Physics Reports 1044 (2023) 1–68

Contents lists available at ScienceDirect

Physics Reports

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

Discovering causal relations and equations from data

Gustau Camps-Valls [a,*], Andreas Gerhardus [b,1], Urmi Ninad [c,b,1], Gherardo Varando [a,1], Georg Martius [d,e], Emili Balaguer-Ballester [f,g], Ricardo Vinuesa [h], Emiliano Diaz [a], Laure Zanna [i], Jakob Runge [b,c]

[a] Universitat de València, València, Spain
[b] German Aerospace Center, Jena, Germany
[c] Technische Universität Berlin, Berlin, Germany
[d] University of Tübingen, Tübingen, Germany
[e] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[f] Bournemouth University, Bournemouth, UK
[g] Medical Faculty Mannheim and Heidelberg University, Mannheim, Germany
[h] FLOW, Engineering Mechanics, KTH Royal Institute of Technology, Stockholm, Sweden
[i] New York University, NY, USA

ARTICLE INFO

Article history:
Received 16 October 2023
Accepted 16 October 2023
Available online 7 November 2023
Editor: Procaccia

Keywords:
Causal inference
Causal discovery
Complex systems
Nonlinear dynamics
Equation discovery
Knowledge discovery
Understanding
Artificial intelligence
Neuroscience
Climate science

ABSTRACT

Physics is a field of science that has traditionally used the scientific method to answer questions about why natural phenomena occur and to make testable models that explain the phenomena. Discovering equations, laws, and principles that are invariant, robust, and causal has been fundamental in physical sciences throughout the centuries. Discoveries emerge from observing the world and, when possible, performing interventions on the system under study. With the advent of big data and data-driven methods, the fields of causal and equation discovery have developed and accelerated progress in computer science, physics, statistics, philosophy, and many applied fields. This paper reviews the concepts, methods, and relevant works on causal and equation discovery in the broad field of physics and outlines the most important challenges and promising future lines of research. We also provide a taxonomy for data-driven causal and equation discovery, point out connections, and showcase comprehensive case studies in Earth and climate sciences, fluid dynamics and mechanics, and the neurosciences. This review demonstrates that discovering fundamental laws and causal relations by observing natural phenomena is revolutionised with the efficient exploitation of observational data and simulations, modern machine learning algorithms and the combination with domain knowledge. Exciting times are ahead with many challenges and opportunities to improve our understanding of complex systems.

Contents

* Correspondence to: Image Processing Laboratory (IPL), E4 building - 4th floor, Parc Científic Universitat de València, C/ Cat. Agustín Escardino Benlloch, 9. 46980 Paterna (València), Spain.
E-mail address: gustau.camps@uv.es (G. Camps-Valls).

[1] These authors contributed equally.

https://doi.org/10.1016/j.physrep.2023.10.005

"As in Mathematics, so in Natural Philosophy, the Investigation of difficult Things by the Method of Analysis ought ever to precede the Method of Composition. This Analysis consists in making Experiments and Observations, and in drawing general Conclusions from them by Induction, and admitting of no Objections against the Conclusions, but such as are taken from Experiments, or other certain Truths ... By this way of Analysis, we may proceed from Compounds to Ingredients, and from Motions to the Forces producing them; and in general, from Effects to their Causes, and particular Causes to more general ones, till the Argument end in the most general. This is the Method of Analysis" (Newton, 1718).

1. Introduction

This paper reviews the recent advances in *causal discovery* and *equation discovery* from data. Both problems are conundrums for scientists and philosophers of science. After all, Science is about studying, discovering, and understanding the structure and behaviour of the physical and natural world through observation and experimentation. Understanding the system's structure involves performing interventions on the systems to evaluate their responses. However, interventional experiments are often not feasible for economic or ethical reasons, so relying on observations, simulations, and domain knowledge must be exploited. In recent decades, discovering causal relations and underlying governing laws from data have emerged as exciting fields of research that promise advancing science.

1.1. Understanding in the physical sciences

A pertinent question arises here; what is understanding? Understanding is the ability to comprehend and make sense of processes to gain a deeper knowledge of a system. Understanding involves analysing information, making (eventually causal) connections, and coming to conclusions. Whether the conclusions should be falsifiable has been the subject of active discussion in the Philosophy of Science [1,2]. We aim to *understand* complex systems by following the *scientific method*; make an observation, ask a scientific question, form a hypothesis, theory, model, or explanation of the phenomena, and make predictions, which are ultimately tested and whose results are used to make new hypotheses or predictions (see Fig. 1).

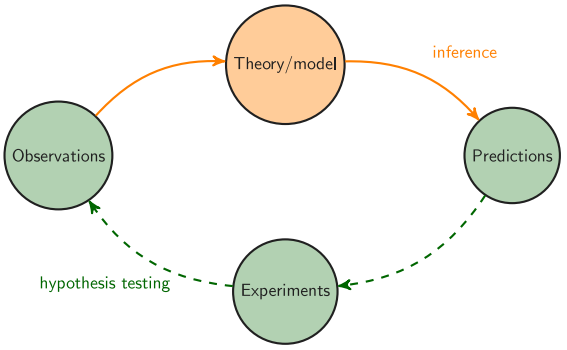


Fig. 1: Standard loop in understanding complex systems following the standard scientific method. Understanding involves experimentation by refining a descriptive mechanistic model. The initial model hypothesis is tested in practice and, through experiments, yields observations that are confronted with the model's predictions. The unexplained processes are then used to improve the model's misspecification and predictions.

Understanding involves reasoning and thinking critically about a subject, a system's behaviour, or a problem. Understanding and explaining how a system works is more complicated than making predictions about the system's behaviour. The Oracle of Delphi gave accurate predictions of the future and the optimal course of action, but the lack of understanding frequently led to disaster. True understanding is about making (truly accurate) predictions and, more importantly, gaining knowledge of the causal chain. Science generally aims to answer causal questions, infer causal relations, and attain mathematical models (mainly laws and equations) that work well in most situations, explain the system and underlying processes, and are invariant across space and time. Without it, we cannot predict the consequences of our actions (*interventions*) or analyse when, where and why things went wrong (*counterfactuals*) [3].

Yet, what do we want to understand? And how do we generally do it? In physical sciences, one typically analyses phenomena and instantiations of the physical world, uses observations, and refines and tests models. For learning about the system, one aims to (1) characterise its complexity in terms of trajectories, persistence, stability and collapse, bifurcations and viability boundaries [4–7], (2) obtain explanatory and causal models of their behaviour [8–10], and (3) discover and formalise general laws, governing equations, and parameterisations [11–13]. These three components allow us to advance science and technology. However, in many systems, governing equations and causal relations are (partially) unknown, and recourse to first principles is untenable. Resorting to algorithms that can discover laws, governing equations, and causal relations from data may thus constitute a paradigm shift that promises to accelerate science.

1.2. Scientific discovery

Scientific discovery is the process or product of successful scientific inquiry. Objects of discovery can be things, events, processes, causes, properties, theories, hypotheses, and their characteristics. Philosophical discussions of scientific discovery vary widely in scope and definition, from the narrowest sense of a "eureka moment" to the broadest sense of a "successful scientific endeavour". The utilisation of datasets to create and test new hypotheses in philosophical

discourse has led to a multifaceted and intricate discussion regarding the precise definition and potential misuse of the term "discovery".

Human nature aims to discover. Always. Since the Bronze Age.[2] Generations have created and discovered new principles, techniques, and operations through millennia. Right after the Neolithic Revolution, the world stopped except for some remarkable technological advances, like the invention of the water wheel (476–221 BC) and the windmill (ca 644 BC). Romans were amazed by stories of what Archimedes (287–212 BC) had been able to do. But, bold as it may sound, one may claim that modern science was invented between 1572 when Tycho Brahe saw a *nova* or new star, and 1704 when Newton published his Opticks [14]. What happened in that period prepared humanity and scientists for a New Era: a research program endorsed with a scientific method that allowed scrutinising new theories and validating or refuting hypotheses and models of the world, and all that in the light of evidence and observations. After fitting many ovoids to observational data, Kepler discovered the laws of planetary motion (1609) and needed four years to discover Mars' orbit was an ellipse. The scientific method was slow but sure. Galilei discovered the law of falling bodies (1638) by dropping two cannonballs of different masses from the tower of Pisa and measuring the effect of mass on the fall rate to the ground. And in 1662, Boyle discovered the law of ideal gases. Only ten years later, in 1672, Newton discovered that white light is a mixture of distinct coloured rays, and in 1687 he formulated the classical mathematical description of the fundamental force of universal gravitation and the three physical laws of motion. The triumph of Newtonianism marks the end of the beginning of scientific discovery.

The history of science is a long and complex narrative punctuated by moments of major scientific revolutions [15]. Kuhn identified a general pattern: A discovery is not a simple act but an extended, complex process that culminates in paradigm changes. The first scientific revolution occurred in the 16th and 17th centuries when the Copernican revolution challenged the traditional Ptolemaic view of the universe [16]. The next major scientific revolution was the Enlightenment during the 18th century, which saw the emergence of the scientific method and the development of modern physics and chemistry, from formulating the laws of motion to discovering electricity. This period also saw the emergence of scientific societies, which helped to propagate and popularise scientific ideas. The 19th century saw the emergence of the theory of evolution, which revolutionised the field of biology [17]. This revolution was followed by the rise of modern genetics, which further expanded our understanding of the evolution of life [18]. The 20th century saw the emergence of the quantum revolution, which revolutionised our understanding of the physical world as it could not be fully explained by classical physics [19]. Revolutions happen only gradually, as it takes time for the scientific community to recognise *"both that something is and what it is"* [15]. Eventually, a new paradigm becomes established, and the strange phenomena become the expected phenomena.

The idea that there is such a thing as 'the Scientific Revolution' and that it took place in the 17th century is thus a fairly recent one [20]; some have argued that it can be seen as the construction of intellectuals looking back from the 20th century [14]. Like the term 'Industrial Revolution', the idea of a scientific revolution brings problems of the multiplication (how many scientific revolutions?) and periodisation (how often?). Some philosophers of science have argued for continuity, others have sought multiple revolutions: the Darwinian revolution, the Quantum revolution, the DNA revolution, and so on, while others claim that the real Scientific Revolution came in the 19th century when science and technology married. Recently, we are witnessing the so-called 4th Industrial Revolution, which conceptualises rapid change to technology, industries, and societal patterns due to increasing interconnectivity, smart automation and the amalgamation of artificial intelligence and automated machines. Yet, is it only a technological revolution, or can machines discover and explain new science? Are we facing the emergence of machine discovery of science?

1.3. Knowledge discovery from data

1.3.1. Discoverability and heuristic strategies

The questions of what and how phenomena and mechanisms can be discovered have been the subject of intense research and philosophical discussion. In the philosophy of science, *discoverability* is the concept that scientific knowledge must be discoverable and verifiable [1,21–24]. This means that hypotheses or theories must be supported by evidence and based on empirical observations and data. Furthermore, scientific knowledge must be available for anyone to see and verify, ensuring that it is not biased or limited to a specific group of people.

Recent advancements in the philosophy of science have seen a revival of interest in *heuristic strategies* to discover knowledge; these strategies are seen as problem-solving activities, whereby a discovery is a solution to a problem. Heuristics-based discovery methodologies are neither completely subjective and intuitive nor algorithmic or formalisable. This view has shifted the scientific researcher from being viewed as a 'puzzle solver' to a 'problem solver' and 'decision maker' in complex, variable, and changing environments [25]. In this paper, we will review mathematical models that address equation discovery and causal discovery by generally formalising the problems as concrete statistical inference tasks; regression, conditional dependence or density estimation.

2 https://en.wikipedia.org/wiki/Timeline_of_scientific_discoveries.

Table 1
Simple taxonomy of models and their level.
Source: Figure partly reproduced from [9].

Model/Level	i.i.d.	Distribution shifts	Counterfactuals	Physical insight	Data-driven
3 - Mechanistic	✓	✓	✓	✓	✓
2 - Structural causal	✓	✓	✗	?	?
1 - Statistical/ML	✓	✗	✗	✗	?

1.3.2. Modern approaches to data-driven discovery

Observational discovery relies on modelling. Yet, what types of models? Table 1, cf. [26], gives a simple categorisation of models from mechanistic/physical models based on first principles and (rigid) equations and laws but with desirable properties of interpretability, invariance and robustness to distribution shifts, to purely statistical (machine learning) models that excel in prediction and are learned from data. In the middle, we have structural causal models, which can answer counterfactual questions but do not necessarily capture physical knowledge. All three models are used in quantitative data-driven science and map to different levels of discovery: learning statistical associations in data streams, identifying causal relations between variables, and discovering equations from data. Note the resemblance to the Ladder of Causation proposed by Pearl and Mackenzie [27] with three rugs: association, intervention, and counterfactual. Discovering causal and physical laws from observations is a paradigm shift in AI and can impact the physical sciences and other disciplines. The fields of discovery of scientific knowledge and causal models of scientific phenomena are intertwined and tightly connected: our scientific endeavour is constantly challenged with causal questions, robust model building, intervention analysis, and hypothesis testing. The fields also share important theoretical challenges, where generalisability, compressibility, robustness, invariance, and extrapolation come into play.

Level 1 — Learning statistical associations. The most rudimentary approach to building association links from multivariate time series data involves computing pairwise Pearson's correlations or mutual information, which capture relationships between variables at lag zero. Networks derived from these measures have found applications in numerous scientific and engineering domains, including climate network analysis [28], financial market network analysis [29], and brain network analysis [30,31]. However, mutual information-based associations at lag zero cannot be interpreted directionally since the term "information" implies a lack of directionality. Other regions or variables may influence the nodes under investigation, or the association could be due to a common driving process. Lagged association measures are commonly used to account for directional links and quantify the time lag of associations.

Lagged correlation analysis has a long history in climate research [32] and neuroscience [33,34], with the delay at the maximum of the cross-correlation function being used to interpret the delay of the underlying physical mechanism coupling two processes. Other lagged measures of association, such as mutual information [35], have been proposed to determine lags in nonlinear processes. In addition to analysing time lags, the magnitude of the cross-correlation is often used as a measure of the impact of one process on another or as a measure of the strength of an association. This aligns with the statistical interpretation of the square of correlation as the proportion of variance in one process that can be linearly represented by another [36,37].

However, relying solely on association measures, even those that account for lags and nonlinearity, cannot uncover directionality, detect the delay of the underlying mechanism, or provide a physically or causally interpretable estimate. The widespread use and abuse of association measures in engineering and science throughout the 20th century have impeded the exploration and development of meaningful causation measures and hindered the discovery of new and alternative explanatory laws from data.

Level 2 — Learning causal relations from observations. A fundamental objective in the scientific enterprise is understanding the causes behind the phenomena we observe [8,9]. This is particularly challenging in disciplines dealing with complex dynamical systems, where experimental interventions are expensive, unethical, or practically impossible. In some fields (e.g., climate sciences, economics, cardiology, and neurosciences), the current alternative is to rely on computationally expensive simulation experiments. Still, those do not adequately represent all relevant physical processes involved. At the same time, a rapidly increasing amount of time series data is generated by observations and also models. How can we use this wealth of data to gain new insights into our fundamental understanding of these systems?

In recent years, rapid progress has been made in computer science, physics, statistics, philosophy, and applied fields to infer and quantify potential causal dependencies from data without intervening in the systems. Although the truism that correlation does not imply causation holds, the key idea shared by several approaches follows Reichenbach's common cause principle [38]: if variables are dependent, then they are either causal to each other (in either direction) or driven by a common driver. To estimate causal relationships among variables, different methods take different, partially strong,

assumptions. Granger [35] addressed this question quantitatively using prediction. At the same time, in the last decades, several complementary concepts emerged, from nonlinear dynamics [39] based on attractor reconstruction to computer science exploiting statistical independence relations in the data [8,40]. More recently, statistics and machine learning research utilised the framework of structural causal models (SCMs) [9] for this purpose. Causal inference from data is becoming a mature scientific approach [8].

Causal inference strives to discover the system's causal structure and quantify causal effects by combining domain knowledge, ML models, and data [9,27,41–43]. Causal inference can leverage observational or (interventional) model output data [41] to learn, understand and evaluate the plausibility of specific causal relations among the considered variables. Causal inference is becoming a mature field of science. Today, many methods and tools are available to address challenges in complex systems [39,44] and many other fields. Causality is pivotal not only for a better academic understanding of processes in science but also for more robust forecasts, attributing the causes of events, and improving the physics embedded in physics models. Many fields of science and engineering are using causal inference/discovery methods, from Earth and climate sciences [10,45–51], neurosciences [52–54], social sciences [55,56], health and epidemiology [57–59], or economics [60,61].

Level 3 — Equation discovery in physical systems. The scientific enterprise distinctly differs from other intellectual endeavours by relying on formal theories, laws, and models to explain and predict observations and using such observations to construct, revise, and evaluate its formal statements [22–24]. Many of these activities have been studied by philosophers of science for over a century. The Logic of Science [1] (or justification) aims to characterise how observational data, simulations, and experiments can collectively support or refute laws, models, or theories.

A common claim was that scientific discovery requires some 'creative spark', which cannot be analysed rationally or logically [62]. Popper, Hempel, and many other philosophers of science maintained that the discovery process was inherently irrational and beyond any formal understanding. The key insight came from Simon [1,62,63], who proposed that scientific discovery, rather than *"depending on some unknown mystical ability, is a variety of problem-solving that involves searching through a space of problem states generated by applying mental operators and guided by heuristics to make the search tractable."* Such observation established the first heuristic programming methods of hypothesis (model) search to automate the creative process and law discovery. Discovering numeric laws from data has been approached by many authors in the past using grammars, logic rules, propositional bases, entailment, and genetic algorithms, to name a few [22,64–75]. Later, [76] proposed an automated algorithm to discover Hamiltonians, Lagrangians, and other geometric and momentum conservation laws without prior knowledge of physics, kinematics, or geometry. A new field was born; learning explicit mathematical laws from observations, which was often referred to as *equation discovery* or *data-driven system identification* [77].

Inspired by earlier work on the DENDRAL system [78], which inferred structural models of organic molecules from their mass spectra, the community developed different systems that created models of other scientific phenomena (e.g., [79]). The field was named computational scientific discovery, and the challenge of automating it has been approached by many researchers since then [80–85]. Efforts in this paradigm differ from mainstream work in machine learning by producing scientific formalisms [23,86,87], ranging from componential models in particle physics [88] to reaction pathways in chemistry and to regulatory models in genetics [89]. Reviews have been edited by Shrager and Langley [90], Dzeroski and Todorovski [91] and Simidjievski et al. [77].

Recently, the field has been approached by scientists in AI, functional analysis and mathematical operators, nonlinear control, and system identification. Modern approaches that we will review in this paper consider: Automated reverse engineering of nonlinear dynamical systems [82], sparse-promoting solutions that identify parsimonious models of nonlinear dynamics; e.g. relevance vector machine, a sparse Bayesian regression method [92] or the SINDy method [12,93], which has been combined with deep neural networks [94], reduced-order models [95] and Koopman operators based on kernel theory and autoencoders [96–100], differentiable networks that can learn the true underlying equation and extrapolate to unseen domains [101], a constrained symbolic regression methodology, named AI Feynman, that enforces desirable constraints in equation learning (compositionality, units, separability, symmetry, smoothness) [102], genetic programming to distil laws of physics [103], or a transformer-based architecture using massive pretraining can predict formulas from data [104].

A relevant challenge in this context is the discovery of state variables from experimental/observational data. Almost in all equation and causal discovery algorithms, the variables are given or assumed, which is impossible when trying to understand new or highly complex systems. Some approaches exist in the literature based on the combination of learning compact and expressive feature representations, manifold learning for the determination of the intrinsic dimensionality of the system representation coordinates, and SINDy as a regulariser that enforces dynamic equations in the state space [13,93,105,106]. These approaches are somewhat related to the discovery of causal relations under noise and latent functions, which is also an active field of research [107–109].

1.3.3. AI for scientific discovery

The debate about AI-based theories of scientific discovery has been ongoing for decades, beginning with whether computers can devise new concepts or merely process the concepts already included in a given computer language.

However, the discussion has been revived with the development of new computational tools for data analysis. It is now largely uncontroversial that machine learning tools can aid discovery, though there is still debate about whether they generate new knowledge or merely speed up data processing. Moreover, there is the question of whether data-intensive science fundamentally differs from traditional research and the ethical implications of "superhuman AI". Philosophers have also focused on the opacity of machine learning, asking whether we can say that humans and machines are "co-developers" of knowledge. Ultimately, the debate about AI-based theories of scientific discovery is still ongoing, with researchers considering both the potential benefits and ethical implications of such tools.

The fields of equation discovery and causal inference that we will review in this paper promise to shed light on the previous fundamental questions: what can an algorithm learn and discover, what can AI explain, and what new science may emerge through a collaborative AI-machine dialogue about science? An integrative approach seems necessary, where domain experts, data and machine work together in a data-driven framework that formulates and answer causal questions and discover new laws [8]. At a more general level, it becomes pertinent to ask if machines can start a new scientific revolution and even if AI itself is the ultimate scientific revolution [110–112].

"With the advent of data-driven methods that learn patterns and relations from data, the tedious human endeavour of scientific discovery (laws, equations and causes of phenomena) is being revolutionised... and accelerated in many fields."

1.4. Outline

This paper aims to review the most important concepts, methods, and previous works on causal inference and discovery in the physical sciences. We use statistical learning techniques to discover causal relations, physical laws, and governing equations from data. Sections 2 and 3 present general frameworks and taxonomies for causal discovery and learning physical laws from data, respectively. Both sections categorise the field, reviewing concepts and methods, their specific characteristics, challenges, and opportunities in the physical sciences. Section 4 provides examples of causal discovery and equation discovery in a wide range of fields of the physical sciences: dynamical systems, neuroscience, classical and quantum systems, fluid mechanics, geosciences and climate sciences. We pay attention to how causality concepts and methods can improve our knowledge of a given physical system from observations. Section 5 outlines the most promising future lines of research in this area of study at the intersection of machine learning and nonlinear physical processes.

2. Causal discovery in the physical sciences

Causal discovery, see for example [9,43] for extended expositions of the topic, has become increasingly popular in the last years as a tool to discover the underlying causal structure of physical systems [10]. There is an abundance and ever-growing number of methods designed to work under different assumptions and tackle other use cases. This section reviews several methods for causal discovery, focusing on methods for time series and their potential use in physical sciences. To this end, in Section 2.1, we first provide a taxonomy for many available causal discovery methods. Section 2.2 discusses causal discovery's challenges in real-world applications. We conclude in Section 2.3 by discussing opportunities for applications of causal discovery in the physical sciences. We would also like to point to other reviews focusing on causal discovery of time series [10,113,114]. In addition, Runge et al. [115] provides a shorter accessible summary of methods for causal discovery and causal effect estimation with practical case studies to illustrate typical challenges, such as contemporaneous causation, hidden confounding and non-stationarity.

2.1. A taxonomy of causal discovery methods

In this section, we structure the zoo of existing causal discovery methods to guide method users in finding a method suitable for their application and guide method developers in identifying open challenges. To this end, we summarise the central formal aspects of the graphical-model-based causal inference framework in Section 2.1.1. Then, in Section 2.1.2, we discuss several characteristics (hereafter referred to as "axes") by which methods can be conceptually distinguished. Lastly, in Section 2.1.3, we present an extensive (but not exhaustive) list of causal discovery methods and characterise these methods according to previously introduced axes.

2.1.1. Preliminaries

At the heart of the graphical-model-based causal inference framework are *structural causal models (SCMs)*, e.g. [3,9,116]. An SCM serves as a causal model for the data-generating process and specifies how the system reacts to *interventions*, that is, to idealised experimental manipulations that deliberately hold fixed a subset of the system's variables while not perturbing the system in any other way.

An SCM for a system described by the set of variables $\mathbf{V} = \{V^1, \ldots, V^n\}$ consists of n so-called *structural assignments*

$$V^i := f^i(pa^i, \epsilon^i) \quad \text{with } 1 \leq i \leq n\,, \tag{1}$$

together with a product distribution $p_\epsilon(\epsilon^1, \ldots, \epsilon^n) = p^1_\epsilon(\epsilon_1) \cdot \ldots \cdot p^n_\epsilon(\epsilon_n)$ of the random variables ϵ^i. Formally, the f^i are measurable functions that depend non-trivially on all of their input arguments, and the $pa^i \subseteq \mathbf{V} \setminus \{V^i\}$ are subsets

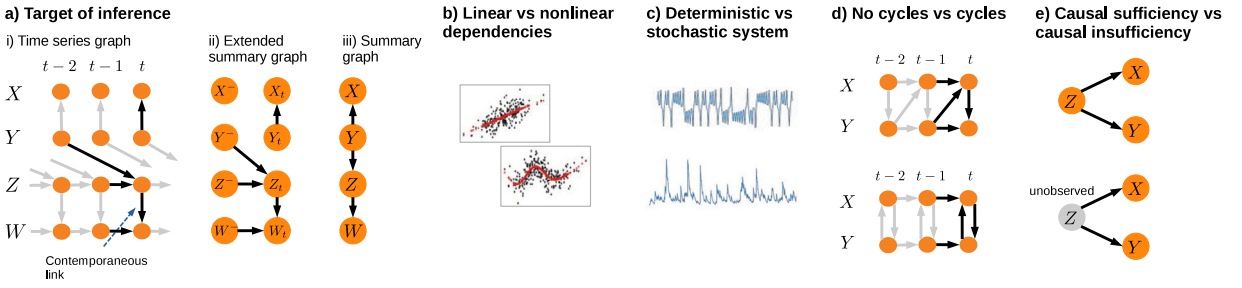


Fig. 2. Illustration of some axes along which causal discovery methods are categorised in this review; see references in the main text to the respective subparts of the figure.

of the system variables $V^1, \ldots, V^n$. The functions f^i are interpreted as the causal mechanisms by which the values of the respective variable V^i are determined from the value of ϵ^i and the values of the variables in pa^i. Consequently, the variables in pa^i are referred to as the *causal parents* of V^i. The random variables ϵ^i are interpreted as noise that summarises all factors that are not modelled explicitly, and the factorisation of $p_\epsilon(\epsilon^1, \ldots, \epsilon^n)$ amounts to the assumption that the $\epsilon^1, \ldots, \epsilon^n$ are jointly independent. This joint independence is motivated by the view that any dependence between the noise variables must be due to a causal relationship between them and that such a dependence should then rather be modelled explicitly by enlarging the set $\mathbf{V}$ of system variables.

The *causal graph* of an SCM with system variables $V^1, \ldots, V^n$ is the directed graph whose vertices are the variables V^i and with a directed edge $V^i \to V^j$ if and only if V^i is a causal parent of V^j, that is, if and only if $V^i \in pa^i$. Consequently, the causal graph of an SCM shows the *qualitative* cause-and-effect relationships as specified by the sets pa^i. If the causal graph is acyclic, that is, if the causal graph is a directed acyclic graph (DAG), then the SCM is said to be *acyclic*.

An SCM obtains causal meaning by asserting how the modelled system reacts to interventions. Formally, an intervention on the variable $V^k \in \mathbf{V}$ is a mapping, conventionally denoted as $do(V^k := v^k)$, that maps the original SCM and a number v_k to a new SCM in which the original structural assignment for V^k is replaced by the new structural assignment $V^k := v^k$ and the noise variables ϵ^k is removed. This new SCM is referred to as an *intervened SCM*, and $do(V^k := v^k)$ is interpreted as an idealised experimental manipulation by which the value of V^k is held fixed at v_k while leaving the system unaltered else. This specification of how the system reacts to interventions is why using the symbol ":=" instead of "=" in (1) is conventional. On the level of causal graphs, $do(V^k := v^k)$ amounts to removing all edges that point into V^k because, in the intervened SCM, the variable V^k has no causal parents. Interventions on subsets of variables are defined similarly.

In an acyclic SCM, the combination of noise distribution p_ϵ and functions f^i uniquely determines a distribution of the system variables $V^1, \ldots, V^n$. This distribution is often referred to as the *entailed distribution of the SCM*. The entailed distribution $p(\cdot)$ of the original SCM (that is, of the SCM that models that system without interventions) is often referred to as the *observational distribution*. The entailed distributions of the intervened SCMs are often referred to as *interventional distributions* and are conventionally often denoted as $p(\cdot|\ do(V^k := v^k))$; and similarly for interventions on subsets of system variables.

When using this notation, it is important to keep in mind that $p(\cdot|\ do(V^k := v^k))$ is, in general, not equal to $p(\cdot|\ V^k = v_k)$. Indeed, $p(\cdot|\ do(V^k := v^k))$ is the distribution of the *intervened* SCM, whereas $p(\cdot|\ V^k = v_k)$ corresponds to *observing* $V^k = v^k$; put differently: Correlation is not equal to causation.

The article [117] discusses in much detail the more complicated case of cyclic SCMs. As shown there, cyclic SCMs need not entail a unique distribution for the system variables. However, [117] defines a restricted class of cyclic SCMs, termed *simple SCMs*, that entail a unique distribution and that are closed under interventions. Moreover, acyclic SCMs are a special case of simple SCMs.

In the time series case, which we are predominantly interested in this paper, (1) can be generalised by putting a time index t on V^i, f^i, pa^i and ϵ^i. The commonly used term *causal stationary* then refers to time-invariance of the qualitative cause-and-effect relationships, that is, to the situation that $pa^i_{t+\Delta t} = \{V^j_{s+\Delta t} \mid V^j_s \in pa^i_t\}$ for all t and Δt.

2.1.2. Axes for categorising causal discovery methods

This section introduces and explains several axes for categorising and distinguishing causal discovery methods. While it would be possible to consider more axes yet, the authors believe that the choice of axes presented here is a reasonable compromise between a sufficiently fine-grained categorisation on the one hand and clarity of exposition on the other hand. Table 2 lists many causal discovery methods and categorises them according to the aforementioned axes. In Fig. 2, we graphically illustrate some of the axes.

Table 2
Taxonomy of methods for causal discovery. The entries in parentheses (·) indicate that there are versions of the algorithm in which the assumption is relaxed. Methods in grey rows are described in more detail in the text. We use the abbreviations 'TSG' for time series graph, 'summary' for summary graph and 'ext. sum. graph' for an extended summary graph.

Method	Target of inference		Approach	Process assumptions				Data assumption
	Bi-/Multi-variate	Graph type	Indep./ Asymm./ Score	Non-/linear	Stoch./Det.	Con- temp.	Cycles	Hidden var.
GC [35]	Bi.	Summary	indep.	Linear	stoch.	✗	Lagged-only	✗
Multi-GC [118]	Multi.	Summary	indep.	Linear	stoch.	✗	Lagged-only	✗
Multi-nonlin-GC [119]	Multi.	Summary	indep.	nonlin.	stoch.	✗	Lagged-only	✗
TE [120]	Bi.	Summary	indep.	nonlin.	stoch.	✗	Lagged-only	✗
Multi-TE [121]	Multi.	Summary	indep.	nonlin.	stoch.	✗	Lagged-only	✗
CCM [39]	Bi.	Summary	indep.	nonlin.	det.	✓	?	Partially
Ext.-CCM [122]	Bi.	Summary	indep.	nonlin.	det.	✓	?	Partially
tsPC [123]	Multi.	TSG	indep.	Both	stoch.	✓	(✓)	✗
PCMCI [124]	Multi.	TSG	indep.	Both	stoch.	✗	Lagged-only	✗
PCMCI$^+$ [123]	Multi.	TSG	indep.	Both	stoch.	✓	(✓)	✗
PCGCE [125]	Multi.	ext. sum. graph	indep.	Both	stoch.	✓	(✓)	✗
FCIGCE [125]	Multi.	ext. sum. graph	indep.	Both	stoch.	✓	(✓)	✓
tsFCI	Multi.	TSG	indep.	Both	stoch.	(✓)	(✓)	✓
SVAR-FCI [125]	Multi.	TSG	indep.	Both	stoch.	✓	(✓)	✓
SVAR-GFCI [126]	Multi.	TSG	Score & indep.	Both	stoch.	✓	(✓)	✓
LPCMCI [127]	Multi.	TSG	indep.	Both	stoch.	✓	(✓)	✓
(F)GES [128–131]	Multi.	Summary	Score	Linear	stoch.	✓	Lagged-only	✗
DYNOTEARS [132]	Multi	TSG	Score	Linear	stoch.	✓	Lagged-only	✗
IDYNO [133]	Multi	TSG	Score	Linear and non-linear	stoch.	✓	Lagged-only	✗
NTS-NOTEARS [134]	Multi	TSG	Score	Linear and non-linear	stoch.	✓	Lagged-only	✗
TiMiNo [135]	Multi.	TSG	indep.	Both	stoch.	✓	Lagged-only	✗
RHINO [136]	Multi.	TSG	indep.	Both	stoch.	✓	Lagged-only	✗
VARLiNGAM [137]	Multi.	TSG	asymm.	Linear	stoch.	✓	Lagged-only	✗

Bivariate vs. multivariate causal discovery. This axis concerns the number of variables that are being considered. Bivariate causal discovery aims to discover the causal relationship between exactly two variables X and Y (in the non-temporal case) or between exactly two component time series X^i and X^j (in the time series case). Multivariate causal discovery aims to discover the causal relationships between any number of variables or component time series, respectively. Bivariate causal discovery often (but not necessarily) assumes *causal sufficiency* (see axis on causal sufficiency below). In the time series case, bivariate causal discovery often (but not necessarily) targets to infer the *summary graph* rather than the *time series graph* or *extended summary graph* (see axis on time series graph discovery below). If time lags are at least partially resolved in the bivariate time series case, that is, if the target of inference is the time series graph or the extended summary graph, then one effectively deals with a multivariate causal discovery problem.

Time series graph discovery vs. summary graph discovery vs. extended summary graph discovery. This axis is specific to the temporal setting and concerns the target of inference. Some methods are designed to learn the *time series graph* [138], also known as *full-time graph* [135] and *time series chain graph* [139], that is, the collection of all causal links $X^i_{t-\tau} \to X^j_t$ including the respective lags τ of these links. Part $(a)(i)$ of Fig. 2 shows an example of a time series graph with four component time series. As indicated by the grey edges, the pattern of edges in this graph is implicitly assumed to repeat both to the left (past) and right (future). Due to this repetitive structure of the edges, a time series graph is uniquely specified by the collection of edges that point into a vertex at an arbitrary reference time step t. Other methods disregard the information about the time lags and instead learn the *summary graph* [9]. In the summary graph, there is exactly one vertex per component time series X^i and an edge $X^i \to X^j$ if and only if there is an edge in the time series graph $X^i_{t-\tau} \to X^j_t$ at any lag τ. Part $(a)(iii)$ of Fig. 2 shows the summary graph associated with the time series graph in part $(a)(i)$ of the same figure. Another option is to learn *extended summary graphs* [125]. These graphs go midway between learning time series and summary graphs by distinguishing between contemporaneous and lagged links but disregarding the information about the specific time lags of lagged links. Specifically, the extended summary graph contains exactly two vertices per component time series X^i, namely the vertex X^i_t for the present time steps and the vertex $X^{i,-}$ for all past time steps. There is an edge $X^i_t \to X^j_t$ if and only if, this same edge is also in the time series graph, there is an edge $X^{i,-} \to X^j_t$ if only if there is at least one $\tau \geq 1$ such that $X^i_{t-\tau} \to X^j_t$ is in the time series graph, and there is no edge between $X^{i,-}$ and $X^{j,-}$. Part $(a)(ii)$ of Fig. 2 shows the extended summary graph associated with the time series graph in part $(a)(i)$ of the same figure. Resolving the lag structure does yield more information but also implies a more complex

target of inference. Learning more complex graphs (e.g., a time series graph vs. a summary graph) is conceptually and statistically more challenging.

Methods for time series causal discovery typically require the user to specify a maximal lag τ_{max} up to which the method is supposed to be sensitive. If the target of inference is the time series graph, then this choice is apparent as the learned graph has exactly $\tau_{max} + 1$ steps. As opposed to that, when learning summary graphs or extended summary graphs the choice of τ_{max} is not apparent from the learned graph.

Methods based on independence, asymmetry, scores and context. This axis distinguishes causal discovery methods by the type of information/signal that they use to learn the causal graphs from data. In this review, as is common in the literature (for example, see [140]), we distinguish the independence-based, asymmetry-based, and score-based approaches. Further, we consider the context-based approach to distinguish those methods that employ the invariance of causal mechanisms across different environments. An exact delineation between these four approaches is not always possible as there are *hybrid* methods that combine more than one approach.

> *"The wide variety of causal discovery methods can be structured into independence-based, asymmetry-based, score-based, and context-based approaches."*

First, *independence-based causal discovery*, sometimes called *constraint-based causal discovery*, utilises marginal and conditional independencies between variables to learn the causal graph or a set of causal graphs consistent with those independencies. Recall that an SCM is defined by a collection of structural assignments for each variable, where each assignment is a function of the variable's parents and a noise term. The collection of noise variables is assumed to be jointly independent. Independence-based causal discovery relies on the fact that, for data generated by an SCM, the structure of the SCM's causal graph imprints some independencies onto the data [3,141,142]. This property is known as *causal Markov condition* [43]. Alternatively, if one does not assume that an SCM generates the data, then the causal Markov condition is not automatically implied but needs to be assumed separately, leading to the so-called *causal Markov assumption*. The *d*-separation criterion [143] allows to graphically determine all independencies that are necessarily implied in a given causal graph [3,142,144]. The basic idea then is to run statistical tests of marginal and conditional independencies on the data and, second, use the results of these tests to constrain the causal graph's structure.

For the second of these two steps to hold, one further needs to make the *causal faithfulness assumption* [43]. This assumption says there are no independencies beyond those necessarily implied by the causal Markov condition in the observed data.

Independence-based causal discovery is non-parametric in that no assumption on the SCM's functional relationships and/or noise distributions needs to be made. However, choosing a particular method for (conditional) independence testing may implicitly impose a parametric assumption. For example, testing for (conditional) independence by (partial) correlation implicitly makes the assumption that the data-generating process is linear Gaussian. Conversely, if a parametric assumption can be made, this assumption might favour specific methods for (conditional) independence testing. For example, suppose one can assume linear Gaussian data. In that case, it is reasonable to use a (partial) correlation instead of more general (conditional) independence tests like, for example, a test based on (conditional) mutual information as given in [145].

Typically, there are multiple graphs that by means of the causal Markov condition, imply the exact same set of (conditional) independencies. For example, the three graphs $X \to Y \to Z$ and $X \leftarrow Y \leftarrow Z$ and $X \leftarrow Y \to Z$ by means of the causal Markov condition all imply exactly the same independence, namely that X and Z are conditionally independent given Y (and no further independencies). Such graphs are said to be *Markov equivalent* to each other and constitute a *Markov equivalence class*. Consequently, independence-based causal discovery algorithms cannot distinguish between Markov equivalent graphs.

Second, *asymmetry-based causal discovery* makes and relies on parametric assumptions on the form of the functional relationships and/or noise distributions of the data-generating SCM [9].

This approach is motivated by the elementary bivariate case, that is, by finding the causal relationship between two variables X and Y. As explained above, with independence-based causal discovery, it is not possible to distinguish the Markov equivalent graphs $X \to Y$ and $X \leftarrow Y$. This impossibility is not a shortcoming of independence-based causal discovery but rather is fundamental unless stronger assumptions are made [9,146]. The proof of the impossibility of distinguishing between $X \to Y$ and $X \leftarrow Y$ works by showing that if the true data-generating SCM goes in the direction $X \to Y$, then one can always construct an alternative SCM in the direction $X \leftarrow Y$ that gives rise to the same data distribution as the true SCM.

The basic idea for removing this fundamental ambiguity is as follows: For certain choices of *restricted* SCMs, defined by certain restricted parametric assumptions, it is impossible to have a restricted SCM in both directions. Hence, given the assumption that the true SCM lies in the restricted class of models, it becomes possible to distinguish the causal and anti-causal direction. A restricted class of SCMs with this property is said to be *identifiable*. This approach to causal discovery relies on the expectation that the SCM in the causal direction generically has lower complexity than any alternative SCM

in the anti-causal direction. As explained in Section 4.1.2 of [9], this expectation can be motivated by the principle of *independence of cause and mechanism* [146,147] There are also asymmetry-based causal discovery methods for learning the causal graph between two or more variables for multivariate causal discovery [9].

Third, *score-based causal discovery* chooses one or multiple best-scoring graphs with respect to a predefined scoring function. This scoring function is typically built on the likelihood of the observed data given a particular graph and an assumed parametric statistical model [9]. This approach requires searching over the space of causal graphs. Even if causal sufficiency (see axis on causal sufficiency) and acyclicity (see axis on cycles below) are assumed, in which case the causal graph is a *directed causal graph (DAG)*, the search space of graphs already grows super-exponentially, e.g. [129]. An exact search is thus infeasible even for a moderate number of variables. Instead, greedy search techniques are often used, e.g. in the famous GES algorithm [129,130] (see below for details on this algorithm) If the assumed statistical model does not yield identifiability beyond the Markov equivalence class, then the scoring function must be chosen such that Markov equivalent graphs have the same score. Hence, one can search over the space of Markov equivalence classes rather than over the space of graphs.

Fourth, *context-based causal discovery* requires access to data of the same system in different contexts. The term *different context* is understood rather broadly: Its meaning ranges from, for example, observing the same physical system at other locations to, for example, observing a system both before and after an intervention. The basic assumption and idea of this approach to causal discovery are that the causal mechanisms, that is, the functional mappings from causes to effects and hence also the conditional distributions of the effects given their causes, remain unchanged across all contexts (unless the effect variable is the target of an intervention in one of the contexts). In contrast, marginal distributions and hence also the conditional distributions of causes given their effects can change [9]. A prime example of a context-based causal discovery method is Invariant Causal Prediction [148] (see below for more details). The *joint causal inference (JCI)* framework [149] proposes to model all contexts with one graph by including one or multiple so-called *context variables* whose values determine the context and subsequently pooling the data from the different contexts into one joint dataset.

Linear or nonlinear dependencies. This axis concerns the form of the functional relationships in the data-generating structural causal model. Broadly, see part (*b*) of Fig. 2, one can distinguish between linear and nonlinear functional relationships. In independence-based and score-based causal discovery, an assumption of linearity can enter implicitly by using partial correlation for testing conditional independence (independence-based approach) or by choice of the statistical model (score-based approach). In asymmetry-based causal discovery, an assumption of linearity is, if made, typically explicit by choice of the functional model. Various asymmetry-based causal discovery methods do not assume linearity but still use restricted functional model classes that do not allow for entirely generic dependencies, for example, the functional model class of nonlinear additive noise models [150]. However, for simplicity, we here only distinguish the methods by whether or not they assume linearity.

Deterministic vs. stochastic systems. This axis concerns an assumption on the type of data-generating process Some methods assume the data are generated by a deterministic process, for example, a deterministic dynamical system. In contrast, other methods make explicit use of the assumption that the data-generating process is inherently stochastic, see part (*c*) of Fig. 2. In the case of stochastic data-generating processes, the stochasticity is interpreted as dynamical noise that arises due to factors outside of the model. Dynamical noise needs to be distinguished from measurement noise: The former is an inherent property of the data-generating process, and the latter arises from uncertainty in the data-collection process The causal inference and discovery frameworks have also been extended to dynamical systems, both deterministic and stochastic, without stable equilibrium distribution [151–153].

Contemporaneous links. This axis is specific to the temporal setting and concerns a connectivity assumption on the causal time series graph. Some methods make the assumption that all causal links in the time series graph are *lagged*, meaning that all causal links are of the form $X^i_{t-\tau} \to X^j_t$ with $\tau > 0$, whereas *contemporaneous* links, that is, links of the form $X^i_t \to X^j_t$ are assumed to be absent. Other methods do not make this assumption. For example, in the time series graph in part $(a)(i)$ of Fig. 2, there are the contemporaneous edges $Z_t \to W_t$ and $Y_t \to X_t$. Consequently, methods that assume the absence of contemporaneous links would, by assumption, disallow this particular time series graph. Contemporaneous causal links correspond to causal influences that act on a time scale shorter than the measurement interval; for example, a causal influence on a time scale of six hours in daily measured data.

Causal cycles. This axis concerns a connectivity assumption on the causal graph Many methods assume the absence of cyclic causal relationships. This assumption means that a variable X^j_t cannot be a causal ancestor of another variable $X^i_{t-\tau}$ if that second variable $X^i_{t-\tau}$ is a causal ancestor of the first variable X^j_t. For example, the lower graph in part (*d*) of Fig. 2 has the causal cycle $X_t \to Y_t \to X_t$. Because causation cannot go backwards in time, the assumption of acyclicity only restricts the contemporaneous section of the causal time series graph. The assumption is thus only relevant for $\tau = 0$.

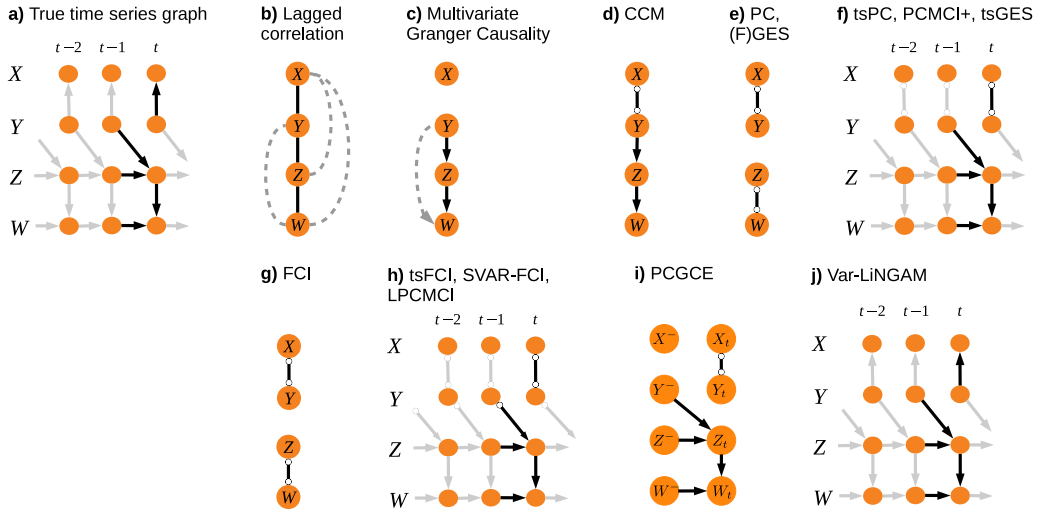


Fig. 3. Figure illustrating a time series graph (TSG) and the respective graphs discovered by applying various causal discovery methods to data generated from an SCM with that time series graph. (a) Time series graph. (b) The discovered undirected graph by considering (lagged) correlations, where spurious correlations are highlighted as dashed grey lines. (c) The directed graph discovered by multivariate Granger causality does not consider contemporaneous links and retains a spurious link from Y to W. (d) Graph discovered by CCM. (e) The graph discovered by applying the plain PC and (F)GES algorithms fail to show lagged links and, in addition, fail to orient a link that the time series adapted algorithms can orient. (f) The time series version of PC (tsPC), PCMCI+, and the time series version of GES (tsGES) discover both lagged and contemporaneous links and orient edges up to the Markov equivalence class. (g) Plain FCI has the same drawbacks as plain PC or (F)GES. (h) FCI-based time series causal discovery algorithms account for latent confounders and thus discover causal arrows up to latent confounding. In particular, the algorithm cannot exclude that the association between Y_{t-1} and Z_t is due to latent confounding rather than a causal relationship. (i) The PCGCE algorithm discovers the extended summary graph up to its Markov equivalence class. (j) Var-LiNGAM discovers all causal relationships correctly if the assumptions of linear relationships and additive non-Gaussian noise are satisfied.

In particular, even with the assumption of acyclicity, it is possible to model temporal feedbacks. For example, although the upper graph in part (*d*) of Fig. 2 is acyclic, it displays a causal influence of time series X on Y (by the edge $X_t \to Y_t$) and a causal influence of time series Y on X (by the edge $Y_{t-1} \to X_t$). There are also causal discovery methods that allow cyclic causal relationships, e.g. [117,154–156]. These methods typically infer less informative graphs than those inferred by methods that do not allow causal cycles. The early work [11] considers the special case of causal discovery in *linear* cyclic systems.

Causal sufficiency. This axis concerns an assumption that can be viewed as an assumption on the data-generating process or the data-collection process. The assumption of *causal sufficiency* [43] says that there are no *latent confounders*, also called *unobserved confounders* or *hidden common causes*. A latent confounder is an unobserved variable that (potentially indirectly through other unobserved variables) causally influences two observed variables $X^i_{t-\tau}$ and X^j_t. For example, in the lower graph in part (*e*) of Fig. 2 the unobserved variable Z acts as a latent confounder of the variables X and Y. Consequently, this graph violates causal sufficiency whereas the upper graph in part (*e*) of Fig. 2 satisfies causal sufficiency. Methods that do not assume causal sufficiency typically infer graphs that are less informative than the graphs inferred by methods that do assume causal sufficiency. For example, consider the elementary non-temporal bivariate case with two variables X and Y. If these variables are dependent and causal sufficiency is assumed, then either X causes Y ($X \to Y$), or Y causes X ($X \leftarrow Y$). If causal sufficiency is not assumed, then there is the third possibility that neither X causes Y nor vice versa but that there rather is an unobserved variable L which causes both X and Y ($X \leftarrow L \to Y$).

2.1.3. Description and categorisation of causal discovery methods

Here, we list and briefly explain several existing causal discovery methods for time series data. We summarise this list and the placement of each method with respect to the axes presented above in Table 2. In Fig. 3, we illustrate and compare an example time series graph and the respective graphical objects obtained by some of the discussed causal discovery algorithms when applied to data generated from an SCM with that time series graph.

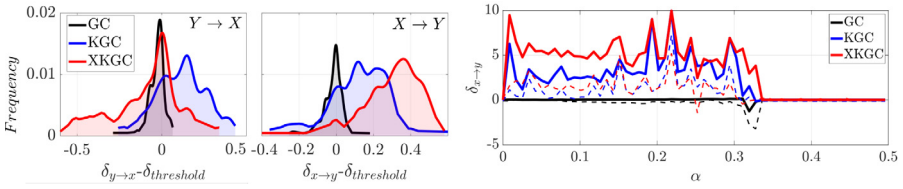
Granger causality. Granger Causality (GC) [35,157] is originally a statistical test to decide whether a time series X_t is a *cause* of another time series Y_t, in the sense that past values of X_t have significant predictive power in forecasting Y_t.

GC is thus, in principle, a simple test of temporal (or lagged) relationship and predictability. Nevertheless, under causal sufficiency and no contemporaneous effects assumptions, it can be formally shown that GC testing detects actual causal links (see e.g. Peters et al. [9] for a formal derivation of these results in the SEM setting).

In a multivariate setting, which is seldom the case, testing if X_t causes Y_t requires controlling for all possible confounders. Therefore the conditional GC [118,157–159], includes in the restricted and full models the past of all other *relevant* time series in the system that are not X_t and Y_t. Classically, GC considers linear models for which standard t-tests or F-tests can be employed, but non-linear extensions have been considered both in econometrics [160–164] and in physical and biological applications [119,165,166].

Linear and nonlinear Granger causality in dynamic systems. A unified view of (nonlinear) GC with kernel methods for the physical sciences was introduced in [119]. Two examples are given here:

1. Bivariate system with coupled non-linear and autoregressive relations given by $x_{t+1} = 3.4x_t(1 - x_t^2)\exp(-x_t^2) + \varepsilon_t^x$ and $y_{t+1} = 3.4y_t(1 - y_t^2)\exp(-y_t^2) + 0.5x_t y_t 2 + \varepsilon_t^y$, where ε is white Gaussian noise with zero mean and variance 0.4. The causal direction is $X \to Y$. The histogram of the estimated causal index δ on the left figure reveals GC's insensitivity to the causal direction, the high false positive rate of KGC [167], and a higher detection power by the XKGC [119].
2. Two logistic maps system defined as $x_{t+1} = 1-1.8x_t^2$ and $y_{t+1} = (1-\alpha)(1-1.8y_t^2)+\alpha(1-1.8x_t^2)$, where $\alpha \in [0, 1]$ controls the coupling strength. The causal relationship implemented is $X \to Y$, and the challenge is to assess the detection power of methods without introducing any external variable. The left figure shows XKGC [119] improves detection power over GC/KGC for any α.



Transfer entropy [120] between Y_t and X_t measures the amount of unique information contained in the past of X_t about the state of Y_t and is defined as $\mathcal{T}_{X\to Y|Z} = H(Y_t|Y-, Z-) - H(Y_t|Y-, X-, Z-)$. Transfer entropy can be considered as the generalisation of GC by extending the implicit conditional independence test to arbitrary orders of dependence. Indeed, Barnett et al. [121] proved that (linear) GC and transfer entropy causality are equivalent under the assumptions of VAR model class and Gaussian error distributions.

CCM.

Convergent cross-mapping (CCM) [39] is based on the simple observation that if data from a deterministic dynamical system is generated by a system of ordinary differential equations (ODEs), then the explicit form of these equations directly defines the causes of each variable in the system: x is a cause of y if the dynamics (any of the derivatives of y) is expressed in terms of the state of x. For example, in the Lorenz attractor system

$$\begin{aligned} x' &= \sigma y - \sigma x \\ y' &= -xz + \rho x - y \\ z' &= xy - \beta z \end{aligned} \tag{2}$$

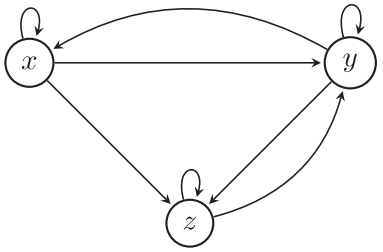


Fig. 4: Summary graph for Lorenz attractor system.

x_t is caused by y_t, y_t is caused by x_t and z_t and z_t by x_t and y_t as summarised by the summary directed graph of Fig. 4. Learning the generating ODE from time series data would allow us to recover the causal relations and summary-directed graph. Nevertheless, relying on Takens' theorem [168], Sugihara et al. [39] concluded that it is not necessary to recover the exact ODE to recover its causal properties: if one has a cause and effect variable within an ODE system, then a qualitative description of the dynamics of the cause based on the dynamics of the effect can be recovered. Surprisingly, while one would need a large number of lags to estimate an ODE where no parametric assumptions have been made, Takens' theorem states that a good enough estimate, i.e. one that retains the causal properties of the ODE, can be made with at most $2d + 1$ lags where d is the number of variables in the ODE.

The CCM pseudo-algorithm for checking if two variables X and Y are causally related:

1. Choose embedding dimension E: number of lags to use with $1 \leq E \leq 2d+1$.
2. Estimate cross-map skill $\rho(l)$ for a sequence of several observations $l_1, \ldots, l_N$ with $l_N \leq L$, L is the maximum number of available observations in the time series. For each l_i:
   (a) Construct shadow manifold M_x: in practice represented by matrix $Y \in R^{l \times E}$ with time series $y_t, y_{t-1}, y_{t-2}, y_{t-E+1}$
   (b) Assume the shadow manifold satisfies Takens' theorem condition and retains the metric properties of manifold M. Thus estimate euclidean distance d_i of E+1 nearest points on manifold M_x to point $(x_t, x_{t-1}, \ldots, x_{t-E+1})$. Denote the time indices corresponding to these points as $t_1, t_2, \ldots, t_{E+1}$.
   (c) Construct estimate of y_t using simplex projection in shadow manifold: weighted average of $E+1$ nearest points (on M_x) with weights determined according to the exponentially weighted distance on M_x of each point (calculated in the previous step):
   $$\hat{y}_t = \sum_{i=1}^{E+1} w_i y_{t_i} \qquad \text{where} \qquad w_i = \frac{\exp(-\frac{d_i}{d_1})}{\sum_i \exp(-\frac{d_i}{d_1})} \tag{3}$$
   (d) Construct cross-map skill $\rho(l) = Corr(y_t, \hat{y}_t | M_x)$
3. Check if cross-map skill $\rho(l)$ converges as l tends to L. As the number of observations used increases, the manifold estimation should be denser, so cross-map skill should improve and converge, provided our assumption that M_x retains the metric properties of M is true.

The algorithm should also be applied symmetrically to establish the convergence of the cross-map skill $\hat{x}_t | M_y$. If both cross-map skills converge, we can establish that both variables belong in the same ODE system, and the causal relations are bi-directional. Note that in step 2c, we only use the shadow manifold M_x to determine which points and with which weights should be used to estimate y_t. If the convergence of the cross-map skill happens in only one direction, the proper conclusion is that a uni-directional causal relationship exists between the two variables. If the cross-map skill of $\hat{y}_t | M_x$ converges, the proper conclusion is that y causes x. This is somewhat counterintuitive, at least from the point of view of more classical causal discovery methods, because to establish that x is an effect of the cause y, we must be able to predict the cause y using the effect x, where for all other methods discussed in this work it is the other way around.

PC-based methods. In the following, we start with an exposition of the PC algorithm [43]. We then explain both a naive (tsPC) and more sophisticated (PCMCI) time series adaption.

1. *PC.* The original PC algorithm (named after Peter and Clark's authorship [43]) was constructed for i.i.d random variables and thus, in particular, for non-time series data. Below, we will also describe an extension to the time series case. The PC algorithm assumes that the underlying causal graph is a (*directed acyclic graph DAG*). A DAG has only directed edges ($\rightarrow$ and $\leftarrow$) and no cycles. As for independence-based algorithms in general, the PC algorithm assumes the causal Markov condition and causal faithfulness to infer d-separations on the causal graph from conditional independencies alone. Moreover, the algorithm assumes causal sufficiency. Consequently, the algorithm cannot distinguish between two graphs with the same set of d-separations. The graphical representation of an equivalence class of DAGs with the same d-separations is known as a (*completed partially-directed acyclic graph CPDAG*), which is the object of discovery of PC (see [169]). As compared to DAGS, CPDAGs can contain undirected edges ($\circ\!\!-\!\!\circ$). These undirected edges signify that both orientations ($\rightarrow$ or $\leftarrow$) are compatible with the set of conditional independencies.

   The PC algorithm starts from a fully connected undirected graph and consists of three phases:
   (a) The *skeleton phase* uses statistical (conditional) independence tests to infer the adjacencies of the underlying causal graph. If two variables X and Y are found to be independent conditional on a (possibly empty) set of variables $\mathbf{Z}$, then the edge between X and Y is removed.
   (b) The *collider orientation phase* then orients all *collider motifs*, that is, motifs of the form $X \rightarrow Y \leftarrow Z$ where X and Z are non-adjacent. These orientations can be inferred because collider motifs impose a particular pattern of (conditional) (in-)dependencies.
   (c) The *orientation phase* finally uses graphical rules [169] to infer the orientation of as many remaining unoriented edges as possible using the acyclicity assumption and the fact that all colliders have been found in the previous step.

   Although the PC algorithm was originally developed for the acyclic case, the work [156] shows that PC is also consistent in the presence of cycles if the learned graph is interpreted in a slightly different way using the so-called σ-separation [117].
2. *tsPC.* The naive extension of the PC algorithm to the time series case is called tsPC, an example implementation is given in [123]. The general idea is to fix an integer $\tau_{\max}$ that is supposed to be equal to or larger than the maximum

time-lag of any edge in the causal time series graph and to learn the finite segment of the time series graph on a time window $[t - \tau_{\max}, t]$. Here, t is an arbitrary reference time step, and samples are created by sliding the time window over all recorded time steps. This approach implicitly assumes that the causal relationships do not change throughout the recorded time steps. As discussed in [123], tsPC suffers from a sub-optimal finite-sample performance due to autocorrelation. This issue is remedied by the *PCMCI* algorithm, explained below See Fig. 3 for an illustration of the graphs learned by PC and tsPC for the time series case example.

3. *PCMCI.* The PCMCI algorithm [170] is a time series causal discovery algorithm that addresses some of the shortcomings of the naive time series adaption of PC, in particular the issue of low detection power. PCMCI assumes the time series graph to have no contemporaneous causal influences. This assumption implies the absence of contemporaneous cycles, but feedback cycles involving time lags are possible. Additionally, PCMCI assumes the time series data are generated by a causally stationary process (that is, the causal relationships are assumed to not change over time).
As discussed in [123,170], two main challenges with time series data hamper the performance of independence-based discovery algorithms in time series. These challenges are related to autocorrelation, a common feature in time series. First, using non-i.i.d. samples (created in a sliding window fashion as explained above) typically leads to ill-calibrated conditional independence tests, that is, uncontrolled type I errors, because the degrees of freedom are reduced and cannot be easily measured. This ill-calibratedness leads to inflated false positives, that is, the discovery of dependence when, in fact, independence is true. Secondly, high autocorrelation implies that there is little new information in the next time step compared to the previous step. Depending on how the conditioning sets in conditional independence tests are selected, this results in low effect sizes leading to low detection power of true links. The effect size of a (conditional) independence test is defined as the absolute value of the population value of the test statistic; for example, in a (partial) correlation test, the effect size is the absolute value of the population value of the (partial) correlation. While there is a trade-off in addressing both of these challenges, the PCMCI algorithm and its generalisation PCMCI$^+$ (see below) algorithms remedy these challenges to an extent by using a particular choice of conditioning sets in the independent tests that decides about the presence versus the absence of an edge between a given pair of variables. We spell out the details below.

The PCMCI algorithm unfolds in two phases:

(a) The first phase, referred to as PC$_1$, is a *condition-selection phase* that aims to infer a superset $\hat{\mathcal{P}}(X_t^j)$ of the parents of each variable X_t^j at time step t. The PC$_1$ algorithm is a variant of the PC and works as follows: Each sub-step of the skeleton phase is indexed by the integer p, starting at $p = 0$ and successively increasing p in increments of one. Within each sub-step, the algorithm tests for independence of $X_{t-\tau}^i$ and X_t^j given a conditioning set that consists of those p potential parents of X_t^j (less $X_{t-\tau}^i$) that have the highest association with X_t^j according to the previous (conditional) independence tests. This particular choice of conditioning sets increases the effect sizes of the (conditional) independence tests, as can be understood information-theoretically [171]. A higher effect size leads to a higher statistical power (equivalently, to a lower probability of a type II error), that is, makes it more likely to detect dependence if dependence is true. However, since the effects of autocorrelation have not been dealt with yet, this phase of PCMCI is affected by the same issues as the naive time series extension of PC in terms of false positives.

(b) The second phase conducts for each pair of variables $X_{t-\tau}^i$ and X_t^j the so-called *momentary conditional independence (MCI) test* that tests the null hypothesis $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \hat{\mathcal{P}}(X_t^j)\backslash\{X_{t-\tau}^i\}, \hat{\mathcal{P}}(X_{t-\tau}^i)$. If this hypothesis is not rejected, the edge between $X_{t-\tau}^i$ and X_t^j is removed. While the condition on $\hat{\mathcal{P}}(X_t^j)$ only would suffice to condition out confounded and indirect connections, the additional conditioning on $\hat{\mathcal{P}}(X_{t-\tau}^i)$ removes auto-dependencies from $X_{t-\tau}^i$ such that the conditional independence tests are well-calibrated and false positives are controlled at the desired level [170]. Note that no orientation phase is required because, by the assumption of no contemporaneous causal influences, all edges are time-lagged and oriented by time order.

The numerical studies in [170] show that in combination with the MCI tests in its second phase, PCMCI improves detection power and false positive control compared to the naive time series adaption. The PCMCI$^+$ algorithm [123] extends PCMCI to the case where contemporaneous edges are allowed by suitably modifying the MCI phase (but still disallowing contemporaneous causal cycles and latent confounders). See Fig. 3 for an illustration of this method.

FCI-based methods. A major success of causal discovery is the development of methods for learning causal relationships without assuming causal sufficiency. One famous example is the FCI algorithm [43,172,173]. We review the standard FCI algorithm and a time series adaption of FCI called tsFCI.

1. *FCI.* The FCI algorithm generalises the PC algorithm (see above) to the causally insufficient case (see [174]). In addition to latent confounders, the algorithm can also deal with *selection variables*. These variables influence whether a given sample point belongs to the observed population. For example, a certain satellite observation might be more likely to be made if the cloud cover is not too dense. Consequently, the statistical dependence relations will be biased (giving it the name *selection bias*) as only one segment of the entire population of possible satellite observations is being considered. Below, we assume the absence of selection variables and explain the specialisation of FCI in this case.

Quick introduction to maximal ancestral graphs. To deal with latent confounders (and selection variables), FCI works with a larger class of graphical models than PC does: Instead of DAGs, FCI works with *maximal ancestral graphs* (*MAGs*) [174]. Maximal ancestral graphs can be interpreted as projections of the underlying DAG (which consists of observed variables, latent confounders, and selection variables) to a graph over the observed variables only, that is, a graph in which the latent confounders and selection variables have been marginalised out. When assuming the absence of selection variables (as we do here), it is sufficient to work with a subset of MAGs that are called *directed maximal ancestral graphs* (*DMAGs*) [156].
Directed maximal ancestral graphs can have two types of edges: directed edges ($\to$) and bidirected edges ($\leftrightarrow$). A directed edge $X \to Y$ says that variable X causally influences variable Y. This causal influence can be direct or indirect through one or multiple unobserved variables. A bidirected edge $X \leftrightarrow Y$ says that X and Y are subject to latent confounding and that, at the same time, neither X causally influences Y nor the other way around. Being subject to latent confounding means that there is an unobserved variable L that (potentially indirectly through other unobserved variables) causally influences both X and Y. In addition, an edge between X and Y (i.e., $X \to Y$ or $X \leftarrow Y$ or $X \leftrightarrow Y$) means that X and Y are not (conditionally) independent given any set of observed variables.
A subtle part of the interpretation of DMAGs is that directed edges $X \to Y$ can "hide" latent confounding. That is to say, while $X \to Y$ does say that X causally influences Y, it is possible that X and Y are also subject to latent confounding.

The FCI algorithm works in a way that is similar to the PC algorithm: First, a sequence of (conditional) independence tests is performed to find the skeleton (that is, the adjacencies) of the graph. Second, several orientation rules are applied to determine the direction of as many links as possible. For these details, we refer to the original works [43,172,173] or to more technical reviews of FCI, for example, see Section S2 in the supplementary material of [127].
As in the case of the PC algorithm, FCI does not learn a unique DMAG but rather a Markov equivalence class of DMAGs. These equivalence classes are graphically represented by *directed partial ancestral graphs* (*DPAGs*) [156,173,175]. In addition to directed ($\to$) and bidirected edges ($\leftrightarrow$), DPAGs can also contain edges of the types $X \circ\!\!\to Y$ and $X \circ\!\!-\!\!\circ Y$. An edge $X \circ\!\!\to Y$ says that Y does not have a causal influence on X while X might or might not have a causal influence on Y, whereas an edge $X \circ\!\!-\!\!\circ Y$ does not make any claim about whether or not X or Y have a causal influence on each other See Fig. 3 for an illustration of the graph that the FCI algorithm discovers when applied to time series data. The work [156] has shown that FCI, originally developed with the assumption of acyclicity, can also be consistently applied to data that is generated by a cyclic SCM with certain regularity conditions.

2. *tsFCI.* The tsFCI algorithm [176] adapts FCI to causally stationary time series. As compared to the FCI algorithm, tsFCI applies the following two conceptual modifications: First, lagged links ($\tau \geq 1$) are by default oriented as $X^i_{t-\tau} \circ\!\!\to X^j_t$. These default orientations are valid because an effect cannot precede its cause. Note that it would not be valid to orient all lagged links as $X^i_{t-\tau} \to X^j_t$ because $X^i_{t-\tau} \leftrightarrow X^j_t$ (i.e., latent confounding) is a possibility. Second, so-called *homologous* edges are by default oriented in the same way. That is if the edge between $X^i_{t-\tau}$ and X^j_t has been found to have a certain orientation (for example, $X^i_{t-\tau} \to X^j_t$ or $X^i_{t-\tau} \leftarrow X^j_t$) and if in addition there is an edge between $X^i_{s-\tau}$ and X^j_s for $s \neq t$, then this latter edge is immediately oriented in the same way as the former edge (for example, oriented as $X^i_{s-\tau} \to X^j_s$ or $X^i_{s-\tau} \leftarrow X^j_s$). This copying of edge orientations is valid because of causal stationarity. In addition to these modifications, tsFCI uses the knowledge of time order and causal stationarity to apply further modifications that are useful from a computational and/or statistical point of view.
There are two versions of tsFCI, both of which have been introduced in the original work [176]: One version does not allow for contemporaneous causal influences in the data-generating process one version in which such influences are allowed. Fig. 3 shows an illustration of the output of tsFCI (its version that allows contemporaneous causal influences) and other FCI-based time series causal discovery algorithms (see below) in the case of an example time series graph.

Greedy Equivalence Search (GES). GES [128–130] is probably the most famous score-based causal discovery method for i.i.d data assuming that the true causal graph is a DAG. It performs greedy steps directly on the CPDAG, thus searching in the Markov equivalent class space Chickering [129] proved that an efficient two-phase greedy search, combined with the BIC score is sufficient to find the true CPDAG in the large sample limit (the Meek conjecture [128]) assuming causal sufficiency. Ramsey et al. [131] developed Fast GES (FGES) an optimised and parallelised version of the GES algorithm, which they were able to scale up to a million variables. Additionally, GES has been improved by bounding polynomially the score evaluations [177]; to obtain statistical efficiency [178]; to obtain finite-sample correction of confidence intervals [179] and to deal with latent variables [180]. Similarly to other methods developed for i.i.d. data, (F)GES can be applied to uniformly-sampled causally stationary time series by simply considering the transformed lagged variables and imposing that the effects cannot precede the causes in time.

Continuous optimisation methods. A recent advance in structure learning has been the development of so-called continuous optimisation methods, particularly score-based methods which avoid the explosion of the discrete space of DAGs by employing continuous optimisation. The first such method is the NOTEARS algorithm proposed by Zheng et al. [181] which proposed $h(\mathbf{W}) = \mathrm{tr}(\exp(\mathbf{W} \circ \mathbf{W})) - d$ as differentiable characterisation of acyclicity for a weighted adjacency matrix $\mathbf{W}$, that is $h(\mathbf{W}) = 0$ if and only if the associated graph is a DAG. Such differentiable characterisations allow

the plug-in use of different continuous optimisation methods and even complex function parameterisations [136,182–187]. Various implementations of the continuous optimisation framework for structure learning from time series are available: (1) DYNOTEARS [132] considers structural linear VAR (SVAR) models, allowing for contemporaneous links and enforcing acyclicity among the instantaneous edges. (2) IDYNO [133] is designed to perform structural discovery from both observational and interventional data. Moreover, both linear and non-linear relationships are considered. (3) NTS-NOTEARS [134] models cause–effect relationships through one-dimensional convolution neural networks (CNN) and allows prior knowledge to be encoded and exploited directly by the optimisation procedure.

(VAR)LiNGAM. Linear non-Gaussian acyclic model (LiNGAM) [137] is a classical method for causal discovery which assume acyclicity, causal sufficiency, linear relationships and non-Gaussian additive independent noises. Under those assumptions, the model is shown to be identifiable thanks to classical results from independent component analysis (ICA) [188]. Specifically, a linear SEM can be represented by,

$$\boldsymbol{X} = \mathbf{B}\boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{X}$ is the vector of system variables, $\boldsymbol{\varepsilon}$ the noise vector and $\mathbf{B}$ is the matrix of coefficients. Thus solving Eq. (4) for $\boldsymbol{X}$ we obtain,

$$\boldsymbol{X} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\varepsilon}.$$

The above equation is an ICA problem, and its theory states that when $\boldsymbol{\varepsilon}$ are non-Gaussian noises variables, the mixing matrix $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$ is identifiable in the large sample limit [188]. LiNGAM-ICA works by first exploiting the ICA results to obtain the mixing matrix $\mathbf{A}$, and secondly by permuting and normalising $\mathbf{A}$ to obtain the appropriate matrix $\mathbf{B}$ and the corresponding causal DAG The LiNGAM can also be solved directly without using ICA [189]. The direct-LiNGAM iteratively selects the variable which is the most independent from the residual (in a linear regression onto the remaining variables) and replaces the original data with the residuals matrix. In the end, the procedure gives a causal order; and ultimately, simple linear regressions (e.g. estimated with least squares) are used to obtain the final estimation of the lower triangular $\mathbf{B}$ matrix.

VAR-LiNGAM [175,190,191] is the extension of the classical LiNGAM model to time series data. It considers a linear structural vector auto-regressive (VAR) model, with possibly acyclic instantaneous relationships, and assumes non-Gaussian disturbances. In particular, the process $\boldsymbol{X}_t$ is assumed to evolve following

$$\boldsymbol{X}_t = \sum_{\tau=0}^{k} \mathbf{B}_\tau \boldsymbol{X}_{t-\tau} + \boldsymbol{\varepsilon}_t.$$

Where $\mathbf{B}_0$ is the matrix of instantaneous relationships and its sparsity pattern corresponds to a DAG while $\mathbf{B}_\tau$ for $\tau > 0$ are the matrices of lagged relationships; moreover the non-Gaussian noise vector $\boldsymbol{\varepsilon}_t$ has independent components and is assumed independent over time. Hyvärinen et al. [191] propose two methods to estimate the coefficients of the model above: (1) a two-stage method which combines least-square estimation of the autoregressive model and classical LiNGAM estimation; (2) a method based on multichannel blind deconvolution. In Fig. 3, we illustrate that if the assumptions are satisfied, VAR-LiNGAM is able to discover and orient all causal links.

Structural equation models for time series, TiMiNo. Peters et al. [135] extended the classical structural equation modelling framework to the time series setting by introducing time series models with independent noise (TiMINo). In particular, a multi-variate time series $X_t = (X_t^i)_{i\in V}$ satisfies a TiMINo if there exists a $p > 0$ and, for every $i \in V$, there are subsets $\mathbf{PA}_0^i \subseteq X^{V\setminus\{i\}}$ and $\mathbf{PA}_k^i \subseteq X^V$ for $k = 1, \ldots, p$ such that,

$$X_t^i = f_i\left((\mathbf{PA}_p^i)_{t-p}, \ldots, (\mathbf{PA}_1^i)_{t-1}, (\mathbf{PA}_0^i)_t, N_t^i\right), \tag{5}$$

where N_t^i are assumed to be jointly independent over i and time and for each i, N_t^i are identically distributed in time. Thus, the assumed model is extremely general, instantaneous relationships are allowed and the noise contribution is not restricted by allowing arbitrarily mixing through the f_i functions. Nevertheless, to prove the identifiability of the full-time graph and the causal summary time graph, additional assumptions have to be made. Specifically, in Peters et al. [135] it is assumed that either (i) Eq. (5) follows an identifiable functional model class (e.g. nonlinear functions with additive Gaussian noise or linear functions with additive non-Gaussian noise [192]) or (ii) each function f_i exhibits a time structure, that is the union of the causal parents of X_t^i contains at least one X_{t-k}^i and moreover the joint distribution is faithful with respect to the full-time graph with the summary time graph being acyclic. Under either assumption (i) or (ii), the complete causal graph is proved to be identifiable in the large sample limit. Practically, to estimate a TiMINo, methods for causal discovery for additive noise models with i.i.d. data [193] can be adapted to the time series setting. As in the direct-LiNGAM, it iteratively selects the causal order between the variables by fitting regression models and evaluating the independence between the variables and the residuals. To fit the regression models for f_i, various methods can be used, such as vector autoregressive models (linear), generalised additive models or Gaussian processes. Moreover, to test independence from the residuals, the HSIC [194] can be applied to all possible shifted time series up to the maximum lag order.

Invariant causal prediction. Invariant causal prediction deals with the setting of independent and identically distributed samples of the random vectors $X = (X_1, X_2, \ldots, X_p)^\top \in \mathbb{R}^p$, $E = (E_1, E_2, \ldots, E_q)^\top \in \mathbb{R}^q$, and $Y \in \mathbb{R}$. For a variable or vector of interest Y, E is a set of environment variables that may be causes of X but are not direct causes or effects (direct or indirect) of Y. ICP assumes that Y is generated *causally* from a subset $S^* \subseteq \{1, \ldots, p\}$ of the p variables considered, so that there is causal sufficiency, and Y is generated from a Structural Causal Model obeying:

$$Y = g(X_{S^*}, \epsilon), \quad \epsilon \sim F, \quad \epsilon \perp\!\!\!\perp X_{S^*}, \tag{6}$$

where g and F are arbitrary functions and distributions, respectively. A set $S \subseteq \{1, \ldots, p\}$ is a generic subset of the full set of candidate causes. We refer to S and X_S interchangeably for brevity. The task of ICP is to infer the set S^* of direct causes of the variable of interest Y.

The Invariance Causal Prediction (ICP) framework [148] is based on the observation that Y is independent of E given X_{S^*}, denoted $Y \perp\!\!\!\perp E | X_{S^*}$. Assuming we have a set of candidate causes X that includes S^*, the causal subset, and an environment variable E that we know does not directly cause Y or is an effect of Y. We can search for S^* by applying a conditional independence test $Y \perp\!\!\!\perp E | X_S$ on Y, E and subsets $S \subseteq \{1, \ldots, p\}$ of X. ICP then selects as causal variables the intersection of all those subsets S where the corresponding conditional independence test is not rejected:

$$\hat{S}^* = \bigcap_{S : p_S > \alpha} S. \tag{7}$$

Here p_S is the p-value associated with the conditional independence test of $Y \perp\!\!\!\perp E | X_S$, with the null hypothesis corresponding to conditional independence. We do not reject conditional independence at significance level α if $p_S > \alpha$.

One way to interpret the problem setting is that causal associations are more robust than other associations. So ICP finds the causes of a variable of interest by investigating which associations are invariant across environments. Another interpretation is that the environment variables define data generated under different interventions to the system (SCM). In physical sciences, regional and temporal variables are good candidates since these often describe changes in environments that alter the conditions under which physical processes occur.

Peters et al. [148] introduce an algorithm to implement ICP that assumes linear relationships between causes X and effects Y and a categorical, univariate variable E. In [195], more general algorithms are presented that allow for nonlinear relationships between cause and effects and for continuous and multivariate environment variables E. Pfister et al. [196] provide an ICP variant for time series data. The proposed time-series ICP relies on the causal invariance assumption across time points, thus removing the requirement of environment knowledge. This setting is also partially robust to hidden confounders, similar to the original ICP [148] framework, and in general, the ICP is expected to be conservative with respect to violations of its assumptions [196].

Causal frameworks for continuous-time systems. As we already reviewed, discrete-time causal systems fit directly as an extension of i.i.d. and DAG framework. When considering discrete-time systems, we can express the value of a variable at a time t as a function of other variables (and itself) observed at past instants, thus the complete causal graph can be seen as a DAG extended (possibly infinitely) in time. Classical methods for causal discovery in the i.i.d. setting can then be adapted to discrete-time systems quite straightforwardly.

Conversely, considering continuous-time systems, raise the issue that a time-extended DAG is not feasible, since the included variables would be uncountable [197]. Nevertheless, modelling continuous dynamical system helps in dealing with non-uniform sampling and extrapolating among different sampling frequencies, two major drawbacks of causal discovery methods for discrete-time systems. The following are the major causal frameworks available for continuous-time systems:

1. *Causal interpretation of ODE and SDE.* Various efforts have been made to describe causal systems with ODEs and SDEs. First, causal discourses around ODE were used to obtain different justifications for the cyclic SEM [198–201]. Rubenstein et al. [153] described Dynamical Structural Causal Models (DSCM) as extensions of SEM where each equation or assignment is a relationship between a set of causal parent trajectories and an effect trajectory. Under some stability conditions, such DSCM can be obtained from ODE systems SDEs have also been studied from a causal perspective [197,201]. The advantage of SDEs in modelling physical systems is that they allow incorporating an inherent source of stochasticity, a common assumption in numerous real-world systems [202]. Graphical parameterisations of SDE equilibrium distributions leading to models allowing for cycles have been investigated and different structure learning algorithms have been proposed [203].
2. *Dynamic Causal Models.* Dynamic Causal Models (DCM) [202,204–207], in short, is a Bayesian framework for fitting and comparing causal models for coupled dynamical systems. DCM was introduced and applied mostly in Neuroscience, and particularly in the problem of estimating the connectivity between brain regions from neuroimaging data, as discussed in detail in Section 4.2.2. Recently, and it has been even employed to model the COVID-19 pandemic dynamics [208,209].
3. *Local independence graphs.* Local independence is a notion of conditional independence for stochastic processes (both discrete and continuous time ones) which can be (in)dependent on each other pasts. In detail, for real-valued stochastic processes $X_t = (X_t^1, \ldots, X_t^p)$ and $A, B, C \subseteq \{1, \ldots, p\}$, we say that X^B is locally independent of X^A given

X^C at time t if the past of X^C until time t provides the same information, to predict $E[X_t^\beta | \mathcal{F}_t^{A\cup C}]$,[3] as the past of $X^{A\cup C}$ until time t, for each $\beta \in B$.

Didelez [211,212] studied graphical representations of local independence with directed graphs together with δ-separation and proved the equivalence of the pairwise and global Markov properties for multivariate counting processes Directed graphs and δ-separation has been then extended to mixed graphs and μ-separation [210] to model partially unobserved systems A constrained-based algorithm has been proposed [213], which is proven to be sound and complete under faithfulness assumption. Local independence graphs can be applied to multivariate processes which are solutions of SDEs (such as the multivariate Ornstein–Uhlenbeck process) or event and counting processes such as Hawkes processes [214].

2.2. Challenges

In this section, we discuss several challenges for causal discovery that are frequent in real-world applications. We distinguish between challenges related to the data-generating process itself (see Section 2.2.1), challenges associated with the available data (see Section 2.2.2), and challenges of statistical or computational nature (see Section 2.2.3). Users should carefully consider the challenges they face in their application and choose a suitable causal discovery method. More generally, how to reason and deal with typical challenges in causal inference is further discussed in a time series context in Runge et al. [115].

2.2.1. Process challenges

> *"Discovering causal relations from observational data is impossible without assumptions about the mechanisms and faces important challenges related to data and statistical characteristics. The field will need to incorporate domain knowledge and post-selection inference."*

Non-linearities pose challenges for causal discovery in both independence-based and score-based methods. Non-linear conditional independence tests, such as the GPDC test [215] and tests based on conditional mutual information [145], are computationally more expensive and tend to have lower statistical power than linear tests. Non-linear functional relationships in score-based methods require more complex score functions, which can decrease finite-sample performance. However, non-linear functional relationships can enhance identifiability in some asymmetry-based causal discovery methods [9]. Overall, nonlinearities increase model complexity and require careful consideration when selecting appropriate causal discovery methods.

Most time series causal discovery methods assume data generated from a causally stationary process with a unique equilibrium distribution. However, many real-world processes are non-stationary. Recently there have been works on devising causal discovery methods that first detect the variables afflicted with non-stationary driving mechanisms and subsequently infer the entire causal graph, including possibly proxy variables corresponding to the driving force of non-stationarity [149,216]. Furthermore, techniques exist to detect regime or context changes in non-stationary data [216–218], but it remains an active area of research. Domain experts may be able to identify the source of non-stationarity in time series data and preprocess the data to remove it.

Time series data is common in physical sciences, and it has a distinctive feature of auto-correlation. Many causal discovery algorithms are not designed for time series data and show decreased performance when applied without modification [123,170]. However, some methods are specifically designed for time series data, reducing the detrimental effect of auto-correlation. See the PCMCI algorithm above for a discussion on the challenges of auto-correlated data. In many domains, space adds to time as well. In principle, one could feed different spatial locations of the same variable as distinct variables into causal discovery methods. However, this naive approach ignores spatial correlations and quickly results in a high-dimensional problem. Another workaround, employed for example in [219,220], is to perform dimension-reduction as a preprocessing step and then perform causal discovery on the dimensionally-reduced space. The development of causal discovery inherently designed for spatio-temporal data is an active area of research; for example, see [221] Dealing with variables whose dynamics operate on different time scales, such as e.g. fast atmospheric and slow oceanic processes, pose important challenges. Granger causality in the frequency domain, e.g. [222–224], and a combination of wavelet analysis with transfer entropy [225] are examples of approaches to deal with the time scales of causal influences.

Finally, it is worth mentioning that many causal discovery methods assume the data-generating SCM to be acyclic (see axis VII in Section 2.1.2). However, in real-world applications, one can often not exclude the existence of feedback that acts on time scales below the measurement interval. Such feedbacks make the time series graph **cyclic**. In the non-temporal setting, one might often not be able to exclude the existence of causal cycles. As discussed and referenced above, there are causal discovery methods that can handle cyclic causal graphs.

[3] Where $\mathcal{F}_t^{A\cup C}$ is a right-continuous and complete filtration which represents the history of the processes $X^{A\cup C}$ see e.g. [210] for a detailed description.

2.2.2. Data challenges

Various data challenges arise when tackling the problem of causal discovery in practice. This is mainly due to the discrepancy between the assumed hypothesis needed from each method or framework and the real-world data. One of the most common assumptions of most causal discovery methods is causal sufficiency, which is the hypothesis that all relevant variables are observed. Unobserved variables are especially problematic when they are possible confounders between system variables since omitting confounders from the causal discovery could lead to learning spurious or wrong relationships. There are some available methods which do not assume causal sufficiency, such as LPCMCI [127], SVAR-FCI [126] and GPS [180] (see Section 2.1.3).

Missing data and selection bias are other common issues in real-world applications, and there have been some efforts in developing causal discovery methods which are resilient to these challenges [226–229]. Not always, especially in the physical realm, the data follow predictable and well-behaved distributions. One such example is zero-inflated data, which is common, for instance, in gene expression data, where single-cell expressions lack detectable values of transcripts that appear abundant on bulk (thousands of cells) gene expression experiments. Recent advances have developed graphical models and causal discovery methods in such scenarios [230,231].

2.2.3. Statistical and computational challenges

The high dimensionality of data in physical systems, such as spatiotemporal data, and small sample sizes are central statistical challenges for causal discovery. On the other hand, large sample sizes raise issues of unaffordable computational time, which can scale up to cubically for kernel methods typically used for independence testing [194,232]. High-dimensional data leads to large conditioning sets in particular algorithms, effectively reducing the sample size available to test the hypothesis.

As noted in Section 2.2.1, non-linearity is a common characteristic of processes in the physical sciences. For the case of independence-based or hybrid casual discovery techniques, this calls for devising non-parametric tests of independence, for instance, tests based on measures of conditional mutual information [145], or on Gaussian process regression or other kernel-based measures on independence [194], or using quantile regression [233] and copula-based methods [234] (also applied to Granger causality), etc. The no-free-lunch theorem of [235] states that no single conditional independence test can have power against all alternatives. Here, the challenge lies in devising and applying (a combination of) conditional independence tests that are the most suited for a particular physical system.

The concept of post-selection inference [236] involves performing statistical inference on a model that was selected based on data-driven methods rather than being pre-selected. While there are some advances in solving the post-selection inference problem for regression and causal effect estimation, few solutions have been proposed for the inference after causal discovery setting [236–238]. One possible solution is sample-splitting, but this is often statistically inefficient. A recent development is the randomised version of the greedy equivalence search (GES) algorithm, which allows for finite-sample correction of classical confidence intervals [179].

2.3. Opportunities for the physical sciences

The field of causal discovery from observational data is still in its infancy but growing in methodologies, theoretical guarantees of performance, and empirical evidence. Causal inference, in general, is a vast field that offers alternatives and scientific opportunities that we review in what follows. Only revisiting the whole body of empirical science based on association would take a village, but the advances would pay off.

2.3.1. Causal hypothesis testing and targeted interventions

> *"Observational causal discovery offers revolutionary opportunities to test hypotheses, evaluate the impact of interventions, attribute extreme events with counterfactuals, and characterise complex systems by deriving causal pathways and robust forecasting models."*

Scientists need a principled way to test different hypotheses against each other. A causal hypothesis is a supposition or theory about how things interact, specifically on whether one thing causes another. Causal studies aim to confirm or reject any given causal hypothesis. The problem is that hypotheses in the physical sciences are often presented as narratives giving a chain of causal factors that lead to the studied phenomenon. Without a causal vocabulary and analytical tools, it is often impossible to precisely state the hypothesis, which leads to several competing hypotheses or, even worse, a false hypothesis, which is accepted as true due to its compelling narrative quality. Testing hypotheses have been conducted in myriad ways [3,9,239].

Causal graphs, as graphical representations of assumed or learned causal relations, provide a more principled way to talk about causal hypotheses. Learned graphs imply causal links and pathways and provide evidence for deciding between rivalling causal hypotheses, in Kretschmer et al. [240], for instance, regarding competing hypotheses of Arctic climate teleconnections.

The conclusiveness and interpretability of discovered causal graphs from purely observational datasets on the often untestable validity of the methods' assumptions and the statistical complexity of the task. But observational causal discovery can help inform more targeted subsequent interventions, which are often too expensive to employ on a large scale [3,239]. Incorporating interventions, if performed meaningfully, could thus make the causal discovery process

much more efficient and robust (discovered DAGs not being confined to the Markov equivalence class). Interestingly, interventions could be differentiable, i.e., 'learnable' from data [241].

2.3.2. Cause–effect estimation

Causal discovery results in qualitative causal graphs, or often Markov equivalence classes of graphs. But often, the target question is a quantitative estimate of a causal effect of one variable X on another variable Y, as pioneered by Pearl [3]. This topic is discussed in a time series context in Runge et al. [115]. The quantity of interest then is the (interventional) distribution of Y given an intervention in X, $p(Y = y \mid do(X = x))$. The fundamental problem is that typically $p(Y = y \mid do(X = x)) \neq p(Y = y \mid X = x)$. Confounders, for example, can introduce a non-causal association between the treatments and the outcome. Randomised experiments would be the gold standard by eliminating the unwanted non-causal associations [3,242,243]. The goal of causal effect estimation is to do so without access to interventions by expressing $p(Y = y \mid do(X = x))$ as a function of the observational distribution $p(\mathbf{x})$:

$$p(Y = y \mid do(X = x)) = \text{function of } p(\mathbf{x})\,. \tag{8}$$

If such a re-expression is possible, one calls the causal effect identifiable and obtains a causal estimand, which involves only the observational distribution. The most well-known method for causal effect estimation from data without parametric assumptions is *covariate adjustment* [3], which refers to de-confounding the causal relationship by adjusting for a set of variables $\mathbf{Z}$. In the general case, the adjustment formula is

$$p(y \mid do(X = x)) = \int p(y \mid x, \mathbf{z}) p(\mathbf{z}) d\mathbf{z}\,. \tag{9}$$

Recent work has focused on finding statistically optimal adjustment sets [244], i.e., for which the estimators have minimal variance. Using the *do*-calculus [40,245–247], it is possible to determine whether a causal effect is, in principle, identifiable from observational data or not. To this end, causal effect estimation requires fully specified causal graphs from causal discovery (with its inherent reliance on further assumptions) or domain expertise that can qualitatively specify a causal graph. For example, it is known that temperature influences ecosystem respiration, but one may want to quantify how much when given a graph of other observed and unobserved confounding variables. The graph then encodes assumptions about the absence and presence of causal relations.

Different variants of causal effects can be defined based on the interventional distribution $p(y \mid do(X = x))$, and an estimate then involves further parametric assumptions. For example, in a linear model, the total causal effect on the expected value of Y when setting X by intervention to x' as opposed to x is given by

$$\Delta_{X\to Y}(x', x) = \Delta x \cdot \beta_{X\to Y}\,, \tag{10}$$

where $\Delta x = x' - x$ and $\beta_{X\to Y}$ can be estimated as the regression parameter of X in the linear regression of Y on $X \cup \mathbf{Z}$.

2.3.3. Causal pathway analysis and mediation

Next to quantifying the overall causal effect of X on Y, a relevant follow-up question is often about the causal pathways: the mechanisms by which this effect propagates. In complex systems, it is often interesting to analyse how perturbations spread throughout the systems and through which subprocesses perturbations are mediated [124,171]. Within the structural causal model framework, mediation formally leads to counterfactual quantities, see, for example, VanderWeele [248] and briefly below in Section 2.3.8. But for linear models, the mediated causal effect (MCE) of X on Y that passes through a mediator M (here $X, Y, M \in \mathbf{V}$) can be computed by summing up the contributions along all paths passing through it:

$$\mathrm{MCE}(X, Y|M) = \sum_{\pi_k^M} \prod_{\lambda_{i\to j}\in\pi_k^M} \beta_{i\to j}\,, \tag{11}$$

where the summand iterates over causal paths π_k^M from X to Y through M and the product is over all links $\lambda_{i\to j}$ on each path. The link coefficient $\beta_{i\to j}$ can be estimated as the regression coefficient of V_t^j in the linear regression of V_t^j on the parents of V_t^j. Mediation analysis can also answer the complementary question: how strong is the direct effect of X on Y.

2.3.4. Identifying causes and pathways leading to anomalies

Anomaly detection [249] is an important problem in engineering and the physical sciences. In Earth sciences, extreme events form a subclass of anomalies and can be structured across different dimensions, such as compound extremes [250]. While detecting anomalies is an important problem, it does not answer the often relevant question of what causes a particular anomaly or, more generally, what causes the anomalous process. In engineering, business science, and healthcare, a related problem is *root cause analysis* [251]. Causal discovery can address the problem of identifying causal drivers (parents) or indirect mediating pathways and facilitate quantitative analyses to analyse the contribution of different physical drivers in causing an extreme.

2.3.5. Causal complex network analysis

Complex systems are often viewed as networks of interacting subprocesses, for example, the human brain [252], or the Earth system [253,254]. Tools of network theory [255] have been used to analyse quantities such as the information flow as it propagates through the system or the stability of subprocesses [256]. A common network measure is the node degree, which quantifies the number of processes linked to a node. A more involved measure is betweenness centrality, which quantifies the number of shortest paths through a particular node. A crucial question is then to define what these paths mean. In works where the networks are based on pairwise correlation or mutual information [253,254], one may associate paths with a transfer of information.

However, there is a difference between information being transferred versus perturbations propagating through the network. Here a question can be to identify how critical individual subprocesses are in spreading and mediating perturbations in such dynamic complex systems. The propagation of perturbations, aka interventions, relates to a causal question requiring a causal definition of network links able to distinguish direct from indirect interactions.

In addition, the toolbox of classical network measures is not rich enough for quantifying gateways and mediators of perturbations. Essentially, these measures—with many originating from the social sciences [257]—are based on a different definition of links, for example, two persons knowing each other, as opposed to dynamical interactions in a complex system. Hence, here measures based on causal pathways on which perturbations propagate in a complex system's interaction network can be utilised, such as those studied in Runge et al. [124] and Runge [171]: Identifying the nodes of the causal graph with the components of the complex system, the average causal effect can be defined as a causal version of the out-degree or closeness centrality, which quantifies by how much an individual component causes any of the remaining components. This serves as a quantitative measure of how much a component is a gateway of perturbations. On the other hand, the average causal susceptibility measures how much a component is changed on average by a perturbation in any of the remaining components as a causal version of the in-degree or in-closeness. Finally, the average mediating causal effect measures how much of the pairwise causal effects between any pair are mediated through a particular variable, which can be seen as a causal version of betweenness centrality. In Runge [171], these measures are generalised in an information-theoretic framework.

2.3.6. Causally robust forecasting models

Forecasting a time series from multivariate predictors constitutes another problem where causal knowledge helps. Even considering the case of forecasting inside the same distribution, that is, assuming a stationary distribution, it can be proven information-theoretically that causal predictors maximise the mutual information with the target variable and, by the Markov property, any further predictors do not add further information. More formally [258], the negative log-likelihood can be decomposed as follows

$$\lim_{n\to\infty} -l = \lim_{n\to\infty} -\frac{1}{n}\sum_{t=1}^{n} \log \widehat{p}\left(X_{t+1} \mid \mathcal{P}; \theta\right) \tag{12}$$

$$= \mathbb{E}\left[\log \frac{p(X_{t+1} \mid \mathcal{P})}{\widehat{p}(X_{t+1} \mid \mathcal{P}; \theta)}\right] + \underbrace{I\left(X_{t+1}; \mathbf{X}^{-}_{t+1} \setminus \mathcal{P} \mid \mathcal{P}\right)}_{=0} + H\left(X_{t+1} \mid \mathbf{X}^{-}_{t+1}\right), \tag{13}$$

where n is the sample size, $\widehat{p}\left(X_{t+1} \mid \mathcal{P}; \theta\right)$ the prediction model for $X_{t+1} \in \mathbf{X}_{t+1}$ of the true underlying $p\left(X_{t+1} \mid \mathcal{P}\right)$ given its causal parents $\mathcal{P}$ and model parameters θ. As shown, the log-likelihood decomposes into the model approximation error given $\mathcal{P}$ (first term), the conditional mutual information between the target and unselected variables $\mathbf{X}^{-}_{t+1} \setminus X_{t+1}$ given $\mathcal{P}$ (second term, zero by the Markov condition), and the irreducible entropy or uncertainty (last term). This is especially relevant for finding optimal sets of predictors in the case where greedy selection strategies do not work because the predictors cause the target variable synergistically, for example, $X_{t+1} = Z^1_t \cdot Z^2_t + \eta_{t+1}$. As shown in Runge et al. [259], an optimal subset selection can be better performed on the smaller subset of causal predictors. In Kretschmer et al. [260] and Di Capua et al. [261], causal pre-selection was used in a climate context Beyond stationary distributions, Huang et al. [262] address the task of causally-informed forecasting under nonstationary environments through state–space models.

2.3.7. Physical simulation model evaluation

Causal graphs and causal effects can be utilised to intercompare the output of physical models and evaluate and validate them against observations at the level of causal dependencies [45,263–265]. One approach in this direction is to compare the causal graph obtained from the observational data to those obtained from simulated data. This procedure has been proposed in the climate sciences to compare climate model simulations and observational data through their corresponding causal graphs derived from PCMCI [265], cf. Section 4. The methodology could be adapted and applied to other physical science problems where one typically has complex datasets and simulations to confirm hypotheses.

2.3.8. Counterfactual causal attribution of extreme events

Counterfactual questions are not about the distributions of a target variable due to possible (future) interventions, but about the distribution of a target variable for an alternative past intervention, given that a particular outcome was observed. Formally, just like interventional causal queries are represented by interventional SCMs, counterfactual queries

are represented in counterfactual SCMs [3,9]. Given an SCM over variables $\mathbf{V}$ and observations $\mathbf{v}$, in the *counterfactual SCM*, the noise distribution is updated such that the $\mathbf{V} = \mathbf{v}$ holds. Then the noise terms may not be independent anymore. Counterfactual queries are then *do*-statements in the counterfactual SCM. One example of a counterfactual distribution query is $p(y'_{x'}|y_x)$, which specifies the probability of observing $Y = y' \neq y$ under the hypothetical past intervention $do(X = x')$ when, in fact, $Y = y$ was observed under the intervention $do(X = x)$. Such queries can be computed in different ways [3,266] and generally require more assumptions about the underlying structural causal model than causal effect questions or causal discovery. Next to counterfactual distributional queries, Halpern [267] discusses causation and counterfactuals regarding single events.

An example of a counterfactual question in climate is the causal attribution of extreme events [268]. The above query $p(y'_{x'}|y_x)$ is one specific type of a counterfactual question and is sometimes called "probability of necessity" (PN), which is typically the quantity of interest in lawsuits. Extreme event attribution requires to study of anthropogenic forcings compared to their absence, that is, solely natural forcings or internal variability of the climate system. If the probability of necessity is high enough, then a human-caused extreme event is established.

2.3.9. Signal tracking for the discovery of proximal causes

Many phenomena, such as extreme events in complex systems, such as El Niño events in the climate system and extreme volatility in the financial system, are caused by an initial anomaly that triggers a travelling cascade of events [269]. This phenomenon is often called the "butterfly effect",[4] characterised by an anomaly in one part of a system having extreme consequences in another space and time. Such cascaded events are challenging to detect, predict, understand and characterise [270,271], and has led to the development of the field of the science surrounding the concept of predictability in complex systems [272–274]. A potential strategy for uncovering the cause of notable events is causal discovery, for instance, by conducting a simulation which begins from the start of the event in question and tracing the initiating signal back to its source. However, this presents difficulty in determining the initiating trigger. It is challenging to have a dialogue about recognising the drivers of intense impacts because the amount of correlated drivers is usually much greater than the number of causally pertinent drivers, which may only have a substantial effect when combined (synergy) [10]. These kinds of relations are hard to portray with a pairwise network.

2.3.10. Causal benchmarks, software and platforms

Method development and comparison require benchmark datasets with known causal ground truth for validation. Ideally, such ground truth comes from expert knowledge of real data or actual experiments that can also be used to falsify causal relationships predicted from observational causal inference methods. Unfortunately, in many fields, such as Earth system sciences, such datasets exist only for expert-labelled causal relations among a few variables (e.g., some bivariate examples [275]). A tractable approach is to generate synthetic data with simple model systems that mimic properties and challenges of data from the system under study but where the underlying ground truth is known. These can then be used to study the performance of causal discovery (and causal inference methods more generally) for different challenges in realistic finite sample situations. From a practitioner's perspective, it is essential to determine which method is best suited for a particular task with particular challenges and for a specific set of assumptions. Synthetic data, adapted to the problem at hand, can be used to choose the suitable method, including method parameters. A list of key methods for causal discovery and the available software and platforms is given in Table 3. An example is the SAVAR model [219] that mimics spatio-temporal features of climate data. The website causeme.net [10,276] aims to provide an open platform with synthetic models mimicking real data challenges on which causal discovery methods can be compared. Next to method comparison, the platform also calls for submissions of actual and modelled data sets where the causal structure is known with high confidence and was used on the Causality 4 Climate NeurIPS competition [276]. That competition sparked the investigation of a particular property of synthetic data and models called *var-sortability*, which led to new insights in causal discovery methods [277].

2.4. Perspectives

This section reviewed causal discovery in the physical sciences, describing the main methods, challenges, and opportunities for future research. We laid out the fundamental elements of the causal discovery framework—SCMs, graphs, and associated distributions—and gave an overview of the methodological concepts of learning qualitative causal graphs [10,115]. We deliberated on commonplace difficulties encountered in the field, such as determining and preprocessing causal variables, addressing non-stationarity, contemporaneous causation and hidden confounding, and selecting parametric models for nonlinear dependencies and non-Gaussian distributions. Section 4 will illustrate causal discovery methods in neurosciences and Earth sciences case studies.

The body of causal inference has traditionally been embedded in several communities, mainly statistics, social sciences, econometrics and health sciences. Irreconcilable positions and long-standing discussions exist [281]. Pearl argues that

[4] The term is attributed to Lorenz when he noted that a weather model failed to reproduce the results of runs with the unrounded initial conditions. However, the idea was earlier recognised by Poincaré and further formalised by Wiener. The analogy became popular and originated the quantitative science of characterising *instability* in complex systems undergoing nonlinear dynamics and deterministic chaos.

Table 3
Methods and open-source software for causal discovery.

Method	Software
Granger causality (GC) [35], kernel GC [167], explicit KGC [119]	causal-learn, statsmodels, KGC, XKGC
CCM [39,278]	rEDM
PC [43,279], FCI [176]	Tetrad, causal-learn, pcalg, Tigramite, MXM, bnlearn, dbnlearn, PyWhy
PCMCI [170], PCMCI$^+$ [123], LPCMCI [127]	Tigramite
DYNOTEARS [132]	Causalnex
TiMINo [135]	R script
VARLiNGAM [191]	causal-learn, lingam, original R code
ICP [148,196,280]	seqICP

it is essential to distinguish between causal and statistical information, as they refer to two separate concepts,[5] and suggests that clear distinctions should be established in the notation used, and each should be subject to different means of calculation [282]. Arguably, nonparametric SCMs (as a natural generalisation of those used by econometricians and social scientists in the 1950–1960s) have developed the field of causality in new mathematical underpinnings: explicate and enumerate causal assumptions, test implications, decide measurements and experiments, recognise and generate equivalent models, recognise instrumental variables, generalise structural equation models and solve the mediation and external validity problems. These tools, methods and solutions help to determine the accuracy and validity of causal claims in the analysis. The machine learning community is approaching the field of causal discovery in innovative ways by leveraging data, assumptions and models collectively. In recent decades, mathematical foundations have been established to address questions of causality in various scientific fields, mainly emerging for statistics and machine learning [3,9,43]. The causal machine learning (CausalML) field has recently introduced [283] as an umbrella for machine learning methods based on SCMs. It aims to advance the field in several directions: causal supervised learning, causal generative modelling, causal explanations, causal fairness, and causal reinforcement learning. Applications of the new methods are vast and promise advances in computer vision, natural language processing, and graph representation learning. Therefore, the field of causal discovery is growing in methods, approaches and impactful applications. A unified agenda for Causal Inference is built and deployed in the wild.

However, despite the significant advances in the last decades, many unresolved philosophical and methodological issues remain for causal discovery from observational data. All such challenges also create avenues of research On the one hand, we identified and discussed algorithmic and data challenges and summarised possible ways to address them in Section 2.2. Indeed, we must develop more effective methods for incorporating (uncertain) expert knowledge, determining the spatio-temporal complexity of the underlying dynamic phenomena, and creating more reliable and statistically efficient algorithms.

On the other hand, perhaps the most critical challenge is the theoretical impossibility of causal discovery from purely observational data [8]. There are, however, ways to tackle the challenge. For example, specifying a causal DAG using domain knowledge can help mitigate the potential inaccuracy of their assumptions of sufficiency and faithfulness. Another possibility to learn about a certain equivalence class may consider incorporating domain knowledge into structure learning algorithms by using "allow lists" and "deny lists" to determine which edges should or should not be included in a DAG or creating a Bayesian prior to assigning varying levels of probability to certain causal relationships [284,285]. This is very much related to using *inductive biases* (such as Occam's razor) [286] and causal invariances (such as parameter modularity and independence of mechanism) [150] to learn structure beyond likelihood-based scores and conditional independence constraints. Finally, if data from natural experiments, such as $do(A = a)$, is available, this intervention information can be incorporated into the algorithm [287] to automate this reasoning process. Incorporating the abundant domain knowledge within the causal discovery routine can address the identifiability and faithfulness assumptions (very much in line with the basis or sparsity priors used in equation discovery, cf. Section 3). By joining forces, both can contribute to resolving pressing scientific issues, ranging from process comprehension to evaluating and upgrading the physics included in physics models.

Many problems in the physical sciences can be framed as causal questions. Yet many researchers in economics and health services, and even many computer scientists in machine learning, have been trained to be reluctant to use the language of causality [57,281,282,288]. This is a cognitive barrier to resolve in the future. Besides, the language barrier between the methodological and domain science communities is a significant challenge in the causal discovery endeavour. Bridging this divide by translating domain questions into actionable and precisely stated causal inference tasks seems reasonable. Additionally, hesitance to adopt causal inference can be attributed to the lack of suitable benchmarks to help choose an appropriate method. A benchmarking platform (https://causeme.net) was introduced that covers the causal discovery problem setting. It is necessary to have more of these benchmark platforms and easily accessible databases to facilitate better collaboration between the two communities. To successfully address a causal inference problem, it is

[5] Statistical information deals with the probability of certain variables being observed. In contrast, causal information deals with hypothetical relationships in new situations.

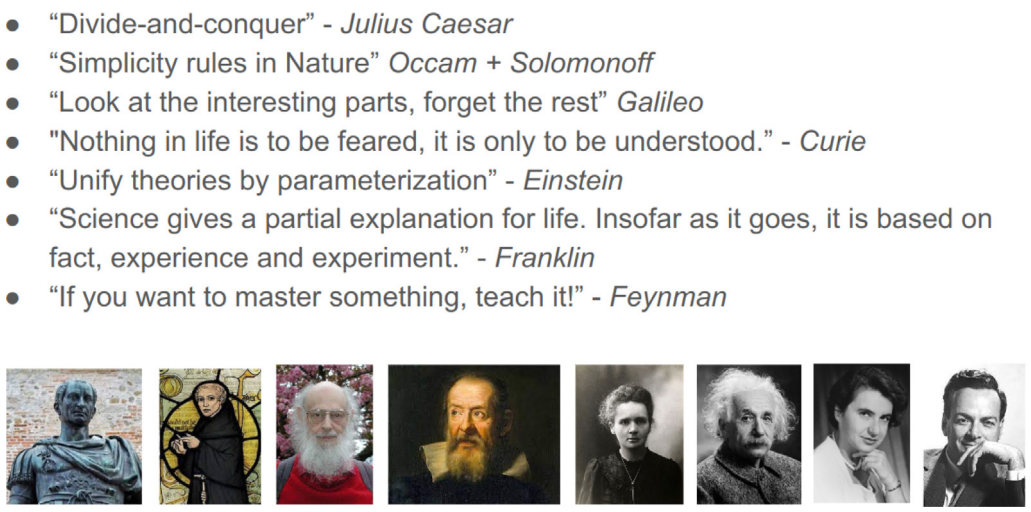


Fig. 5. Historical Inspirations/Motivations in Law/Equation Discovery.

essential for the domain scientist and computer scientist to work together; assumptions must be discussed and formalised, data characteristics must be jointly analysed, and conclusions must be assessed from both perspectives.

It is commonly believed that most research questions in science can be interpreted as causal inference problems. Sentences such as "we find that X [increases/decreases/lags/leads/affects/drives/impacts] Y" are often found in papers, creating an impression of causality. Nevertheless, scientists should be more explicit and transparent when making assumptions which lead to causal conclusions. This is not only sensible but is also necessary when analysing complex systems like the Earth, the Brain, or the Economy, since the outcomes of such research may have significant economic, environmental, and social implications. The field of Causal Inference provides a comprehensive arsenal of tools grounded in rigorous mathematical principles and a vibrant interdisciplinary milieu to confront the challenge at hand successfully.

3. Learning physical laws from data

"For centuries, scientists observed Nature to extract simple laws and equations to explain the world mechanisms, to anticipate and predict behaviours and gain faith in their interventions/actions."

Distilling mechanistic models of the world is how physicists have successfully understood and explained the natural world. The prototypical process starts with an experiment or observation. A mathematical model is hypothesised, which can predict a new experiment's outcome. The observations will either support or falsify the hypothesis, leading to more experiments and refined models.

For many complex systems, we have poor models because certain interaction terms are unknown. Another case is when we know some microscopic interaction laws. Still, the emergent properties at a larger scale do not directly follow, such that predictions at the larger scale need new coarse-scale interactions. To cope with these situations and make sense of the sheer amount of data produced by modern instruments, researchers have been looking into automating the processes involved in model building and creating new insights. Many motivations and inspirations have been adopted to guide the scientific method development (see Fig. 5): e.g. a physical law or equation describing the system should be compositional, thus following the "divide and conquer" rule, should be as simple as possible (but not simpler) thus following Occam's razor, further developed and formalised by Solomonoff, eventually focus on the interesting parts of the system and disregard the rest, create understanding, generalise and unify different working theories and models, and the learned representation (and its parameters) should be self-explainable, amenable and intuitive.

This section reviews the state-of-the-art in equation discovery from data. Unlike in traditional law discovery *à la Kepler* where trial-and-error was dominant, modern statistics and machine learning techniques exploit the regularities found in the data to discover plausible, simple and explainable equations, to learn feature representations that describe (typically dynamic) systems. We will first consider the *explicit* discovery of equations that describe observed data, also called *Symbolic Regression*. Second, we look into *implicit* discovery through dimensionality reduction techniques and transfer operators. We finish the section by discussing the main challenges and research opportunities.

3.1. Explicit equation discovery with symbolic regression

Symbolic Regression (SR) refers to a class of machine learning techniques that aim to discover mathematical relationships and patterns in data. SR aims to find a compact, human-readable mathematical expression that accurately reflects the

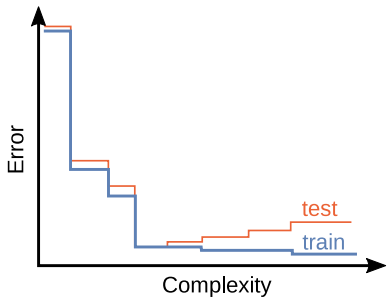


Fig. 6. Illustration of a Pareto curve of solutions to an SR problem.

underlying relationships in the data. This approach is particularly useful in cases where the relationships between variables are complex or unknown, and traditional statistical methods may not provide adequate explanations. SR operates under the assumption that the underlying data-generating mechanism can be described by a sparse and algebraic input–output relationship.

There is a major divide in the method to achieve that. One class of methods use genetic algorithms or other discrete search methods to find mathematical terms, typically represented by a graph of mathematical operations. Another class of methods uses continuous space search methods and solves a relaxation of the discrete problem of finding compact equations. A third and most recent addition to the family of symbolic regression methods uses massive amounts of synthetic data for pretraining a system that can quickly guess a suitable expression at test time.

More formally, we are trying to solve the following optimisation problem:

$$\arg\min_{f} \mathbb{E}_{(x,y)\sim\mathcal{D}} \|f(\boldsymbol{x}) - \boldsymbol{y}\|^2 + \lambda C(f) \tag{14}$$

where we attempt to find a low-complexity function (equation) f that best[6] maps the inputs $\boldsymbol{x} \in \mathbb{R}^n$ to their corresponding outputs $\boldsymbol{y} \in \mathbb{R}^m$ in the data distribution $\mathcal{D}$, $C(f)$ refers to a measure of complexity of f, and λ is the weighting factor. For instance, the complexity measure could be the number of terms in the equation.

A central problem when performing symbolic regression is selecting an appropriate weighting factor. More generically, the question is which level of complexity is right.

There is probably no definite answer to this question. Instead, we consider the solution to a symbolic regression problem as a family of Pareto-optimal solutions:

$$f^{(c)} = \arg\min_{f} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}} \|f(\mathbf{x}) - \mathbf{y}\|^2 \qquad \text{s.t.} \quad C(f) = c \tag{15}$$

where $f^{(c)}$ refers to the best fitting expression with complexity c. A more complex expression will be able to fit the given data at least as well as a less complex expression. Fig. 6 illustrates a typical Pareto curve along the optimisation objectives: goodness of fit (error reduction) and function complexity. In addition to the training error, which monotonically decreases with function complexity, the illustration also shows a hypothetical test error that shows the overfitting of too complex functions.

We will now look more closely into different methods that have been proposed to solve the optimisation problem Eq. (15) in practice.

3.1.1. Symbolic regression using discrete search methods

The problem of symbolic regression is, at its core, a search for suitable functions f in (15). Since those functions should have low complexity, it is natural to attempt to perform a search for functions. The first attempt to do that was proposed by Cramer [289] by inventing Genetic Programming, which got popularised and applied through Koza [290,291]. The idea is simple: search for computer programs to solve a particular problem by iteratively creating many random programs and selecting the best fit, and create a new pool of candidates by recombination and random modification. This mimics the biological evolution process of nature to create the genetic material of living organisms. Applied to symbolic regression: The functions are represented as a graph of input variables, operators and basic algebraic functions.

In the paper by Schmidt and Lipson [76], this approach was refined and applied to the discovery of physical laws. As the method was indeed able to discover Lagrangian and Hamiltonian formulations from data, it stimulated a growing interest in symbolic regression and sparked the development of many methods. These general search methods are also referred to as *evolutionary algorithms*. The approach of Schmidt and Lipson [76] is illustrated in Fig. 7. The general method for symbolic regression was implemented in a tool called *Eureqa* [292] that is now only available as an online service [293].

[6] We use the squared error here for simplicity, but other notions of distance are possible.

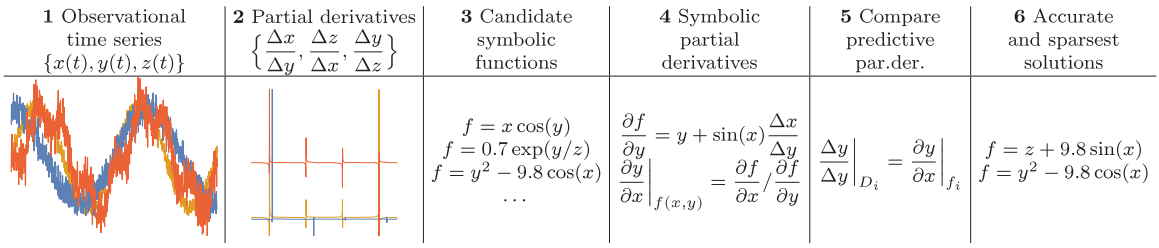


Fig. 7. Schematic view of the symbolic regression method for discovering physical laws in [76]. Starting from observational data (1), partial derivatives are computed numerically for all pairs of variables (2). A set of candidate symbolic functions f is derived (3), whose symbolic partial derivatives are computed (4) and compared to the predictive ones (5). The process 3–5 is iterated until, finally, a small set of the most accurate and simple equations is returned (6).

There are several publicly available and open-source implementations, such as the PySR [294], gplearn [295], Glyph [296] and Operon [297]. A more detailed overview of genetic algorithm-based methods and their combination with gradient descent can be found in Kommenda et al. [298].

Feynman AI. An approach that exploits physical knowledge such as units and makes reasonable assumptions for equations in physics is Feynman AI [102,299]. The method augments the genetic algorithm searching for expressions by enforcing fitting physical units, decomposing the problem using symmetries and checking separability. To check for symmetries, a neural network is trained on the data to allow accessing whether the underlying function is symmetric. It is worth noting that a large amount of data is used here. From a set of 100 equations taken from the Feynman Lectures, the method was able to recover all of them whereas Eureqa only solves 71.

Search with deep reinforcement learning. A method that uses Deep Reinforcement Learning to search for a suitable solution to the symbolic regression problem is Deep Symbolic Regression (DSR) [300]. The key idea is to treat the search for expressions as an exploration problem in reinforcement learning (RL). The functional expressions are represented as a sequence of tokens corresponding to a depth-first graph traversal and are generated by a recurrent neural network. Numerical constants are fitted using the BFGS optimiser. This generative network is trained on the given dataset using RL to find a highly fitting solution. An interesting contribution is a formulation of a risk-seeking policy gradient that tries to optimise for the best-case scenario (a good solution can be found) rather than the typical average case. The method was able to solve 83% of the standard Nguyen-1 dataset.

3.1.2. Sparse linear regression and neural network approach

The symbolic regression problem can also be tackled via traditional regression methods. In contrast to the search in a discrete set of functions, the search is performed in a dense set, typically represented by a real-valued parameterised function. So (14) is solved by choosing a large enough function class described by f_w with $\boldsymbol{w} \in \mathbb{R}^p$. The optimisation is then performed over the space of parameter-values w. Linear regression is a special case, where the function $f_w(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x}^\top$.

What about the complexity regularisation term $C(f_w)$ in (14)? Ideally, the term should count the number of non-zero parameters in $\boldsymbol{w}$, expressed as $|\boldsymbol{w}|_0$ and referred to as L_0 norm. However, this term jeopardises the efficient solution of the regression problem because it is non-linear and non-differentiable. One practical alternative is to use the L_1 norm instead, i.e. the sum of absolute values, which also leads to sparse solutions (see Fig. 8). In the case of linear regression, this is termed LASSO regression [301].

The methods differ in the function class f_w, the regularisation term and the optimisation method used.

SINDy: Sparse identification of dynamical systems. In some cases, the class of building blocks that might occur as summands in the analytical description of the data are known. Then a rather simple but effective method can be employed that is called *sparse identification of dynamical systems*, SINDy for short. It was proposed in Brunton et al. [12] to find differential equations of dynamical systems from observations. For the general symbolic regression problem, the FFX method by McConaghy [302] was already earlier proposing the same idea. The input data is passed through a predefined library of base functions and interaction terms. Then the resulting high-dimensional representation is fit to the data using sparse linear regression. All relevant terms keep a non-zero weight and constitute the final expression.

Let us unpack this in more detail for a dynamical system of n variables described by the system of ordinary differential equations $\frac{d}{dt}\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x})$, where $\boldsymbol{x}(t) \in \mathbb{R}^n$. Each component of $\boldsymbol{g}$ can now be substituted by a linear combination of library

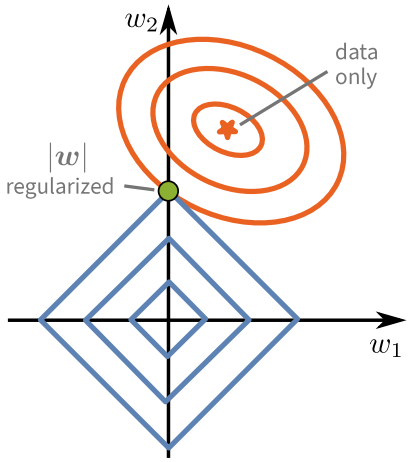


Fig. 8. L_1 regularisation typically leads to sparse solutions. The lines show the isolines of quadratic loss (red) and $|w|$ (blue). Instead of the data-only solution (red star), a sparse solution (green) is found.

functions:

$$\begin{aligned}
\frac{d}{dt}x_1 &= g_1(x_1, x_2, \ldots, x_n) &= w_{11}l_1(x_1, \ldots, x_n) + \cdots + w_{1m}l_m(x_1, \ldots, x_n)\\
\frac{d}{dt}x_2 &= g_2(x_1, x_2, \ldots, x_n) &= w_{21}l_1(x_1, \ldots, x_n) + \cdots + w_{2m}l_m(x_1, \ldots, x_n)\\
&\vdots\\
\frac{d}{dt}x_n &= g_n(x_1, x_2, \ldots, x_n) &= w_{n1}l_1(x_1, \ldots, x_n) + \cdots + w_{nm}l_m(x_1, \ldots, x_n),
\end{aligned}$$

where $l_1, \ldots, l_m$ is the *predefined finite library of candidate functions* and w_{ij} are the scalar coefficients to be learned following our objective (14). To obtain a sparse solution, the L_1 regularisation (LASSO) can be used, i.e. $C(f) = |\boldsymbol{w}|_1$, as described above. Alternatively, the (squared) L_2 norm of the weights $\|\boldsymbol{w}\|_2^2$ can be used, corresponding to classical ridge regression that permits a closed-form solution. However, an iterative pruning of small weights must be used to obtain a sparse solution (see [303]).

Illustration of sparse identification of dynamical systems (SINDy) [12] using a synthetic dataset of the well-known Lotka–Volterra system, as shown in Adsuara et al. [303]. The Lotka–Volterra system models the interaction between prey and its predator in ecology and is given by the following equations:

$$\begin{aligned}
\frac{d}{dt}x_1 &= \alpha x_1 - \beta x_1 x_2\\
\frac{d}{dt}x_2 &= -\gamma x_2 + \delta x_1 x_2
\end{aligned}$$

with the coefficients, α and γ, being the intrinsic growth/decrease rates of x_1 and x_2, and β and δ are cross terms taking into account the interaction between species. In our particular case, we will set $\alpha = 3/2$, $\beta = 1$, $\gamma = 3$, and $\delta = 1/2$. We show the results of the identification of these parameters using SINDy in the table below for two levels of additive white Gaussian noise of the signal-to-noise ratio of 5 (high noise level) and 40 dB (low noise level). As usual, the created data was split into train/test data (75%, 25%), respectively. The ODE coefficients are recovered sufficiently well to achieve a high correlation coefficient R but are generally underestimated due to the sparsity regularisation.

Library functions	**Learned Coefficients** 40 dB		5 dB		**True Coefficients**	
	$\frac{d}{dt}x_1$	$\frac{d}{dt}x_2$	$\frac{d}{dt}x_1$	$\frac{d}{dt}x_2$	$\frac{d}{dt}x_1$	$\frac{d}{dt}x_2$
x_1	1.3822	0	1.1404	0	1.5	0
x_2	0	−2.9123	0	−2.7946	0	−3
x_1x_2	−0.9797	0.4849	−0.9520	0.4710	−1	0.5
x_2^3	0	0	0	−0.0001	0	0
R	0.9999		0.8674			

For fitting dynamical systems, the temporal derivatives need to be computed. Finite differences are often too sensitive to noise such that kernel regression (aka Gaussian processes), which allow for explicit derivative computation, are preferred [304,305]. A more recent approach is to solve noise estimation and model identification in one joint optimisation

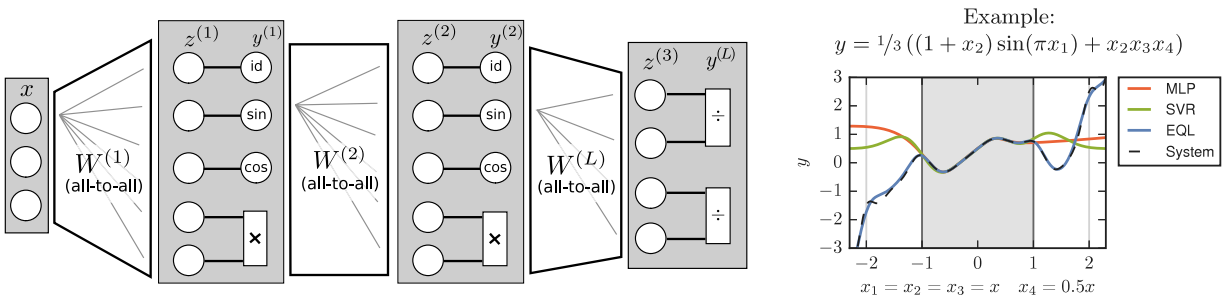


Fig. 9. Equation Learning Architecture and example equation. Left: Illustration of the EQL Architecture, reproduced from [101], a feed-forward neural network with special *activation* functions (sin, multiplication etc.). Note that each unit type will occur many times. Right: Example system with four inputs $x_{1,2,3,4}$ and one output y. Training is only in the $[-1, 1]^4$. EQL recovers the equation and extrapolates [101].

procedure [306]. Intuitively, for every data point, the corruption by noise is estimated. Since this optimisation problem is highly underdetermined, an additional constraint is used, namely that when integrating the estimated dynamical system model a small error should occur. This trick separates noise from the signal and leads to an improved estimation quality.

Neural network approach: Equation learner. Enlarging the function class f is possible using neural networks. Probably the first work in this direction is the Equation Learner (EQL) introduced in Martius and Lampert [307] that uses a neural network with algebraic base functions and a particular regularisation scheme to solve (14) and (15). The function f is represented by a neural network, modified only to contain elementary operations that should appear in a potential solution. Fig. 9(left) shows the architecture of the Equation Learner (EQL) in a simplified form.

The input variables are mapped with a dense layer to multiple instances of trigonometric functions, identity, multiplication and division, but more base functions, such as squares or exponentials, are possible. The resulting values are again mapped with a matrix to another layer of elementary functions, and so forth, until the last layer corresponds to the output (containing only division operators in the picture).

The network is trained using stochastic gradient descent (e.g. Adam) on the mean squared error loss and L_1 regularisation on the weights to induce sparsity:

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}} \|f_W(\boldsymbol{x}) - \boldsymbol{y}\|^2 + \lambda|\boldsymbol{W}| \tag{16}$$

where f_W denotes the neural network with parameters $\boldsymbol{W}$. Note that the system needs to be differentiable for training, such that a pure complexity term, such as L_0 regularisation that would count the number of non-zero weights, does not work out of the box. Although methods have been developed since then [308], the L_1 regularisation in (16) is effective but creates an undesired trade-off between error and sparsity. Something that we also encountered in the illustrating example when applying SINDy. The EQL method introduces an additional regularisation phase after converging with L_1 that clamps all $|W_i| < \epsilon \ll 1$ to zero and optimises without regularisation. This yields a practical approach to optimising for sparsity without trade-offs. In Kim et al. [309] an alternative to L_1 with $L_{0.5}$ was used.

The reader may wonder how the system is successfully trained with elementary functions such as division or square root. Indeed, a naive application would fail due to exploding values or gradients. In Sahoo et al. [101] and Werner et al. [310], suitable parameterisations and training steps are proposed. Choosing different λ (Eq. (16)) will create differently sparse resulting networks. Each represents a particular symbolic expression resembling the Pareto curve illustrated in Fig. 6. In Fig. 9 (right), a synthetic example system is shown. The training data is only generated in the $[-1, 1]^4$ hypercube. The correct equation was discovered, and perfect extrapolation is possible in this case, see Sahoo et al. [101] for details.

Instead of manually selecting a particular solution, which might be a good procedure when structural insights are to be obtained when investigating some unknown phenomenon, one can also use several or all solutions along the Pareto curve to estimate uncertainty about the predictions for extrapolation, as proposed in Werner et al. [311]. Fig. 10 illustrates this approach. For each found equation, a Laplace approximation allows to approximate the uncertainty due to parameter estimation errors and yields a Gaussian posterior. Combining these using a weight based on the validation error and the complexity yields the estimated density. Note how the uncertainty in extrapolation shows clearly the structure of the discrete set of automatically generated hypotheses.

3.1.3. Learning to solve symbolic regression

All methods so far treat every symbolic regression problem in isolation — the search or optimisation algorithm was applied to a new dataset from scratch. We are now looking into the idea of learning to solve a particular problem quickly by using data from a whole class of symbolic regression instances. Generally, the idea is to approximate the inverse mapping from data to a suitable equation. The Dreamcoder paper by Ellis et al. [312] showed the first instantiation of this idea. Provided with the language of algebraic expressions (arithmetic operations, variables, base functions) and a *simulator* to generate data for a particular equation instance, the method learns a probabilistic mapping from data to equation terms

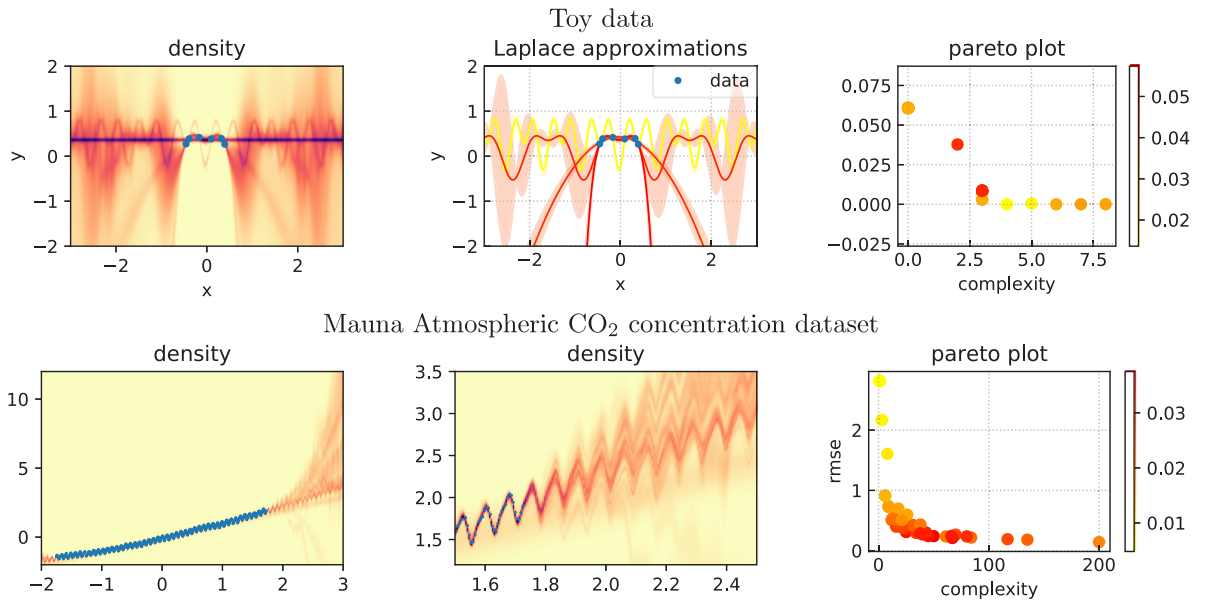


Fig. 10. Illustration of uncertainty estimates using a mixture of Laplace approximations of learned equations. Top row: toy example $y = 0.8\cos x - 0.4 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.03^2)$ with just 6 datapoints. Bottom row: Atmospheric CO_2 concentration at Mauna Loa Observatory [315] (concentration vs. time, both in arbitrary units). The left panels show the predictive distributions. The panel in the middle shows individual local Laplace approximations with 2σ (shaded area) for the toy data and a zoomed density for the Mauna dataset. The colour represents the weight and aligns with the Pareto plots on the right side, showing RMSE over the complexity of each equation.
Source: Reproduced from [311].

and a library of common equation building blocks. Given a particular instance of data, the system can relatively quickly guess and verify suitable explaining equations. Developing the idea of pretraining further and specialising it for symbolic regression was done by the following method.

NeSymReS. The approach in Biggio et al. [104] is to use a high-capacity transformer model pretrained to solve the symbolic regression problem. The method is called *Neural Symbolic Regression that Scales* (NeSymReS). The method uses a large set of symbolic regression problems to approximate the inverse mapping from data to equations. After this pretraining phase, given new data, the inverse mapping can generate likely candidate equations. Intuitively, an experienced data scientist might solve the problem similarly: looking at the data and postulating a particular functional form that might explain it, testing it, and potentially trying a different plausible hypothesis. Let us look closer at the method. As visualised in Fig. 11, the core is a transformer[7] architecture [314] that can generate algebraic expressions symbol-by-symbol given a set of data points.

The input is represented as a set of $(\boldsymbol{x}, \boldsymbol{y})$ pairs (1024 in Biggio et al. [104]), which are processed through a set-encoder. The latter is invariant to permutations of the data points. The output of the transformer are tokens that correspond to the typical symbols of input variables, base functions, operators and constant placeholders, which resemble the skeleton of the predicted function. Importantly, the transformer does not have to guess the right constants, just their location in the expression, as these constant placeholders are fit to the data using non-linear optimisation (here BFGS).

Trained on millions of synthetically generated pairs of random expressions with corresponding data, the transformer does a remarkable job guessing likely equations. Importantly, the prediction is not deterministic but allows sampling of possible functions. Thus, new potential solutions can be generated and validated when new data is presented at test time until a sufficiently good fit is found or the Pareto.

Fig. 12 shows the accuracy of different SR methods for unseen equations from the Feynman and Nguyen benchmarks. The performance is presented in dependence on wall-clock time. NeSymReS is remarkably fast at finding a well-fitting expression for the data.

As a downside, the method was only shown for three input variables, and it remains to be seen how much it can be scaled in this respect. Also, the dataset used to guess an equation at test time cannot be big (currently in the order of 1000 data points) because the set transformer encoder cannot yet handle larger sets well.

[7] The transformer architecture is the basic building block of many large-scale machine learning systems, such as GPT-3 [313].

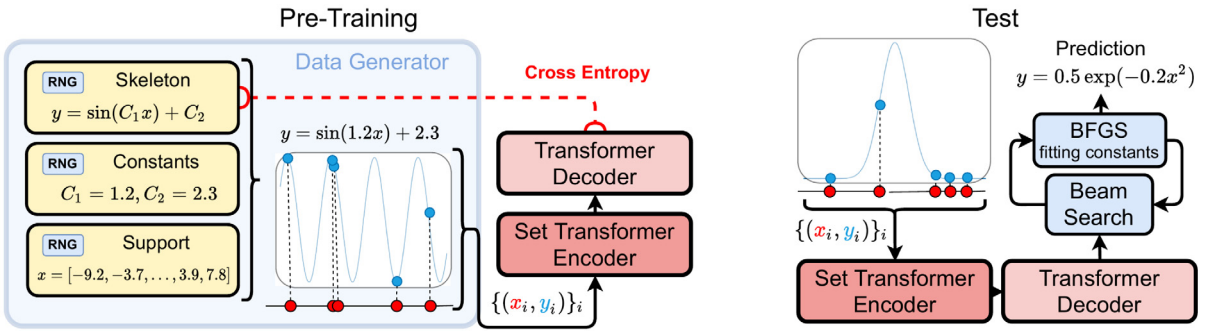


Fig. 11. Overview of the NeSymReS method. Left: step with randomly generated training data. Right: inference of candidate equations for unseen data.
Source: Reproduced from [104].

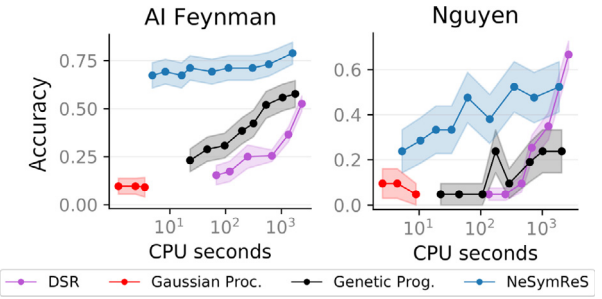


Fig. 12. Performance of NeSymReS, DSR, classical SR (using gplearn [295]), and Gaussian processes on the AI Feynman and Nguyen datasets (equations unseen during training).
Source: Reproduced from [104].

3.1.4. Comparison

As there are quite a number of methods, we aim to discuss their differences, strengths and weaknesses by comparing them along a set of axes. We start with using domain knowledge, as we seldom face a completely uninformed setting in physics. We continue with aspects of the embedding, scaling, speed and usability, summarised in Table 4.

Using domain knowledge. A common form of domain knowledge is the base functions and their approximate frequency of occurrence in describing the system under consideration. In standard symbolic regression with genetic algorithms, the number and kind of base functions are very flexible, and each term can have its individual penalty in terms of complexity. More specific domain knowledge, such as monotonicity, function image constraints and derivative constraints, can also be considered, as presented in Kronberger et al. [316]. Another recent work is presented in Cornelio et al. [317] that allows to incorporate axiomatic constraints.

In FFX/SINDy, the library of functions is the prime way to specify domain knowledge. Relative preferences could be implemented by different regularisation strengths.

For the EQL framework, the choice of base functions is a bit more complicated, as the systems need to remain optimisable with gradient descent. In Werner et al. [310], a suitable relaxation for functions with divergences (in function value or derivative) is proposed, and a way to specify preferences among base functions is analysed. The control of the relative frequency of used terms is possible but less direct than in genetic algorithm-based methods. In NeSymReS, domain knowledge can be embedded by selecting/generating the training set with appropriate synthetic problems, although this was not explicitly demonstrated.

Scaling. Most SR methods are for small-scale problems with a few hundred to a few thousand data points and low-dimensional problems, i.e. 1–10 input and output variables. Classical search methods scale unfavourably with dimensionality as the search space grows exponentially. That is why most SR methods are good at finding relatively small and compact equations for low-dimensional systems but fail for both high-dimensional systems or those where larger equations are the most compact solution.

FFX/SINDy can handle large output dimensions easily, using some form of ridge regression. However, it also suffers from large input dimensions as the library bank becomes exponentially large unless a factorisation or other simplifying structure is known, for instance, a strong locality assumption in PDEs.

Table 4
Comparison of symbolic regression methods. See Section 3.1.4 for more details.

	Embeddable	Scaling	Speed	Restriction	Domain knowledge
Genetic Programming [76]	✗	✗	Slow	For small systems	Base-functions, complexity of terms
AI Feynman [102]	✗	✗	Slow	For physical systems in canonical form	Physic: units, symmetries
DSR [300]	✗	✗	Medium	Small input dim	Training domain
FFX [302], SINDy [12]	✓	✓[a]	Blazing	Needs known library	Training domain
EQL [101]	✓	✓	Slow	Base functions limited, sometimes less concise	Base-functions, complexities
NeSymReS [104]	✗	✗	Fast	Small input dim	Training set

[a] It scales well to large output sizes; for high-dimensional input strong structural assumptions are required.

NeSymReS is also limited to several variables and small datasets. Although, the limitation of the dataset size can be lifted by sampling a smaller subset of data points for guessing the skeletons and using all data for the parameter tuning with BFGS.

EQL is the only scalable method, as gradient descent works on all dimensions simultaneously, and machine learning methods are developed to scale. A larger initial neural network should be used for larger systems, but no specific adaptations are required.

Differentiability and embeddability. An interesting feature is whether the SR system can be used as a module in a larger computational pipeline. For instance, if observations are available as images and the causal variables have to be first extracted from the images before they can be used for a concise SR prediction module. An example using SINDy inside an autoencoder architecture [94] learns the coordinate frame and dynamics equations at the same time, as detailed in Section 3.2.4 below. EQL is conceptually easy to embed into larger architectures, such as deep networks, as it is end-to-end differentiable. An example is discovering PDEs [318], or learning an energy function given observations of the dynamics in the context of density functional theory [319], discussed in more detail in Section 4.5. The other methods are more difficult to embed.

Speed. As shown in Fig. 12, symbolic regression methods are best compared in performance per compute-time, because search methods can, in principle, find the global optimum given enough time (although this time might be longer than the age of the universe), so just the "final" performance is difficult to measure. SINDy is not in this comparison because it requires knowledge of the base functions. However, it would be the fastest method, followed by NeSymReS, DSR and classic GPs. EQL is likely the slowest method on small systems because it requires a long training time. However, the time does not significantly increase with system size and amount of data.

When to use which method? For systems where we have a good idea about the occurring modules of the functional form, FFX and SINDy are probably the method of choice. They are simple and effective. For dynamical systems SINDy is the most specialised method. The other SR methods are good if the functional building blocks are unknown or nested structures are expected. Genetic Programming based methods generally shine on small problem settings. Modern implementations can also fit constants, but they result is sometimes complex nested structures. DSR and NeSymReS are faster than standard SR methods and can yield potentially less complex equations. However, less software is built for them, and they are less easy to use. For high-dimensional data or when high fitting accuracy is required, EQL might be the right choice. Also, when SR should be embedded, only FFX, SINDy and EQL are practically useable.

3.2. Implicit equation discovery with dimensionality reduction and transfer operators

This section reviews state-of-the-art methods for recovering implicit feature representations of systems from data. We will review connections among methods and emphasise the role and examples in the broad discipline of physics. Unlike in the previous section, an explicit equation is not discovered but rather an operator that encapsulates the system's characteristics (typically spatio-temporal dynamics). The field is tightly related to dimensionality reduction and feature extraction in machine learning and signal processing [320], but also to transfer operators in functional analysis [321].

3.2.1. Reduced-order models

In many domains, the goal is to study system dynamics from model simulations. In this case, equations are encapsulated in the model itself, but large-scale, high-fidelity nonlinear models can be challenging to simulate and require significant computational power. In such cases, reduced order models (ROMs) can simplify analysis and control design by trading off model accuracy for computational complexity reduction. ROM can combine complex component-level simulation models into system-level simulations used for control analysis and design.

Two main classes of techniques for building ROMs: model-based and data-driven. Model-based methods rely on a mathematical or physical understanding of the underlying model and are designed for specific PDE-based models. In contrast, data-driven methods use input–output data from the original high-fidelity first-principles model to construct a ROM that accurately represents the underlying system. Data-driven ROMs can be either static or dynamic models. Static ROMs can be developed using techniques such as curve fitting and lookup tables (LUT), while dynamic ROMs can be developed using deep learning techniques such as LSTM, feedforward neural nets, and neural ODEs. The obtained ROM ideally contains the essential physical mechanisms of the original system while exhibiting simpler dynamics that can enhance interpretability. In addition to the simpler physics, the ROM would be much cheaper from the computational point of view than the numerical integration of the governing equations, which can help obtain more efficient control and optimisation techniques.

Developing a ROM typically requires two steps. The first step is to find a set of coordinates where the original dynamics can be expressed in a compact form. This is typically done in terms of *modes*, associated with coherent features in the original system [322,323]. In the second step, it is necessary to find a set of differential equations governing the temporal evolution of the amplitudes of the aforementioned modes. These equations, which constitute a dynamical system, enable shedding light on the physics of the (reduced) phenomenon under study. A widely-used approach to perform the modal decomposition is the so-called proper-orthogonal decomposition (POD) [324], which is also known as principal component analysis (PCA) in statistics and empirical orthogonal function (EOF) analysis in meteorology [320], and is closely connected with the singular-value decomposition (SVD) [93].

Proper-orthogonal decomposition (POC), aka PCA or EOF [320,324] decomposes a dataset or a high-dimensional (spatio-temporal) field into a set of orthogonal basis functions called modes or eigenfunctions, which capture the dominant features of the data. The first few modes explain most of the variability in the data, while the later modes explain smaller and smaller amounts of variability. By truncating the number of modes, one can obtain a low-dimensional representation of the data that preserves the essential features of the original system.

Given a set of data snapshots $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$, where m is the dimension of the data and n is the number of snapshots, we seek to decompose the data into a set of r orthogonal modes $\{\mathbf{u}_i\}_{i=1}^r$, such that $\mathbf{X} \approx \mathbf{U}\Sigma\mathbf{V}^\top$, where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \in \mathbb{R}^{m\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix containing the singular values, and $\mathbf{V}^\top \in \mathbb{R}^{r\times n}$ is the matrix of temporal coefficients. The modes $\{\mathbf{u}_i\}_{i=1}^r$ can be computed by performing a singular value decomposition (SVD) of the data matrix $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ where σ_i is the ith singular value, and $\mathbf{v}_i$ is the ith right singular vector. The modes are then given by $\mathbf{u}_i = \frac{1}{\sqrt{\sigma_i}}\mathbf{X}\mathbf{v}_i$.

The POD/PCA framework enables decomposing spatio-temporal data (e.g. flow velocities, weather or climate variables which depend on the spatial coordinates and time) into a set of spatial modes (which only depend on the spatial coordinates), multiplied by their temporal coefficients (which define their change of amplitude with time). POD/PCA ensures that components are orthogonal and optimality with respect to the variance explained by a reduced number of modes, cf. Fig. 13 for an example of PCA on a toy spatio-temporal data flow. Other alternative multivariate methods, like partial least squares (PLS) or canonical correlation analysis (CCA) seek projections that maximise covariance or correlation, respectively [320]. Still, all these projection methods are linear and thus cannot cope with nonlinear spatio-temporal feature relations and complex dynamics. This can be addressed with kernel machines [320]. Oblique and nonlinear transformations can also be learned by embedding Varimax in Reproducing Kernel Hilbert Spaces (RKHS) explicitly [119,325]. Other ways to obtain non-linear transformations from observation space to ROM space, e.g. using neural networks, will be discussed below.

3.2.2. Transfer operators for learning nonlinear dynamics

Transfer operators are related to the abovementioned methods and allow the characterisation and modelling of complex dynamic systems. These operators' eigenfunctions can decompose a system given by an ergodic Markov process into fast and slow dynamics and identify modes of the stationary measure called metastable sets.

The Koopman operator is a linear operator that describes the dynamics of a system by lifting the state variables into an infinite-dimensional Hilbert space. Thus, it enables us to effectively linearise complex temporal trajectories and hence is a compelling approach in dynamical systems research [321]. Its application is expanding in both theoretical [321,326], and practical domains from molecular dynamics and fluid dynamics, atmospheric sciences, and control theory [327–329].

The advantages of the Koopman operator are numerous. First, it is a powerful tool for analysing and predicting the behaviour of a system over time. By lifting the state variables into a higher dimensional space, the Koopman operator can identify patterns in a system's behaviour that may otherwise be difficult to detect. This can be especially useful for uncovering hidden dynamical structures in chaotic systems.

Second, the Koopman operator allows us to develop data-driven models of dynamical systems. Using the operator's eigenfunctions as basis functions, it is possible to develop models of dynamical systems operating on these summarised coordinates, without solving or even understanding the underlying equations of motion. This makes the Koopman operator an attractive tool for model-based control and optimisation. Third, the Koopman operator is useful to discern key

properties of highly nonlinear dynamical systems. In short, with the expansion of original state variables to infinite dimensions, the operator can uncover subtle nonlinear behaviour in a system that would otherwise be difficult or impossible to detect [321].

Koopman operator [321] The Koopman operator is a linear operator that describes the evolution of an observable function of a dynamical system. Let $\mathcal{M}$ be a manifold of dimension n and $f : \mathcal{M} \to \mathbb{R}$ be a real-valued function. The Koopman operator, denoted by $\mathcal{K}$, is defined as an infinite-dimensional linear operator that acts on the space of observable functions f such that for any $f \in L^2(\mathcal{M})$,

$$\mathcal{K}f(\mathbf{x}) = f(T(\mathbf{x})),$$

where $T : \mathcal{M} \to \mathcal{M}$ is the evolution operator that maps each point $\mathbf{x} \in \mathcal{M}$ to its next iterate in time. Now, let $f : \mathcal{M} \to \mathbb{R}$ be a bounded, measurable observable and $\mathcal{K}$ be the Koopman operator associated with a dynamical system. Then, there exists a sequence of eigenfunctions $\psi_j : \mathcal{M} \to \mathbb{C}$ and a corresponding sequence of eigenvalues $\lambda_j \in \mathbb{C}$ such that

$$\mathcal{K}\psi_j(\mathbf{x}) = \lambda_j \psi_j(\mathbf{x}).$$

The Koopman operator preserves the linear structure of the space of observables, provides a linear representation of nonlinear dynamics, and its eigenfunctions provide a useful basis for approximating dynamical systems.

There are, however, some drawbacks associated with the Koopman operator. First, it is intrinsically an infinite-dimensional operator, and although there are efficient finite-dimensional approximations available, the accurate computation of their eigenvalues and eigenfunctions can be computationally expensive [96]. This can be a problem for real-time applications, such as model-based control [330]. Second, related to this drawback, the operator's eigenfunctions are often difficult to interpret, hampering the capacity of the learned representations to explain the underlying system's dynamics. Finally, the Koopman operator assumes that a locally-linear behaviour can represent nonlinearities sufficiently accurately. Thus, the Koopman operator is an important theoretical and applied research tool for understanding and predicting the behaviour of complex dynamical systems. Its data-driven approach to model-based control and optimisation has opened up new possibilities for real-time applications [330]. Moreover, its ability to uncover subtle nonlinear behaviour in chaotic systems has made it invaluable for studying chaotic dynamical systems [331].

More specifically, given a space in which the dynamics is linear, a successful approach to approximate transfer operators (such as the Koopman operator) from the data is called dynamic-mode decomposition (DMD) [327]. It is also possible to obtain ROMs using DMD [327], which is also based on concepts from linear algebra and assumes that the system's state can be advanced in time via a linear operator $\mathbf{A}$. While the POD modes are orthogonal in space, the DMD ones are orthogonal in time, and each mode is associated with a particular frequency and a growth rate. Therefore, DMD may help to identify temporal patterns in the data more clearly than POD. In contrast, POD may lead to a more compact low-order representation of the original system due to its optimality property. See an illustrative example in Fig. 13.

Dynamic-mode decomposition (DMD) [327] is a technique used to approximate the normal modes and eigenvalues of a linear system. Additionally, these modes can be associated with a damped or driven sinusoidal behaviour in time. DMD is useful for identifying a system's frequency and decay/growth rate. Let us define a dynamical process formulated as $\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, t, \mu)$, where $\mathbf{x}$ defines a measurement, t is a time, μ is a parametric dependence, and f indicates an unspecified system but from which we obtain many data. Therefore, the complex dynamical system f can be approximated as follows $\frac{d\mathbf{x}}{dt} \approx \mathbf{A}\mathbf{x}$, where $x \in \mathbb{R}^n$, $n \gg 1$ and $\mathbf{A}$ defines a linear dynamical system. Then its general solution is the 'exponential solution' defined as $\mathbf{x} = \mathbf{v}e^{\lambda t}$, where $\mathbf{v}$ and λ are eigenvectors and eigenvalues of the linear system $\mathbf{A}$. The problem of finding the eigenvectors $\mathbf{v}$ and the eigenvalues λ is a eigenvalue problem defined as $\lambda\mathbf{v} = \mathbf{A}\mathbf{v}$.

Yet, we are interested in obtaining $\mathbf{A}$, not its eigendecomposition. This is what the so-called 'exact DMD' does. DMD uses observations/measurements $x_j = \mathbf{x}(t_j)$, defined at a time point j to construct two matrices: the first concatenating the data from the first snapshot to $(m-1)$th snapshot, and the second with the shifted-by-1-time-step samples, $\mathbf{X}$ and $\mathbf{Y}$, respectively. The goal is thus building a linear dynamical system A fitted with $\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}$, and thus *learn* the linear dynamical system $\mathbf{A}$ that takes the data $\mathbf{x}$ from current state $(j-1)$ to future state (j), that is $\mathbf{Y} = \mathbf{A}\mathbf{X}$. The linear dynamical system A can be extracted using a pseudo-inverse $\mathbf{X}^\dagger$ of $\mathbf{X}$, that is $\mathbf{A} = \mathbf{Y}\mathbf{X}^\dagger$. Intuitively, the linear dynamical system $\mathbf{A}$ performs a least-square fitting from the current state $\mathbf{X}$ to the future state $\mathbf{Y}$.

Over the last decades, different numerical methods have been introduced: Ulam's method [332], extended dynamic-mode decomposition (EDMD) [333–335], and the variational approach of conformation dynamics (VAC) [336,337]. The advantage of purely data-driven methods is that they can be applied to simulation and observational data. Hence, information about the underlying system itself is not required. An overview and comparison of such methods can be found in [329]. Applications and variants of these methods are also described in [338–340], while kernel-based reformulations of the methods above have been proposed before in [334,341]. Note that the framework of higher-order dynamic-mode decomposition (HODMD) [342] enables relaxing the linear assumption by including several temporal snapshots to build the operator by exploiting Takens' delay-embedding theorem [343]. The HODMD approach requires additional hyper-parameter tuning, but it has led to very insightful results, for instance, in the context of complex turbulent flows, where this method has enabled identifying the coherent structures responsible for the concentration of pollutants in cities [344]. Another relevant application of HODMD includes cardiovascular flows [345].

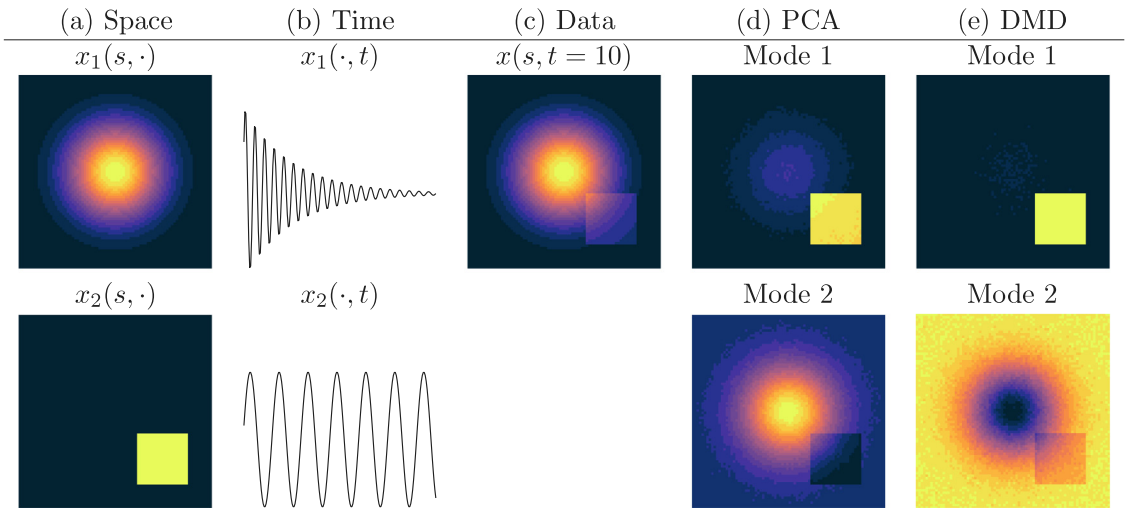


Fig. 13. Comparison between DMD and PCA with synthetic spatio-temporal data. The signal under analysis $x(s,t)$ (c) is the sum of two generative signals (a,b): $x_1(s)$ is a Gaussian that decays exponentially, and $x_2(s,t)$ is a square that oscillates at a lower frequency. The projections onto the two top components of PCA (d) and DMD (e) show that DMD extracts cleaner spatial coherence patterns from the data.

3.2.3. Dynamic modes in neural-network latent spaces

Despite the interesting properties of POD and DMD, their inherent linearity typically leads to the requirement of very large numbers of modes to reconstruct most of the variance of the original signal, for example, in three-dimensional turbulent flows [346]. Neural networks, especially autoencoders (AEs), have been proposed to obtain a reduced-order nonlinear representation of the original data. AEs exploit non-linear activation functions to produce significantly more compact representations in the latent space than those with, e.g. POD [347]. Fig. 14[a–b] shows the use of AEs for learning (dynamic) feature representations.

AEs have been used in fluid mechanics to obtain compact modal decompositions of the flow around a two-dimensional cylinder [348] and in more complex turbulent flows, e.g. the flow in a simplified urban environment [349]. Interestingly, when restricting neural networks to linear activation functions, one recovers the POD modes, as shown by Milano and Koumoutsakos [350] with a multilayer perceptron (MLP) in turbulent channel flow. Shallow NNs have been used for flow reconstruction, in this case from sparse measurements, as illustrated by Erichson et al. [351] for several flow cases. A more general illustration of the potential of AEs based on convolutional neural networks (CNNs) was presented by Lee and Carlberg [352], and an application to spectral submanifolds was developed by Cenedese et al. [353]. The reader is referred to Refs. [354,355] for a survey of classical methods applicable to linear subspaces.

Despite the superior compression performance of AEs compared with POD, the former does not have two very interesting properties of the latter, namely the optimality and orthogonality of the resulting modes. These are important properties due to their connection with interpretable and parsimonious ROMs. Regarding optimality, Fukami et al. [356] proposed an interesting approach based on hierarchical autoencoders (HAEs). They first trained a CNN-based AE fixing the dimension of the latent space to just one, obtaining one latent vector. Then, they trained another CNN-AE with a latent dimension of two and fixed the first latent variable to the one obtained in the previous NN, thus obtaining a second latent vector. Through this recursive strategy, they obtained a sequence of latent vectors exhibiting progressively less contribution to the reconstruction of the original signal, allowing them to establish a ranking in the resulting modes. This approach was tested in the flow around a two-dimensional cylinder, although it is important to note that the resulting modes were not orthogonal. This was addressed by Eivazi et al. [349], who used β-variational autoencoders (β-VAEs), which enable introducing stochasticity in the latent space to impose orthogonality in the resulting AE modes, a phenomenon that was explained in Rolinek et al. [357] among a connection of β-VAEs to PCA. Also in the case of the β-VAEs, the modes can also be ranked in terms of their contribution to the reconstruction.

3.2.4. Equation discovery in latent representations

Perhaps the biggest challenge in data-driven model discovery is balancing model efficiency with descriptive capabilities. Parsimonious models with the fewest terms required to capture essential interactions promote interpretability and generalisability. However, obtaining parsimonious models is linked to the coordinate system in which the dynamics are measured. The previous methods based on dimensionality reduction, e.g. ROM, DMD or AE, extract expressive components without simultaneously discovering coordinates. In [94], AEs were trained for data reconstruction and to recover a parsimonious dynamical system model through sparse regression using SINDy (see Section 3.1.2). See Fig. 14[c]. The joint goal of discovering models and coordinates is critical for understanding many modern systems. Using SINDy as an explicit equation discovery regulariser in the latent space balances simplifying coordinate transformations and nonlinear dynamics to identify coordinate transformations where only a few nonlinear terms are present.

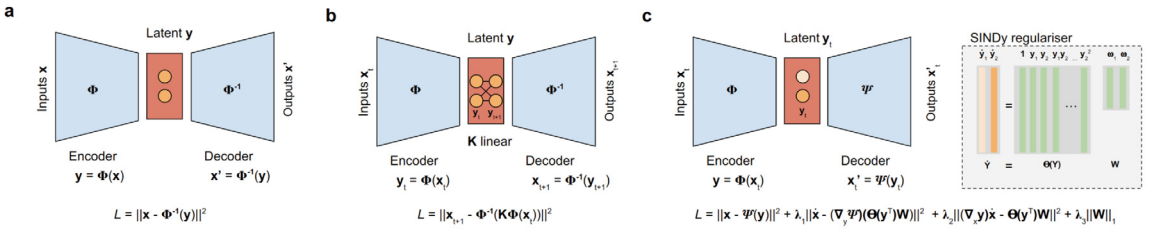


Fig. 14. Approximating Koopman operator with explicit nonlinear mappings using autoencoders. (a) An autoencoder neural network learns a mapping $\boldsymbol{\Phi}$ compressed latent representation $\mathbf{y}$ from input data $\mathbf{x}$ by minimising the reconstruction error L. (b) One can incorporate the Koopman linear operator $\mathbf{K}$ operating in the latent representation $\mathbf{y}_t$, which can then be used for prediction $\mathbf{x}_{t+1}$ from the transformed $\mathbf{y}_{t+1}$ [326]. (c) Those representations are not necessarily physically consistent. This can be addressed by enforcing equation dictionaries using SINDy in the loss for the simultaneous discovery of coordinates and parsimonious dynamics [94]. The loss now accounts for the reconstruction of the input data, as in a regular autoencoder, and the temporal dynamics (gradients ∇_x, ∇_y) of $\mathbf{x}$ and $\mathbf{y}$ projected onto SINDy bases.

3.2.5. Discovering fundamental variables

Despite advances in equation discovery (either through implicit or explicit representations), the main core problem is identifying state variables. The discovery typically refers to the identification of the governing equations, not the identification of the physical forces or variables. The vast majority of data-driven models of discovery rely, however, on pre-existing knowledge of the state variables, e.g., the position and velocity of a rigid body object. This relies on deep domain knowledge and strong assumptions. In addition, such assumptions cannot work properly for new physical systems or when those state variables cannot be measured. The work [13] proposed a principle for determining the number and identity of state variables in a system from high-dimensional data and demonstrated high effectiveness using video recordings of physical systems. The algorithm discovered the intrinsic dimension of the observed dynamics and could identify candidate sets of state variables without prior knowledge of the underlying physics. Alternatively, other studies sought to identify the fundamental state variables via manifold learning in ambient RKHS (termed *Diffusion* maps, e.g., [358,359], see also Section 4.1.2).

In short, the field of *variable discovery* is filled with many opportunities in the physical, biological and chemical sciences [105,106,358], as well as many challenges [13]. Finally, and interestingly, we want to emphasise that variable discovery is intimately related to revealing latent confounders in the field of causal inference [107–109].

3.3. Perspectives

Let us indulge ourselves with a brief overview of the main challenges (both conceptual and technical) and the opportunities for future research in the field of equation discovery for the physical sciences.

3.3.1. Challenges

The field of equation discovery from data is very prolific and is situated at the intersection of many communities: statistics, machine learning, computational fluid dynamics, Bayesian inference, dynamical systems and control theory, functional analysis and causal inference. The field has occupied scientists for centuries at all levels. The quest for optimal and automated solutions has traditionally considered moving in one or several subspaces in the sparsity-extrapolation-generalisation space, i.e., models should be simple, generalisable/robust, and capable of extrapolating outside the sample space (Fig. 15). These are very ambitious goals, implying both theoretical and practical challenges.

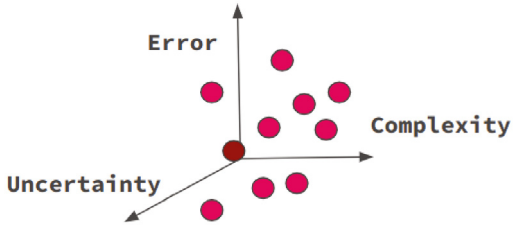


Fig. 15: On the quest for the optimal model in the sparsity-extrapolation-generalisation space.

Theoretical challenges. From a more theoretical perspective, both the *identifiability* of the system's equations [360,361] and the role (or preference) for *sparsity* (simplicity) have been questioned [362,363]. In addition, there is a long-standing debate on the *evaluation* of the obtained solution, where many criteria can be adopted. Is it only about invariances and robustness in space and time? Is Nature always simple and compositional, such that compactness and sparsity rule in natural systems? The issue of model's (i.e. hypotheses) intercomparison and evaluation also speaks to the more elusive question of how to reconcile solutions offered by different equation discovery methods.

Another important theoretical challenge is related to the fact that, very often, one (1) assumes that all involved state variables are given/observed, which resembles the sufficiency assumption in causal inference, and (2) selects a subset of representative states, assumes a particular basis to express the solution, or can operate on a manifold

subspace [84,93,94,364]. Both are strong assumptions that challenge the process of discovering equations and raise many questions. How do we choose the right variables to include in the equation discovery method? How much does the solution change when a variable is omitted or added? Here, *identifiability* issues and *latent confounders* play a substantial role. And foremost, what if we cannot measure the underlying state variables? Can we identify them automatically? Several methods have arguably proposed to discover the latent variables [13], and other efforts exploit the link between RKHS techniques and a plausible master equation underlying the spatiotemporal evolution of the data probability density function [365] to automatically learn a set of fundamental coordinates of the system, cf. Section 3.2.5. The inferred subspace has shown effectiveness in identifying the dynamics of (partial) differential equations [358,359], see examples in Section 4.1.2. However, like other explicit methods discussed in this section, it requires prior guidelines on the differential equation to model and incorporates a range of heuristics during its processing pipeline [358].

Practical challenges. Important practical challenges are related to the *model development* and *data characteristics*: (1) *high dimensionality*, (2) *nonlinear relationships* and (3) *risk of overfitting*. In many cases, the number of variables and parameters can be very large, making it difficult to find the most relevant features and relationships. This is the scenario where non-identifiability arises. Another challenge is that real-world systems often exhibit highly nonlinear relationships between input and output variables. This makes finding a good representation of the underlying dynamics difficult, and the search space for equations can become very large. While several nonlinear methods that capture complex dynamics exist (kernels, neural networks), performance evaluation and hyperparameter tuning are still important challenges. Finally, symbolic regression approaches can also suffer from overfitting, where the model fits the training data well but performs poorly on new data, which speaks to the trade-off between the model's accuracy and complexity.

When working with spatiotemporal data, it is relatively unclear how to incorporate information about time-lagged relations and interventional data. Both the explicit and the implicit approximations show particular challenges, though: On the one hand, symbolic regression techniques (such as SINDy) face significant problems in defining the basis functions, working in high-dimensional problems, and the impact of (even a limited) amount of noise. Other methods, like those based on AI Feynman, even if they incorporate sensible criteria to guide the equation discovery (like compositionally, reversibility or physical unit consistency), almost predict completely different equations when changing constants in the true equation [366]. On the other hand, implicit methods (like DMD or Koopman operators) do not provide an explicit equation but a latent feature representation to explain system dynamics. These methods struggle with nonstationarities, nonlinearities, and gaps noise, which remain unresolved problems in the literature [367]. Note that similar challenges to those in causal discovery remain here, cf. Section 2.2. Besides, DMD methods fail to generalise outside the training data and violate basic physical laws. To alleviate this, integrating domain knowledge (such as symmetries, invariances and conservation laws) in DMD has been recently introduced as an effective, robust approach [368].

3.3.2. Opportunities

While symbolic regression presents challenges, it offers exciting research opportunities for the physical sciences. Three main opportunities can be identified: (1) *model interpretability*, (2) *model compression and evaluation*, and (3) *model selection*. Equation discovery is a step forward in the system's understanding. The field leverages fully interpretable models, and unlike causal models, equation discovery (symbolic regression) models are directly applicable predictive models. The discovered equations from data can provide insights into the underlying dynamics of a system, thus helping researchers better understand how different factors interact and contribute to a particular outcome. Even with implicit latent representations, interpretability can be accomplished with interventional analysis. Another interesting opportunity is model compression, as symbolic regression models can also be used to compress large datasets into simple equations that capture the system's essential features; this can make it easier to analyse and visualise the data and make it more computationally efficient to work with. Another practical opportunity of symbolic regression models is that they offer a reduced set of possible solutions typically ranked in amenable Pareto fronts, which of course, trigger difficulties in choosing the right model but also fruitful scientific discussions about the plausibility of identified relations. This can save researchers time and effort and help identify unexpected patterns and relationships in the data.

Model interpretability and intervention analysis. The discovered explicit models are interpretable in nature. However, when complexity cannot be traded for accuracy or whenever an implicit feature representation is learned, intervention (or sensitivity) analysis offers opportunities for interpretability. For example, one can (1) *ignore or simplify the problem* by performing small perturbations away from real-world dynamics, which might help identify the proper relationship between variables; (2) *intervene on exogenous variables* (e.g. wind or solar irradiation, mixing coefficients, initial conditions in climate sciences, or targeted, direct brain stimulation in neurosciences) which is equivalent to collecting more data; (3) *create a library of trajectories* under different conditions and select those which match the desired intervention; and (4) *intervene in the learned latent space* and decode the intervention back to input space, thus allowing us to generate interventions that follow the system's natural trajectories. Further analysis methods, e.g. for studying interventions of the learned ODE, might be fruitful, as they can access long-term dynamical properties, such as how distortion in the eigenvalues affects the system's stability as the phase space changes.

Model compression and evaluation. Scientists frequently use metrics to evaluate new ideas or distinguish between competing hypotheses. As we have seen before, a governing equation should be simple but not simpler, accurate for prediction, robust under distortions and changes, and invariant in space and time. Equation discovery offers a *direct* way to learn plausible models and an *indirect* way to contrast and evaluate derived models. For this, one typically assesses (1) the *predictive accuracy* when answering how well the (simplest) hypothesis explains the data; (2) the model's *invariance* under distributional shift to account for the causal mechanisms; and (3) the *robustness* under interventions to study how the proposed process (or descriptive equation –representation) is consistent with interventions on the model dynamics, such as deactivation of components or targeted modification of exogenous variables. The latter is the rarest form of validation due to its high computational cost and difficulty in experimental design.

Model selection. Enforcing sparsity in model selection can lead to unrealistically too simple models. That is why methods that can provide solutions along the Pareto line (Fig. 6) are needed to capture complex relationships and offer subsets of plausible model solutions. Alternative regularisation schemes will likely be important alongside profound estimates of uncertainty and extrapolation indicators.

4. Case studies in the physical sciences

This section gives concrete examples of applying different data-driven causal and equation discovery in important fields of the physical sciences: neuroscience, Earth and climate sciences, and fluid and mechanical dynamics, cf. Table 5.

Table 5
Case studies presented and the main methods used in this section.

	Neuroscience	Earth & climate	Fluid dynamics
Causal discovery	Causal connectivity (DCM, GC, TE, SCM)	Carbon-water interactions Climate model comparison (CCM, PCMCI)	–
Equation discovery	Learning trajectories (kFDA, GP, Variational Bayes RNN, Diffusion Maps)	Ocean Mesoscale closures (RVM, DMD, SINDy)	Turbulence understanding Vortex shedding (SINDy, Genetic Programming)

4.1. Neuroscientific applications of physics-based machine learning

4.1.1. Overview of parsimonious models for neural population dynamics

Neuroscientific modelling falls within the remit of the field known as computational or theoretical neuroscience, which studies the transmission of information in the nervous system at multiple spatiotemporal scales (ranging from neuronal to whole-brain levels) in relation to perception, cognition, and behaviour (e.g., [369,370]). Thus, an ongoing challenge in computational neuroscience is to link biophysically detailed models operating at microscopic levels with meso/macroscopic theories of cortical processing [371]. This enterprise is often addressed by deducting low-dimensional systems of partial differential equations or maps describing *coarse-grained* variables derived from collective neural responses. Such different neurobiologically plausible simplifications are commonly termed ensemble, population, neural-mass, or simply *firing-rate* models (see, for instance, [372,373]).

These synthesis efforts are intimately connected with empirically discovering a reduced dynamical system generating the observed neuronal activity. However, neurocomputational modelling traditionally focused on analytical, deductive approaches mapping realistic cortical networks to *tissue-level* descriptions, as opposed to data-driven model discovery, reviewed in Section 3. Thus, ensemble models are typically principles-based, often hinged on assumptions about dynamical interactions arising within homogeneous pools of neurons (e.g., [374–377]).

Early neural ensemble models stemmed from applying statistical mechanical principles to the interaction of homogeneous pools of (excitatory and inhibitory) populations [378,379]. Later, physics formalisms like the Fokker–Planck approach for describing the spatiotemporal evolution of the probability distribution of neuronal activity enabled theoretical neuroscientists to take a more holistic approach to identify mean-field approximations of networks of spiking neurons (e.g., [369,373,380]). These and other nonlinear dynamical systems tools [381] provided closed-from, exact solutions for the collective behaviour of neural populations [375,382,383] capable of an extensive dynamical repertoire, although strongly dependent on universal theoretical assumptions, given their deductive nature (see Section 1). Alternatively, a Laplacian assumption on this probability distribution resulted in neural mass descriptions [205], recently proposed as building blocks for whole-brain models with translational applications [384].

Overall, these chiefly deductive approaches rendered compact models of differential equations based on *a priori* assumptions about neural and synaptic variables, fostering the interpretability of high-complex neuronal networks. By contrast, inferential approaches in neuroscience have been typically utilised to empirically identify neural dynamics underlying cognition and behaviour, as discussed next.

4.1.2. Empirical reconstruction of neuronal trajectories

There is an increasing focus on applying classic and deep machine learning approaches to reconstruct attracting and transient dynamics of cerebral cortex responses [385]. A neural trajectory $T \ \ (n \times d)$ is often defined as the sequence of n neural response vectors $\boldsymbol{x}(t)$ embedded in a d-dimensional state–space (the *ambient* space), spanned by neural ensemble activity or proxies thereof (e.g., firing-rates, electromagnetic potentials), their lags and nonlinear transformations [385–387].

Traditionally, standard dimensionality reduction techniques (e.g., PCA, Multi-dimensional Scaling, Discriminant analysis etc., see Section 3.2) were directly applied for the visualisation of high-dimensional neural trajectories, showcasing coarse-grained aspects of firing-rate dynamics concerning, for example, cognitive decisions or motor functions [388,389]. More recently, Gaussian processes provided a flexible approach to derive a low-dimensional manifold representing the dynamical systems generating the observed activity. They just require a reasonable hypothesis on temporal correlations between observations (the prior covariance function [215]). For instance, Gaussian process-based factor analysis (GPFA) [390], and other latent-variable methods [391], provide such low-dimensional subspace while simultaneously approximating the probability of spiking — without the compelling need for probability density estimation [390,392]. These and related approaches can identify latent neural trajectory manifolds in prefrontal and motor cortices underlying decision-making [391,393,394]. Specifically, recent GPFA variants were able to discern between competing models for the contribution of upstream areas to recurrent dynamics supporting decision-making in the monkey prefrontal cortex [386].

Covariance (kernel) function methods can also recreate salient facets of cortical dynamics like attracting sets. To this end, they leverage delay-embedding techniques in RKHS spanned by neuronal correlations and their temporal structure [387,395–397] for identifying compact manifolds mapping animal's choice with attracting sets of ensemble trajectories [396,397]. Fig. 16 presents an illustrative example of these RKHS techniques for recreating neuronal trajectories underlying the effect of dopamine at the circuit level. This approach facilitated the evaluation of mechanistic theories of dopamine modulation during decision-making, which was challenging given the limitations of direct experimental manipulations [396]. The figure shows the flow field of trajectories derived from the activity of neuronal constellations in the rodent anterior cingulate cortex. Interestingly, the dynamic landscape depicted during working memory tasks in an optimal RHKS can be approximately described as transients connecting multiple attracting sets mapping spatial choices (Fig. 16a). This robust multi-stable scenario is completely disrupted by high doses of amphetamine (a well-known trigger of dopamine release, Fig. 16b), while it is enhanced by low doses (see [396]), in line with long-standing theoretical predictions of biophysical models [398].

Further facets of brain dynamics, such as chaotic attractors, have been recently addressed with recurrent, piecewise-linear architectures, amenable to optimisation via back-propagation variants or Bayesian variational inference [399,400]. In these approaches, tractability is promoted by leveraging units' linearisation for approximating trajectory inference in a parsimonious, transparent fashion. Empowered by these characteristics, such linearised recurrent networks could infer pathological whole-brain dynamics from functional magnetic resonance imaging (fMRI) recordings [401]. Moreover, recent developments of these methods embody biophysically-inspired computations such as dendritic processing, fostering their reconstruction capabilities of nonlinear dynamical systems [400].

More broadly, classic and deep architectures can facilitate the Bayesian inference of the optimal range of biophysically realistic model parameters. This is a challenging task given the potentially high sensitivity of realistic networks to different parametrisations [402]. Along these lines, approximate Bayesian computation (ABC) has been combined with connectionist approaches to identify parameters in models operating at multiple spatial scales, ranging from microscopic-level Hodgkin–Huxley-type single neurons [403] to macroscopic, cognitive-level decision-making models [404].

For instance, Sequential Neural Posterior Estimation (an ABC method) alternates between deep learners and variational Bayes approaches for parameter approximation. First, a standard (non-biophysical plausible) classifier is used to constrain the range of initial parameters θ generated from the prior $p(\theta)$ by predicting their suitability. Subsequently, a deep learner operating on multivariate data $\boldsymbol{x}$ and parameters sampled from such constrained prior $\hat{p}(\theta)$ estimates the likelihood $\hat{p}(\boldsymbol{x}|\theta)$, enabling a progressively more refined Bayesian computation of the posterior over neuronal and synaptic parameters $p(\theta|\boldsymbol{x})$ [405]. These approaches were able, e.g., to discern between neuronal model configurations, essentially indistinguishable in observed activity, the ones metabolically optimal in the pyloric network in crustacean [406]; or to infer reaction times and choices in classic (drift–diffusion-type), descriptive models of decision-making [404].

Alternatively, neural trajectory reconstruction of Hodgkin–Huxley ensembles has been recently tackled with nonlinear manifold learning in RKHS [358,359]. These methods, like common dimensionality reduction-based techniques (see Section 3.2), identify first an optimally reduced set of coordinates from an original higher-dimensional ambient space embedding the time series $\boldsymbol{y}$. However, by contrast with other approaches, components spanning the low-dimensional representation are typically lead eigenvectors of a discretised Laplace operator governing the spatiotemporal evolution of the underlying $p(\boldsymbol{y}(\boldsymbol{x}, t))$, where $\boldsymbol{x}$ is homologous to a spatial coordinate. Thus, the reduced subspace spanned by the main non-redundant eigenvectors is often termed a Diffusion Map [365] (see also Section 3.2.5) given by the set of ith emergent coordinates $\{\phi(\boldsymbol{x})_i\}$.

In a subsequent stage, a fully connected (deep/shallow) architecture learns the dynamical system on the diffusion map, in other words, estimates the function f that maps the temporal flow of the time series $\boldsymbol{y}$ to its derivatives w.r.t. emerging coordinates, for instance, in a one-dimensional diffusion map ϕ_1,

$$\frac{\partial y(\phi_1, t)}{\partial t} \approx f\left(y, \frac{\partial y}{\partial \phi_1}, \frac{\partial^2 y}{\partial \phi_1^2}, \ldots, \frac{\partial^n y}{\phi_1^n}, \gamma\right), \tag{17}$$

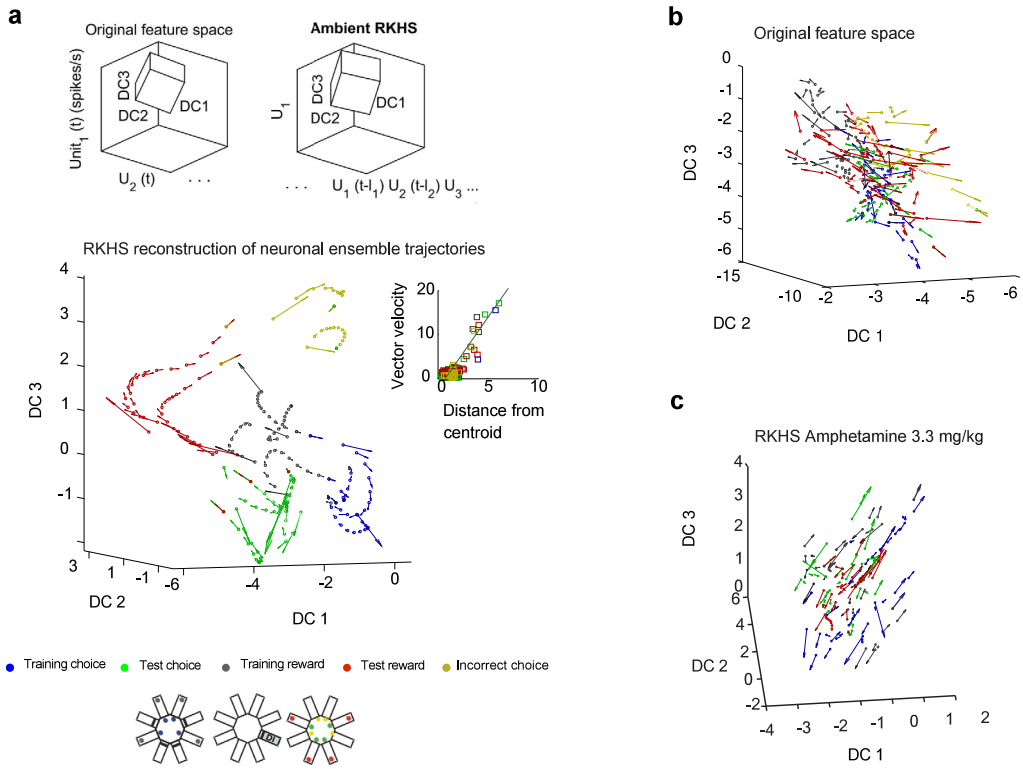


Fig. 16. Example of the reconstruction of neural trajectories in an RKHS derived from rodent anterior cingulate cortex (ACC) multi-array recordings, taken from [396]. **a**. The flow field stems from projecting an ACC ensemble firing rates onto the three main coordinates of a discriminant subspace (DC1–DC3, computed here by kernel-fisher discriminant analysis orthogonalised for a faithful representation, details in, e.g., [387,395–397]). The DC1–DC3 is embedded into an optimal ambient high-dimensional RKHS (top schematics), spanned by neuronal ensemble firing rate and its higher-order correlations (up to 3rd order in this example) operating on a delay-coordinate map. The colour code (bottom) corresponds to rat choices in this experiment (the schematic of the radial-arm maze used in this experiment is taken from [387,395]), which occupy distinct regions of the subspace. The flow field indicates faster shifts (large vector lengths) during the transition points, while it slows downs nearer the centroids of the clusters, suggesting an attracting-like dynamic landscape. The inset quantifies this uneven distribution of flow field speeds as a function of the distance to the centroid, further supporting this observation on flow convergence. **b**. This ordered phase-space structure cannot be achieved in the original feature space or with delay-coordinate maps. **c**. It is also destroyed when the animal receives a high dose of amphetamine, even in an optimal ambient space.
Source: Figure adapted from [396] and from [387] with publishers' permission (the Society for Neuroscience and the Public Library of Science).

where γ is a set of parameters which enables the learned map to reproduce bifurcations.

Interestingly, when the argument of f contains no derivatives and incorporates additive Gaussian noise, the system's dynamics reduces to the well-known Langevin stochastic differential equation, stemming from a Fokker–Planck process governing the temporal evolution of a probability distribution [407]. Thus, by approximating a discretised Fokker–Planck operator, it is possible to empirically infer parameters of the stochastic process leveraging conventional likelihood estimation techniques [407,408]. Langevin dynamics fit many natural phenomena and is of special interest in high-level decision-making models in neuroscience. This formalism was recently used in [408] for model discovery based on a one-dimensional Langevin equation and an additional stochastic spike generator,

$$\begin{aligned} \frac{dx(t)}{dt} &\approx D \cdot F(x) + \sqrt{2\,D} \cdot \psi(t;\,0,\,1), \\ y &\sim Poisson(x;\,\lambda(t)), \end{aligned} \tag{18}$$

where $x(t)$ is a latent trajectory, D is the diffusion constant, ψ is white normal noise, F is the deterministic map to be inferred, and $\lambda(t)$ is the parameter of an inhomogeneous Poisson process generating the observed spike train time series $y(t)$. This approach is capable of identifying a parsimonious model (that is, a set of $\{D, F(x), \lambda(t)\}$) from the spike train time series. Thus, it was used to discern between competing models of perceptual decision-making by comparing probability distributions underpinning such alternative parameter sets via standard Kullback–Leibler divergence [408].

In short, neuroscientific studies typically conceive the discovery of biophysical laws as the inference of deterministic dynamics embedded in essentially stochastic neural processes. This goal has been interpreted either as empirically reconstructing attracting and transient components of neural activity (disentangled from coupled noise e.g., [358,387,393]); or

as identifying parameters of parsimonious, *a priori* model shapes [403,408]. These approaches provide valuable insights on the latent neural dynamical landscape.

However, a liking theme in such inferential methods is that, despite their advances in tractability (e.g., [400,408]) and interpretability (e.g., [358,397]), they are often not designed to empirically discern a unique set of differential equations, as it is popular in other areas of physics (e.g., [12,409,410], see examples in Section 4). This is at odds with the chief goal for deductive approaches, outlined in Section 4.1.1. Key reasons for this shortcoming might be found in the high-noise levels arising in intrinsically stochastically-dominated neural processes, in which most hidden variables are not experimentally accessible (especially in *in-vivo*) [387,408]. This challenging scenario hinders the direct application of approaches common in other fields and poses an intriguing question for future research endeavours.

4.2. Learning causally interacting brain regions from neurophysiological recordings

4.2.1. Causality in the connected brain

The debate on the causal role of brain connectivity has a long-standing tradition (see e.g. [411]). The classic view of functional segregation (mapping functions to physical brain regions) veered to connectionism, that is, brain functions result from interactions between neurocomputational units [411]. Consistently, the focus gradually shifted from functional segregation (the study of regionally-specific brain activation) to functional integration (the study of the connectivity between cortical areas [411]).

Historically, connectivity studies establish the distinction between structural (the anatomical location of white matter, axonal tracts), functional and *effective* connectivity (and sometimes with normative connectivity, in contrast to individual-specific connectomes) [53]. This classification is relevant to determine the type of causality questions that can—or *cannot*—be addressed [52,412]. Typically, functional connectivity methods estimate statistical dependencies such as spatiotemporal correlations or coherence measures between brain ensembles. In contrast, a subset of these methods, commonly termed effective connectivity approaches, refer to the quantification of directed interactions between brain circuits (e.g., [52,53,413,414]). In this arena, the quest for demonstrating causality relationships in neuroscience has attracted much attention over the recent decades [53,412], and its plausibility has been widely debated [412,415–418]. For instance, in neuroimaging, multiple issues such as confounding factors [419,420] and varying temporal delays (intrinsic to, for instance, fMRI) challenge estimates of network information flow directionality (e.g., [53,415,418,421,422] among many others).

Methodologically, a key characteristic of causal approaches—in difference with conventional probabilistic modelling—is the need for predicting how the system reacts under interventions [415]; in other words, for defining counterfactual models (see Section 2.3.8). Problematically, a large amount of interventional data is necessary to falsify the wide range of causal hypotheses in a high-dimensional system like the cerebral cortex [415]. For instance, *targeted* brain interventions via intracranial electrical stimulation (iES) in conscious patients is typically a robust approach for testing causality [53], but large-scale datasets using this experimental protocol are scarce, given ethical and experimental limitations of invasive techniques [415]. However, comprehensive, high-quality interventional data would be fundamental to falsify as many competing causal scenarios as possible. This is especially important in cognitive neuroscience given the lack of experimental access to some fundamental variables, which increases the number of plausible causal models underlying observable behaviour [415].

This shortage of comprehensive targeted lesion/stimulation datasets, and the improvement of whole-brain registration techniques, led to the development of analytical methods (or adaptations of existing ones) to better understand causality in cortical circuits. Most notably, Granger Causality and related approaches (GC, originated in the field of Economics [35]), Structural Causal Modelling (SCM [3,200]), and Dynamic Causal Modelling (DCM [204]) have been extensively used, as will be discussed next.

4.2.2. Causal methods in neuroscience

GC and an extension of this concept, Transfer Entropy (TE), are perhaps the most common *model-free* methods for assessing causal relations in neuroscience. These two generalist approaches estimate the direction of causality between interacting neural populations by analysing the time series derived from brain responses [207,417,423]. They are regular statistical tools for studying orchestrated interactions between brain regions via magneto/electroencephalography (M/EEG) and fMRI recordings (e.g., [54,207,424]). At microscopic levels, they have also been applied to detect synaptic connections between neurons [425]. Specifically, GC is based on the assumption that time series prediction leveraging its past values significantly improves by inputting historical values from another, causally connected time series (see details in Section 2.1.3). Thus, the presence of causal relationships is detected by testing the hypothesis that one time series autocorrelations have predictive power for the other time series [423].

TE expands this idea to accommodate broader types of nonlinear temporal interactions by computing the amount of information that one time series *transfers* to another. Similarly to GC, it conjectures that the current value of one time series can be better estimated by conditioning the predictive probability to past values of both itself and another time series, inferring causality direction [121]. Alternatively, SCM and its recent variants are Bayesian approaches for assessing plausible causal graphs in brain networks. They have been applied, for instance, to foster interpretability in behavioural

decoding approaches [426]. However, their use in cognitive neuroscience is still challenging, given the key difficulties discussed in Section 4.2.1, and the indirect nature of most neuroimaging measurements (reviewed in [415]).

Accompanying these model-free approaches, perhaps the most standard model-based technique for connectivity inference between brain regions is Dynamic Causal Modelling (DCM). DCM is a Bayesian method incorporating different degrees of *a priori* biological plausibility for understanding mechanisms underlying neuroimaging data (see e.g., [205,206,427,428]). It has been employed to study neural pathways of effective connectivity in e.g., motor control, attention, learning, decision-making, emotion, and other higher cognitive functions [429]; and even to model EEG seizure activity dynamics in epilepsy [428].

In general, whole-brain modelling methods like DCM or other more recent models [430] can provide a more nuanced understanding of the underlying mechanisms of brain function than model-free approaches [431]. However, the need for large datasets w.r.t. the complexity of the range of alternative models hampers the interpretation of the estimated connectivity [53,205,206,427]. Indeed, classic DCMs [432] have been criticised for the difficulty in falsifying their model selection approach [421] and perhaps for this reason, they were not extensively tested in clinical settings [427]. Specifically [421] suggested the ambiguity of DCM inference in generating a unique optimal connectivity map due to, e.g., known challenges in model fitting and selection in such a large space of possible architectures [421,433].

These caveats of DCM as a robust approach for causality assessment led to the development of variants such as spectral DCMs, the canonical microcircuit DCM — introducing higher degrees of laminar-specific, biophysical detail towards more informative priors for E/MEG modelling-, or the stochastic dynamic causal model, sDCM (see a review in [429]). sDCM incorporates random processes to the basic DCM equations, enhancing its fitting capability to hemodynamic responses and hence alleviating excessive dominance of priors in Bayes model selection [434]. In a classic DCM for fMRI data, the neural state $\boldsymbol{x}(t) \in \{1, n\}$ (for n interacting brain regions) corresponding to a single task-based input $u(t)$, is determined using the simple first-order differential equation $\frac{d\boldsymbol{x}(t)}{dt} = (\boldsymbol{A} + u(t) \cdot \boldsymbol{B})\,\boldsymbol{x}(t) + u(t) \cdot \boldsymbol{c}$; where the matrix $\boldsymbol{A}$ encodes (endogenous) connections between brain regions, $\boldsymbol{B}$ the strength in which inputs modulate each connection (*modulatory* inputs) and $\boldsymbol{c}$ the gain of the *driving* inputs to each region. sDCM expands this approach by adding intrinsic $\boldsymbol{\beta}(t)$ and extrinsic $\gamma(t)$ stochastic fluctuations to account for the incomplete observability of both states and inputs to brain areas relevant to the cognitive task:

$$\begin{aligned} \frac{d\boldsymbol{x}(t)}{dt} &= (\boldsymbol{A} + \nu(t) \cdot \boldsymbol{B})\,\boldsymbol{x}(t) + \nu(t) \cdot \boldsymbol{c} + \boldsymbol{\beta}(t), \\ \nu(t) &= u(t) + \gamma(t), \\ \boldsymbol{y}(t) &= g(\theta) * \boldsymbol{x}(t) + \boldsymbol{\epsilon}(t), \end{aligned} \tag{19}$$

where $\nu(t)$ is a *hidden* input cause masked by fluctuations (univariate here for simplicity), and the last equation represents a hemodynamic model (present in all DCMs variants) of non-neural parameters $\{\theta, \boldsymbol{\epsilon}(t)\}$. Finally, their convolution with the neural state $\boldsymbol{x}(t)$, yields the observed fMRI blood-oxygenation level-dependent (BOLD) response $\boldsymbol{y}(t)$ in relevant brain areas (termed regions of interest, ROI) [429].

Fig. 17 summarises an illustrative reliability study from [434], specially designed for assessing sDCM robustness. In this example, a large sample of participants ($n = 180$) was recruited from three different geographical locations. fMRI recordings were obtained from healthy subjects from the same age range while performing a classic 2-Back working memory task (to recall numbers shown two trials before). This N-Back task activates the dorsolateral prefrontal cortex-hippocampal formation (DLFC-HF) network connectivity, which is abnormal in schizophrenia patients [435]. Bernal-Casas et al. [434] used sDCM to identify the DLPFC-HF effective connectivity and compared the consistency of the models in this multi-centre setting (Fig. 17). Three *a priori* likely mechanisms to explain the BOLD responses to this task were implemented in three different families of models: with only driving inputs to the two regions ($B \equiv 0$, Fig. 17a, left), only connectivity modulation ($C \equiv 0$, Fig. 17a, centre) and both mechanisms combined (Fig. 17a, right). Noticeably, the random effects Bayes Model selection process strongly favours a specific connectivity model belonging to the driving inputs family (Fig. 17b) over all the rest, consistently for the three independent locations. Specifically, the DLPFC-HF connectivity parameters were statistically indistinguishable across datasets (Fig. 17b), supporting the reliability of sDCM results.

In line with these whole-brain analyses, other studies showcased the consistency of causality methods at microscopic levels. As a representative example, [436] recently proposed the effectiveness of GC in inferring information directionality in zebrafish motor circuits from single-cell calcium imaging signals. Causally strong, interventional data was inaccessible in this setting. Despite this, results were in full agreement with the known physiology of this species. In addition, and besides these standard methods (GC, TE and DCM), recent approaches have also addressed the causality robustness question from different angles, for instance, by focusing on changes in information *reversibility* as a sign of aberrant resting-state brain dynamics — which could subserve as a biomarker of Alzheimer's disease [437].

These and other even more indirect causality measures (like standard statistical approaches [52]) have provided useful insights when operating on neurophysiological recordings with high temporal precision. The ideal recording modalities are thus those capable of directly recording local (electrical) field potentials (LFP), such as intracranial electroencephalography (iEEG) or neuronal-level techniques (like in Chen et al. [436]). However, when indirect causality measures are estimated from other modalities—especially from functional imaging—the multiple confounders discussed earlier rank them in a

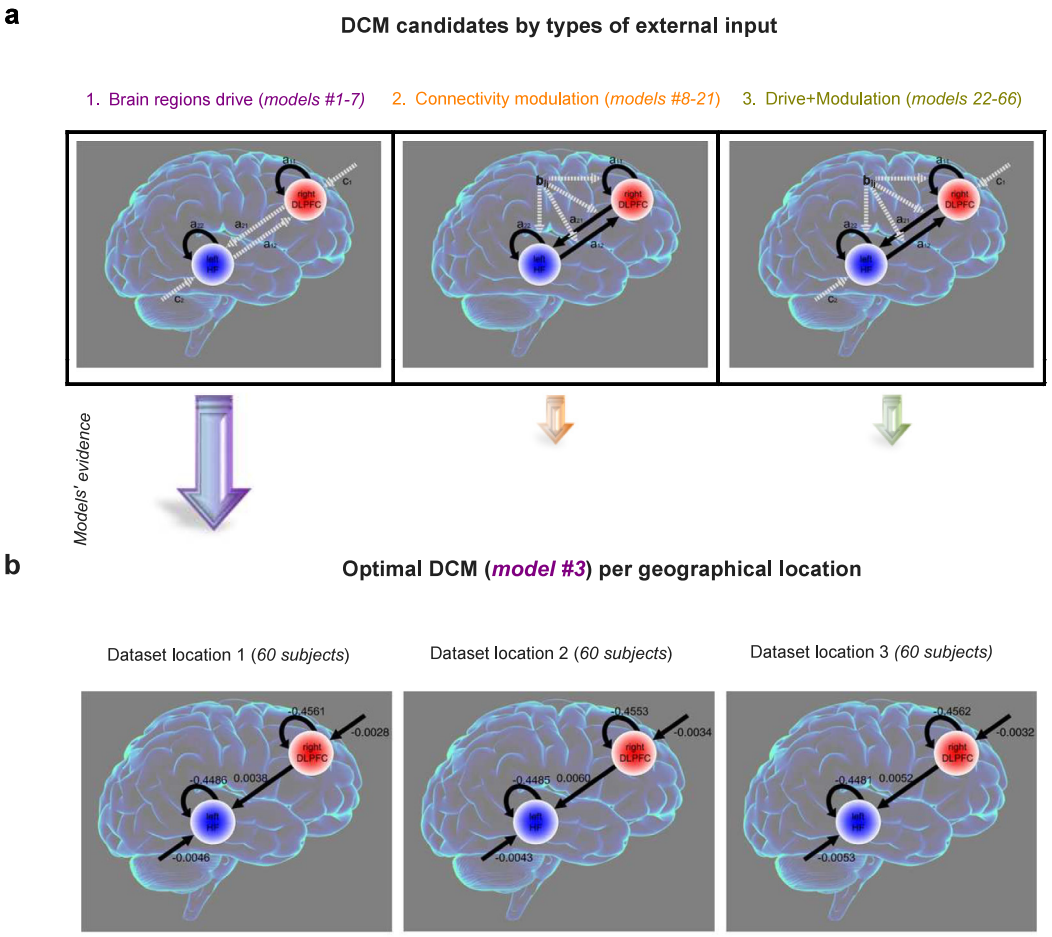


Fig. 17. Example of consistency assessment for stochastic DCM, taken from [434]. **a**. Connectivity hypotheses associated with the performance of a 2-Back working memory task by healthy human participants [434]. Left: input fluctuations to the two ROIs (right DLPFC and left HF; 7 distinct model combinations). Centre: input variance modulates the connections themselves (14 models). Right: combinations of both mechanisms (44 models). Connectivity hypotheses are tested on independent datasets collected from three locations (Bonn, Berlin and Mannheim, 60 subjects each). Model evidence (log-likelihood marginalised over model's free parameters [427]) is much stronger for a model of the first family (#3, random effects Bayes factor 98%+ in favour of this model). **b**. Remarkably, connectivity parameters do not differ across sites (Friedman non-parametric test, $p > 0.2$), and interaction between sites and model parameters were not found ($p > 0.8$), supporting the model's robustness. Figure adapted from [434] with the publisher's permission (Elsevier).

weak position in a causality scale when compared with approaches based on interventions [53]. Therefore, their capability for providing a reliable indication of causality interactions is highly disputed [53,421,438].

Nevertheless, limitations for assessing causal relationships in neuroimaging do not preclude indirect analytical approaches to constrain the universe of plausible causal graphs for a specific scenario [52,53]. Thus, there is a reasonable consensus in considering them as valuable contributors to strengthen causality claims, provided they are combined with more direct measures of causality based on interventional data [52,53]. Indeed, the ideal scenario from a causality perspective occurs when its inference is consistent throughout different approaches; in other words, when different methods having complementary views provide synergistic evidence [53]. For instance, converging evidence between a causal fMRI/EEG model, a targeted lesion, and the stimulation of a specific cortical circuit would score high on a causality scale than either of these methods alone; since the counterfactual could be established by focal stimulation [439] combined with real-time neuroimaging recordings [53], enriching the conclusions of the lesion study.

This optimal coalescence of multiple causal approaches for effective connectivity inference was termed Convergent Causal Mapping and is the recommended approach for designing new experiments [53]. Thus, from this concerted perspective, studies considering a single approach in isolation—especially if it is not based on interventions—should not make strong translational claims, like suggesting direct therapeutic applications [53,439]. In addition, future works should consider testing robustness to different environments [415] (like in the example shown in Fig. 17 [434]) for further reinforcing the credibility of the inferred causal flow.

4.3. Learning causal graphs of carbon and water fluxes

4.3.1. Introduction

The Earth is a highly complex, dynamic, and networked system where very different physical, chemical and biological processes interact in and across several spheres. Land and atmosphere are tightly coupled systems interacting at different spatial and temporal scales [122]. The main challenge to quantifying such relations globally comes from the lack of sufficient in-situ measurements and the fact that some of these variables are latent and not directly observable with remote sensing systems. One can, for example, measure SM but not GPP directly. As an alternative, many studies have relied on model simulations to investigate SM-precipitation [440], GPP-SM [441] and ET-SM relations [442,443], to name just a few. However, assuming a model implies assuming the knowledge of the causal mechanisms and relations governing the system. This is not necessarily a correct assumption, especially in model misspecification, non-linearities and non-stationarities. Discovering such relations from data is of paramount relevance in these cases. In the following, we review the performance of two standard methods of causal discovery from time series data to learn the relationships between environmental factors and carbon and heat and energy fluxes at the local (site, flux tower) level and the global (planetary, product-derived) level. At the local level, we exploit data acquire by eddy-covariance instruments estimating fluxes exchange. At the global level, we exploit Earth observation data from satellite observations.

4.3.2. Clustering of biosphere–atmosphere causal graphs at the site level

The atmosphere and terrestrial ecosystems constitute another closely interconnected complex system where processes interact across a range of temporal and spatial scales. Further, causal relations also depend on vegetation types, climatic regions, and the season. Fortunately, measurement campaigns of the past decades have resulted in good coverage of measurement sites, available in the FLUXNET database [444], a collection of long-term global observations of biosphere–atmosphere fluxes measured via the eddy covariance method. Runge et al. [115] discuss a similar case study in-depth.

Here we review the study of Krich et al. [445] that analysed causal networks for different seasons at eddy covariance flux tower observations in the FLUXNET network and how they depend on meteorological conditions. Fig. 18 explains the methodological setup. From a selection of 119 FLUXNET sites (Fig. 18(a)) daily time series data of the following variables were considered (see Fig. 18(b) for one site): short-wave downward radiation (or global radiation, Rg), air temperature (T), net ecosystem exchange (NEE) (inverted), vapour pressure deficit (VPD), sensible heat (H), latent heat flux (LE), gross primary productivity (GPP), precipitation (P), and soil water content (SWC). For details on data processing, we refer to Krich et al. [445].

Causal networks were then estimated with PCMCI [170] (time lags from 0 to 5 days) in sliding windows of 3 months to capture the temporal evolution of biosphere–atmosphere interactions. Based on findings in Krich et al. [446], a smoothed seasonal mean was subtracted to remove the common driver influence of the seasonal cycle. This results in 10.038 networks for the different months and sites (an example network is shown in Fig. 18(c)). Node colours indicate the level of autocorrelation (auto-MCI-partial correlation [170]), and link colours the cross-link strength (cross-MCI); time lags are indicated by small labels, and straight edges are contemporaneous. Since the strongest and most consistent links are contemporaneous, further analysis focused on these 15 links.

A previous study [446] discussed individual networks in more detail; the scope of Krich et al. [445] was to apply a dimension reduction, here t-distributed stochastic neighbour embedding (t-SNE [447]) which considers each of the causal graphs as an observation in a high-dimensional space of the contemporaneous MCI partial correlation values (Fig. 18(d)). t-SNE allows projecting this high-dimensional space onto two dimensions (Fig. 18(e, left)) that are the dominant features of transitions between different states of biosphere–atmosphere interactions. The coloured clusters in Fig. 18(e, left) are based on the OPTICS approach [448], and the four corners indicate the four archetypes of network connectivity and the networks' underlying meteorological conditions (averages taken over the sliding windows in Fig. 18(b)). Finally, Fig. 18(e, right) shows the convex hulls of clusters and their average network.

Each point of the low-dimensional embedding represents a specific ecosystem's biosphere–atmosphere interactions at a specific time and allows us to investigate their behaviour. A main finding of Krich et al. [445] was that ecosystems from different climate zones or vegetation types have similar biosphere–atmosphere interactions if their meteorological conditions are similar. For example, temperate and high-latitude ecosystems feature similar networks to tropical forests during peak productivity. During droughts, both ecosystems behave more like typical Mediterranean ecosystems during their dry season Such meta-analyses of causal networks allow for another perspective on understanding ecosystems, including an analysis of anomalous changes in network structure as indicators of ecosystem shifts (see Section 2.3.4).

4.3.3. Causal relations at global scale

As an alternative to Granger causality, the work [39] presented the convergent cross-mapping (CCM) method, which may deal with the issues of non-stationary and nonlinear processes and deterministic relations in dynamic systems with weak to moderate cause–effect variable coupling. CCM assesses the reconstruction of a variable's state space using time embeddings to determine if $X \to Y$. This method has been extended to account for causal relations operating at different time lags and applied to various research areas. However, it is sensitive to noise levels, hyperparameter selection, and false detections in strong, unidirectional variable coupling cases. To address these issues, the robust CCM (RCCM) [122]

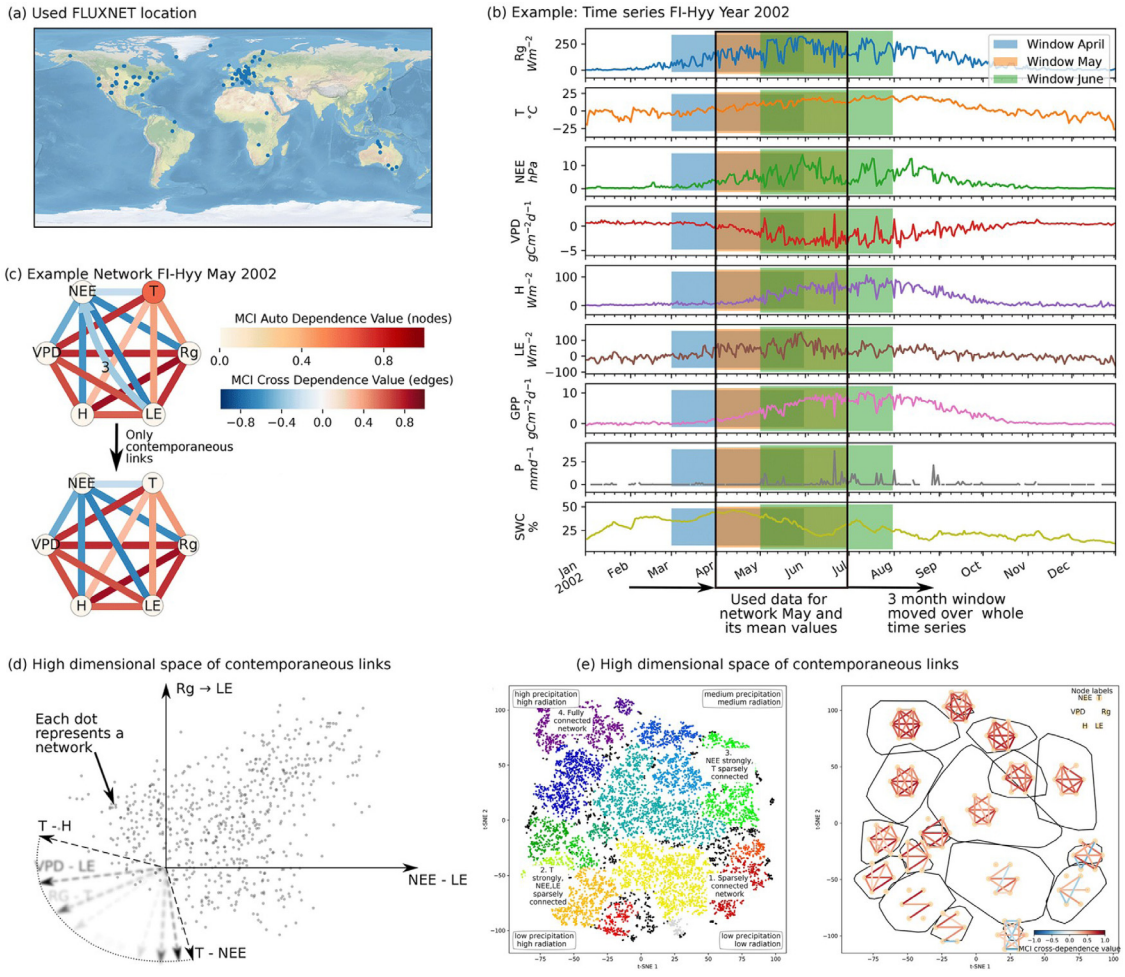


Fig. 18. Clustering causal graphs of local measurement stations of biosphere–atmosphere interactions (adapted from Krich et al. [445]). See the main text for explanations.

alternatively relies on bootstrap resampling through time and the derivation of more stringent cross-map skill scores. The method also exploited the information-geometric causal inference (IGCI) method in [449] to infer weak and strong causal relationships between variables and estimate the embedding dimension to derive global maps of causal relations.

Let us exemplify the RCCM method to discover interactions of three key variables in the carbon cycle: moisture, photosynthesis and air temperature (Tair). For that, we use data compiled in the Earth System Data Lab (ESDL), which contains harmonised products with a spatial resolution of 0.25° and a temporal resolution of 8 days, spanning over 11 years from 2001 to 2011. The RCCM method is applied in each grid cell, which allows us to infer spatial patterns of causal relations between several key variables of the carbon and water cycles.

Fig. 19 shows GPP drives Tair mostly in cold ecosystems due to changes in land surface albedo. Results show GPP is an important forcing of local temperature in many areas. Recent studies have found temperature is an important factor of GPP, driven by radiative factors in cold climates and turbulent energy fluxes in warmer, drier ecosystems. SM and Tair are closely linked, limiting evaporation and raising Tair under dry conditions. This could explain the significant impact of Tair in high latitudes. GPP is mainly influenced by Tair in water-limited regions, especially in high northern latitudes where cold temperatures limit photosynthesis and plant growth. GPP and ET are tightly related as carbon assimilation in plants is linked with water losses through transpiration [450]. Low water availability reduces GPP and ET, causing increased air and surface temperatures and a drier atmosphere. SM being stronger than GPP is mostly seen in transitional wet/dry climates [451]. No strong forcings in tropical rainforest areas indicate GPP is mostly driven by solar radiation and affected by high VPD values [452].

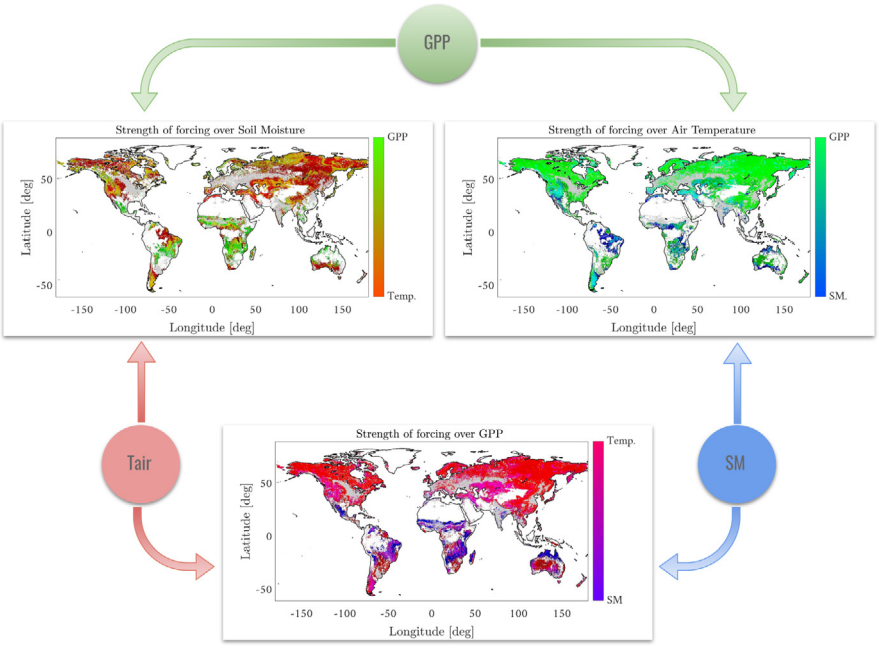


Fig. 19. Applying RCCM in [122] to discover causal relations between GPP, Tair and SM. GPP drives Tair in cold ecosystems; Tair controls SM in water-limited areas; GPP dominates SM. Croplands were masked to avoid interference from human activity.

4.4. Causal climate model intercomparison

As introduced in Section 2.3.7, causal inference can help to assess the output of physical models and evaluate and compare them against observations at the level of causal dependencies [45,263–265]. Climate models [453] provide short-term predictions and future climate projections under given anthropogenic emission scenarios and are the basis for climate-related decision-making. As models can only provide an approximation to the real system, it is essential to evaluate them against observations. Such climate model evaluation is largely based on means, climatologies, or spectral properties [263,454]. Here the problem of equifinality may occur: even though a particular model might well fit descriptive statistics of the data, the model might not well simulate the causal physical mechanisms that produce this statistic, given that multiple model formulations and parameterisations, even when wrong, can fit the observations equally well. The issue is that such models would lead to erroneous future projections — a causal problem of out-of-distribution prediction. Causal model evaluation [10] can evaluate the ability of models to simulate the causal interdependencies of its subprocesses in a process-based model evaluation framework [455].

Here we briefly summarise one approach in this direction [265]. The author aimed to compare causal networks among regional subprocesses in sea-level pressure between observations and climate models of the CMIP ensemble [453]. Fig. 20**a–d** illustrates the method's steps. First, the regional subprocesses were constructed from gridded climate time series (daily-mean sea level pressure from the NCEP-NCAR reanalysis [456]) using Varimax principal component analysis (PCA) to obtain a set of regionally confined climate modes of variability (Fig. 20**b**). The Varimax-PCA weights were then applied to the pressure data from each climate model (the regional weights' cores are indicated in red). Each component is associated with a time series (3-day averaged) and is one of the causal network nodes. Then the causal discovery method PCMCI [170] was applied to these time series to reconstruct the lagged time series graph among these nodes, which constitute characteristic causal fingerprints (Fig. 20**c,d**) for the observational data as well as the individual models. Node colours indicate the level of autocorrelation (auto-MCI-partial correlation [170]), and link colours the cross-link strength (cross-MCI); time lags are indicated by small labels. Only the around 200 most significant links are shown.

These causal fingerprints can then be used for model evaluation and intercomparison. Fig. 20**e** depicts a comparison of models among each other, that is, the matrix of average F1-scores for pair-wise network comparisons between ensemble members of 20 climate models (labelled following CMIP-nomenclature in capital letters) for simulations spanning approximately the historical period from 1948 to 2017 and two surrogate models (Random, Independent). The rows show the models taken as references in each case, and the columns indicate the models compared to these references. Higher scores imply a better agreement between networks, i.e., that the two models are more similar regarding their causal fingerprint. One can see that causal fingerprints from different ensemble members of the same model (diagonal in Fig. 20**e**) are more consistent than networks estimated from two different models (off-diagonal). The blocks are consistent with different models sharing a common development background. In Fig. 20**f**, the models' causal fingerprints are each

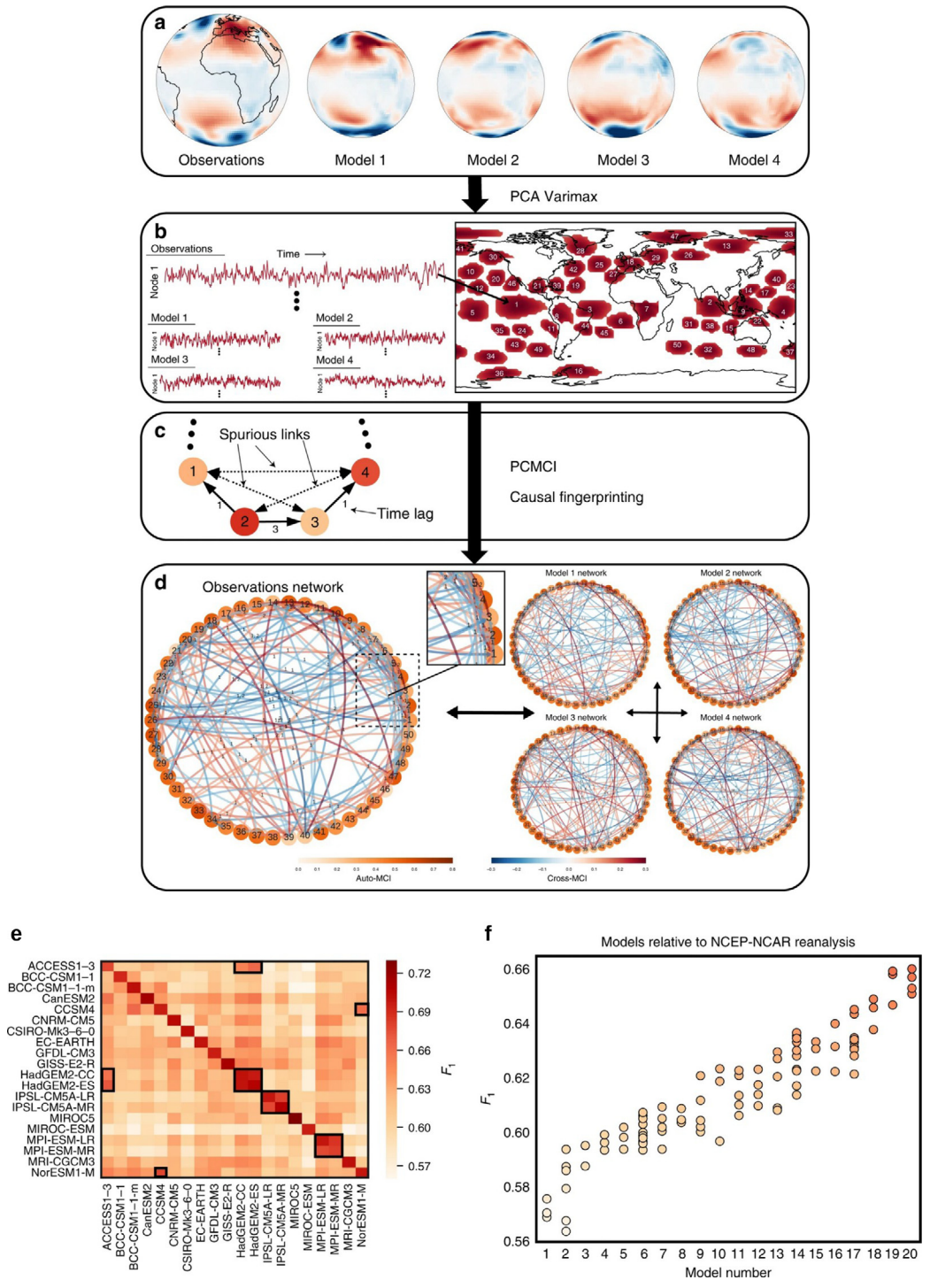


Fig. 20. Causal climate model evaluation (adapted from Nowack et al. [265]). See the main text for explanations.

compared to the fingerprint of the observational data (ordered F1-scores). The result is a continuum of more- and less-similar models (but models have significantly different causal fingerprints). The networks can be further investigated to analyse which regional interactions the models differ more from observations.

Causal model evaluation can provide important information to model developers on where their models can be improved. Furthermore, Nowack et al. [265] show that more realistic fingerprints also affect projected changes in land

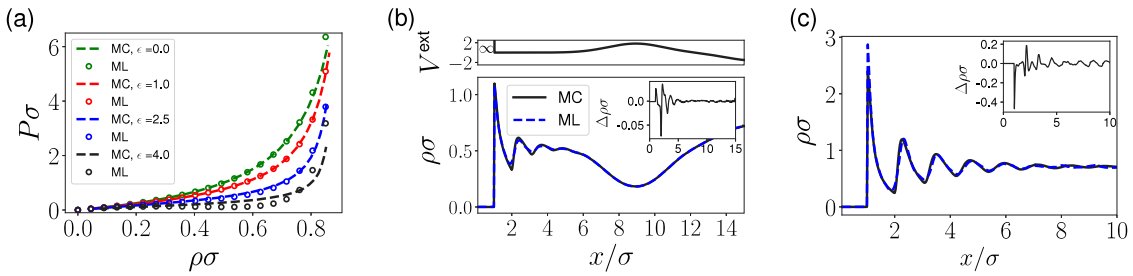


Fig. 21. Results for learning the density functional for Lennard-Jones fluids. (a) shows the equation of state $P(\rho)$ (pressure) for different interaction strengths ϵ comparing Monte-Carlo simulations (MC) with the Machine learning (ML) results. (b) density profile for $\epsilon = 1.25$, $\mu = \ln(1.15)$ inside the training region, but V is not in the training data. (c) density profile at a hard wall for $\epsilon = 1.9$, $\mu = \ln(1.9)$ (outside the training region $\epsilon \in [0.5, 1.5]$). Dark solid lines are simulation profiles, and blue dashed lines are ML results. Insets in (b) and (c) show $\Delta\rho = \rho^{\mathrm{mc}} - \rho^{\mathrm{ml}}$.

surface precipitation. Hence, causal model analyses could be used to constrain climate change projections. The assumption is that the underlying physical processes (e.g., large-scale circulation) lead to dynamical coupling mechanisms captured in the causal fingerprints. One may now argue that high modelling skill on historical data is also relevant for modelling future changes if the physical processes remain important under future climate change.

4.5. Learning density functionals

Being able to describe many-body systems is exciting and important for many applications. Density functional theory Evans [457] (DFT) is an approach to creating a description for classical and quantum many-body systems in equilibrium. The aim is to find a unique (free) energy functional that gives rise to the particle density profile. The analytical form of the (free) energy functional is generally unknown, except for a handful of particular model systems. One way to treat more complex systems is to perform computer simulations and learn the energy functional via machine learning. The first attempts in classical DFT used a convolutional network [458], which does not allow much theoretical insight.

In Lin et al. [319], the above-mentioned symbolic regression method, EQL [101], was adapted to represent part of the energy function. This is an interesting application, as the problem contains known parts of the computational pipeline that we do not want to replace and other parts that should be replaced via the data-driven approach. The fact that EQL can be embedded into any differentiable computational structure is crucial here.

The problem can be formulated as a self-consistency equation:

$$\rho(x) = \exp\left(\mu - \frac{\delta F(\rho(x))}{\delta\rho}\bigg|_{\rho=\rho^{\mathrm{eq}}} - V\right), \tag{20}$$

where ρ is the particle density, F is the external free energy functional that needs to be learned, and μ and V are chemical and external potential, respectively. Notice that the derivative of F (which is represented by an EQL network) is used in the equation. An analytical description for F can be obtained using symbolic regression on simulation data. In Lin et al. [319], for the case of hard rod particles and Lennard-Jones fluids, solutions were found that extrapolate well to unseen situations (different external potential or mean density), as shown in Fig. 21. It is a promising approach to gain more theoretical insights when applied to less studied systems.

4.6. Discovering and assessing governing equations in boundary-layer transition to turbulence

A classical approach to discovering the governing equations of a reduced-order model (ROM) describing a particular phenomenon, for which the governing partial differential equations (PDEs) are known, is to perform Galerkin projection [459,460]. In Galerkin projection, a set of orthogonal basis modes (obtained, for instance, via POD) are used to develop a ROM of the system from data. Then, the governing PDEs are projected onto these modes, transforming the PDEs into a system of ordinary differential equations (ODEs) governing the dynamics of the temporal coefficients associated with those modes [93].

For incompressible fluid flows, the spatiotemporal velocity vector $\boldsymbol{u}(\boldsymbol{x}, t)$ (where $\boldsymbol{x}$ are the spatial coordinates and t time) can be expressed as follows after performing POD:

$$\boldsymbol{u}(\boldsymbol{x}, t) \simeq \boldsymbol{u}_0(\boldsymbol{x}) + \sum_{k=1}^{r} a_k(t)\boldsymbol{u}_k(\boldsymbol{x}), \tag{21}$$

where $\boldsymbol{u}_0$ is the mean flow, $\boldsymbol{u}_k$ are the spatial modes, $a_k(t)$ are the temporal coefficients and r is the number of retained modes in the ROM. The expansion (21) is then substituted into the governing PDEs, *i.e.* the incompressible Navier–Stokes equations, taking advantage that the POD modes are linear combinations of the instantaneous flow realisations (thus

satisfying the boundary conditions) and are solenoidal (, *i.e.* divergence-free, due to the incompressibility condition). It is then possible to take an inner product in space with $\boldsymbol{u}_i(\boldsymbol{x})$. Since the POD modes are orthogonal, a set of ODEs can be obtained for the time derivatives of the temporal coefficients $\mathrm{d}a_i(t)/\mathrm{d}t$ as a function of the spatial modes and also $a_i(t)$. Despite being a widely-used method, it has the limitations of requiring knowledge of the underlying PDEs, and it also may exhibit convergence problems in more challenging scenarios. As discussed in Section 3.2, another alternative to produce a ROM for physical systems in a purely data-driven way is dynamic-mode decomposition (DMD), in which the obtained modes are orthogonal in time [461]; note that this approach, in its compressive version [462], shares similarities with the eigensystem-realisation algorithm (ERA) [463]. In this sense, DMD and its connections with the Koopman operator [464] were exploited by Eivazi et al. [465] to reproduce the dynamics of a near-wall model of turbulence [466] by using external forcing to reproduce the nonlinear behaviour of the system [321,326].

In addition to the approaches mentioned above, other techniques enable learning the equations of motion just from data, as discussed in the early work by Crutchfield et al. [467]. These approaches typically rely on a library of candidate functions to build the resulting governing equation and solve an optimisation problem to obtain the expression that best represents the data. Note that it is essential to use any knowledge on the physical properties of the analysed data to inform the library (e.g. whether non-linearities, periodicities, etc. are present in the system that produced the data under study), as well as to define the relevant state variables, sampling rate, the initial set of parameters defining relevant trajectories, etc. Embedding prior physical information into the obtained model is crucial for the success of these approaches, and failing to do so may lead to rate and even incorrect models [468]. Furthermore, these approaches typically suffer from the curse of dimensionality [469], making it even more important to make the right choices in the library of candidate functions to ensure convergence; thus, being able to define the best basis functions to reduce the dimensionality of the system while retaining the most relevant physics is also is critical. Generally, only after solving the optimisation problem is it possible to assess which terms in the library are necessary and which ones may be combined, a fact that complicates *a-priori* equation discovery.

SINDy has been successfully applied to boundary-value problems [470] using forcing functions and performs well even with noise. Furthermore, SINDy has produced very successful results in a wide variety of fluid-mechanics problems, ranging from thermal convection [410], chaotic electroconvection [409], the so-called "fluidic-pinball" problem [471], turbulent wakes [472] and ROM development [473]. Interestingly, SINDy has also been successfully combined with autoencoders to discover low-dimensional latent spaces [94], benefiting from the non-linear data-compression capabilities of the latter and the interpretability of the former. This is certainly a promising direction to discover hidden complex relations in fluid-flow data and other high-dimensional physical systems, which requires further investigation, particularly when obtaining deeper insight into the interpretation of the latent variables.

Besides the methods above based on discovering nonlinear dynamical systems, other strategies exist to obtain equations from data. For instance, gene-expression programming (GEP), a branch of evolutionary computing [474], is based on having a population of candidate functions to build the solution that best approximates the data and progressively improving this population by the survival of the fittest. The main advantage of this approach is that it leads to closed-form equations, even for data where the governing equation is unknown. In principle, it leads to interpretable solutions (although, in some cases, the resulting equations are so convoluted that interpretability is complicated). GEP has been used to model turbulence [475], particularly in the context of the so-called Reynolds-averaged Navier–Stokes (RANS) equations. In short, the RANS equations are obtained after decomposing the instantaneous velocity into a mean and a fluctuation component (Reynolds averaging [476]), and although this simplifies the flow-simulation process (RANS approaches are widely used in industry), the so-called closure problem emerges [477,478]. This problem is associated with the unknown impact of turbulent fluctuations on the mean flow. All the existing models for these stresses are empirical, which precludes RANS simulations from producing accurate results for arbitrary flow cases. In this context, Weatheritt and Sandberg [479] used GEP to obtain general expressions for these turbulent stresses in various cases, including turbulent ducts [480], which are challenging for RANS models due to the presence of secondary flows. They achieved quite successful RANS models for the secondary flows. GEP effectively obtained more general expressions for the turbulent stresses than those in the classical literature [481], a critical step for RANS models to produce a reasonable performance for complex flows [482].

Finally, we conclude with a technique that is not aimed at discovering equations from data but rather focuses on identifying the dominant terms in the equations for various geometrical regions in the domain under study, given the available data. This method is based on data-driven balance models [483]. It can help improve our system's physical interpretation by understanding the most relevant terms defining various mechanisms in the data, particularly in non-asymptotic cases where the negligible terms are not obvious. Using unsupervised learning, the authors sought clusters of points in the domain with negligible covariance in directions that represent terms with a negligible contribution to the physics, a condition equivalent to stating that the equation is satisfied by a few dominant terms within the cluster. In particular, they used Gaussian-mixture models (GMMs) [484] to cluster the data. Then they obtained a sparse approximation in the direction of maximum variance using sparse principal-component analysis (SPCA) [485]. Callaham et al. [483] show the applicability of this framework to a wide range of problems, including turbulence transition, combustion, nonlinear optics, geophysical fluids, and neuroscience. In all these cases, they obtained relevant insight into the governing equations, which can help uncover novel and unexpected physical relations. In particular, their application to the case of transition to turbulence is very illustrative, as shown in Fig. 22. This figure shows that starting from high-fidelity turbulence data, the RANS equations [477,478] mentioned above are obtained and their terms analysed.

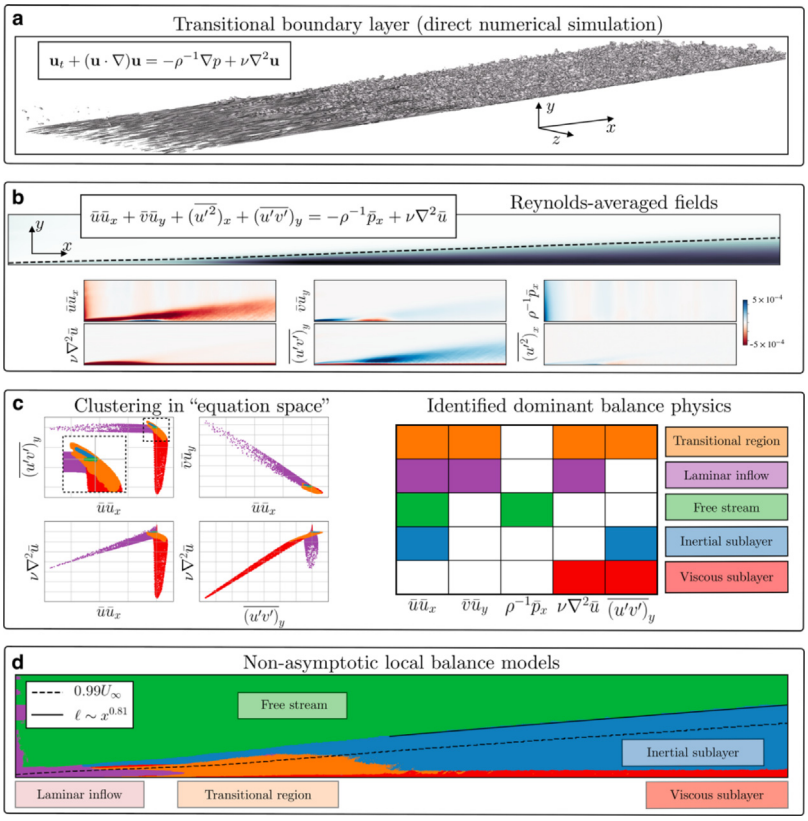


Fig. 22. Data-driven balance model by Callaham et al. [483] applied to a boundary layer undergoing transition to turbulence. (a) Instantaneous data from high-fidelity simulations [486] and (b) terms in the RANS equations obtained from the turbulence statistics. (c) Covariance of the various terms grouped into clusters, labelled based on their physical meaning. (d) Representation of the various clusters in the flow field, together with various boundary-layer quantities.
Source: Figure reproduced from Ref. [483] with permission of the publisher (Springer Nature).

Visualisation of the clustering in equation space reveals some interesting relations, such as the high covariance of the so-called viscous and Reynolds shear-stress-gradient terms, $\nu\nabla^2\overline{u}$ and $\overline{(u'v')}_y$ respectively, which identify the viscous sublayer in the domain. Note that subscripts here denote partial derivatives with respect to the corresponding spatial variable, the overbar indicates time averaging, and the prime is used for fluctuating quantities. The inertial sublayer, which would correspond to the turbulent region in this boundary-layer flow, would be dominated by the convection of the mean flow $\overline{u}\,\overline{u}_x$ and $\overline{(u'v')}_y$, which are again correctly identified through a strong covariance and highlighted in the corresponding region of the domain.

4.7. Learning reduced-order models for vortex shedding behind an obstacle

In this section, we illustrate the possibility of learning ROMs in the case of flow around an obstacle, focusing on the wake. One possibility is to perform a modal decomposition, for instance, based on POD, and then carry out Galerkin projection of the governing Navier–Stokes equations onto the POD modes, as discussed in Section 4.6. This would lead to differential equations governing the temporal evolution of the POD coefficients associated with the spatial modes. This approach may exhibit two main problems in the case of turbulent flows, namely the possible numerical challenges of performing Galerkin projection and the need for many modes to reconstruct a significant fraction of the flow energy. As stated above, autoencoders can provide a compressed version of the original data by exploiting non-linearities, thus exhibiting the great potential to express high-dimensional turbulence data in a few non-linear modes. As shown by Eivazi et al. [349], it is possible to learn a reduced representation of the original data where the latent vectors expressed in physical space exhibit orthogonality. This is achieved by promoting the learning of a latent space with disentangled latent vectors, which also enables learning parsimonious latent spaces. This is done by regularising the loss function, where the associated hyperparameter β gives the name to the β-VAE framework discussed in Section 3.2. Larger values of β give more weight to the term in the loss responsible for learning statistically-independent latent variables, therefore, when $\beta = 0$ one obtains the standard reconstruction loss function. In contrast, larger values of β lead to higher orthogonality of

the learned modes. At the same time, larger β values will yield a worse reconstruction for the set number of latent vectors in the model. Based on this trade-off, it is possible to obtain a good balance between reconstruction and orthogonality. Eivazi et al. [349] illustrated this on the turbulent flow around two wall-mounted obstacles and showed that with only 5 AE modes, it is possible to reconstruct around 90% of the turbulent kinetic energy (TKE) with over 99% orthogonality of the modes. In comparison, 5 POD modes only reconstruct around 30% of the TKE. This is very interesting because the β-VAE, which, unlike other AE-based methods, produces orthogonal modes, yields a reduced representation that can be interpreted from a physical point of view. The first AE and POD modes are very similar, identifying the shear layers around the obstacles and the wake shedding. However, the AE modes exhibit a broader range of scales, incorporating additional higher-frequency turbulent fluctuations into the basic identified features (similar to those in the POD results). Consequently, there is great potential for this type of method to shed light on the physics of complex turbulent flows, in particular when using novel data-driven methods, such as transformers [487,488], to predict the dynamics of the latent space.

Another linear approach discussed above to obtain low-dimensional representations of the flow is dynamic-mode decomposition (DMD), which is based on building a linear operator connecting the instantaneous snapshots in time. Unlike POD, the DMD modes are orthogonal in time, *i.e.* they are associated with a single frequency, which helps identify temporal features in fluid flows. HODMD enables establishing more complex relationships among snapshots, and although it requires additional hyper-parameter tuning, it can help to identify more detailed patterns in the flow. Martínez-Sánchez et al. [489] used HODMD to study the turbulent flow in a simplified urban environment, emphasising the structures behind a wall-mounted obstacle. In this type of flow, a number of flow features emerge around the obstacle [490], where a very important feature is the so-called arch vortex. This vortex, where the legs exhibit wall-normal and the spanwise roof vorticities, is responsible for the high concentration of pollutants in urban environments; therefore, understanding its formation mechanisms can have important implications for urban sustainability. In this context, HODMD enabled identifying two types of modes, namely vortex-generator and vortex-breaker features. The former is associated with low frequency, whereas the latter exhibits higher frequency, and both play important dynamic roles in flow physics. Another extension of the HODMD method also applied to the flow around a wall-mounted obstacle, was proposed by Amor et al. [491]. This study featured the so-called on-the-fly version of HODMD. The data is analysed dynamically as the simulation is run, without storing massive amounts of data for post-processing. Furthermore, more refined criteria for convergence of the modal decomposition were proposed, thus yielding a more effective way to analyse the data. Consequently, this on-the-fly approach reduces up to 80% in memory requirements compared with the traditional offline method. This is a big advantage when applied to large-scale numerical databases.

Causality maps, discussed in Section 2, have been used to study the dynamic interactions present in turbulent flows, focusing on the physical roles of various features. In particular, Lozano-Durán et al. [492] studied the time series of the first Fourier modes in a turbulent channel. They found the following strong causal relations among modes: i) wall-normal modes causing streamwise modes, a phenomenon very closely connected with the well-known lift-up mechanism [493,494] in near-wall turbulence; (ii) wall-normal modes causing spanwise modes, which is associated with the roll generation, also connected with the lift-up process and the incompressibility of the flow; (iii) streamwise modes causing spanwise ones, and spanwise modes causing wall-normal ones; both phenomena are connected with the mean-flow instability, including spanwise meandering and breakdown of the streaks [495,496]. These causal relations were also identified [497] in other simplified models of near-wall turbulence, such as the nine-equation model by Moehlis et al. [466], a fact that confirms the robustness of the causality framework utilised to study turbulence phenomena. Regarding the flow around a wall-mounted obstacle, the various modes discussed above and their connection with the arch vortex were assessed by Martínez-Sánchez et al. [497] also using causality analysis. As can be observed in Fig. 23 (left), the flow under consideration exhibits large-scale separation at the sharp edges of the obstacle and very prominent vortical structures in the wake. Fig. 23 (right) exhibits the vortex-generator and breaker modes discussed above (associated with low and high frequencies, respectively), as well as an additional type of mode of intermediate frequency, denoted as hybrid mode. Clear causal relations are identified between the vortex-breaker and hybrid modes, closely connected with developing vortex-generator modes. This is of great interest because these causal relations define a sequence of events required for the production of the arch vortex (and the subsequent accumulation of pollutants in urban environments); thus, being able to control and inhibit this sequence of events may lead to novel sustainability solutions in cities (as well as to a deeper physical understanding of these complex turbulent flows).

4.8. Uncovering new physical understanding in wall-bounded turbulence

Turbulent flow is one of the most elusive areas of study within fluid mechanics. The wide range of spatial and temporal scales present in turbulence, and the highly non-linear behaviour that characterises it, significantly complicate the possibilities of gaining a deep physical understanding of the main mechanisms within turbulence; this becomes even more complicated in the case of wall-bounded turbulence, which is ubiquitous in science and engineering. Turbulence is characterised by coherent structures, three-dimensional regions that instantaneously satisfy certain physical properties. Note that this term sometimes refers to the features extracted by modal analysis. Still, we will consider the above definition in this work's context. A very important coherent structure in wall-bounded turbulence is the near-wall streak, extensively studied in the 1960s by Kline et al. [499]. As reported by Kim et al. [500], the near-wall production of

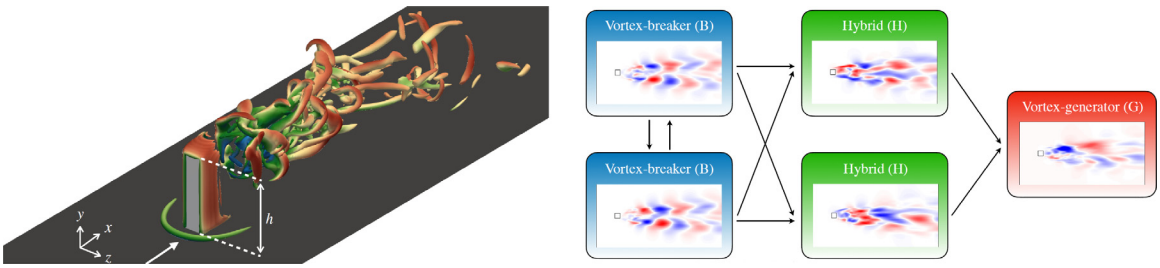


Fig. 23. (Left) Instantaneous snapshot of the flow around a wall-mounted square cylinder, where the vortex clusters are identified with the Q criterion [498]. The structures are coloured by their streamwise velocity, ranging from (dark blue) negative to (dark red) positive values. (Right) Schematic representation of the causal relations among modes, where two vortex-breaker (B), two hybrids (H), and one vortex-generator (G) modes are shown.
Source: Figure adapted from Ref. [497].

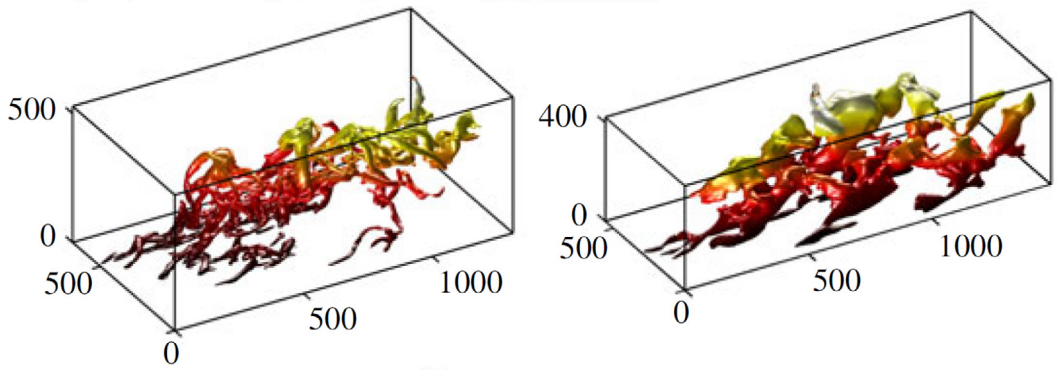


Fig. 24. Coherent structures in a turbulent channel flow. We show (left) a vortex cluster and (right) an intense Reynolds-stress event.
Source: Figure adapted from Ref. [504] with permission from the publisher (Cambridge University Press).

turbulence is very closely connected with the dynamics of these streaks. Another important quantity in wall turbulence is the Reynolds shear stress, briefly introduced in Section 4.6. This quantity is essentially a correlation between streamwise (u') and wall-normal (v') fluctuations and is responsible for the wall-normal momentum transport. Studying the coherent structures most relevant to the development of the Reynolds stresses is a critical goal for reaching a deeper understanding of turbulence. Several studies in the 1970s [501,502] focused on the quadrant analysis to carry out this task; in this analysis, different near-wall events are classified in terms of the sign of their fluctuations, where the most dominant events are the so-called sweeps ($u' > 0$, $v' < 0$) and ejections ($u' < 0$, $v' > 0$). More recently, del Álamo et al. [503] have studied vortex clusters in turbulent channels, and Lozano-Durán et al. [504] have analysed extreme Reynolds-stress events in the same flow case. The latter is defined as the three-dimensional connected regions satisfying:

$$|u'v'| > Hu_{\mathrm{rms}}v_{\mathrm{rms}}, \tag{22}$$

where the subscript 'rms' denotes root-mean-squared quantities, and H is an empirical threshold denoted as a hyperbolic hole. In Fig. 24, we show both types of coherent structure in a turbulent channel flow. Additional insight into the role of both types of structures can be obtained by tracking their evolution in time, such that the various structure interactions (advection, merges, splits, and dissipation) can be assessed [505]. Convolutional neural networks (CNNs) have been used to predict the temporal evolution of the structures in turbulent channels [506], an approach that enables a deeper understanding of their dynamic behaviour. In turbulence, there is a direct cascade of energy from the larger, energy-containing structures towards the smaller dissipative ones; however, there is also an energy path in the opposite direction [507]. This picture, observed in homogeneous isotropic turbulence, becomes even more complicated in the case of wall-bounded turbulence [508]. Each wall-normal location has a different energy cascade because the wall segregates the flow by introducing wall-normal inhomogeneity. A comprehensive review of coherent structures in turbulence was provided by Jiménez [509], who highlighted the potential and challenges of this perspective on turbulence. A very interesting open question raised in this work is the multi-scale organisation and interaction among the various individual structures and how they can dynamically produce the underlying physics of the flow.

Despite the extensive body of work on coherent structures in turbulence, there are still a number of open questions regarding the objective identification of the structures which play the most important role in the dynamics of turbulent

flows. This fundamental question has implications for the theoretical knowledge of the physics of turbulence and the potential of flow-control strategies. If it is possible to identify these structures and they can be suppressed, there may be potential for novel and effective drag-reduction techniques. The vortex clusters and Reynolds-stress structures were defined based on historical reasons and physical intuition. Although they play an important role in the flow, it is unclear whether these are the most relevant motions. A new type of structure that maximises the momentum transfer in a turbulent channel was identified by Jiménez [510], and he reported significant differences between these and the Reynolds stresses. An extension of this idea was implemented in two-dimensional decaying turbulence by removing subregions of the domain and assessing their relative influence in the future evolution of the flow [511]. The idea is to quantify the "significance" of the various regions, and the result confirmed the initial physical intuition regarding this case: the most significant regions were vortices. The least significant ones exhibited high strain. In this direction, Cremades et al. [512] proposed an approach to exploit the explainability of neural networks to assess the relevance of the coherent structures in turbulent flows. In this study, the SHapley Additive exPlanations (SHAP) framework [513,514] was used on the coherent structures identified in a turbulent channel; more concretely, the intense Reynolds stresses were first identified, and then a CNN was used to predict the location of those structures in the next time step [506]. The SHAP technique allows for identifying the impact of each of the features in the input (in this case, the three-dimensional Reynolds-stress events) on the prediction of the next step, thus enabling an assessment of their relevance to the future evolution of the flow. This framework could be used to find new ways of objectively identifying coherent regions in the flow. Another way to gain insight into the detailed mechanisms of turbulence via neural networks is to perform flow estimation, e.g. from the quantities measured at the wall to the turbulent fluctuations above [515,516]. After training a neural network to make this prediction, detailed knowledge of the connection between the scales at the wall and the ones above can be gained through neural-network interpretability [517]. This approach allows us to discover a symbolic equation that can reproduce the predictive capabilities of the network. This can be achieved through the methodology developed by Cranmer et al. [518], which relies on symbolic regression (e.g. based on genetic programming) to obtain the equation relating input and output; see Section 4.6 for a related discussion. By analysing such an equation, it is possible to identify the characteristics of the scales relevant to this wall-normal interaction in wall-bounded turbulent flows.

4.9. Discovery of ocean mesoscale closures

The closure problem, described above in RANS equations, apply to many ocean and atmosphere modelling. In climate modelling, we must resolve (spatial) scales from metres to thousand kilometres. However, due to computational limitations, we need to truncate the spatial spectrum at a given scale — equivalent to the grid spacing of the numerical climate model. Therefore, all processes occurring below the spatial scales need to be approximated — this is the so-called parameterisation or closure problem for subgrid processes. The closure problem, described above in RANS equations, apply to many ocean and atmosphere modelling.

While RANS separates terms into time-averaged and fluctuating components, the most common approach is based on Large Eddy Simulation (LES), in which the filtering separates into a resolved scale and a sub-grid scale. The LES decomposition is based on the self-similarity of small-scale turbulent structures. The resolved scales are defined using a convolution integral with associated physical width, usually the grid cell size. Commonly used filters are box filters, normalised Gaussian, or a combination of both filters. Applying the filtering to the governing equations of the fluid (momentum and buoyancy) gives rise to a set of equations for the resolved scale, with a term—coined subgrid scale forcing—which depends on the fine scale. For the momentum equation, this term subgrid term would be expressed as

$$\mathbf{S} = \begin{pmatrix} S_x \\ S_y \end{pmatrix} = (\overline{\mathbf{u}} \cdot \overline{\nabla})\overline{\mathbf{u}} - \overline{(\mathbf{u} \cdot \nabla)\mathbf{u}}, \tag{23}$$

where ∇ is the horizontal 2D gradient operator, and the horizontal velocity $\mathbf{u} = (u, v)$, and the overline denotes the filtered (hence resolved) velocity on the grid.

Therefore $\mathbf{S}$ must be approximated with only resolved scales $\overline{\mathbf{u}}$ since the total variable $\mathbf{u}$ is not available to the model. Typically turbulence subgrid closures in a fluid are ad-hoc, such as Smagorinsky-type closures, in which the form of the closure is based on some physical argument that is assumed to hold across scale and regimes. This is rarely the case.

For ocean and atmosphere problems, the closure idea can also be boiled down to finding an expression for multiscale interaction that only depends on the resolved scale of the fluid. Similarly to traditional fluid problems, closures or parameterisations in the ocean and atmosphere modelling is often empirical and a source of error in simulations. Instead, equation discovery algorithms, as discussed in previous sections, can be used to uncover relationships between variables. For the closure problem, these algorithms can be applied to derive equations that describe the behaviour of the subgrid scales using resolved variables based on simulated data. The goal is to find the simplest (in some sense) mathematical relationship that accurately captures the behaviour of the subgrid-scale model, which can then be used for prediction. The main advantage of equation discovery algorithms is that they can uncover relationships that may not be immediately apparent, reducing the need for expert knowledge and human intuition in model building, as typically done for parameterisation in ocean and atmospheric modelling.

Zanna and Bolton [519] used Relevance Vector Machine (RVM), a sparse Bayesian regression algorithm, to find closure models for momentum and buoyancy subgrid forcing. The RVM algorithm finds the most relevant input features—functions of the resolved scales—that will describe the subgrid-scale model. The RVM starts with many basis functions

and iteratively removes irrelevant basis functions, arriving at a compact set of basis functions that best represent the data. Compared to other methods, the RVM algorithm has the advantage of handling noisy and redundant data and high-dimensional input spaces. Finally, it provides a probabilistic output, which can be a useful measure of uncertainty.

Below is a closure found by Zanna and Bolton [519], using data from an ocean primitive equation model

$$\mathbf{S} \approx \kappa_{BT} \overline{\nabla} \cdot \begin{pmatrix} \zeta^2 - \zeta D & \zeta \tilde{D} \\ \zeta \tilde{D} & \zeta^2 + \zeta D \end{pmatrix}, \tag{24}$$

where $\zeta = \overline{v}_x - \overline{u}_y$, $D = \overline{u}_y + \overline{v}_x$, $\tilde{D} = \overline{u}_x - \overline{v}_y$, the short-hands $()_{x,y} \equiv \frac{\partial}{\partial x,y}$ are used for spatial derivatives, ζ is the relative vorticity, and D and $\tilde{D}$ are the shearing and stretching deformation of the flow field, respectively. The authors were able to relate the found expression to energy transfer across scales, which mimics the impact of unresolved scales on large-scale energetics.

However, sparse linear regression entails trade-offs between the size and expressiveness of the feature library and the complexity and cost of sparse regression, as discussed in Zanna and Bolton [519] and above. If we wish to include a deep library of functions, the number of different expressions needed will grow exponentially and might be limited by accurately taking derivatives of functions. Finally, many expressions might be highly correlated, preventing convergence [520].

As discussed above, genetic programming (GP) [291] is an alternative approach. GP algorithms, unlike sparse regression, do not require a defined library of functions. [521] used GP with some modifications, including building spatial derivatives in spectral space and combining them with sparse regression to find robust expressions in turbulent datasets generated by idealised simulations Focusing on results from [521], they look for the missing subgrid forcing for potential vorticity, q — a variable that combines momentum and buoyancy effects in geophysical flows, and related to $\nabla \times \mathbf{u}$. In the first few iterations, the algorithm discovered quadratic expressions proportional to $(\overline{\mathbf{u}} \cdot \nabla)\overline{q}$, similarly to previous theoretical studies [522,523]. Often these expressions cannot be used as standalone parameterisations implemented in coarse-resolution models due to numerical stability constraints. The next few iterations of the GP-sparse regression algorithm led to eddy-viscosity models that dissipate energy at small scales, $\nabla^4\overline{q}$ and redistribute energy to larger scales, i.e. kinetic energy backscatter $\nabla^6\overline{q}$ [524]. Additional terms, which are cubic in model variables and contain a double-advection operator, $(\overline{\mathbf{u}}\cdot\nabla)^2$, can ensure dissipation of enstrophy [525], helping with model stability. In addition, there were additional terms that we were not discovered previously. In summary, our discovered closure contains elements of existing subgrid parameterisations, which have pros and cons when used as standalone ones but, when combined, could capture all necessary properties for stable implementation and accurate representation of momentum, energy and enstrophy fluxes missing at coarse resolution.

To test our discovered closure, we implement it in a coarse-resolution simulation (see Fig. 25). The goal is to improve the physics of the coarse-resolution model (panel B) relative to the high-resolution model simulation (panel A). To this end, we run the coarse-resolution simulation with a physics-based (empirical parameterisation; panel C), with data-driven parameterisation learned using a convolutional neural network (panel D), and with the equation-discovery parameterisation (panel E). All simulations are improving the flow, and some aspects of the statistics are also improved. However, generalisation is vastly different without retraining the data-driven driven parameterisations or tuning the physics-based parameterisations. We test our parameterisations in the same model in which we changed the rotation rate in order to form jets and less isotropic turbulence (panels F for high resolution and G for low resolution without any parameterisation). The physics-based parameterisation has little impact on the flow (panel H), but the implementation of the deep learning parameterisation has a very detrimental effect on the flow, most trying to make the flow more isotropic (panel I). On the other hand, the implementation of the equation discovery-based parameterisation substantially improves the flow (panel J) — reinforcing the need to discover relationships from data that encapsulate the necessary laws of physics to mimic the scale interactions which are internal to the fluid and not dependent on the configuration of the simulations.

This symbolic parameterisation includes up to the seventh spatial derivative of $\overline{q}$, which may be unrealistic to implement into a climate model. However, it might be more realistic than a fully non-local approach, such as the convolutional neural network parameterisations considered in other studies or extremely local physics-based parameterisations (such as anti-viscosity). Most importantly, the sparse model can generalise well without retraining, while the neural network-based parameterisations perform poorly.

5. Concluding remarks

The fields of causal and equation discovery have emerged in recent years as important research areas that apply artificial intelligence and machine learning to analyse complex systems [26,83,115]. The fields respectively aim to identify causal relationships and discover equations that can be used to predict the behaviour of the system, including the effects of interventions.

In this paper, we have reviewed the state-of-the-art in both fields and discussed their respective approaches and techniques. Causal discovery aims to discover the qualitative cause-and-effect relationships between the variables in a system. In order to achieve this task using non-experimental data, causal discovery employs certain enabling assumptions (see Section 2). Among these assumptions is that the data-generating process can be described by a structural causal model

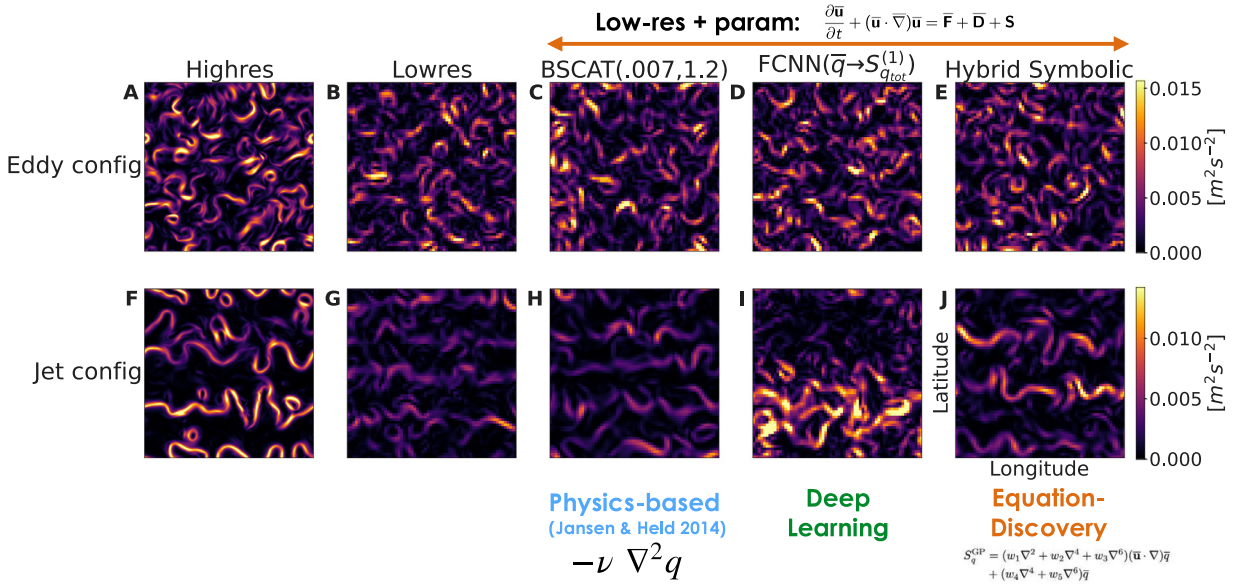


Fig. 25. Snapshot of potential vorticity in two different simulations: Top = Eddy (mostly isotropic turbulence), Bottom = Jet (some elongated sharp features mixed with isotropic turbulence features). A, F: High-Resolution simulations; B, G: Coarse Resolution; C, H: Coarse Resolution with physics-based parameterisations; D, I: Coarse Resolution with a Neural Network-based Dramatisation; E, J: Coarse Resolution with parameterisation discovered with symbolic regression. The data-driven parameterisations are trained on eddy configuration, and only the equation-discovery lead to robust generalisation in different regimes without retraining [521].

and the corresponding causal graph. Methods for causal discovery are manifold and can be partitioned into constraint-based, score-based, asymmetry-based, and context-based methods. Data from the physical world typically comes in the form of time series with autocorrelation and potentially non-stationary behaviour. Autocorrelation and potential non-stationarity pose statistical challenges for many causal discovery methods as they are typically designed for i.i.d. data. In addition, the true functional relationship between variables can be highly non-linear, and the variables can be high-dimensional, both increasing model complexity and affecting the efficiency of causal discovery methods. All these challenges are compounded by the fact that the data acquired from real-world processes are often far from ideal, with problems such as missing data and inherent selection bias that might lead to the observed data not being representative of the process underlying it. These and many other challenges are avenues for future research in causal discovery and many of its sister fields, such as Bayesian networks and conditional independence testing, to name a few.

In equation discovery, the focus is on understanding the structure of a system by discovering equations, state variables and laws that can be used to predict (and, more importantly, to understand) its behaviour (see Section 3). The main techniques used in this field are symbolic regression, evolutionary algorithms, and deep learning. These methods offer the potential to discover both linear and nonlinear equations but suffer from the need for large datasets and the difficulty of finding accurate equations in complex systems. More relevant challenges have to do with identifiability issues and the impossibility of evaluating the generality of the equations or even the criteria to select the most general ones. Broader (and perhaps more philosophical) questions need to be addressed, such as compressibility or sparsity, confronted with expressive power, the role of physical units and modularity, to name a few.

Thus, causality studies and equation inference approaches have synergistic goals. Both fields have made significant advances in recent years and offer considerable promise for further research. In particular, techniques from both fields can be combined to create hybrid models capable of uncovering causal relationships and equations. Additionally, developing more efficient algorithms and better methods for dealing with the challenge of overfitting could lead to further progress in both fields. More specifically, the question remains as to which current approaches provide stronger guarantees for the uniqueness/equifinality of the discovered equation or inferred causal graph. This key aspect in the inferential discovery of physical models should receive more attention in future research enterprises.

A wide range of case studies in many areas of interest in the physical sciences (neurosciences, Earth and climate sciences, fluid mechanics) has illustrated the performance of causal discovery and equation discovery algorithms (see Section 4). We noted that specific methods and techniques reside in particular fields and do not permeate to others, mainly because of the needed assumptions and data characteristics. Yet, as has been the case for centuries, there is a lack of transdisciplinary in science. Directly stemming from this review, it is evident that analysing complex systems requires an inter/trans-disciplinary approach that combines method and domain expertise. Techniques from artificial intelligence, machine learning, and control theory can be combined to better understand a system's behaviour and make

accurate predictions. As such, future research in causal and equation discovery should consider the potential benefits of a more integrative and fused approach to analysing complex systems. This is perhaps especially important for recent developments designed for answering critical questions in specific areas but which, given their fundamental nature, have a wider appeal. For instance, the progress in the empirical inference of transfer operators in fluid mechanics or chemical reaction pathways has unexplored implications in understanding the metastable dynamics of neuronal network responses.

Overall, the fields of causal discovery and equation discovery are rapidly advancing, and there is a growing synergy between them. Despite the remaining challenges, researchers have made great strides in uncovering the underlying structure of complex systems. With continued research and development, we can look forward to further advances in both fields and unlocking complex systems' mysteries.

CRediT authorship contribution statement

Gustau Camps-Valls: Conceptualised and defined the structure of the paper, wrote the introduction and conclusions and participated in writing all sections, especially on equation discovery, contributed to the case studies on the discovery of causal relations among carbon and water fluxes. **Andreas Gerhardus, Urmi Ninad and Gherardo Varando** wrote the causal discovery section. **Georg Martius:** Wrote the equation discovery section and contributed to the case study on learning density functionals. **Emili Balaguer-Ballester:** Wrote the neuroscientific case studies. **Ricardo Vinuesa:** Wrote the equation discovery section, and the case studies related to fluid mechanics. **Emiliano Diaz:** Wrote the causal discovery section, and contributed to the case studies on the discovery of causal relations among carbon and water fluxes. **Laure Zanna:** Wrote the ocean mesoscale closures. **Jakob Runge:** Conceptualised and defined the structure of the paper, wrote the causal discovery section, and contributed to the case studies on the discovery of causal relations among carbon and water fluxes, and on climate model intercomparison.

Declaration of competing interest

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

Acknowledgements

G.C-V. received funding from the European Research Council (ERC) under the ERC Synergy Grant USMILE (grant agreement 855187), the Fundación BBVA with the project 'Causal inference in the human-biosphere coupled system (SCALE)', the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100 (IMIRACLI), the GVA PROMETEO AI4CS project on 'AI for complex systems' (2022–2026) with CIPROM/2021/056, and the European Union's Horizon 2020 research and innovation program within the project 'XAIDA: Extreme Events - Artificial Intelligence for Detection and Attribution', under grant agreement No 101003469. G.V. and E.D. were partly supported by the ERC USMILE project (grant agreement 855187).

J.R. has received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112), from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003469 (XAIDA), from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100 (IMIRACLI), and by the Helmholtz AI project CausalFlood (grant no. ZT-I-PF-5-11). J.R., A.G., and L.Z. were supported in part by the National Science Foundation under Grant No. NSF PHY-1748958. U.N. was supported by grant no. 948112 Causal Earth of the European Research Council (ERC).

G.M. is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645 He acknowledges the support from the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039B). G.M. thanks Gabriel Kornberger for discussions on symbolic regression.

E.B-B. received funding from the Royal Society International Exchanges programme 2021 Round 1 (Ref. IES\R1\211062, "Validating Cortical Network Models at the Edge of Asynchrony"), and from the EU H2020 Research and Innovation Programme under the Grant Agreement No. 945539 (Human Brain Project SGA3), via the Voucher awarded to the Partnering Project "Async-Neuromorph". Funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

R.V. received funding from the European Research Council (ERC) under the ERC Consolidator Grant DEEPCONTROL ("2021-CoG-101043998").

L.Z. received M^2LInES research funding by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. LZ acknowledges funding from the NSF Science and Technology Center, Center for Learning the Earth with Artificial Intelligence and Physics (LEAP).

All authors read, revised, commented and approved the manuscript.

References

[1] K. Popper, The Logic of Scientific Discovery, Routledge, 2005.
[2] P. Munz, Our Knowledge of the Growth of Knowledge: Popper or Wittgenstein? Routledge, 2014.
[3] J. Pearl, Causality: Models, Reasoning, and Inference, second ed., Cambridge University Press, Cambridge, UK, 2009.
[4] J.D. Sterman, Learning in and about complex systems, Syst. Dyn. Rev. 10 (2–3) (1994) 291–330.
[5] J. Kwapień, S. Drożdż, Physical approach to complex systems, Phys. Rep. 515 (3–4) (2012) 115–226.
[6] S. Salcedo-Sanz, D. Casillas-Pérez, J. Del Ser, C. Casanova-Mateo, L. Cuadra, M. Piles, G. Camps-Valls, Persistence in complex systems, Phys. Rep. 957 (2022) 1–73.
[7] R.M. May, Will a large complex system be stable? Nature 238 (5364) (1972) 413–414.
[8] J. Pearl, Causality: Models, Reasoning and Inference, second ed., Cambridge University Press, New York, NY, USA, 2009.
[9] J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference: Foundations and Learning Algorithms, MIT Press, Cambridge, MA, USA, 2017.
[10] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M.D. Mahecha, J. Muñoz-Marí, et al., Inferring causation from time series in Earth system sciences, Nat. Commun. 10 (1) (2019) 1–13.
[11] T. Richardson, A discovery algorithm for directed cyclic graphs, in: Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence, Uai '96, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996, pp. 454–461.
[12] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. 113 (15) (2016) 3932–3937.
[13] B. Chen, K. Huang, S. Raghupathi, I. Chandratreya, Q. Du, H. Lipson, Automated discovery of fundamental variables hidden in experimental data, Nat. Comput. Sci. 2 (7) (2022) 433–442.
[14] D. Wootton, The Invention of Science: A New History of the Scientific Revolution, Penguin UK, 2015.
[15] T.S. Kuhn, The Structure of Scientific Revolutions, University of Chicago Press, 1962.
[16] N. Copernicus, M.-P. Lerner, A.P. Segonds, J.-P. Verdet, C. Luna, D. Savoie, M. Toulmonde, De Revolutionibus Orbium Coelestium, Vol. 1, Johnson Reprint Corporation, 1965.
[17] C. Darwin, On the Origin of Species, John Murray, London, 1859.
[18] J.D. Watson, F.H. Crick, Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid, Nature 171 (4356) (1953).
[19] W. Heisenberg, Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen, Z. Phys. 33 (3) (1925) 879–893.
[20] H. Butterfield, The Origins of Modern Science, Vol. 90507, Simon and Schuster, 1965.
[21] C.J. Ducasse, Whewell's philosophy of scientific discovery. II, Philos. Rev. 60 (2) (1951) 213–234.
[22] P. Langley, Scientific discovery, causal explanation, and process model induction, Mind Soc. 18 (1) (2019) 43–56.
[23] P. Langley, H.A. Simon, G.L. Bradshaw, J.M. Zytkow, Scientific Discovery: Computational Explorations of the Creative Processes, MIT Press, 1987.
[24] D. Klahr, H.A. Simon, Studies of scientific discovery: Complementary approaches and convergent findings, Psychol. Bull. 125 (5) (1999) 524.
[25] W.C. Wimsatt, W.K. Wimsatt, Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality, Harvard University Press, 2007.
[26] J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference: Foundations and Learning Algorithms, The MIT Press, 2017.
[27] J. Pearl, D. Mackenzie, The Book of Why: The New Science of Cause and Effect, Basic books, New York, 2018.
[28] J.F. Donges, Y. Zou, N. Marwan, J. Kurths, Complex networks in climate dynamics, Eur. Phys. J. Spec. Top. 174 (1) (2009) 157–179.
[29] P. Fiedor, Networks in financial markets based on the mutual information rate, Phys. Rev. E 89 (5) (2014) 052801.
[30] H. Johnson, G. Harris, K. Williams, BRAINSFit: Mutual information registrations of whole-brain 3D images, using the insight toolkit, Insight J. (2007).
[31] K. Takagi, Principles of mutual information maximization and energy minimization affect the activation patterns of large scale networks in the brain, Front. Comput. Neurosci. 13 (2020).
[32] G.T. Walker, Correlation in seasonal variations of weather, VIII: A preliminary study of world weather, Mem. Indian Meteorol. Dep. 24 (4) (1923) 75–131.
[33] J.D. Medaglia, M.-E. Lynall, D.S. Bassett, Cognitive network neuroscience, J. Cogn. Neurosci. 27 (8) (2015) 1471–1491.
[34] J. Richiardi, S. Achard, H. Bunke, D. Van De Ville, Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience, IEEE Signal Process. Mag. 30 (3) (2013) 58–70.
[35] C.W. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica (1969) 424–438.
[36] H. Von Storch, F.W. Zwiers, Statistical Analysis in Climate Research, Cambridge University Press, 2002.
[37] C. Chatfield, The Analysis of Time Series: Theory and Practice, Springer, 2013.
[38] H. Reichenbach, The Direction of Time, Vol. 65, Univ of California Press, 1991.
[39] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, S. Munch, Detecting causality in complex ecosystems, science 338 (6106) (2012) 496–500.
[40] J. Pearl, Causal diagrams for empirical research, Biometrika 82 (4) (1995) 669–688.
[41] J. Pearl, Causal inference in statistics: An overview, Stat. Surv. 3 (2009) 96–146.
[42] J. Pearl, M. Glymour, N.P. Jewell, Causal Inference in Statistics: A Primer, John Wiley & Sons, 2016.
[43] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, MIT Press, Boston, 2000.
[44] A. Wagner, Causality in complex systems, Biol. Philos. 14 (1999) 83–101.
[45] A. Pérez-Suay, G. Camps-Valls, Causal inference in geoscience and remote sensing from observational data, IEEE Trans. Geosci. Remote Sens. 57 (3) (2019) 1502–1513.
[46] I. Ebert-Uphoff, Y. Deng, Causal discovery in the geosciences—Using synthetic data to learn how to interpret results, Comput. Geosci. 99 (2017) 50–60.
[47] F. Raia, Causality in complex dynamic systems: A challenge in earth systems science education, J. Geosci. Educ. 56 (1) (2008) 81–94.
[48] F. Reitsma, Geoscience explanations: Identifying what is needed for generating scientific narratives from data models, Environ. Model. Softw. 25 (1) (2010) 93–99.
[49] D. Niemeijer, R.S. de Groot, Framing environmental indicators: moving from causal chains to causal networks, Environ. Dev. Sustain. 10 (1) (2008) 89–106.
[50] T.G. Shepherd, Storyline approach to the construction of regional climate change information, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 475 (2225) (2019) 20190013.
[51] A.E. Goodwell, P. Jiang, B.L. Ruddell, P. Kumar, Debates—Does information theory provide a new paradigm for Earth science? Causality, interaction, and feedback, Water Resour. Res. 56 (2) (2020) e2019WR024940.
[52] A. Reid, D. Headley, R. Mill, R. sanchez romero, L. Uddin, D. Marinazzo, D. Lurie, P. Valdés-Sosa, S. Hanson, B. Biswal, V. Calhoun, R. Poldrack, M. Cole, Advancing functional connectivity research from association to causation, Nature Neurosci. 22 (2019) 1–10.
[53] S.H. Siddiqi, K.P. Kording, J. Parvizi, M.D. Fox, Causal mapping of human brain function, Nat. Rev. Neurosci. 23 (6) (2022) 361–375.

[54] P.A. Stokes, P.L. Purdon, A study of problems encountered in Granger causality analysis from a neuroscience perspective, Proc. Natl. Acad. Sci. 114 (34) (2017) E7063–e7072.
[55] M.M. Marini, B. Singer, Causality in the social sciences, Sociol. Methodol. 18 (1988) 347–409.
[56] F. Russo, Causality and Causal Modelling in the Social Sciences, Springer, 2010.
[57] M.A. Hernán, The C-word: scientific euphemisms do not improve causal inference from observational data, Am. J. Public Health 108 (5) (2018) 616–619.
[58] T.A. Glass, S.N. Goodman, M.A. Hernán, J.M. Samet, Causal inference in public health, Annu. Rev. Public Health 34 (2013) 61–75.
[59] M. Hernan, J. Robins, Causal Inference: What if, Chapman & Hill/CRC, 2020.
[60] J. Hicks, et al., Causality in Economics, Australian National University Press, 1980.
[61] S. LeRoy, Causality in Economics, London School of Economics, Centre for Philosophy of Natural and Social Sciences, 2004.
[62] H.A. Simon, et al., The scientist as problem solver, in: Complex Information Processing: The Impact of Herbert A. Simon, 1989, pp. 375–398.
[63] C.G. Hempel, The Philosophy of Carl G. Hempel: Studies in Science, Explanation, and Rationality, Oxford University Press, 2001.
[64] B. Falkenhainer, R. Michalski, The structure mapping engine: Algorithm and examples, Artificial Intelligence 32 (1) (1986) 1–63.
[65] M. Kokar, Knowledge acquisition: A realization of new artificial intelligence, Artificial Intelligence 32 (1986) 251–290.
[66] J. Żytkow, R. Michalski, R. Stepp, Representation and learning of categorical structures, Mach. Learn. 5 (1) (1990) 7–48.
[67] C. Schaffer, Constructing explanations for propositional knowledge bases, Mach. Learn. 4 (4) (1990) 321–353.
[68] K. Nordhausen, P. Langley, Inverse entailment and progol, Mach. Learn. 5 (1) (1990) 25–38.
[69] P. Moulet, Learning rules from structured data, Mach. Learn. 8 (1) (1992) 47–75.
[70] A. Gordon, A. Moore, A. Carlson, Using genetic algorithms to discover good representations, Mach. Learn. 15 (1) (1994) 239–263.
[71] T. Murata, K. Tanaka, A constructive induction algorithm incorporating prior knowledge, Mach. Learn. 14 (1) (1994) 71–96.
[72] S. Džeroski, L. Todorovski, Reliable induction of recursive production rules, Mach. Learn. 20 (3) (1995) 229–256.
[73] T. Washio, H. Motoda, Inductive inference of first-order rules with non-linear structures, Mach. Learn. 27 (2) (1997) 153–172.
[74] P. Bradley, S. Gold, S. Silverman, Constructive induction from incomplete data: A comparative study, Mach. Learn. 42 (1) (2001) 7–48.
[75] J. Koza, F. Bennett, D. Andre, M. Keane, Nonlinear genetic programming: Automatic discovery of reusable programs, Mach. Learn. 42 (1) (2001) 185–223.
[76] M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data, Science 324 (5923) (2009) 81–85.
[77] N. Simidjievski, L. Todorovski, J. Kocijan, S. Džeroski, Equation discovery for nonlinear system identification, IEEE Access 8 (2020) 29930–29943.
[78] E. Feigenbaum, B. Buchanan, J. Lederberg, The DENDRAL project, AI Mag. 2 (1971) 37–46.
[79] P. Langley, H. Simon, G. Bradshaw, Scientific discovery: Computational explorations of the creative process, AI Mag. 8 (3) (1987) 30–44.
[80] J. Evans, A. Rzhetsky, Machine science, Science 329 (5990) (2010) 399–400.
[81] S. Fortunato, C.T. Bergstrom, K. Börner, J.A. Evans, D. Helbing, S. Milojević, A.M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al., Science of science, Science 359 (6379) (2018) eaao0185.
[82] J. Bongard, H. Lipson, Automated reverse engineering of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA 104 (24) (2007) 9943–9948.
[83] M.D. Schmidt, R.R. Vallabhajosyula, J.W. Jenkins, J.E. Hood, A.S. Soni, J.P. Wikswo, H. Lipson, Automated refinement and inference of analytical models for metabolic networks, Phys. Biol. 8 (5) (2011) 055011.
[84] D. Waltz, B.G. Buchanan, Automating science, Science 324 (5923) (2009) 43–44.
[85] R.D. King, J. Rowland, W. Aubrey, M. Liakata, M. Markham, L.N. Soldatova, K.E. Whelan, A. Clare, M. Young, A. Sparkes, et al., The robot scientist Adam, Computer 42 (8) (2009) 46–54.
[86] P. Langley, H. Simon, G. Bradshaw, Automated discovery in the physical sciences, AI Mag. 23 (3) (2002) 11–28.
[87] P. Langley, H. Simon, G. Bradshaw, Scientific discovery and the future of AI, AI Mag. 23 (3) (2002) 29–39.
[88] A. Kocabas, A genetic programming system for automated discovery in the physical sciences, Mach. Learn. 7 (3–4) (1991) 295–314.
[89] R. King, P. Langley, H. Simon, Automated discovery in the biological sciences, AI Mag. 25 (3) (2004) 21–36.
[90] J. Shrager, P. Langley, Computational Models of Scientific Discovery and Theory Formation, Morgan Kaufmann, 1990.
[91] S. Dzeroski, L. Todorovski, Inductive Logic Programming: Techniques and Applications, Springer Science+Business Media, 2007.
[92] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (Jun) (2001) 211–244.
[93] S.L. Brunton, J.N. Kutz, Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control, Cambridge University Press, 2022.
[94] K. Champion, B. Lusch, J.N. Kutz, S.L. Brunton, Data-driven discovery of coordinates and governing equations, Proc. Natl. Acad. Sci. 116 (45) (2019) 22445–22451.
[95] I. Mezić, Spectral properties of dynamical systems, model reduction and decompositions, Nonlinear Dynam. 41 (1) (2005) 309–325.
[96] E. Kaiser, J.N. Kutz, S.L. Brunton, Data-driven approximations of dynamical systems operators for control, in: The Koopman Operator in Systems and Control, Springer, 2020, pp. 197–234.
[97] E. Kaiser, J.N. Kutz, S.L. Brunton, Data-driven discovery of Koopman eigenfunctions for control, Mach. Learn.: Sci. Technol. 2 (3) (2021) 035023.
[98] V. Kostic, P. Novelli, A. Maurer, C. Ciliberto, L. Rosasco, M. Pontil, Learning dynamical systems via Koopman operator regression in reproducing kernel Hilbert spaces, in: NeurIPS 2022, 2022, pp. 1–9.
[99] S. Klus, I. Schuster, K. Muandet, Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces, J. Nonlinear Sci. 30 (1) (2020) 283–315.
[100] B. Lusch, J.N. Kutz, S.L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, Nature Commun. 9 (1) (2018).
[101] S.S. Sahoo, C.H. Lampert, G. Martius, Learning equations for extrapolation and control, in: J. Dy, A. Krause (Eds.), Proc. 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, Vol. 80, Pmlr, 2018, pp. 4442–4450.
[102] S.-M. Udrescu, M. Tegmark, AI Feynman: A physics-inspired method for symbolic regression, Sci. Adv. 6 (16) (2020) eaay2631.
[103] M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data, Science 324 (5923) (2009) 81–85.
[104] L. Biggio, T. Bendinelli, A. Neitz, A. Lucchi, G. Parascandolo, Neural symbolic regression that scales, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 139, Pmlr, 2021, pp. 936–945.
[105] A. Pukrittayakamee, M. Malshe, M. Hagan, L. Raff, R. Narulkar, S. Bukkapatnum, R. Komanduri, Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks, J. Chem. Phys. 130 (13) (2009) 134101.
[106] K.T. Schütt, F. Arbabzadah, S. Chmiela, K.R. Müller, A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, Nat. Commun. 8 (1) (2017) 13890.
[107] O. Stegle, D. Janzing, K. Zhang, J.M. Mooij, B. Schölkopf, Probabilistic latent variable models for distinguishing between cause and effect, in: J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (Eds.), Advances in Neural Information Processing Systems, Vol. 23, Curran Associates, Inc., 2010.
[108] R.P. Monti, I. Khemakhem, A. Hyvarinen, Autoregressive flow-based causal discovery and inference, 2020, arXiv preprint: 2007.09390.
[109] A. Diaz, J. Johnson, G. Varando, G. Camps-Valls, Learning latent functions for causal discovery, 4, (3) 2023, p. 035004,
[110] S. Russell, Human-compatible artificial intelligence, in: Human-Like Machine Intelligence, Oxford University Press, Oxford, 2021, pp. 3–23.

[111] M.A. Boden, Creativity and artificial intelligence: A contradiction in terms, in: The Philosophy of Creativity: New Essays, Oxford University Press, New York, 2014, pp. 224–246.
[112] D. Gillies, Artificial intelligence and scientific method, Mind 107 (428) (1998).
[113] C.K. Assaad, E. Devijver, E. Gaussier, Survey and evaluation of causal discovery methods for time series, J. Artificial Intelligence Res. 73 (2022) 767–819.
[114] R. Moraffah, P. Sheth, M. Karami, A. Bhattacharya, Q. Wang, A. Tahir, A. Raglin, H. Liu, Causal inference for time series analysis: Problems, methods and evaluation, Knowl. Inf. Syst. 63 (2021) 3041–3085.
[115] J. Runge, A. Gerhardus, G. Varando, V. Eyring, G. Camps-Valls, Causal inference for time series, Nat. Rev. Earth Environ. 10 (2023) 2553.
[116] K.A. Bollen, Structural Equations with Latent Variables, John Wiley & Sons, New York, NY, USA, 1989.
[117] S. Bongers, P. Forré, J. Peters, J.M. Mooij, Foundations of structural causal models with cycles and latent variables, Ann. Statist. 49 (5) (2021) 2885–2915.
[118] J. Geweke, Measurement of linear dependence and feedback between multiple time series, J. Amer. Statist. Assoc. 77 (378) (1982) 304–313.
[119] D. Bueso, M. Piles, G. Camps-Valls, Explicit Granger causality in kernel Hilbert spaces, Phys. Rev. E 102 (2020) 062201.
[120] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2) (2000) 461.
[121] L. Barnett, A.B. Barrett, A.K. Seth, Granger causality and transfer entropy are equivalent for Gaussian variables, Phys. Rev. Lett. 103 (23) (2009) 238701.
[122] E. Diaz, J. Adsuara, A. Moreno-Martinez, M. Piles, G. Camps-Valls, Inferring causal relations from observational long-term carbon and water fluxes records, Sci. Rep. 12 (2022) 1610.
[123] J. Runge, Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, in: J. Peters, D. Sontag (Eds.), Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), in: Proceedings of Machine Learning Research, vol. 124, Pmlr, 2020, pp. 1388–1397.
[124] J. Runge, V. Petoukhov, J.F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, J. Kurths, Identifying causal gateways and mediators in complex spatio-temporal systems, Nat. Commun. 6 (1) (2015) 1–10.
[125] C.K. Assaad, E. Devijver, E. Gaussier, Discovery of extended summary graphs in time series, in: J. Cussens, K. Zhang (Eds.), Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, in: Proceedings of Machine Learning Research, vol. 180, Pmlr, 2022, pp. 96–106.
[126] D. Malinsky, P. Spirtes, Causal structure learning from multivariate time series in settings with unmeasured confounding, in: T.D. Le, K. Zhang, E. Kıcıman, A. Hyvärinen, L. Liu (Eds.), Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery, in: Proceedings of Machine Learning Research, vol. 92, Pmlr, London, UK, 2018, pp. 23–47.
[127] A. Gerhardus, J. Runge, High-recall causal discovery for autocorrelated time series with latent confounders, Adv. Neural Inf. Process. Syst. 33 (2020) 12615–12625.
[128] C. Meek, Graphical Models: Selecting Causal and Statistical Models (Ph.D. thesis), Carnegie Mellon University, 1997.
[129] D.M. Chickering, Optimal structure identification with greedy search, J. Mach. Learn. Res. 3 (Nov) (2002) 507–554.
[130] D.M. Chickering, Learning equivalence classes of Bayesian-network structures, J. Mach. Learn. Res. 2 (2002) 445–498.
[131] J. Ramsey, M. Glymour, R. Sanchez-Romero, C. Glymour, A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images, Int. J. Data Sci. Anal. 3 (2) (2017) 121–129.
[132] R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, K. Georgatzis, P. Beaumont, B. Aragam, DYNOTEARS: Structure learning from time-series data, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 108, Pmlr, 2020, pp. 1595–1605.
[133] T. Gao, D. Bhattacharjya, E. Nelson, M. Liu, Y. Yu, IDYNO: Learning nonparametric DAGs from interventional dynamic data, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 162, Pmlr, 2022, pp. 6988–7001.
[134] X. Sun, O. Schulte, G. Liu, P. Poupart, NTS-NOTEARS: Learning nonparametric DBNs with prior knowledge, in: The 26th International Conference on Artificial Intelligence and Statistics (AISTATS), 2023.
[135] J. Peters, D. Janzing, B. Schölkopf, Causal inference on time series using restricted structural equation models, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, Vol. 26, Curran Associates, Inc., 2013.
[136] W. Gong, J. Jennings, C. Zhang, N. Pawlowski, Rhino: Deep causal temporal relationship learning with history-dependent noise, 2022.
[137] S. Shimizu, P.O. Hoyer, A. Hyvärinen, A. Kerminen, A linear non-Gaussian acyclic model for causal discovery, J. Mach. Learn. Res. 7 (72) (2006) 2003–2030.
[138] J. Runge, J. Heitzig, V. Petoukhov, J. Kurths, Escaping the curse of dimensionality in estimating multivariate transfer entropy, Phys. Rev. Lett. 108 (2012) 258701.
[139] R. Dahlhaus, M. Eichler, Causality and graphical models in time series analysis, Oxford Stat. Sci. Ser. 27 (2003).
[140] C. Glymour, K. Zhang, P. Spirtes, Review of causal discovery methods based on graphical models, Front. Genet. 10 (2019).
[141] T. Verma, J. Pearl, Causal networks: Semantics and expressiveness, in: R.D. Shachter, T.S. Levitt, L.N. Kanal, J.F. Lemmer (Eds.), Uncertainty in Artificial Intelligence, in: Machine Intelligence and Pattern Recognition, vol. 9, North-Holland, 1990, pp. 69–76.
[142] D. Geiger, T. Verma, J. Pearl, Identifying independence in Bayesian networks, Networks 20 (5) (1990) 507–534.
[143] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
[144] T. Verma, J. Pearl, Equivalence and synthesis of causal models, in: P.P. Bonissone, M. Henrion, L.N. Kanal, J.F. Lemmer (Eds.), Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, Uai '90, Elsevier Science Inc., New York, NY, USA, 1990, pp. 255–270.
[145] J. Runge, Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information, in: A. Storkey, F. Perez-Cruz (Eds.), International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 84, Pmlr, 2018, pp. 938–947.
[146] J.M. Peters, Restricted Structural Equation Models for Causal Inference (Ph.D. thesis), ETH Zurich and MPI for Intelligent Systems, 2012.
[147] P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, B. Schölkopf, Inferring deterministic causal relations, in: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Uai '10, AUAI Press, Arlington, Virginia, USA, 2010, pp. 143–150.
[148] J. Peters, P. Bühlmann, N. Meinshausen, Causal inference by using invariant prediction: identification and confidence intervals, J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (5) (2016) 947–1012.
[149] J.M. Mooij, S. Magliacane, T. Claassen, Joint causal inference from multiple contexts, J. Mach. Learn. Res. 21 (99) (2020) 1–108.
[150] P. Hoyer, D. Janzing, J.M. Mooij, J. Peters, B. Schölkopf, Nonlinear causal discovery with additive noise models, in: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (Eds.), Advances in Neural Information Processing Systems, Vol. 21, Curran Associates, Inc., 2008.
[151] J.M. Mooij, D. Janzing, B. Schölkopf, From ordinary differential equations to structural causal models: the deterministic case, 2013, arXiv preprint arXiv:1304.7920.

[152] S. Bongers, J.M. Mooij, From random differential equations to structural causal models: The stochastic case, 2018, arXiv preprint arXiv:1803.08784.
[153] P.K. Rubenstein, B. Bongers, S. Bernhard, J.M. Mooij, From deterministic ODEs to dynamic structural causal models, in: Thirty-Fourth Conference on Uncertainty in Artifical Intelligence, AUAI Press, Corvallis, Oregon, 2018.
[154] P. Forré, J.M. Mooij, Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders, in: A. Globerson, R. Silva (Eds.), Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI-18), AUAI Press, 2018.
[155] E.V. Strobl, A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias, Int. J. Data Sci. Anal. 8 (1) (2019) 33–56.
[156] J. M. Mooij, T. Claassen, Constraint-based causal discovery using partial ancestral graphs in the presence of cycles, in: J. Peters, D. Sontag (Eds.), Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), in: Proceedings of Machine Learning Research, vol. 124, Pmlr, 2020, pp. 1159–1168.
[157] C.W. Granger, Testing for causality: a personal viewpoint, J. Econom. Dynam. Control 2 (1980) 329–352.
[158] Y. Chen, G. Rangarajan, J. Feng, M. Ding, Analyzing multiple nonlinear time series with extended granger causality, Phys. Lett. A 324 (1) (2004) 26–35.
[159] A.B. Barrett, L. Barnett, A.K. Seth, Multivariate Granger causality and generalized variance, Phys. Rev. E 81 (4) (2010) 41907.
[160] D. Bell, J. Kay, J. Malley, A non-parametric approach to non-linear causality testing, Econom. Lett. 51 (1) (1996) 7–18.
[161] C. Hiemstra, J.D. Jones, Testing for linear and nonlinear Granger causality in the stock price-volume relation, J. Finance 49 (5) (1994) 1639–1664.
[162] A. Abhyankar, Linear and nonlinear granger causality: Evidence from the UK stock index futures market, J. Futures Mark. (1986-1998) 18 (5) (1998) 519.
[163] A. Warne, Causality and Regime Inference in a Markov Switching VAR, Technical Report, Sveriges Riksbank Working Paper Series, 2000.
[164] C. Diks, V. Panchenko, A new statistic and practical guidelines for nonparametric Granger causality testing, J. Econom. Dynam. Control 30 (9–10) (2006) 1647–1669.
[165] N. Ancona, D. Marinazzo, S. Stramaglia, Radial basis function approach to nonlinear Granger causality of time series, Phys. Rev. E 70 (5) (2004) 056221.
[166] D. Marinazzo, M. Pellicoro, S. Stramaglia, Kernel method for nonlinear Granger causality, Phys. Rev. Lett. 100 (2008) 144103.
[167] D. Marinazzo, M. Pellicoro, S. Stramaglia, Kernel-Granger causality and the analysis of dynamical networks, Phys. Rev. E 77 (5) (2008) 056215.
[168] F. Takens, Detecting strange attractors in turbulence, in: Dynamical Systems and Turbulence, Warwick 1980, in: Lecture Notes in Mathematics, vol. 898, Springer, Berlin, 1981, pp. 366–381.
[169] C. Meek, Causal inference and causal explanation with background knowledge, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Uai '95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 403–410.
[170] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, D. Sejdinovic, Detecting and quantifying causal associations in large nonlinear time series datasets, Sci. Adv. 5 (11) (2019) eaau4996.
[171] J. Runge, Quantifying information transfer and mediation along causal pathways in complex systems, Phys. Rev. E 92 (6) (2015) 062829.
[172] P. Spirtes, C. Meek, T. Richardson, Causal inference in the presence of latent variables and selection bias, in: P. Besnard, S. Hanks (Eds.), Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Uai '95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 499–506.
[173] J. Zhang, On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias, Artificial Intelligence 172 (16) (2008) 1873–1896.
[174] T. Richardson, P. Spirtes, Ancestral graph Markov models, Ann. Statist. 30 (4) (2002) 962–1030.
[175] J. Zhang, Causal reasoning with ancestral graphs, J. Mach. Learn. Res. 9 (47) (2008) 1437–1474.
[176] D. Entner, P.O. Hoyer, On causal discovery from time series data using FCI, in: P. Myllymäki, T. Roos, T. Jaakkola (Eds.), Proceedings of the 5th European Workshop on Probabilistic Graphical Models, Helsinki Institute for Information Technology HIIT, Helsinki, FI, 2010, pp. 121–128.
[177] D.M. Chickering, C. Meek, Selective greedy equivalence search: finding optimal Bayesian networks using a polynomial number of score evaluations, in: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, 2015, pp. 211–219.
[178] M. Chickering, Statistically efficient greedy equivalence search, in: J. Peters, D. Sontag (Eds.), Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), in: Proceedings of Machine Learning Research, vol. 124, Pmlr, 2020, pp. 241–249.
[179] P. Gradu, T. Zrnic, Y. Wang, M. Jordan, Valid inference after causal discovery, in: NeurIPS 2022 Workshop on Causality for Real-World Impact, 2022.
[180] T. Claassen, I.G. Bucur, Greedy equivalence search in the presence of latent confounders, in: J. Cussens, K. Zhang (Eds.), Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, in: Proceedings of Machine Learning Research, vol. 180, Pmlr, 2022, pp. 443–452.
[181] X. Zheng, B. Aragam, P.K. Ravikumar, E.P. Xing, DAGs with NO TEARS: Continuous optimization for structure learning, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018.
[182] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, E. Xing, Learning sparse nonparametric DAGs, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 108, Pmlr, 2020, pp. 3414–3425.
[183] I. Ng, S. Lachapelle, N. Rosemary Ke, S. Lacoste-Julien, K. Zhang, On the convergence of continuous constrained optimization for structure learning, in: G. Camps-Valls, F.J.R. Ruiz, I. Valera (Eds.), Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 151, Pmlr, 2022, pp. 8176–8198.
[184] S. Lachapelle, P. Brouillard, T. Deleu, S. Lacoste-Julien, Gradient-based neural DAG learning, in: International Conference on Learning Representations, 2020.
[185] I. Ng, A. Ghassami, K. Zhang, On the role of sparsity and DAG constraints for learning linear DAGs, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 17943–17954.
[186] Y. Yu, T. Gao, N. Yin, Q. Ji, DAGs with no curl: An efficient DAG structure learning approach, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 139, Pmlr, 2021, pp. 12156–12166.
[187] K. Bello, B. Aragam, P. Ravikumar, DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization, in: Advances in Neural Information Processing Systems, 2022.
[188] H. Aapo, K. Juha, E. Oja, Independent Component Analysis, 2001.
[189] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P.O. Hoyer, K. Bollen, DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model, J. Mach. Learn. Res. 12 (2011) 1225–1248.
[190] A. Hyvärinen, S. Shimizu, P.O. Hoyer, Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 424–431.
[191] A. Hyvärinen, K. Zhang, S. Shimizu, P.O. Hoyer, Estimation of a structural vector autoregression model using non-Gaussianity, J. Mach. Learn. Res. 11 (May) (2010) 1709–1731.

[192] J. Peters, J.M. Mooij, D. Janzing, B. Schölkopf, Identifiability of causal graphs using functional models, in: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Uai '11, AUAI Press, Arlington, Virginia, USA, 2011, pp. 589–598.
[193] J. Mooij, D. Janzing, J. Peters, B. Schölkopf, Regression by dependence minimization and its application to causal inference in additive noise models, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 745–752.
[194] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, A. Smola, A kernel statistical test of independence, Adv. Neural Inf. Process. Syst. 20 (2007).
[195] C. Heinze-Deml, M.H. Maathuis, N. Meinshausen, Causal structure learning, Annu. Rev. Stat. Appl. 5 (2018) 371–391.
[196] N. Pfister, P. Bühlmann, J. Peters, Invariant causal prediction for sequential data, J. Amer. Statist. Assoc. 114 (527) (2019) 1264–1276.
[197] N. Hansen, A. Sokol, Causal interpretation of stochastic differential equations, Electron. J. Probab. 19 (none) (2014) 1–24.
[198] A. Hyttinen, F. Eberhardt, P.O. Hoyer, Learning linear cyclic causal models with latent variables, J. Mach. Learn. Res. 13 (109) (2012) 3387–3439.
[199] J.M. Mooij, D. Janzing, B. Schölkopf, From ordinary differential equations to structural causal models: The deterministic case, in: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, Uai '13, AUAI Press, Arlington, Virginia, USA, 2013, pp. 440–448.
[200] S. Bongers, T. Blom, J.M. Mooij, Causal modeling of dynamical systems, 2018.
[201] J. Peters, S. Bauer, N. Pfister, Causal models for dynamical systems, in: Probabilistic and Causal Inference: The Works of Judea Pearl, 2022, pp. 671–690.
[202] G. Abbati, P. Wenk, M.A. Osborne, A. Krause, B. Schölkopf, S. Bauer, Ares and mars adversarial and mmd-minimizing regression for sdes, in: International Conference on Machine Learning, Pmlr, 2019, pp. 1–10.
[203] G. Varando, N.R. Hansen, Graphical continuous Lyapunov models, in: Conference on Uncertainty in Artificial Intelligence, Pmlr, 2020, pp. 989–998.
[204] K.J. Friston, L. Harrison, W. Penny, Dynamic causal modelling, Neuroimage 19 (4) (2003) 1273–1302.
[205] A.C. Marreiros, K.E. Stephan, K.J. Friston, Dynamic causal modeling, Scholarpedia 5 (7) (2010) 9568.
[206] K.E. Stephan, W.D. Penny, R.J. Moran, H.E. den Ouden, J. Daunizeau, K.J. Friston, Ten simple rules for dynamic causal modeling, Neuroimage 49 (4) (2010) 3099–3109.
[207] K. Friston, R. Moran, A.K. Seth, Analysing connectivity with Granger causality and dynamic causal modelling, Curr. Opin. Neurobiol. 23 (2) (2013) 172–178.
[208] K.J. Friston, T. Parr, P. Zeidman, A. Razi, G. Flandin, J. Daunizeau, O.J. Hulme, A.J. Billig, V. Litvak, R.J. Moran, et al., Dynamic causal modelling of COVID-19, Wellcome Open Res. 5 (2020).
[209] K.J. Friston, G. Flandin, A. Razi, Dynamic causal modelling of COVID-19 and its mitigations, Sci. Rep. 12 (1) (2022) 12419.
[210] S.W. Mogensen, N.R. Hansen, Markov equivalence of marginalized local independence graphs, Ann. Statist. 48 (1) (2020) 539–559.
[211] V. Didelez, Graphical models for marked point processes based on local independence, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1) (2008) 245–264.
[212] V. Didelez, Asymmetric separation for local independence graphs, in: 23rd Annual Conference on Uncertainty in Artifical Intelligence, 2006.
[213] S.W. Mogensen, D. Malinsky, N.R. Hansen, Causal learning for partially observed stochastic dynamical systems, in: Thirty-Fourth Conference on Uncertainty in Artifical Intelligence, AUAI Press, Corvallis, Oregon, 2018.
[214] S.W. Mogensen, Equality constraints in linear hawkes processes, in: B. Schölkopf, C. Uhler, K. Zhang (Eds.), Proceedings of the First Conference on Causal Learning and Reasoning, in: Proceedings of Machine Learning Research, vol. 177, Pmlr, 2022, pp. 576–593.
[215] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.
[216] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, B. Schölkopf, Causal discovery from heterogeneous/nonstationary data, J. Mach. Learn. Res. 21 (1) (2020) 3482–3534.
[217] B. Huang, K. Zhang, B. Schölkopf, Identification of time-dependent causal model: A Gaussian process treatment, in: Proceedings of the 24th International Conference on Artificial Intelligence, Ijcai '15, AAAI Press, 2015, pp. 3561–3568.
[218] E. Saggioro, J. de Wiljes, M. Kretschmer, J. Runge, Reconstructing regime-dependent causal relationships from observational time series, Chaos 30 (11) (2020) 113115.
[219] X.-A. Tibau, C. Reimers, A. Gerhardus, J. Denzler, V. Eyring, J. Runge, A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections, Environ. Data Sci. 1 (2022) e12.
[220] D. Bueso, M. Piles, G. Camps-Valls, Explicit Granger causality in kernel Hilbert spaces, Phys. Rev. E 102 (2020) 062201.
[221] R. Christiansen, M. Baumann, T. Kuemmerle, M.D. Mahecha, J. Peters, Toward causal inference for spatio-temporal data: Conflict and forest loss in Colombia, J. Amer. Statist. Assoc. 117 (538) (2022) 591–601.
[222] S.L. Bressler, A.K. Seth, Wiener–Granger causality: a well established methodology, Neuroimage 58 (2) (2011) 323–329.
[223] D. Chicharro, On the spectral formulation of Granger causality, Biol. Cybernet. 105 (5) (2011) 331–347.
[224] L. Faes, S. Erla, G. Nollo, Measuring connectivity in linear multivariate processes: definitions, interpretation, and practical analysis, Comput. Math. Methods Med. 2012 (2012).
[225] M. Lungarella, A. Pitti, Y. Kuniyoshi, Information transfer at multiple scales, Phys. Rev. E 76 (5) (2007) 056117.
[226] E.V. Strobl, S. Visweswaran, P.L. Spirtes, Fast causal inference with non-random missingness by test-wise deletion, Int. J. Data Sci. Anal. 6 (2018) 47–62.
[227] A. Gain, I. Shpitser, Structure learning under missing data, in: International Conference on Probabilistic Graphical Models, Pmlr, 2018, pp. 121–132.
[228] R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, K. Zhang, Causal discovery in the presence of missing data, in: The 22nd International Conference on Artificial Intelligence and Statistics, Pmlr, 2019, pp. 1762–1770.
[229] P. Versteeg, J. Mooij, C. Zhang, Local constraint-based causal discovery under selection bias, in: B. Schölkopf, C. Uhler, K. Zhang (Eds.), Proceedings of the First Conference on Causal Learning and Reasoning, in: Proceedings of Machine Learning Research, vol. 177, Pmlr, 2022, pp. 840–860.
[230] A. McDavid, R. Gottardo, N. Simon, M. Drton, Graphical models for zero-inflated single cell gene expression, Ann. Appl. Stat. 13 (2) (2019) 848.
[231] S. Yu, M. Drton, A. Shojaie, Directed graphical models and causal discovery for zero-inflated data, 2020, arXiv preprint arXiv:2004.04150.
[232] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, USA, 2008.
[233] L. Petersen, N.R. Hansen, Testing conditional independence via quantile regression based partial copulas, J. Mach. Learn. Res. 22 (70) (2021) 1–47.
[234] T. Bouezmarni, J.V. Rombouts, A. Taamouti, Nonparametric copula-based test for conditional independence with applications to Granger causality, J. Bus. Econom. Statist. 30 (2) (2012) 275–287.
[235] R.D. Shah, J. Peters, The hardness of conditional independence testing and the generalised covariance measure, Ann. Statist. 48 (3) (2020) 1514–1538.
[236] R. Berk, L. Brown, A. Buja, K. Zhang, L. Zhao, Valid post-selection inference, Ann. Statist. 41 (2) (2013) 802–837.
[237] A. Belloni, V. Chernozhukov, C. Hansen, Inference on treatment effects after selection among high-dimensional controls, Rev. Econom. Stud. 81 (2) (2014) 608–650.

[238] A. Rinaldo, L. Wasserman, M. G'Sell, Bootstrapping and sample splitting for high-dimensional, assumption-lean inference, Ann. Statist. 47 (6) (2019) 3438–3469.
[239] M. Robins, M. Hernan, Causal inference: what if, Found. Agnostic Stat. (2020) 235–281.
[240] M. Kretschmer, D. Coumou, J.F. Donges, J. Runge, Using causal effect networks to analyze different Arctic drivers of midlatitude winter circulation, J. Clim. 29 (11) (2016) 4069–4081.
[241] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, A. Drouin, Differentiable causal discovery from interventional data, Adv. Neural Inf. Process. Syst. 33 (2020) 21865–21877.
[242] R.A. Fisher, The Design of Experiments, Hafner Press, 1935.
[243] G.W. Imbens, D.B. Rubin, Causal Inference in Statistics, Social, and Biomedical Sciences, Cambridge University Press, 2015.
[244] J. Runge, Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems 34 (NeurIPS 2021), 2021.
[245] Y. Huang, M. Valtorta, Pearl's calculus of intervention is complete, in: R. Dechter, T. Richardson (Eds.), Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, Uai '06, AUAI Press, Arlington, Virginia, USA, 2006, pp. 217–224.
[246] I. Shpitser, J. Pearl, Identification of conditional interventional distributions, in: R. Dechter, T. Richardson (Eds.), Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, Uai '06, AUAI Press, Arlington, Virginia, USA, 2006, pp. 437–444.
[247] I. Shpitser, J. Pearl, Complete identification methods for the causal hierarchy, J. Mach. Learn. Res. 9 (2008) 1941–1979.
[248] T. VanderWeele, Explanation in Causal Inference: Methods for Mediation and Interaction, Oxford University Press, 2015.
[249] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. (CSUR) 41 (3) (2009) 1–58.
[250] J. Zscheischler, O. Martius, S. Westra, E. Bevacqua, C. Raymond, R.M. Horton, B. van den Hurk, A. AghaKouchak, A. Jézéquel, M.D. Mahecha, et al., A typology of compound weather and climate events, Nat. Rev. Earth Environ. 1 (7) (2020) 333–347.
[251] B. Andersen, T. Fagerhaug, Root Cause Analysis: Simplified Tools and Techniques, Quality Press, 2006.
[252] E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nat. Rev. Neurosci. 10 (3) (2009) 186–198.
[253] J. Donges, Y. Zou, N. Marwan, J. Kurths, The backbone of the climate network, Epl 87 (2009) 48007.
[254] J. Ludescher, M. Martin, N. Boers, A. Bunde, C. Ciemer, J. Fan, S. Havlin, M. Kretschmer, J. Kurths, J. Runge, et al., Network-based forecasting of climate phenomena, Proc. Natl. Acad. Sci. 118 (47) (2021) e1922872118.
[255] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. 424 (4–5) (2006) 175–308.
[256] A. Gozolchiani, S. Havlin, K. Yamasaki, Emergence of El Niño as an autonomous component in the climate network, Phys. Rev. Lett. 107 (14) (2011) 148501.
[257] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry (1977) 35–41.
[258] G. Brown, A. Pocock, M.-J. Zhao, M. Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.
[259] J. Runge, R.V. Donner, J. Kurths, Optimal model-free prediction from multivariate time series, Phys. Rev. E 91 (5) (2015) 052909.
[260] M. Kretschmer, J. Runge, D. Coumou, Early prediction of extreme stratospheric polar vortex states based on causal precursors, Geophys. Res. Lett. 44 (16) (2017) 8592–8600.
[261] G. Di Capua, M. Kretschmer, J. Runge, A. Alessandri, R. Donner, B. van Den Hurk, R. Vellore, R. Krishnan, D. Coumou, Long-lead statistical forecasts of the Indian summer monsoon rainfall based on causal precursors, Weather Forecast. 34 (5) (2019) 1377–1394.
[262] B. Huang, K. Zhang, M. Gong, C. Glymour, Causal discovery and forecasting in nonstationary environments with state-space models, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 97, Pmlr, 2019, pp. 2901–2910.
[263] V. Eyring, L. Bock, A. Lauer, M. Righi, M. Schlund, B. Andela, E. Arnone, O. Bellprat, B. Brötz, L.-P. Caron, et al., Earth System Model Evaluation Tool (ESMValTool) v2. 0–an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, Geosci. Model Dev. 13 (7) (2020) 3383–3438.
[264] V. Eyring, P.M. Cox, G.M. Flato, P.J. Gleckler, G. Abramowitz, P. Caldwell, W.D. Collins, B.K. Gier, A.D. Hall, F.M. Hoffman, et al., Taking climate model evaluation to the next level, Nature Clim. Change 9 (2) (2019) 102–110.
[265] P. Nowack, J. Runge, V. Eyring, J.D. Haigh, Causal networks for climate model evaluation and constrained projections, Nat. Commun. 11 (1) (2020) 1–11.
[266] J. Correa, S. Lee, E. Bareinboim, Nested counterfactual identification from arbitrary surrogate experiments, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Vol. 34, 2021, pp. 6856–6867.
[267] J.Y. Halpern, Actual Causality, MiT Press, 2016.
[268] A. Hannart, J. Pearl, F.E. Otto, P. Naveau, M. Ghil, Causal counterfactual theory for the attribution of weather and climate-related events, Bull. Am. Meteorol. Soc. 97 (1) (2016) 99–110.
[269] S.J. Press, A compound events model for security prices, J. Bus. (1967) 317–335.
[270] J. Zscheischler, S. Westra, B.J. Van Den Hurk, S.I. Seneviratne, P.J. Ward, A. Pitman, A. AghaKouchak, D.N. Bresch, M. Leonard, T. Wahl, et al., Future climate risk from compound events, Nature Clim. Change 8 (6) (2018) 469–477.
[271] L. Menzly, T. Santos, P. Veronesi, Understanding predictability, J. Polit. Econ. 112 (1) (2004) 1–47.
[272] E. Grunberg, F. Modigliani, The predictability of social events, J. Polit. Econ. 62 (6) (1954) 465–478.
[273] W. Bialek, I. Nemenman, N. Tishby, Predictability, complexity, and learning, Neural Comput. 13 (11) (2001) 2409–2463.
[274] G. Boffetta, M. Cencini, M. Falcioni, A. Vulpiani, Predictability: a way to characterize complexity, Phys. Rep. 356 (6) (2002) 367–474.
[275] J.M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schölkopf, Distinguishing cause from effect using observational data: methods and benchmarks, J. Mach. Learn. Res. 17 (1) (2016) 1103–1204.
[276] J. Runge, X.-A. Tibau, M. Bruhns, J. Muñoz-Marí, G. Camps-Valls, The causality for climate competition, in: NeurIPS 2019 Competition and Demonstration Track, Pmlr, 2020, pp. 110–120.
[277] A.G. Reisach, C. Seiler, S. Weichwald, Beware of the simulated DAG! causal discovery benchmarks may be easy to game, in: Advances in Neural Information Processing Systems 34 (NeurIPS), 2021.
[278] H. Ye, E.R. Deyle, L.J. Gilarranz, G. Sugihara, Distinguishing time-delayed causal interactions using convergent cross mapping, Sci. Rep. 5 (1) (2015) 1–9.
[279] P. Spirtes, C. Glymour, An algorithm for fast recovery of sparse causal graphs, Soc. Sci. Comput. Rev. 9 (1) (1991) 62–72.
[280] C. Heinze-Deml, J. Peters, N. Meinshausen, Invariant causal prediction for nonlinear models, J. Causal Inference 6 (2) (2018).
[281] B.E. Dowd, Separated at birth: statisticians, social scientists, and causality in health services research, Health Serv. Res. 46 (2) (2011) 397–420.
[282] J. Pearl, Statistics and causality: Separated to reunite——Commentary on bryan dowd's "separated at birth", 2011.
[283] J. Kaddour, A. Lynch, Q. Liu, M.J. Kusner, R. Silva, Causal machine learning: A survey and open problems, 2022, arXiv preprint arXiv:2206.15475.
[284] R. Castelo, A. Siebes, Priors on network structures. Biasing the search for Bayesian networks, Internat. J. Approx. Reason. 24 (1) (2000) 39–57.

[285] R.O. Ness, K. Sachs, P. Mallick, O. Vitek, A Bayesian active learning experimental design for inferring signaling networks, in: Research in Computational Molecular Biology: 21st Annual International Conference, RECOMB 2017, Hong Kong, China, May 3-7, 2017, Proceedings 21, Springer, 2017, pp. 134–156.
[286] D. Janzing, On causally asymmetric versions of Occam's Razor and their relation to thermodynamics, 2007, arXiv preprint arXiv:0708.3411.
[287] G.F. Cooper, C. Yoo, Causal discovery from a mixture of experimental and observational data, 2013, arXiv preprint arXiv:1301.6686.
[288] B. Schölkopf, F. Locatello, S. Bauer, N.R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Toward causal representation learning, Proc. IEEE 109 (5) (2021) 612–634.
[289] N.L. Cramer, A representation for the adaptive generation of simple sequential programs, in: J.J. Grefenstette (Ed.), Proceedings of an International Conference on Genetic Algorithms and the Applications, Carnegie-Mellon University, Pittsburgh, PA, USA, 1985, pp. 183–187.
[290] J.R. Koza, Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems, Technical Report, Stanford University, Stanford, CA, USA, 1990.
[291] J.R. Koza, Genetic programming as a means for programming computers by natural selection, Stat. Comput. 4 (2) (1994) 87–112.
[292] R. Praksova, Eureqa: Software review, Genet. Program. Evol. Mach. 12 (1) (2011) 173–178.
[293] DataRobot Inc, Eureqa as part of DataRobot's service, 2023, https://www.datarobot.com/nutonian/.
[294] M. Cranmer, PySR: Fast & parallelized symbolic regression in Python/Julia, 2020.
[295] T. Stephens, gplearn: Genetic Programming in Python with a scikit-learn inspired API, 2022, https://gplearn.readthedocs.io/en/stable.
[296] M. Quade, J. Gout, M. Abel, Glyph: Symbolic regression tools, J. Open Res. Softw. (2019).
[297] B. Burlacu, G. Kronberger, M. Kommenda, Operon C++: An efficient genetic programming framework for symbolic regression, in: Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion, Gecco '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1562–1570.
[298] M. Kommenda, B. Burlacu, G. Kronberger, M. Affenzeller, Parameter identification for symbolic regression using nonlinear least squares, Genet. Program. Evol. Mach. 21 (3) (2020) 471–501.
[299] S.-M. Udrescu, A. Tan, J. Feng, O. Neto, T. Wu, M. Tegmark, AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 4860–4871.
[300] B.K. Petersen, M.L. Larma, T.N. Mundhenk, C.P. Santiago, S.K. Kim, J.T. Kim, Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients, in: International Conference on Learning Representations, 2021.
[301] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1996) 267–288.
[302] T. McConaghy, FFX: Fast, scalable, deterministic symbolic regression technology, in: Genetic Programming Theory and Practice IX, in: Genetic and Evolutionary Computation, Springer, Ann Arbor, USA, 2011, pp. 235–260.
[303] J.E. Adsuara, A. Pérez-Suay, Á. Moreno-Martínez, G. Camps-Valls, G. Kraemer, M. Reichstein, M. Mahecha, Discovering differential equations from earth observation data, in: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 2020, pp. 3999–4002.
[304] G. Camps-Valls, J. Verrelst, J. Muñoz-Marí, V. Laparra, F. Mateo-Jimenez, J. Gomez-Dans, A survey on Gaussian processes for earth observation data analysis: A comprehensive investigation, IEEE Geosci. Remote Sens. Mag. (6) (2016).
[305] J.E. Johnson, V. Laparra, G. Camps-Valls, Disentangling derivatives, uncertainty and error in Gaussian process models, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 4051–4054.
[306] K. Kaheman, S.L. Brunton, J.N. Kutz, Automatic differentiation to simultaneously identify nonlinear dynamics and extract noise probability distributions from data, Mach. Learn.: Sci. Technol. 3 (1) (2022) 015031, Publisher: IOP Publishing.
[307] G. Martius, C.H. Lampert, Extrapolation and learning equations, 2016, https://arxiv.org/abs/1610.02995.
[308] C. Louizos, M. Welling, D.P. Kingma, Learning sparse neural networks through L_0 regularization, in: International Conference on Learning Representations, 2018.
[309] S. Kim, P.Y. Lu, S. Mukherjee, M. Gilbert, L. Jing, V. Čeperić, M. Soljačić, Integration of neural network-based symbolic regression in deep learning for scientific discovery, IEEE Trans. Neural Netw. Learn. Syst. 32 (9) (2021) 4166–4177.
[310] M. Werner, A. Junginger, P. Hennig, G. Martius, Informed equation learning, 2021.
[311] M. Werner, A. Junginger, P. Hennig, G. Martius, Uncertainty in equation learning, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO), Association for Computing Machinery, 2022, pp. 2298–2305.
[312] K. Ellis, C. Wong, M. Nye, M. Sablé-Meyer, L. Morales, L. Hewitt, L. Cary, A. Solar-Lezama, J.B. Tenenbaum, DreamCoder: bootstrapping inductive program synthesis with wake-sleep library learning, in: Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, in: Pldi 2021, Association for Computing Machinery, New York, NY, USA, 2021, pp. 835–850.
[313] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.
[314] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.
[315] Mauna Loa Observatory, 2020, https://www.climate.gov/teaching/resources/atmospheric-co2-mauna-loa-observatory.
[316] G. Kronberger, F.O. de Franca, B. Burlacu, C. Haider, M. Kommenda, Shape-constrained symbolic regression—Improving extrapolation with prior knowledge, Evol. Comput. 30 (1) (2022) 75–98.
[317] C. Cornelio, S. Dash, V. Austel, T.R. Josephson, J. Goncalves, K.L. Clarkson, N. Megiddo, B.E. Khadir, L. Horesh, Combining data and theory for derivable scientific discovery with AI-Descartes, Nature Commun. 14 (2023) 1777.
[318] Z. Long, Y. Lu, B. Dong, PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network, J. Comput. Phys. 399 (2019) 108925.
[319] S.-C. Lin, G. Martius, M. Oettel, Analytical classical density functionals from an equation learning network, J. Chem. Phys. 152 (2) (2020) 021102.
[320] J. Arenas-García, K. Petersen, G. Camps-Valls, L. Hansen, Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods, IEEE Signal Process. Mag. 30 (4) (2013) 16–29.
[321] M. Khodkar, P. Hassanzadeh, A. Antoulas, A koopman-based framework for forecasting the spatiotemporal evolution of chaotic dynamics with nonlinearities modeled as exogenous forcings, 2019, Preprint arXiv:1909.00076.
[322] K. Taira, S.L. Brunton, S. Dawson, C.W. Rowley, T. Colonius, B.J. McKeon, O.T. Schmidt, S. Gordeyev, V. Theofilis, L.S. Ukeiley, Modal analysis of fluid flows: An overview, Aiaa J. 55 (12) (2017) 4013–4041.
[323] C.W. Rowley, S.T. Dawson, Model reduction for flow analysis and control, Annu. Rev. Fluid Mech. 49 (2017) 387–417.
[324] J.L. Lumley, The structure of inhomogeneous turbulence, in: A.M. Yaglom, V.I. Tatarski (Eds.), Atmospheric Turbulence and Wave Propagation, Nauka, Moscow, 1967, pp. 166–178.

[325] D. Bueso, M. Piles, G. Camps-Valls, Nonlinear PCA for spatio-temporal analysis of earth observation data, IEEE Trans. Geosci. Remote Sens. 58 (8) (2020) 5752–5763.
[326] S. Brunton, B. Brunton, J. Proctor, E. Kaiser, J. Kutz, Chaos as an intermittently forced linear system, Nature Commun. 8 (2017) 19.
[327] P.J. Schmid, Dynamic mode decomposition of numerical and experimental data, J. Fluid Mech. 656 (2010) 5–28.
[328] S.L. Brunton, B.W. Brunton, J.L. Proctor, J.N. Kutz, Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control, PLoS ONE 11 (2016).
[329] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, F. Noé, Data-driven model reduction and transfer operator approximation, J. Nonlinear Sci. 1010 (2018) 9437.
[330] E. Kaiser, J.N. Kutz, S.L. Brunton, Data-driven discovery of Koopman eigenfunctions for control, Mach. Learn.: Sci. Technol. 2 (3) (2021) 035023.
[331] N. Takeishi, Y. Kawahara, T. Yairi, Learning Koopman invariant subspaces for dynamic mode decomposition, Adv. Neural Inf. Process. Syst. 30 (2017).
[332] S.M. Ulam, A Collection of Mathematical Problems, Interscience Publisher, NY, 1960.
[333] M.O. Williams, I.G. Kevrekidis, C.W. Rowley, A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition, J. Nonlinear Sci. 25 (2015) 1307–1346.
[334] M.O. Williams, C.W. Rowley, I.G. Kevrekidis, A kernel-based method for data-driven Koopman spectral analysis, J. Comput. Dyn. 2 (2015) 247–265.
[335] S. Klus, P. Koltai, C. Schütte, On the numerical approximation of the Perron–Frobenius and Koopman operator, J. Comput. Dyn. 3 (2016) 51–79.
[336] F. Noé, F. Nüske, A variational approach to modeling slow processes in stochastic dynamical systems, Multiscale Model. Simul. 11 (2013) 635–655.
[337] F. Nüske, B.G. Keller, G. Pérez-Hernández, A.S.J.S. Mey, F. Noé, Variational approach to molecular kinetics, J. Chem. Theory Comput. 10 (2014) 1739–1752.
[338] C.W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, D.S. Henningson, Spectral analysis of nonlinear flows, J. Fluid Mech. 641 (2009) 115–127.
[339] J.H. Tu, C.W. Rowley, D.M. Luchtenburg, S.L. Brunton, J.N. Kutz, On dynamic mode decomposition: Theory and applications, J. Comput. Dyn. 1 (2014).
[340] R.T. McGibbon, V.S. Pande, Variational cross-validation of slow dynamical modes in molecular kinetics, J. Chem. Phys. 142 (2015).
[341] C.R. Schwantes, V.S. Pande, Modeling molecular kinetics with tICA and the kernel trick, J. Chem. Theory Comput. 11 (2015) 600–608.
[342] S. Le Clainche, J.M. Vega, Higher order dynamic mode decomposition, SIAM J. Appl. Dyn. Syst. 16 (2017) 882–925.
[343] F. Takens, Detecting strange attractors in turbulence, in: D.A. Rand, L.-S. Young (Eds.), Dynamical Systems and Turbulence, in: Lecture Notes in Mathematics, vol. 898, 1981, pp. 366–381.
[344] E. Lazpita, Á. Martínez-Sánchez, A. Corrochano, S. Hoyas, S. Le Clainche, R. Vinuesa, On the generation and destruction mechanisms of arch vortices in urban fluid flows, Phys. Fluids 34 (2022) 051702.
[345] N. Groun, M. Villalba-Orero, E. Lara-Pezzi, E. Valero, J. Garicano-Mena, S. Le Clainche, Higher order dynamic mode decomposition: from fluid dynamics to heart disease analysis, 2022, Preprint arXiv:2201.03030.
[346] W.J. Baars, C. Tinney, Proper orthogonal decomposition-based spectral higher-order stochastic estimation, Phys. Fluids 26 (2014) 055112.
[347] G.E. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
[348] T. Murata, K. Fukami, K. Fukagata, Nonlinear mode decomposition with convolutional neural networks for fluid dynamics, J. Fluid Mech. 882 (2020) A13.
[349] H. Eivazi, S. Le Clainche, S. Hoyas, R. Vinuesa, Towards extraction of orthogonal and parsimonious non-linear modes from turbulent flows, Expert Syst. Appl. 202 (2022) 117038.
[350] M. Milano, P. Koumoutsakos, Neural network modeling for near wall turbulent flow, J. Comput. Phys. 182 (2002) 1–26.
[351] N.B. Erichson, L. Mathelin, Z. Yao, S.L. Brunton, M.W. Mahoney, J.N. Kutz, Shallow neural networks for fluid flow reconstruction with limited sensors, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 476 (2020) 20200097.
[352] K. Lee, K.T. Carlberg, Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders, J. Comput. Phys. 404 (2020) 108973.
[353] M. Cenedese, J. Axås, B. Bäuerlein, K. Avila, G. Haller, Data-driven modeling and prediction of nonlinearizable dynamics via spectral submanifolds, Nature Commun. 13 (2022) 872.
[354] P. Benner, S. Gugercin, K. Willcox, A survey of projection-based model reduction methods for parametric dynamical systems, SIAM Rev. 57 (4) (2015) 483–531.
[355] K. Carlberg, M. Barone, H. Antil, Galerkin v. least-squares Petrov–Galerkin projection in nonlinear model reduction, J. Comput. Phys. 330 (2017) 693–734.
[356] K. Fukami, T. Nakamura, K. Fukagata, Convolutional neural network based hierarchical autoencoder for nonlinear mode decomposition of fluid field data, Phys. Fluids 32 (2020) 095110.
[357] M. Rolinek, D. Zietlow, G. Martius, Variational autoencoders pursue PCA directions (by accident), in: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12406–12415.
[358] F. Kemeth, T. Bertalan, T. Thiem, F. Dietrich, S. Moon, C. Laing, Y. Kevrekidis, Learning emergent partial differential equations in a learned emergent space, Nature Commun. 13 (2022).
[359] T.N. Thiem, M. Kooshkbaghi, T. Bertalan, C.R. Laing, I.G. Kevrekidis, Emergent spaces for coupled oscillators, Front. Comput. Neurosci. 14 (2020) 36.
[360] O. Fajardo-Fontiveros, I. Reichardt, H.R. De Los Ríos, J. Duch, M. Sales-Pardo, R. Guimerà, Fundamental limits to learning closed-form mathematical models from data, Nature Commun. 14 (1) (2023) 1043.
[361] G. Antonelli, S. Chiaverini, P. Di Lillo, On data-driven identification: Is automatically discovering equations of motion from data a chimera? Nonlinear Dynam. (2022) 1–12.
[362] R. Fuentes, R. Nayek, P. Gardner, N. Dervilis, T. Rogers, K. Worden, E. Cross, Equation discovery for nonlinear dynamical systems: A Bayesian viewpoint, Mech. Syst. Signal Process. 154 (2021) 107528.
[363] R. Guimerà, I. Reichardt, A. Aguilar-Mogas, F.A. Massucci, M. Miranda, J. Pallarès, M. Sales-Pardo, A Bayesian machine scientist to aid in the solution of challenging scientific problems, Sci. Adv. 6 (5) (2020) eaav6971.
[364] B.C. Daniels, I. Nemenman, Automated adaptive inference of phenomenological dynamical models, Nat. Commun. 6 (2015) 8133.
[365] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 5–30, Special Issue: Diffusion Maps and Wavelets.
[366] S.S. Suseela, Y. Feng, K. Mao, A comparative study on machine learning algorithms for knowledge discovery, in: 2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV), Ieee, 2022, pp. 131–136.
[367] Z. Wu, S.L. Brunton, S. Revzen, Challenges in dynamic mode decomposition, J. R. Soc. Interface 18 (185) (2021) 20210686.
[368] P.J. Baddoo, B. Herrmann, B.J. McKeon, J. Nathan Kutz, S.L. Brunton, Physics-informed dynamic mode decomposition, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 479 (2271) (2023) 20220576.
[369] W. Gerstner, W.M. Kistler, R. Naud, L. Paninski, Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition, Cambridge University Press, 2014.

[370] C. Koch, Biophysics of Computation: Information Processing in Single Neurons, in: Computational Neuroscience Series, Oxford University Press, Inc., Usa, 2004.
[371] M.I. Rabinovich, P. Varona, Discrete sequential information coding: Heteroclinic cognitive dynamics, Front. Comput. Neurosci. 12 (2018).
[372] A. Byrne, J. Ross, R. Nicks, S. Coombes, Mean-field models for EEG/MEG: from oscillations to waves, Brain Topogr. 35 (2021).
[373] N. Brunel, Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons, J. Comput. Neurosci. 8 (3) (2000) 183–208.
[374] S. Amari, Dynamics of pattern formation in lateral-inhibition type neural fields, Biol. Cybernet. 27 (2) (1977) 77–87.
[375] A. Byrne, R.D. O'Dea, M. Forrester, J. Ross, S. Coombes, Next-generation neural mass and field modeling, J. Neurophysiol. 123 (2) (2020) 726–742, Pmid: 31774370.
[376] A. Tabas, M. Andermann, V. Schuberth, H. Riedel, E. Balaguer-Ballester, A. Rupp, Modeling and MEG evidence of early consonance processing in auditory cortex, PLoS Comput. Biol. 15 (2) (2019) 1–28.
[377] H.R. Wilson, J.D. Cowan, Evolution of the Wilson-Cowan equations, Biol. Cybernet. 115 (6) (2021) 643–653.
[378] H.R. Wilson, J.D. Cowan, Excitatory and inhibitory interactions in localized populations of model neurons, Biophys. J. 12 (1972) 1–24.
[379] R. Potthast, Amari model, in: D. Jaeger, R. Jung (Eds.), Encyclopedia of Computational Neuroscience, Springer New York, New York, NY, 2013, pp. 1–6.
[380] M. Mattia, P. Del Giudice, Population dynamics of interacting spiking neurons, Phys. Rev. E 66 (2002) 051917.
[381] M.I. Rabinovich, R. Huerta, P. Varona, V.S. Afraimovich, Transient cognitive dynamics, metastability, and decision making, PLoS Comput. Biol. 4 (5) (2008) 1–9.
[382] E. Montbrió, D. Pazó, A. Roxin, Macroscopic description for networks of spiking neurons, Phys. Rev. X 5 (2015) 021028.
[383] M. Mattia, M. Biggio, A. Galluzzi, M. Storace, Dimensional reduction in networks of non-Markovian spiking neurons: Equivalence of synaptic filtering and heterogeneous propagation delays, PLoS Comput. Biol. 15 (10) (2019) 1–35.
[384] M. Schirner, L. Domide, D. Perdikis, P. Triebkorn, L. Stefanovski, R. Pai, P. Prodan, B. Valean, J. Palmer, C. Langford, A. Blickensdörfer, M. van der Vlag, S. Diaz-Pier, A. Peyser, W. Klijn, D. Pleiter, A. Nahm, O. Schmid, M. Woodman, L. Zehl, J. Fousek, S. Petkoski, L. Kusch, M. Hashemi, D. Marinazzo, J.-F. Mangin, A. Flöel, S. Akintoye, B.C. Stahl, M. Cepic, E. Johnson, G. Deco, A.R. McIntosh, C.C. Hilgetag, M. Morgan, B. Schuller, A. Upton, C. McMurtrie, T. Dickscheid, J.G. Bjaalie, K. Amunts, J. Mersmann, V. Jirsa, P. Ritter, Brain simulation as a cloud service: The virtual brain on EBRAINS, NeuroImage 251 (2022) 118973.
[385] L. Duncker, M. Sahani, Dynamics on the manifold: Identifying computational dynamical activity from neural population recordings, Curr. Opin. Neurobiol. 70 (2021) 163–170, Computational Neuroscience.
[386] A.R. Galgali, M. Sahani, V. Mante, Residual dynamics resolves recurrent contributions to neural computation, Nature Neurosci. 26 (2) (2023) 326–338.
[387] E. Balaguer-Ballester, C.C. Lapish, J.K. Seamans, D. Durstewitz, Attracting dynamics of frontal cortex ensembles during memory-guided decision-making, PLoS Comput. Biol. 7 (5) (2011) e1002057.
[388] J.P. Cunningham, B.M. Yu, Dimensionality reduction for large-scale neural recordings, Nature Neurosci. 17 (11) (2014) 1500–1509.
[389] J.M. Hyman, L. Ma, E. Balaguer-Ballester, D. Durstewitz, J.K. Seamans, Contextual encoding by ensembles of medial prefrontal cortex neurons, Proc. Natl. Acad. Sci. USA 109 (13) (2012) 5086–5091.
[390] B.M. Yu, J.P. Cunningham, G. Santhanam, S.I. Ryu, K.V. Shenoy, M. Sahani, Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity, J. Neurophysiol. 102 (1) (2009) 614–635, Pmid: 19357332.
[391] M.C. Aoi, V. Mante, J.W. Pillow, Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making, Nature Neurosci. 23 (11) (2020) 1410–1420.
[392] E. Gokcen, A.I. Jasper, J.D. Semedo, A. Zandvakili, A. Kohn, C.K. Machens, B.M. Yu, Disentangling the flow of signals between populations of neurons, Nat. Comput. Sci. 2 (8) (2022) 512–525.
[393] V. Rutten, A. Bernacchia, M. Sahani, G. Hennequin, Non-reversible Gaussian processes for identifying latent dynamical structure in neural data, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 9622–9632.
[394] L. Duncker, M. Sahani, Temporal alignment and latent Gaussian process factor inference in population spike trains, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018.
[395] E. Balaguer-Ballester, A. Tabas-Diaz, M. Budka, Can we identify non-stationary dynamics of trial-to-trial variability? PLoS ONE 9 (4) (2014) 1–13.
[396] C.C. Lapish, E. Balaguer-Ballester, J.K. Seamans, A.G. Phillips, D. Durstewitz, Amphetamine exerts dose-dependent changes in prefrontal cortex attractor dynamics during working memory, J. Neurosci. 35 (28) (2015) 10172–10187. EB–B and CCL contributed equally..
[397] E. Balaguer-Ballester, R. Nogueira, J.M. Abofalia, R. Moreno-Bote, M.V. Sanchez-Vives, Representation of foreseeable choice outcomes in orbitofrontal cortex triplet-wise interactions, PLoS Comput. Biol. 16 (6) (2020) 1–30.
[398] D. Durstewitz, J.K. Seamans, The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia, Biol. Psychiat. 64 (9) (2008) 739–749, Neurodevelopment and the Transition from Schizophrenia Prodrome to Schizophrenia.
[399] D. Durstewitz, A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements, PLoS Comput. Biol. 13 (6) (2017) 1–33.
[400] M. Brenner, F. Hess, J.M. Mikhaeil, L.F. Bereska, Z. Monfared, P.-C. Kuo, D. Durstewitz, Tractable dendritic RNNs for reconstructing nonlinear dynamical systems, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 162, Pmlr, 2022, pp. 2292–2320.
[401] G. Koppe, H. Toutounji, P. Kirsch, S. Lis, D. Durstewitz, Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI, PLoS Comput. Biol. 15 (8) (2019) 1–35.
[402] D.G. Barrett, A.S. Morcos, J.H. Macke, Analyzing biological and artificial neural networks: challenges with opportunities for synergy? Curr. Opin. Neurobiol. 55 (2019) 55–64, Machine Learning, Big Data, and Neuroscience.
[403] J.-M. Lueckmann, P.J. Gonçalves, G. Bassetto, K. Öcal, M. Nonnenmacher, J.H. Macke, Flexible statistical inference for mechanistic models of neural dynamics, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 1289–1299.
[404] J. Boelts, J.-M. Lueckmann, R. Gao, J.H. Macke, Flexible and efficient simulation-based inference for models of decision-making, eLife 11 (2022) e77220.
[405] P.J. Gonçalves, J.-M. Lueckmann, M. Deistler, M. Nonnenmacher, K. Öcal, G. Bassetto, C. Chintaluri, W.F. Podlaski, S.A. Haddad, T.P. Vogels, D.S. Greenberg, J.H. Macke, Training deep neural density estimators to identify mechanistic models of neural dynamics, eLife 9 (2020) e56261.
[406] M. Deistler, J.H. Macke, P.J. Gonçalves, Energy-efficient network activity from disparate circuit parameters, Proc. Natl. Acad. Sci. 119 (44) (2022) e2207632119.
[407] M. Genkin, T.A. Engel, Moving beyond generalization to accurate interpretation of flexible models, Nat. Mach. Intell. 2 (11) (2020) 674–683.

[408] M. Genkin, O. Hughes, T.A. Engel, Learning non-stationary Langevin dynamics from stochastic observations of latent trajectories, Nature Commun. 12 (1) (2021) 5986.
[409] Y. Guan, S.L. Brunton, I. Novosselov, Sparse nonlinear models of chaotic electroconvection, R. Soc. Open Sci. 8 (8) (2021) 202367.
[410] J.-C. Loiseau, Data-driven modeling of the chaotic thermal convection in an annular thermosyphon, Theor. Comput. Fluid Dyn. 34 (4) (2020) 339–365.
[411] A. Razi, K.J. Friston, The connected brain: causality, models, and intrinsic dynamics, IEEE Signal Process. Mag. 33 (3) (2016) 14–35.
[412] L.N. Ross, Dynamical models and explanation in neuroscience, Philos. Sci. 82 (1) (2015) 32–54.
[413] E. Tognoli, J.S. Kelso, The metastable brain, Neuron 81 (1) (2014) 35–48.
[414] E. Balaguer-Ballester, R. Moreno-Bote, G. Deco, D. Durstewitz, Editorial: Metastable dynamics of neural ensembles, Front. Syst. Neurosci. 11 (2018).
[415] S. Weichwald, J. Peters, Causality in cognitive neuroscience: concepts, challenges, and distributional robustness, J. Cogn. Neurosci. 33 (2) (2021) 226–247.
[416] D.L. Barack, E.K. Miller, C.I. Moore, A.M. Packer, L. Pessoa, L.N. Ross, N.C. Rust, A call for more clarity around causality in neuroscience, Trends Neurosci. (2022).
[417] L. Barnett, A.B. Barrett, A.K. Seth, Misunderstandings regarding the application of Granger causality in neuroscience, Proc. Natl. Acad. Sci. 115 (29) (2018) E6676–e6677.
[418] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, F. Bießmann, On the interpretation of weight vectors of linear models in multivariate neuroimaging, NeuroImage 87 (2014) 96–110.
[419] A. Woolgar, P. Golland, S. Bode, Coping with confounds in multivoxel pattern analysis: What should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013, NeuroImage 98 (2014) 506–512.
[420] M.T. Todd, L.E. Nystrom, J.D. Cohen, Confounds in multivariate pattern analysis: Theory and rule representation case study, NeuroImage 77 (2013) 157–165.
[421] G. Lohmann, K. Erfurth, K. Müller, R. Turner, Critical comments on dynamic causal modelling, NeuroImage 59 (3) (2012) 2322–2329.
[422] T. Davis, K.F. LaRocque, J.A. Mumford, K.A. Norman, A.D. Wagner, R.A. Poldrack, What do differences between multi-voxel and univariate analysis mean? how subject-, voxel-, and trial-level variance impact fMRI analysis, NeuroImage 97 (2014) 271–283.
[423] M. Ding, Y. Chen, S.L. Bressler, Granger causality: basic theory and application to neuroscience, in: Handbook of Time Series Analysis: Recent Theoretical Developments and Applications, Wiley Online Library, 2006, pp. 437–460.
[424] D.S. Bassett, O. Sporns, Network neuroscience, Nature Neurosci. 20 (3) (2017) 353–364.
[425] A. Sheikhattar, S. Miran, J. Liu, J.B. Fritz, S.A. Shamma, P.O. Kanold, B. Babadi, Extracting neuronal functional network dynamics via adaptive Granger causality analysis, Proc. Natl. Acad. Sci. 115 (17) (2018) E3869–e3878.
[426] S. Weichwald, T. Meyer, O. Özdenizci, B. Schölkopf, T. Ball, M. Grosse-Wentrup, Causal interpretation rules for encoding and decoding models in neuroimaging, NeuroImage 110 (2015) 48–59.
[427] W.D. Penny, K.E. Stephan, A. Mechelli, K.J. Friston, Comparing dynamic causal models, Neuroimage 22 (3) (2004) 1157–1172.
[428] G.K. Cooray, B. Sengupta, P. Douglas, K. Friston, Dynamic causal modelling of electrographic seizure activity using Bayesian belief updating, Neuroimage (2015).
[429] K. Friston, K.H. Preller, C. Mathys, H. Cagnan, J. Heinzle, A. Razi, P. Zeidman, Dynamic causal modelling revisited, NeuroImage 199 (2019) 730–744.
[430] J. Cabral, F. Castaldo, J. Vohryzek, V. Litvak, C. Bick, R. Lambiotte, K. Friston, M.L. Kringelbach, G. Deco, Metastable oscillatory modes emerge from synchronization in the brain spacetime connectome, Commun. Phys. 5 (1) (2022) 184.
[431] K. Friston, A theory of cortical responses, Philos. Trans. R. Soc. B 360 (1456) (2005) 815–836.
[432] W.D. Penny, K.E. Stephan, J. Daunizeau, M.J. Rosa, K.J. Friston, T.M. Schofield, A.P. Leff, Comparing families of dynamic causal models, PLoS Comput. Biol. 6 (3) (2010) e1000709.
[433] D. Chicharro, S. Panzeri, Algorithms of causal inference for the analysis of effective connectivity among brain regions, Front. Neuroinform. 8 (2014).
[434] D. Bernal-Casas, E. Balaguer-Ballester, M.F. Gerchen, S. Iglesias, H. Walter, A. Heinz, A. Meyer-Lindenberg, K.E. Stephan, P. Kirsch, Multi-site reproducibility of prefrontal–hippocampal connectivity estimates by stochastic DCM, Neuroimage 82 (2013) 555–563.
[435] A.S. Meyer-Lindenberg, R.K. Olsen, P.D. Kohn, T. Brown, M.F. Egan, D.R. Weinberger, K.F. Berman, Regionally specific disturbance of dorsolateral prefrontal–hippocampal functional connectivity in schizophrenia, Arch. Gen. Psychiatry 62 (4) (2005) 379–386.
[436] X. Chen, F. Ginoux, M. Carbo-Tano, T. Mora, A.M. Walczak, C. Wyart, Granger causality analysis for calcium transients in neuronal networks, challenges and improvements, eLife 12 (2023) e81279.
[437] J. Cruzat, R. Herzog, P. Prado, Y. Sanz-Perl, R. Gonzalez-Gomez, S. Moguilner, M.L. Kringelbach, G. Deco, E. Tagliazucchi, A. Ibañez, Temporal irreversibility of large-scale brain dynamics in Alzheimer's disease, J. Neurosci. 43 (9) (2023) 1643–1656.
[438] D.M.A. Mehler, K.P. Kording, The lure of causal statements: Rampant mis-inference of causality in estimated connectivity, 2018.
[439] W.-J. Neumann, A. Horn, A.A. Kühn, Insights and opportunities for deep brain stimulation as a brain circuit intervention, Trends Neurosci. (2023).
[440] R.D. Koster, Y. Sud, Z. Guo, P.A. Dirmeyer, G. Bonan, K.W. Oleson, E. Chan, D. Verseghy, P. Cox, H. Davies, et al., GLACE: the global land–atmosphere coupling experiment. Part I: overview, J. Hydrometeorol. 7 (4) (2006) 590–610.
[441] J.K. Green, S.I. Seneviratne, A.M. Berg, K.L. Findell, S. Hagemann, D.M. Lawrence, P. Gentine, Large influence of soil moisture on long-term terrestrial carbon uptake, Nature 565 (7740) (2019) 476–479.
[442] P. Milly, Potential evaporation and soil moisture in general circulation models, J. Clim. 5 (3) (1992) 209–226.
[443] M. Jung, M. Reichstein, P. Ciais, S.I. Seneviratne, J. Sheffield, M.L. Goulden, G. Bonan, A. Cescatti, J. Chen, R. De Jeu, et al., Recent decline in the global land evapotranspiration trend due to limited moisture supply, Nature 467 (7318) (2010) 951–954.
[444] D. Baldocchi, Measuring fluxes of trace gases and energy between ecosystems and the atmosphere – the state and future of the eddy covariance method, Global Change Biol. 20 (12) (2014) 3600–3609.
[445] C. Krich, M. Migliavacca, D.G. Miralles, G. Kraemer, T.S. El-Madany, M. Reichstein, J. Runge, M.D. Mahecha, Functional convergence of biosphere–atmosphere interactions in response to meteorological conditions, Biogeosciences 18 (7) (2021) 2379–2404.
[446] C. Krich, J. Runge, D.G. Miralles, M. Migliavacca, O. Perez-Priego, T. El-Madany, A. Carrara, M.D. Mahecha, Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach, Biogeosciences 17 (4) (2020) 1033–1061.
[447] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).
[448] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: Ordering points to identify the clustering structure, ACM Sigmod Rec. 28 (2) (1999) 49–60.
[449] D. Mønster, R. Fusaroli, K. Tylén, A. Roepstorff, J.F. Sherson, Causal inference from noisy time-series data—Testing the Convergent Cross-Mapping algorithm in the presence of noise and external influence, Future Gener. Comput. Syst. 73 (2017) 52–62.
[450] C.B. Field, R.B. Jackson, H.A. Mooney, Stomatal responses to increased CO2: implications from the plant to the global scale, Plant Cell Environ. 18 (10) (1995) 1214–1225.

[451] R.D. Koster, P.A. Dirmeyer, Z. Guo, G. Bonan, E. Chan, P. Cox, C. Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, et al., Regions of strong coupling between soil moisture and precipitation, Science 305 (5687) (2004) 1138–1140.
[452] N. Madani, N.C. Parazoo, J.S. Kimball, A.P. Ballantyne, R.H. Reichle, M. Maneta, S. Saatchi, P.I. Palmer, Z. Liu, T. Tagesson, Recent amplified global gross primary productivity due to temperature increase is offset by reduced productivity due to water constraints, AGU Adv. 1 (4) (2020) e2020AV000180.
[453] V. Eyring, S. Bony, G.A. Meehl, C.A. Senior, B. Stevens, R.J. Stouffer, K.E. Taylor, Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev. 9 (5) (2016) 1937–1958.
[454] T.F. Stocker, D. Qin, G. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. Midgley, Climate Change 2013: The Physical Science Basis. Intergovernmental Panel on Climate Change, Working Group I Contribution to the IPCC Fifth Assessment Report (AR5), New York, 2013.
[455] D. Maraun, T.G. Shepherd, M. Widmann, G. Zappa, D. Walton, J.M. Gutiérrez, S. Hagemann, I. Richter, P.M. Soares, A. Hall, et al., Towards process-informed bias correction of climate change simulations, Nature Clim. Change 7 (11) (2017) 764–773.
[456] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, et al., The NCEP/NCAR 40-year reanalysis project, Bull. Am. Meteorol. Soc. 77 (3) (1996) 437–472.
[457] R. Evans, The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids, Adv. Phys. 28 (2) (1979) 143–200.
[458] S.-C. Lin, M. Oettel, A classical density functional from machine learning and a convolutional neural network, SciPost Phys. 6 (2019) 025.
[459] Z. Wang, I. Akhtar, J. Borggaard, T. Iliescu, Proper orthogonal decomposition closure models for turbulent flows: a numerical comparison, Comput. Methods Appl. Mech. Engrg. 237 (2012) 10–26.
[460] B.R. Noack, K. Afanasiev, M. Morzynski, G. Tadmor, F. Thiele, A hierarchy of low-dimensional models for the transient and post-transient cylinder wake, J. Fluid Mech. 497 (2003) 335–363.
[461] J. Kutz, S. Brunton, B. Brunton, J. Proctor, Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems, Siam, 2016.
[462] Z. Bai, E. Kaiser, J. Proctor, J. Kutz, S. Brunton, Dynamic mode decomposition for compressive system identification, Aiaa J. 58 (2020) 561.
[463] J. Juang, R. Pappa, An eigensystem realization algorithm for modal parameter identification and model reduction, J. Guid. Control Dyn. 8 (1985) 620.
[464] C. Rowley, I. Mezic, S. Bagheri, P. Schlatter, D. Henningson, Spectral analysis of nonlinear flows, J. Fluid Mech. 641 (2009) 115–127.
[465] H. Eivazi, L. Guastoni, P. Schlatter, H. Azizpour, R. Vinuesa, Recurrent neural networks and Koopman-based frameworks for temporal predictions in a low-order model of turbulence, Int. J. Heat Fluid Flow 90 (2021) 108816.
[466] J. Moehlis, H. Faisst, B. Eckhardt, A low-dimensional model for turbulent shear flows, New J. Phys. 6 (2004) 56.
[467] J. Crutchfield, B. McNamara, Equations of motion from a data series, Complex Systems 1 (1987) 417–452.
[468] G. Antonelli, S. Chiaverini, P. Di Lillo, On data-driven identification: Is automatically discovering equations of motion from data a Chimera? Nonlinear Dyn. (2022).
[469] P. Gelß, S. Klus, J. Eisert, C. Schütte, Multidimensional approximation of nonlinear dynamical systems, J. Comput. Nonlinear Dyn. 14 (2019) 061006.
[470] D. Shea, S. Brunton, J. Kutz, SINDy-BVP: Sparse identification of nonlinear dynamics for boundary value problems, Phys. Rev. Res. 3 (2021) 023255.
[471] N. Deng, B.R. Noack, M. Morzynski, L.R. Pastur, Low-order model for successive bifurcations of the fluidic pinball, J. Fluid Mech. 884 (2020) A37.
[472] J.L. Callaham, G. Rigas, J.-C. Loiseau, S.L. Brunton, An empirical mean-field model of symmetry-breaking in a turbulent wake, Sci. Adv. 8 (2021) eabm4786.
[473] J.L. Callaham, S.L. Brunton, J.-C. Loiseau, On the role of nonlinear correlations in reduced-order modeling, J. Fluid Mech. 938 (2022) A1.
[474] J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, 1992.
[475] J. Weatheritt, R.D. Sandberg, A novel evolutionary algorithm applied to algebraic modifications of the RANS stress-strain relationship, J. Comput. Phys. 325 (2016) 22–37.
[476] O. Reynolds, On the dynamical theory of incompressible viscous fluids and the determination of the criterion, Phil. Trans. R. Soc. A 186 (1895) 123–164.
[477] S.B. Pope, Turbulent Flows, Cambridge University Press, 2000.
[478] H. Tennekes, J.L. Lumley, A First Course in Turbulence, MIT press, 1972.
[479] J. Weatheritt, R.D. Sandberg, The development of algebraic stress models using a novel evolutionary algorithm, Int. J. Heat Fluid Flow 68 (2017) 298–318.
[480] R. Vinuesa, P. Schlatter, H.M. Nagib, Secondary flow in turbulent ducts with increasing aspect ratio, Phys. Rev. Fluids 3 (2018) 054606.
[481] J.V. Boussinesq, Théorie Analytique de la Chaleur: Mise en Harmonie Avec la Thermodynamique et Avec la Théorie Mécanique de la Lumière T. 2, Refroidissement et Échauffement par Rayonnement Conductibilité des Tiges, Lames et Masses Cristallines Courants de Convection Théorie Mécanique de la Lumière, Gauthier-Villars, 1923.
[482] P.R. Spalart, Strategies for turbulence modelling and simulations, Int. J. Heat Fluid Flow 21 (2000) 252–263.
[483] J.L. Callaham, J.V. Koch, B.W. Brunton, J.N. Kutz, S.L. Brunton, Learning dominant physical processes with data-driven balance models, Nature Commun. 12 (1) (2021) 1–10.
[484] C. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[485] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (1) (2012) 265–286.
[486] J. Lee, T.A. Zaki, Detection algorithm for turbulent interfaces and large-scale structures in intermittent flows, Comput. & Fluids 175 (1) (2018) 142–158.
[487] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural. Inf. Process. Syst. 30 (2017).
[488] M.Z. Yousif, M. Zhang, L. Yu, R. Vinuesa, H. Lim, A transformer-based synthetic-inflow generator for spatially-developing turbulent boundary layers, J. Fluid Mech. 957 (2023) A6.
[489] Á. Martínez-Sánchez, E. Lazpita, A. Corrochano, S. Le Clainche, S. Hoyas, R. Vinuesa, Data-driven assessment of arch vortices in urban flows, Int. J. Heat Fluid Flow 100 (2023) 109101.
[490] B. Monnier, B. Neiswander, C. Wark, Stereoscopic particle image velocimetry measurements in an urban type boundary layer: insight into flow regimes and incidence angle effect, Bound.-Layer Meteorol. 135 (2010) 243—268.
[491] C. Amor, P. Schlatter, R. Vinuesa, S. Le Clainche, Higher-order dynamic mode decomposition on-the-fly: A low-order algorithm for complex fluid flows, J. Comput. Phys. 475 (2023) 111849.
[492] A. Lozano-Durán, H.J. Bae, M.P. Encinar, Causality of energy-containing eddies in wall turbulence, J. Fluid Mech. 882 (2020) A2.
[493] W.M. Orr, The stability or instability of the steady motions of a perfect liquid and of a viscous liquid. Part II. A viscous liquid, Math. Proc. R. Irish Acad. 27 (1907) 69–138.
[494] M.T. Landahl, M.T. Landahlt, Wave breakdown and turbulence, SIAM J. Appl. Math. 28 (1975) 735–756.

[495] J.D. Swearingen, R.F. Blackwelder, The growth and breakdown of streamwise vortices in the presence of a wall, J. Fluid Mech. 182 (1987) 255–290.
[496] F. Waleffe, Hydrodynamic stability and turbulence: beyond transients to a self-sustaining process, Stud. Appl. Math. 95 (1995) 319–343.
[497] Á. Martínez-Sánchez, E. López, S. Le Clainche, A. Lozano-Durán, A. Srivastava, R. Vinuesa, Causality analysis of large-scale structures in the flow around a wall-mounted square cylinder, 2022, Preprint arXiv:2209.15356.
[498] J.C.R. Hunt, A.A. Wray, P. Moin, Eddies, streams, and convergence zones in turbulent flows, in: Center for Turbulence Research (CTR) Proceedings of Summer Program, 1998.
[499] S.J. Kline, W.C. Reynolds, F.A. Schraub, P.W. Runstadler, The structure of turbulent boundary layers, J. Fluid Mech. 30 (1967) 741–773.
[500] H.T. Kim, S.J. Kline, W.C. Reynolds, The production of turbulence near a smooth wall in a turbulent boundary layer, J. Fluid Mech. 50 (1971) 133–160.
[501] J.M. Wallace, H. Eckelman, R.S. Brodkey, The wall region in turbulent shear flow, J. Fluid Mech. 54 (1972) 39–48.
[502] S.S. Lu, W.W. Willmarth, Measurements of the structure of the Reynolds stress in a turbulent boundary layer, J. Fluid Mech. 60 (1973) 481–511.
[503] J.C. del Álamo, J. Jiménez, P. Zandonade, R.D. Moser, Self-similar vortex clusters in the turbulent logarithmic region, J. Fluid Mech. 561 (2006) 329–358.
[504] A. Lozano-Durán, O. Flores, J. Jiménez, The three-dimensional structure of momentum transfer in turbulent channels, J. Fluid Mech. 694 (2012) 100–130.
[505] A. Lozano-Durán, J. Jiménez, Time-resolved evolution of coherent structures in turbulent channels: characterization of eddies and cascades, J. Fluid Mech. 759 (2014) 432–471.
[506] D. Schmekel, F. Alcántara-Ávila, S. Hoyas, R. Vinuesa, Predicting coherent turbulent structures via deep learning, Front. Phys. 10 (2022) 888832.
[507] J.I. Cardesa, A. Vela-Martín, J. Jiménez, The turbulent cascade in five dimensions, Science 357 (2017) 782–784.
[508] J. Jiménez, Cascades in wall-bounded turbulence, Annu. Rev. Fluid Mech. 44 (2012) 27–45.
[509] J. Jiménez, Coherent structures in wall-bounded turbulence, J. Fluid Mech. 842 (2018) P1.
[510] J. Jiménez, Optimal fluxes and Reynolds stresses, J. Fluid Mech. 809 (2016) 585–600.
[511] J. Jiménez, Machine-aided turbulence theory, J. Fluid Mech. 854 (2018) R1.
[512] A. Cremades, S. Hoyas, P. Quintero, M. Lellep, M. Linkmann, R. Vinuesa, Explaining wall-bounded turbulence through deep learning, 2023, Preprint arXiv:2302.01250.
[513] E. Winter, The Shapley value, in: Handbook of Game Theory with Economic Applications, Vol. 3, 2002, pp. 2025–2054.
[514] S. Lun-Chau, R. Hu, J. Gonzalez, D. Sejdinovic, RKHS-SHAP: Shapley values for kernel methods, 2022, Preprint arXiv:2110.09167v2.
[515] L. Guastoni, A. Güemes, A. Ianiro, S. Discetti, P. Schlatter, H. Azizpour, R. Vinuesa, Convolutional-network models to predict wall-bounded turbulence from wall quantities, J. Fluid Mech. 928 (2021) A27.
[516] A. Güemes, S. Discetti, A. Ianiro, B. Sirmacek, H. Azizpour, R. Vinuesa, From coarse wall measurements to turbulent velocity fields through deep learning, Phys. Fluids 33 (2021) 075121.
[517] R. Vinuesa, B. Sirmacek, Interpretable deep-learning models to help achieve the Sustainable Development Goals, Nat. Mach. Intell. 3 (2021) 926.
[518] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, S. Ho, Discovering symbolic models from deep learning with inductive biases, in: Advances in Neural Information Processing Systems (NeurIPS 2020), 2020.
[519] L. Zanna, T. Bolton, Data-driven equation discovery of ocean mesoscale closures, Geophys. Res. Lett. 47 (17) (2020) e2020GL088376.
[520] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity, in: Monographs on Statistics and Applied Probability, vol. 143, 2015, p. 143.
[521] A.S. Ross, Z. Li, P. Perezhogin, C. Fernandez-Granda, L. Zanna, Benchmarking of machine learning ocean subgrid parameterizations in an idealized model, J. Adv. Modelling Earth Syst. 15 (2023) e2022MS003258.
[522] C. Meneveau, J. Katz, Scale-invariance and turbulence models for large-eddy simulation, Annu. Rev. Fluid Mech. 32 (1) (2000) 1–32.
[523] J.A. Anstey, L. Zanna, A deformation-based parametrization of ocean mesoscale eddy Reynolds stresses, Ocean Model. 112 (2017) 99–111.
[524] M.F. Jansen, I.M. Held, Parameterizing subgrid-scale eddy effects using energetically consistent backscatter, Ocean Model. 80 (2014) 36–48.
[525] D.P. Marshall, A.J. Adcroft, Parameterization of ocean eddies: Potential vorticity mixing, energetics and Arnold's first stability theorem, Ocean Model. 32 (3–4) (2010) 188–204.