further developed to embrace the reality and complexity of biology; (ii) the community needs to share and have open access to high-quality data and models; and (iii) we need a change in both language and mindset.

The methodology of causal inference warrants further research. DAGs, though very powerful, have intrinsic limitations when used to model biological systems. Edges in DAGs must be acyclic; that is, we are never trapped in a closed loop by following the directions of edges. DAGs are therefore not suited for analysing reciprocal causal relationships, though they are prevalent in biology across scales from molecular interactions[129] to consciousness.[130] To model systems with such relationships, one can either take advantage of longitudinal data to identify causal relationships[131], or model the net output of nodes involved in reciprocal relationships as a variable instead of modelling individual nodes, or employ computational models other than DAGs, such as graph neural networks and other representation learning techniques.[132] Both the theory and the software required for performing such analyses need further research and development. Despite recent progress discussed above,[26] it is still challenging to perform causal inference with high-dimensional data, for systems with multiple complex traits, that is, many factors with small effect sizes,[133–135] and for complex adaptive systems where feedback loops abound.[136–138]

Open contribution and access to improve biological models is essential for the community to refine and leverage causal models. Resources of causal models are emerging quickly,[139] and community-wide efforts such as DREAM challenges have been hosted to identify causal relationships from *omic* datasets.[140] Nevertheless, the causal language remains foreign to many researchers; we do not always distinguish correlation from causation explicitly, and, in many cases, scientific findings are not represented as causal models. Although it remains a complex and unsolved problem of how to populate the causal language and how to encourage researchers to share data and causal models, we are cautiously optimistic about further development because of the potential gain. The Linus's Law formulated by Eric S. Raymond,[141] *Given enough eyeballs, all bugs are shallow,* is a motivation for all of us to adopt the causal language and share the causal models: To share and to refine causal models in a community is to understand how biology and drugs work collectively.

Several reasons may make practitioners of drug discovery and development, including many colleagues we have interviewed, wary of causal inference. We need the front-up payment of bringing up knowledge and hypothesis, often exposing our naivety and lack of knowledge. In a real-world setting, setting up a model is particularly challenging because we often have incomplete information and inaccurate knowledge about human biology and the pharmacology and toxicology of drugs, not to mention the abundance of invisible or unquantifiable quantities and events that affect the efficacy and safety of drugs. We have to give up the hope about 'learning from data alone' and 'hypothesis-free approaches.' We know that machine-learning

models such as deep neural networks and graphical neural networks can learn patterns if they are trained with enough data. Isn't it better to accumulate more data and let the machine find out the causality? Why do we stick to and propose causal inference?

We welcome these questions and doubts. They lead us to believe that we need a change in both language and mindset to think and talk about causality. We argue that (i) causality exists beyond data, and it is not possible to learn causality from data alone without prior knowledge and hypotheses; and (ii) the central task of drug discovery and development is to discover and validate multiscale causal relationships. The true causal network, in contrast to causal DAG models, can contain feedback loops and therefore cannot be subject to the analysis introduced here as a whole. However, the network can be separated into smaller subnetworks, each verified by the six-step cycle of causal inference, and the causality propagates between the scales. The network connects the drug molecule with disease outcome via its interactions with biological molecules. The interactions emerge as pharmacology and toxicology on cellular, organ-and-system levels. Furthermore, the network can be enriched with individual traits such as genetics, medical history, lifestyle, and environmental factors to allow population analysis.

Concluding remarks

Causal inference offers a principled approach to model- and data-driven predictive modelling and decision-making. It complements statistical and mechanistic models to empower an understandable synthesis and integration of knowledge and data across scales. We foresee that further development of its methodology, open-accessible causal models and high-quality data, and a broader adoption of the causal language and mindset will fuel both forward and reverse translation. In short, causal inference empowers us to seek and to prove causality with new drugs.

Funding

T.M. is supported by grants from the Research Council of Norway (grant numbers 312045 and 331725). J.D.Z.'s work is funded by F. Hoffmann-La Roche Ltd.

Declarations of interest

No interests are declared.

Data availability

We have shared the data as supplementary information.

Acknowledgements

Both authors thank colleagues, students, and team members for discussion, input, and feedback for the manuscript. J.D.Z. wishes to explicitly thank input from the members of the Predictive Modelling and Data Analytics (PMDA) chapter, especially Zhiwen Jiang, Milad Adibi, Juliane Siebourg-Polster, Tony Kam-Thong, Balazs Banfai, Sarah Morillo Leonardo, Martin Ebeling, Cheikh Diack, Anja Lieberherr, Benjamin Ribba, and Julia Pletz, as well as visitors of the weekly *Bioinformatics Club*. We also thank Kaspar Rufibach, Denis Engemann, Gonzalo Christian Duran-Pacheco, and Nicoleta Spinu for sharing their insights, criticism, and feedback.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.drudis.2023.103737.

References

1. Beebee H, Hitchcock C, Menzies PC, eds. *The Oxford Handbook of Causation*. Oxford University Press; 2009.
2. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295–300.
3. Williamson J. Establishing causal claims in medicine. *Int Stud Philos Sci*. 2019;32:33–61. https://doi.org/10.1080/02698595.2019.1630927.
4. Stewart AK. How thalidomide works against cancer. *Science*. 2014;343:256–257. https://doi.org/10.1126/science.1249543.
5. Vargesson N. Thalidomide-induced teratogenesis: history and mechanisms. *Birth Defects Res*. 2015;105:140–156. https://doi.org/10.1002/bdrc.21096.
6. Ito T et al. Identification of a primary target of thalidomide teratogenicity. *Science*. 2010;327:1345–1350. https://doi.org/10.1126/science.1177319.
7. Békés M, Langley DR, Crews CM. PROTAC targeted protein degraders: the past is prologue. *Nat Rev Drug Discov*. 2022;21:181–200. https://doi.org/10.1038/s41573-021-00371-6.
8. Deshaies RJ. Multispecific drugs herald a new era of biopharmaceutical innovation. *Nature*. 2020;580:329–338. https://doi.org/10.1038/s41586-020-2168-1.
9. Tummino TA et al. Drug-induced phospholipidosis confounds drug repurposing for SARS-CoV-2. *Science*. 2021;373:541–547. https://doi.org/10.1126/science.abi4708.
10. Fischer H, Kansy M, Bur D. CAFCA: a novel tool for the calculation of amphiphilic properties of charged drug molecules Accessed 23 October 2019.. *Chim Int J Chem*. 2000 https://www.ingentaconnect.com/content/scs/chimia/2000/00000054/00000011/art00006.
11. Berkson J. Limitations of the application of fourfold table analysis to hospital data (reprint). *Int J Epidemiol*. 2014;43:511–515. https://doi.org/10.1093/ije/dyu022.
12. Sackett DL. Bias in analytic research. *J Chronic Dis*. 1979;32:51–63. https://doi.org/10.1016/0021-9681(79)90012-2.
13. Reichenbach H, Reichenbach M. *The Direction of Time*. Dover; 1999.
14. Hitchcock C, Rédei M. Reichenbach's common cause principle Accessed 31 July 2023.. In: Zalta EN, ed. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University; 2021.
15. Shmueli G. To explain or to predict? *Stat Sci*. 2010;25:289–310. https://doi.org/10.1214/10-STS330.
16. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1:206–215. https://doi.org/10.1038/s42256-019-0048-x.
17. Gupta R, Bugide S, Wang B, Green MR, Johnson DB, Wajapeyee N. Loss of BOP1 confers resistance to BRAF kinase inhibitors in melanoma by activating MAP kinase pathway. *Proc Natl Acad Sci*. 2019;116:4583–4591. https://doi.org/10.1073/pnas.1821889116.
18. Imbens GW, Rubin DB. *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge University Press; 2015.
19. Cinelli C, Forney A, Pearl J. *A crash course in good and bad controls*. *Soc Sci Res Netw*. 2020. https://doi.org/10.2139/ssrn.3689437.
20. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Method*. 2008;8:70. https://doi.org/10.1186/1471-2288-8-70.
21. Suzuki E, Shinozaki T, Yamamoto E. Causal diagrams: pitfalls and tips. *J Epidemiol*. 2020;30:153–162. https://doi.org/10.2188/jea.JE20190192.
22. Laubach ZM, Murray EJ, Hoke KL, Safran RJ, Perng W. A biologist's guide to model selection and causal inference. *Proc R Soc B Biol Sci*. 1943;2021:20202815. https://doi.org/10.1098/rspb.2020.2815.

23. Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. *Appl Inform*. 2016;3:3. https://doi.org/10.1186/s40535-016-0018-x.
24. Guo R, Cheng L, Li J, Hahn PR, Liu H. A survey of learning causality with data: problems and methods. *ACM Comput Surv*. 2020;53:75:1–75:37. https://doi.org/10.1145/3397269.
25. Luo Y, Peng J, Ma J. When causal inference meets deep learning. *Nat Mach Intell*. 2020;2:426–427. https://doi.org/10.1038/s42256-020-0218-x.
26. Lopez R, Hütter JC, Pritchard JK, Regev A. *Large-Scale Differentiable Causal Discovery of Factor Graphs*. Published online 15 June 2022. https://doi.org/10.48550/arXiv.2206.07824.
27. Vasudevan RK, Ziatdinov M, Vlcek L, Kalinin SV. Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality. *npj Comput Mater*. 2021;7:1–6. https://doi.org/10.1038/s41524-020-00487-0.
28. McElreath R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. CRC Press: Taylor and Francis; 2020.
29. National Research Council (U.S.), National Research Council (U.S.), National Academies Press (U.S.), eds. *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press; 2010.
30. Little RJ. Missing data assumptions. *Annu Rev Stat Appl*. 2021;8:89–107. https://doi.org/10.1146/annurev-statistics-040720-031104.
31. Bornkamp B, Bermann G. Estimating the treatment effect in a subgroup defined by an early post-baseline biomarker measurement in randomized clinical trials with time-to-event endpoint. *Stat Biopharm Res*. 2020;12:19–28. https://doi.org/10.1080/19466315.2019.1575280.
32. Bornkamp B et al. Principal stratum strategy: potential role in drug development. *Pharm Stat*. 2021;20:737–751. https://doi.org/10.1002/pst.2104.
33. Heinzmann D, Kong S. Principal stratum strategy to investigate anti-drug antibody impact on outcome in randomized controlled trials. Presented at: https://bbs.ceb-institute.org/wp-content/uploads/2020/09/06-Heinzmann.pdf
34. Kong S, Heinzmann D, Lauer S, Lu T. Weighted approach for estimating effects in principal strata with missing data for a categorical post-baseline variable in randomized controlled trials. *ArXiv210104263 Stat*. Published online January 11, 2021. Accessed 18 November 2021. https://arxiv.org/abs/2101.04263
35. Eichler H et al. 'Threshold-crossing': a useful way to establish the counterfactual in clinical trials? *Clin Pharmacol Ther*. 2016;100:699–712. https://doi.org/10.1002/cpt.515.
36. Ho M et al. The current landscape in biostatistics of real-world data and evidence: causal inference frameworks for study design and analysis. *Stat Biopharm Res*. 2021;15:43–56. https://doi.org/10.1080/19466315.2021.1883475.
37. Hampson LV, Degtyarev E, Tang (Sammi) R, Lin J, Rufibach K, Zheng C. Comment on 'Biostatistical considerations when using RWD and RWE in clinical studies for regulatory purposes: a landscape assessment'. *Stat Biopharm Res*. 2021;15:23–26. https://doi.org/10.1080/19466315.2021.1994459.
38. Lazic SE. *Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility*. Cambridge University Press; 2016.
39. Zhang W, Ramezani R, Naeim A. Causal Inference in medicine and in health policy, a summary. *arXiv:210504655v4*. Published online 10 May 2021. https://doi.org/10.48550/arXiv.2105.04655
40. Rosenbaum PR. *Design of Observational Studies*. Springer; 2010.
41. Scutari M, Denis JB. *Bayesian Networks: With Examples in R*. 2nd ed. CRC Press; 2021.
42. Fabrigar LR, Wegener DT. *Exploratory Factor Analysis*. Oxford University Press; 2012.
43. Lopez R, Gayoso A, Yosef N. Enhancing scientific discoveries in molecular biology with deep generative models. *Mol Syst Biol*. 2020;16. https://doi.org/10.15252/msb.20199198 e9198.
44. Sharma A, Kiciman E. DoWhy: An end-to-end library for causal inference. *ArXiv201104216 Cs Econ Stat*. Published online 9 November 2020. Accessed 29 September 2021. https://arxiv.org/abs/2011.04216
45. Razavi S et al. The future of sensitivity analysis: an essential discipline for systems modeling and policy support. *Environ Model Softw*. 2021;137. https://doi.org/10.1016/j.envsoft.2020.104954 104954.
46. Lipkovich I, Ratitch B, Mallinckrodt CH. Causal inference and estimands in clinical trials. *Stat Biopharm Res*. 2020;12:54–67. https://doi.org/10.1080/19466315.2019.1697739.
47. Prosperi M et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell*. 2020;2:369–375. https://doi.org/10.1038/s42256-020-0197-y.
48. Plenge RM. Disciplined approach to drug discovery and early development. *Sci Transl Med*. 2016;8. https://doi.org/10.1126/scitranslmed.aaf2608 349ps15.
49. Plenge RM. Priority index for human genetics and drug discovery. *Nat Genet*. 2019;51:1073–1075. https://doi.org/10.1038/s41588-019-0460-5.
50. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov*. 2013;12:581–594. https://doi.org/10.1038/nrd4051.
51. Lv BM, Quan Y, Zhang HY. Causal inference in microbiome medicine: principles and applications. *Trends Microbiol*. 2021. https://doi.org/10.1016/j.tim.2021.03.015.
52. Lazic SE. Using causal models to distinguish between neurogenesis-dependent and -independent effects on behaviour. *J R Soc Interface*. 2012;9:907–917. https://doi.org/10.1098/rsif.2011.0510.
53. Gündner AL et al. Path mediation analysis reveals GBA impacts Lewy body disease status by increasing α-synuclein levels. *Neurobiol Dis*. 2019;121:205–213. https://doi.org/10.1016/j.nbd.2018.09.015.
54. Yap CX et al. Autism-related dietary preferences mediate autism-gut microbiome associations. *Cell*. 2021. https://doi.org/10.1016/j.cell.2021.10.015.
55. Tosun D et al. Contribution of Alzheimer's biomarkers and risk factors to cognitive impairment and decline across the Alzheimer's disease continuum. *Alzheimers Dement*. 2021;18:1370–1382. https://doi.org/10.1002/alz.12480.
56. Rockman MV. Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature*. 2008;456:738–744.
57. Schadt EE, Friend SH, Shaywitz DA. A network view of disease and compound screening. *Nat Rev Drug Disc*. 2009;8:286–295.
58. Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature*. 2009;461:218–223.
59. Li Y, Tesson BM, Churchill GA, Jansen RC. Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends Genet*. 2010;26:493–498.
60. Uffelmann E et al. Genome-wide association studies. *Nat Rev Methods Primer*. 2021;1:59. https://doi.org/10.1038/s43586-021-00056-9.
61. Daly AK. Genome-wide association studies in pharmacogenomics. *Nat Rev Genet*. 2010;11:241–246. https://doi.org/10.1038/nrg2751.
62. Nelson MR et al. The support of human genetic evidence for approved drug indications. *Nat Genet*. 2015;47:856–860. https://doi.org/10.1038/ng.3314.
63. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet*. 2019;15:e1008489.
64. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32:1–22.
65. Evans DM, Davey SG. Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Annu Rev Genomics Hum Genet*. 2015;16:327–350. https://doi.org/10.1146/annurev-genom-090314-050016.
66. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. 2008;9:477–485. https://doi.org/10.1038/nrg2361.
67. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res*. 2007;16:309–330.
68. Zhu Z et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. 2016;48:481–487. https://doi.org/10.1038/ng.3538.
69. Reay WR, Cairns MJ. Advancing the use of genome-wide association studies for drug repurposing. *Nat Rev Genet*. 2021;22:658–671. https://doi.org/10.1038/s41576-021-00387-z.
70. Bretherick AD et al. Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. *PLoS Genet*. 2020;16:e1008785.
71. Pietzner M et al. Mapping the proteo-genomic convergence of human diseases. *Science*. 2021;374. https://doi.org/10.1126/science.abj1541 eabj1541.
72. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol*. 2013;37:658–665. https://doi.org/10.1002/gepi.21758.
73. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet*. 2001;17:388–391.
74. Schadt EE et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. 2005;37:710–717.
75. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. *BMC Genet*. 2009;10:23.
76. Millstein J, Chen GK, Breton CV. cit: hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics*. 2016;32:2364–2365.

77. Badsha M, Fu AQ. Learning causal biological networks with the principle of Mendelian randomization. *Front Genet*. 2019;10:460.
78. Chen LS, Emmert-Streib F, Storey JD. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol*. 2007;8:R219.
79. Wang L, Michoel T. Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PLoS Comput Biol*. 2017;13 e1005703.
80. Ludl AA, Michoel T. Comparison between instrumental variable and mediation-based methods for reconstructing causal gene networks in yeast. *Mol Omics*. 2021;17:241–251.
81. Friedman N. Inferring cellular networks using probabilistic graphical models. *Science*. 2004;308:799–805.
82. Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press; 2009.
83. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press; 2009.
84. Neto EC, Keller MP, Attie AD, Yandell BS. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann Appl Stat*. 2010;4:320.
85. Hageman RS, Leduc MS, Korstanje R, Paigen B, Churchill GA. A Bayesian framework for inference of the genotype–phenotype map for segregating populations. *Genetics*. 2011;187:1163–1170.
86. Zhu J et al. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res*. 2004;105:363–374.
87. Zhu J et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet*. 2008;40:854–861.
88. Tasaki S, Sauerwine B, Hoff B, Toyoshiba H, Gaiteri C, Neto EC. Bayesian network reconstruction using systems genetics data: comparison of MCMC methods. *Genetics*. 2015;199:973–989.
89. Wang L, Audenaert P, Michoel T. High-dimensional Bayesian network inference from systems genetics data using genetic node ordering. *Front Genet*. 2019;10:1196. https://doi.org/10.3389/fgene.2019.01196.
90. Scutari M, Howell P, Balding DJ, Mackay I. Multiple quantitative trait analysis using Bayesian networks. *Genetics*. 2014;198:129–137.
91. Zhang B et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*. 2013;153:707–720.
92. Beckmann ND et al. Multiscale causal networks identify VGF as a key regulator of Alzheimer's disease. *Nat Commun*. 2020;11:3942. https://doi.org/10.1038/s41467-020-17405-z.
93. Talukdar H et al. Cross-tissue regulatory gene networks in coronary artery disease. *Cell Syst*. 2016;2:196–208.
94. Koplev S et al. A mechanistic framework for cardiometabolic and coronary artery diseases. *Nat Cardiovasc Res*. 2022;1:85–100. https://doi.org/10.1038/s44161-021-00009-1.
95. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol*. 2007;1:54.
96. Seldin MM et al. A strategy for discovery of endocrine interactions with application to whole-body metabolism. *Cell Metab*. 2018;27:1138–1155.
97. van der Wijst MGP, de Vries DH, Brugge H, Westra HJ, Franke L. An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome Med*. 2018;10:96. https://doi.org/10.1186/s13073-018-0608-4.
98. Kang HM et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*. 2018;36:89–94. https://doi.org/10.1038/nbt.4042.
99. Mandric I et al. Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat Commun*. 2020;11:5504. https://doi.org/10.1038/s41467-020-19365-w.
100. Bryois J et al. Cell-type specific cis-eQTLs in eight brain cell-types identifies novel risk genes for human brain disorders. *Nat Neurosci*. 2022;25:1104–1112.
101. LifeLines Cohort Study et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat Genet*. 2018;50:493–497. https://doi.org/10.1038/s41588-018-0089-9.
102. Neavin D et al. Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biol*. 2021;22:76. https://doi.org/10.1186/s13059-021-02293-3.
103. Yazar S et al. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science*. 2022;376. https://doi.org/10.1126/science.abf3041 eabf3041.
104. Nathan A et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*. 2022;606:120–128. https://doi.org/10.1038/s41586-022-04713-1.
105. Jiang Y, Zhang NR, Li M. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol*. 2017;18:74. https://doi.org/10.1186/s13059-017-1200-8.
106. Mohammadi P, Castel SE, Brown AA, Lappalainen T. Quantifying the regulatory effect size of *cis*-acting genetic variation using allelic fold change. *Genome Res*. 2017;27:1872–1884. https://doi.org/10.1101/gr.216747.116.
107. Subramanian A et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;171:1437–1452.
108. Dixit A et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*. 2016;167:1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.
109. Adamson B et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*. 2016;167:1867–1882.e21. https://doi.org/10.1016/j.cell.2016.11.048.
110. Jaitin DA et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*. 2016;167:1883–1896.e15. https://doi.org/10.1016/j.cell.2016.11.039.
111. Datlinger P et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods*. 2017;14:297–301. https://doi.org/10.1038/nmeth.4177.
112. Schraivogel D et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat Methods*. 2020;17:629–635. https://doi.org/10.1038/s41592-020-0837-5.
113. Markowetz F, Kostka D, Troyanskaya OG, Spang R. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*. 2007;23:i305–i312. https://doi.org/10.1093/bioinformatics/btm178.
114. Tresch A, Markowetz F. Structure learning in nested effects models. *Stat Appl Genet Mol Biol*. 2008;7. https://doi.org/10.2202/1544-6115.1332.
115. Meinshausen N, Hauser A, Mooij JM, Peters J, Versteeg P, Bühlmann P. Methods for causal inference from gene perturbation experiments and validation. *Proc Natl Acad Sci*. 2016;113:7361–7368. https://doi.org/10.1073/pnas.1510493113.
116. Wang Y, Solus L, Yang KD, Uhler C. Permutation-based causal inference algorithms with interventions. *ArXiv170510220 Stat*. Published online 4 November 2017. Accessed 16 December 2021. https://arxiv.org/abs/1705.10220
117. Moffat JG. Turning the light on in the phenotypic drug discovery black box. *Cell Chem Biol*. 2017;24:545–547. https://doi.org/10.1016/j.chembiol.2017.05.005.
118. Pearl J, Bareinboim E. External validity: from Do-Calculus to transportability across populations. *Stat Sci*. 2014;29. https://doi.org/10.1214/14-STS486.
119. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci*. 2016;113:7345–7352. https://doi.org/10.1073/pnas.1510507113.
120. Ewart L, Roth A. Opportunities and challenges with microphysiological systems: a pharma end-user perspective. *Nat Rev Drug Discov*. 2021;20:327–328. https://doi.org/10.1038/d41573-020-00030-2.
121. Lukonin I et al. Phenotypic landscape of intestinal organoid regeneration. *Nature*. 2020;586:275–280. https://doi.org/10.1038/s41586-020-2776-9.
122. Fleck JS et al. Inferring and perturbing cell fate regulomes in human cerebral organoids 2021:2021.08.24.457460. *bioRxiv*. 2021. https://doi.org/10.1101/2021.08.24.457460.
123. Zhang JD, Sach-Peltason L, Kramer C, Wang K, Ebeling M. Multiscale modelling of drug mechanism and safety. *Drug Discov Today*. 2020;25:519–534. https://doi.org/10.1016/j.drudis.2019.12.009.
124. Xie L et al. Towards structural systems pharmacology to study complex diseases and personalized medicine. *PLoS Comput Biol*. 2014;10:e1003554.
125. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet. *Drug Discov Today*. 2020. https://doi.org/10.1016/j.drudis.2020.12.009.
126. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov Today*. 2021;26:1040–1052. https://doi.org/10.1016/j.drudis.2020.11.037.
127. Bica I, Alaa AM, Lambert C, van der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharmacol Ther*. 2021;109:87–100. https://doi.org/10.1002/cpt.1907.

128. Jung J, Concannon C, Shroff R, Goel S, Goldstein DG. Simple rules for complex decisions. *ArXiv170204690 Stat*. Published online April 2, 2017. Accessed 26 November 2019. https://arxiv.org/abs/1702.04690
129. Alon U. *An introduction to systems biology: design principles of biological circuits*. Chapman & Hall/CRC; 2007.
130. Bartels A. Consciousness: what is the role of prefrontal cortex? *Curr Biol*. 2021;31:R853–R856. https://doi.org/10.1016/j.cub.2021.05.012.
131. Shojaie A, Fox EB. Granger causality: a review and recent advances. *Annu Rev Stat Appl*. 2022;9:289–319. https://doi.org/10.1146/annurev-statistics-040120-010930.
132. Schölkopf B, et al. Towards causal representation learning. Published online 22 February 2021. Accessed 5 August 2021. https://arxiv.org/abs/2102.11107v1
133. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol*. 2020;38:56–65. https://doi.org/10.1038/s41587-019-0315-8.
134. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169:1177–1186. https://doi.org/10.1016/j.cell.2017.05.038.
135. Liu X, Li YI, Pritchard JK. Trans effects on gene expression can drive omnigenic inheritance. *Cell*. 2019;177:1022–1034.e6. https://doi.org/10.1016/j.cell.2019.04.014.
136. Peters J, Bauer S, Pfister N. Causal models for dynamical systems. *ArXiv200106208 Math Stat*. Published online 17 January 2020. Accessed 18 November 2021. https://arxiv.org/abs/2001.06208
137. Peters J, Janzing D, Schölkopf B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press; 2017.
138. Holland JH. Complex adaptive systems. *Daedalus*. 1992;121:17–30.
139. Status of causality in biological databases: data resources and data retrieval possibilities to support logical modeling. Brief Bioinform 2021;22:bbaa390. https://academic.oup.com/bib/article/22/4/bbaa390/6055722
140. Cell-to-cell and type-to-type heterogeneity of signaling networks: insights from the crowd. Mol Syst Biol 2021;17(10):e10402. https://doi.org/10.15252/msb.202110402
141. Raymond ES. *The cathedral & the bazaar: musings on linux and open source by an accidental revolutionary*. 1st ed. O'Reilly; 1999.
142. Thoemmes F. Ausgewählte Kausalitätstheorien im Vergleich. *Gesundheitswesen*. 2011;73:880–883.
143. Wright S. The relative importance of heredity and environment in determining the piebald pattern of Guinea-Pigs. *Proc Natl Acad Sci*. 1920;6:320–332. https://doi.org/10.1073/pnas.6.6.320.
144. Splawa-Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat Sci*. 1990;5:465–472. https://doi.org/10.1214/ss/1177012031.
145. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91:444–455. https://doi.org/10.1080/01621459.1996.10476902.
146. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688–701. https://doi.org/10.1037/h0037350.
147. Pearl J, Glymour M, Jewell NP. *Causal inference in statistics: a primer*. Wiley; 2016.
148. Pearl J. The seven tools of causal inference, with reflections on machine learning. *Commun ACM*. 2019;62:54–60. https://doi.org/10.1145/3241036.
149. Shipley B. *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press; 2016.
150. Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *Int J Epidemiol*. 2016;45:1887–1894. https://doi.org/10.1093/ije/dyw341.