systems

MDPI

Opinion

System Science Can Relax the Tension Between Data and Theory

Alessandro Giuliani

Environment and Health Department, Istituto Superiore di Sanità, 00161 Rome, Italy; alessandro.giuliani@iss.it

Abstract: The actual hype around machine learning (ML) methods has pushed the old epistemic struggle between data-driven and theory-driven scientific styles well beyond the academic realm. The potential consequences of the widespread adoption of ML in scientific work have fueled a harsh debate between opponents predicting the decay of basic curiosity-driven science and enthusiasts hoping for the advent of a 'theory-free' objective science. In this work, I suggest how the system science style of reasoning could drastically de-potentiate this (sometimes deceptive) opposition through the generation of multi-purpose relational theoretical frames stemming from the network paradigm. The recognition of the virtual non-existence of purely 'theoryfree' approaches and the need for a careful balancing of theoretical and empirical contributions is the main claim of the present work.

Keywords: machine learning; data science; complexity; relational systems theory; Hopfield networks; deep learning

1. Introduction

As aptly stressed by Sui Huang in [1], the last decade witnessed a deep epistemic shift from 'theory-driven' to 'data-driven' science. This shift began in the field of biomedical sciences with the rapid onset of 'omics' sciences. The term 'omics' designates different high-throughput techniques allowing the measurement of thousands of different variables in a single sample (e.g., different gene expression levels and metabolite or protein concentrations) [2]. This fact, together with unprecedented and cheap computational power, turned upside down the time-honored basis of the classical statistical approach, in which statistical units (samples) are supposed to outnumber their descriptors (variables).

In the classical case, the choice of variables to analyze is strictly hypothesis-driven: the information content of an experiment stems from a scientific question stated in terms of the relations between a few empirical observables.

Until some years ago, students were discouraged from pursuing projects seeking broad data analyses without the support of a strong hypothesis or question. The 'broad approaches' were considered 'fishing expeditions', producing a plethora of chance correlations [1]. This caveat stems from the mathematical definition of statistical significance. Actually, obtaining a p-value of < 0.05 as a result of an empirical study corresponds to an estimated 5% probability (under certain distributional assumptions) that the results are due to pure chance. If we take into account a dozen variables and we can build a credible narrative, even by a single significant result, the probability of chance correlation is unbearably high. This problem is common in any mathematical modeling of experimental data [3] and pushes scientists to focus on the (few) descriptors predicted to have the maximal information content for the problem at hand.

The overlooking of this caveat provoked a recognized knowledge crisis in biomedical sciences [4,5]. It is totally out of scope to apply a classical inferential approach (despite any smart statistical correction procedure) to face a data set made up of fifty thousand variables attached to ten to thirty samples, as occurs in high-throughput gene expression (transcriptomic) experiments. Notwithstanding this, students are no longer discouraged but encouraged to undertake largely hypothesis-free CConfirmmics projects that are by far the most popular (and fund-attracting) types of research in biomedicine [1] that now

Citation: Giuliani, A. System Science Can Relax the Tension Between Data and Theory. *Systems* **2024**, *12*, 474. https://doi.org/10.3390/systems12110474

Academic Editor: Maurice Yolles

Received: 17 September 2024
Revised: 29 October 2024
Accepted: 5 November 2024
Published: 6 November 2024

(thanks to a different statistical approach with respect to the classical inferential paradigm) no longer have the stigma of 'fishing expeditions'.

There are two main possible ways to get rid of the curse of high dimensionality. The first one originates from machine learning (ML) and implies a drastic change in the kind of 'scientific question': delving deeper into causative mechanisms and/or testing a general theory, with the focus shifting toward 'directly applicable' goals like clinical diagnosis [6] or the elucidation of the structures of biopolymers [7]. In these cases, high dimensionality is no longer a source of chance correlations, provided a scientist can rely upon both a training set from which to generate a prediction model and one (or more) independent test set to check its accuracy.

The other approach to very-high-dimensional data is related to statistical physics and can turn the curse of dimensionality into a blessing [8] for basic research. This approach implements a change in the scale of the scientific questions. Instead of pursuing the impossible dream of a deterministic model allowing one to climb up the different layers of an organization from the bottom (e.g., molecules) to the top (e.g., an entire organism), the focus shifts toward the mesoscopic scale, maximizing the correlations linking different organizational levels [9]. In the example of transcriptomes mentioned above, this corresponds to forgetting the identity and idiosyncratic roles of single genes and focusing on the trajectories of the entire genome, considering it as an integrated dynamical system [10]. Along this pathway, many different data analysis techniques come into play, ranging from time-honored multidimensional statistical techniques, like principal component analysis, to complex network and non-linear dynamics-inspired approaches [11,12], with all these methodologies redounding around the same concept of correlation.

In the following, I will describe the main philosophical–methodological premises of ML and statistical physics-inspired approaches and how the system science style can help reconstruct the useful synthesis between data and theory that has inspired centuries of modern science and is now showing some signs of increasing tension.

The main goal of this work is to describe the nature of this (apparently inescapable) polarization and how a systemic style of conducting science can reconcile data-driven and theory-driven attitudes. This reconciliation can take place by the recognition of the coding/decoding dynamics by which the correlation structure between variables, naturally emerging from data (the formal system) with no need for strong a priori hypotheses, can be 'decoded' in terms of the actual interactions structuring the investigated phenomenon (the real system).

Section 2 gives a proof-of-concept of the epistemological impossibility of a 'theory-free' scientific investigation by means of a coarse-grain description of computational approaches endowed with different levels of transparency (explainability). Some basic tenets of the scientific method are briefly described to demonstrate how any data structure encompasses more or less explicit theoretical choices.

Section 3 deals with the coding/decoding process inspired by relational systems theory and the link between this theory and network thermodynamics. The clarification of this link passes through a brief description of Hopfield networks, physics-inspired computational tools in which the stored memories (the formal system) correspond to dynamical attractors (the real system).

Section 4 describes a case study in which the network paradigm (the main ingredient of relational systems theory), in the form of the configuration of pairwise contacts between amino acid residues of a protein molecule, allows us to explain, in the language of biochemistry, the recognition of amino acid residues responsible for allosteric effects. This chapter is, thus, a practical example of the coding/decoding process.

Section 5 (the conclusion) summarizes the motivations of the 'data' and 'theory' reconciliation proposal set forth by this opinion paper. The need to re-state the time-honored principle of the complementary roles played by data and theory originates from the somewhat exaggerated hype surrounding so-called 'artificial intelligence'.

2. Machine Learning and the Dream of a 'Theory-Free' Science

A recent monograph by the UK Royal Society [13] details the avenues of the impact of ML on all fronts of scientific research, from epidemiology to materials science. The basic claim of the 108-page monograph is that the adoption of ML in the scientific workflow will deeply transform knowledge generation in the way it automatically extracts and learns features from raw data [14].

Before critically analyzing the hype around 'ML disruptiveness' in science, it is worth going in-depth into the nature of ML, taking as a paradigm the 'deep learning' algorithms that are at the forefront of machine learning research.

In a broad sense, any fitting procedure that is able to reproduce, with sufficient accuracy, a given property of interest (a dependent variable in statistical jargon, usually denoted by the letter Y) by means of the knowledge of a set of descriptors (independent variables, usually denoted by the letter X) can be considered an ML approach.

ML methods can be ordered based on their degree of 'explainability' [15] from 'black-box' (e.g., multi-layer neural networks) to transparent (e.g., linear discriminant analysis) approaches. The concept of 'explainability' is crucial to understanding the nature of the tension between data and theory; thus, in order to correctly set the problem, let us restrict Y to the case of class prediction, with Y being a binary variable (with only two possible outcomes). As the first step, the machine learner must be trained with a data set (the training set) in which the class labels of the samples are known (the golden standard). The goal is to identify (by suitable metrics of error minimization) rules and feature sets that allow the differentiation of the class labels. For example, if the goal is to predict biological gender, then the class labels (Y variable) could be 'Male' and 'Female', and the feature set (X variables) may comprise both quantitative (e.g., height) and qualitative variables (e.g., 'presence of beard'). If the samples and variables of the training set are both sufficiently representative and informative, then the classifier's predictions in new samples not present in the original training set (test set) will be (almost always) correct [15]. In a 'content-agnostic' situation in which it is out of scope to go in-depth into the nature of between-class differences, prediction accuracy is the only relevant metric for judging an ML tool. While limiting prediction accuracy is perfectly legitimate in many practical situations, e.g., think of the detection of a weapon in the baggage of a traveler, it is a largely defective strategy in scientific work.

In the case of science, limiting the focus to prediction accuracy, with no possibility of going in-depth into the constellation of mutual relations between X variables allowing for such accuracy, is only a very preliminary step in the 'solution'. In order to obtain a real 'piece of novel knowledge', we must go beyond accuracy metrics [15]. We need to not only predict but also explain a given phenomenon: we need a theory. 'Theory' is, here, intended in a soft and broad sense as a representation of the studied problem incorporating domain knowledge, even at a largely metaphorical level [14]. I will go more in-depth into this definition of 'theory' in the following chapters.

Theoretical hints are of no use for further explaining the results originating from black-box deep learning ML methods based on multi-layer neural networks [16].

'Deep learning' usually indicates network models composed of different processing layers able to solve recognition tasks with multiple levels of abstraction. These methods discover intricate structures by navigating very large data sets by the backpropagation algorithm, indicating how the system should change its internal parameters (the 'synaptic weights' across the neurons of the network located in different layers) to optimize the recognition task [16]. In mathematical terms, these systems correspond to 'universal approximators' [17] that, in any case, are able to achieve perfect accuracy in their training set. On the contrary, test set prediction is not for granted and stems from factors that are independent of computational power and deeply 'theory-laden' [14]. These (often-overlooked) factors falsify the 'disruptive' claim of the possibility of a 'perfectly objective' theory-free science fostered by the widespread application of the ML approach and damning the hypothesis-driven scientific method to obsolescence [18].

The basic point is that 'theory-free' perfectly objective data do not exist; every step of the process to generate any data set involves theory-informed choices, the most obvious being the following:

1. The choice of the descriptors (variables) of the statistical units composing the data set.
2. The choice of the inclusion criteria for a given statistical unit to make part of the training set.
3. The choice of an operative rule to define a test set as sufficiently homogeneous with the training set and the consequent definition of the boundary conditions of ML predictions.
4. The choice of the best 'structure-preserving' dimension-reducing approach for eliminating redundancies possibly biasing the ML procedure.
5. The choice of an error-free gold standard (Y variable) that remains stable across different data sets, acting as a 'class label' that univocally defines the statistical units.

The above-mentioned issues (and other more subtle but equally important ones) derive from the very basic notions of the scientific method, forming part of ordinary statistics and scientific methodology academic courses; thus, it seems unreasonable that many philosophers and scientists are unaware of them. My personal opinion is that the increasing fragmentation of scientific culture, together with the impressive progress of computational power, has provoked the emergence of 'data scientists' with a prevailing informatics background and only a surface knowledge of both statistics and experimental sciences. Data scientists tend to focus their interest on procedural issues (e.g., the efficiency and correctness of algorithms) while not paying much attention to the use of the algorithms to solve actual problems that are left to the so-called 'experts in the field'. This is a completely natural way of operating. The problem is that the evident success of technology has generated the pernicious idea that there is no difference between data and reality.

It is worth remarking that the use of low-transparency ML methods does not imply any 'theory-free' epistemology that only pertains to specific scholars (that do not, by chance, in the majority of cases, have a philosophical, more than scientific, background). The emphasis on this point comes from the errors arising from skipping some methodological points due to a too-simplistic focus on computational power and the almost perfect accuracy of the obtained results.

3. The Statistical Physics of Data from a System Science Perspective

In his visionary paper [19], Donald Mikulecky introduces a neat discrimination between 'constitutive' and 'network' principles, stating the following (emphasis added):

The network thermodynamic model of a system has two complementary, but distinct, contributions. Their explicit formal independence and strict complementarity are one of the most striking aspects of the formalism. These two intertwined facets are the constitutive laws for the network elements and the network topology. The use of constitutive laws for the network elements is the way the physical character of each network element is represented abstractly. It is a common feature of the material world. The topology or connected pattern of these elements in a network is an independent reality about the system.

The formalism of network thermodynamics stems from non-equilibrium thermodynamics and goes hand in hand with applications in chemistry and biology [19]. The adoption of this approach implies a shift in the basis of the unity of nature from the statement 'any entity is made of the same basic bricks' to the statement 'any entity can be thought as a set of interacting parts' [20]. This shift allows setting, in a correct way, the interaction between formal and natural systems [19].

A natural system (NS) is encoded into a formal system (FS) for the purpose of mimicking a causal event in the natural world by an implication in the formal system. Then, decoding is performed to see if the system commutes. A commuting modeling relation is a model of the real world.

In other words, a formal model generates proper scientific knowledge if it is able to give accurate predictions of the modeled system behavior while, at the same time, allowing one to translate (decode) the wiring of the correlation network of the formal representation into actual (real world) interactions within the system at hand. This 'encoding'/'decoding' commutation process is the goal of system science [21], which traditionally privileges the 'network' over constitutive laws. Figure 1 (modified by [19]) gives a pictorial example of this style of reasoning.

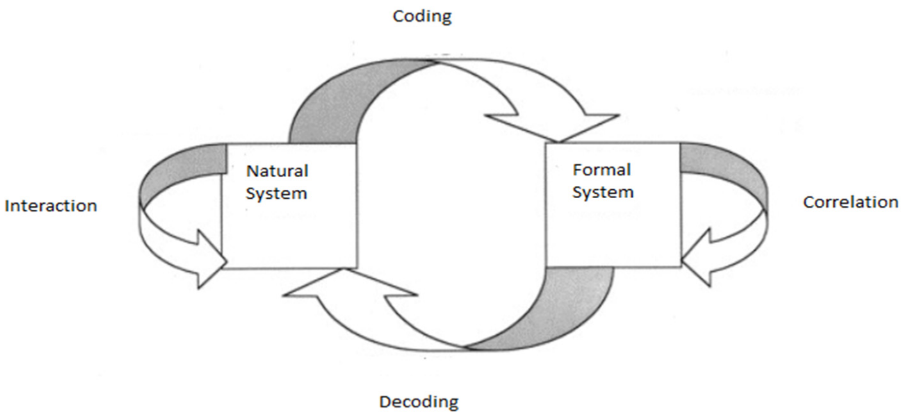


Figure 1. The analysis of the formal system (i.e., the data set with the different statistical units described as many component feature vectors) generates a set of empirical correlations between features. If the 'coding phase' is realistic (i.e., the adopted features catch the relevant properties of the natural system), then the correlation structure emerging from the formal system can be 'decoded' in terms consistent with the natural systems whose interactions between parts can be traced back to the correlations observed in the formal system.

The above strategy stems from Robert Rosen's concept of 'relational biology', which constitutes the theoretical background linking ML and modeling strategies [22] mediated by systems science. Rosen's basic claim can be summarized as 'when studying a complex system, one can forget the matter (constitutive laws of network elements) and focus on the organization (topological) laws'. In the case of biology, Rosen derived this claim by the consideration of organisms as metabolism–repair (M,R) systems. The material counterparts of metabolism and repair are catalysts (enzymes) and RNA molecules, respectively. Such MR systems generate closed causative loops: RNA molecules code for enzymes while, at the same time, needing enzymes to exert their actions. This circular causality rules out any strict mechanistic process in the form of quasi-deterministic pathways. Pathways (when considered in isolation and not embedded into closed networks) derive from Newtonian dynamics, in which there is a hierarchy of cause–effect relations that implies a sort of 'regressio ad infinitum' toward an initial cause placed in the most basic organization layer. On the contrary, pathways are only partial views of a network system that, in turn, at odds with Newtonian dynamics, implies a circular causality and asks for a different epistemology [19,22].

Besides biological considerations, what is important for the generation of suitable models of real systems from a mainly data-driven hypothesis-generating perspective is the possibility of decoding a formal correlation structure in terms of actual interactions between system elements. In the case of complex systems in which (like in thermodynamics) complete knowledge of the whole set of microscopic agents is impossible, we must abandon the differential equations style [23]. In a proper relational approach, the interaction network emerges from the empirical correlations present in the data set [24]. The separation between constitutive and relational laws makes the same relational principles able to explain the behavior of widely different systems, like a protein and a social organization. The cross-disciplinary portability of such relational models defines the relational approach as 'theory-free', but this freedom from theory must be intended only in the very limited sense of independence from constitutive, field-specific, microscopic-level theories, as actually happens in thermodynamics.

Thermodynamics is usually a significant part of every physics textbook, but at odds with other physical theories, thermodynamics focuses on system properties that are independent of mechanisms [19,22]. It focuses on the actual state of a system and the state changes attained by different (largely unknown) mechanisms. Rather than focusing on the details of the dynamics of the system's parts, the relations between the parts are the center of attention. Relational thinking is an extension of thermodynamic reasoning; it says very little about mechanisms, and the emphasis is on a system's function. This implies a given system is described in terms of its functional components independently from its material parts.

This attitude allows discovering the same organizational principles in systems as different as a social network and a protein molecule.

The relational style of reasoning imposes a careful choice of ML approaches privileging so-called PINNs (physics-informed neural networks) [25] that incorporate explicit physical principles in both their structure and behavior. Hopfield networks are the most classic example of PINNs, allowing for a straightforward commutation between emerging correlations between network elements and the real world [26].

A Hopfield network is made up of a set of elements (neurons) whose mutual interaction is modulated by synaptic weights, mathematically equivalent to (not necessarily linear) correlation coefficients. The actual state of the network consists of the values of the components of an N-dimensional vector of features. The evolution in time of a state vector follows an energy function whose local minima are K 'memory' vectors, acting as attractors of the dynamics and corresponding to the patterns that the network stores in its weights. A Hopfield network is, thus, associative memory that can act as ML when presented with an initial prompt that resembles one of the memory vectors. The energy descent gradient finds the most similar pattern among the set of stored patterns, thus performing a recognition task [27] corresponding to the reaching of an equilibrium state in the network.

This strategy allows the generation of a physically motivated explanation of very complex phenomena, like the multiscale organization of biological tissues [27,28].

It is worth noting how, in this case, we enter into the realm of the 'simulation' approach, in which the systems science approach is able to generate theories focused on the network topology principles independent from the (often unknown) constitutive laws of the single elementary players.

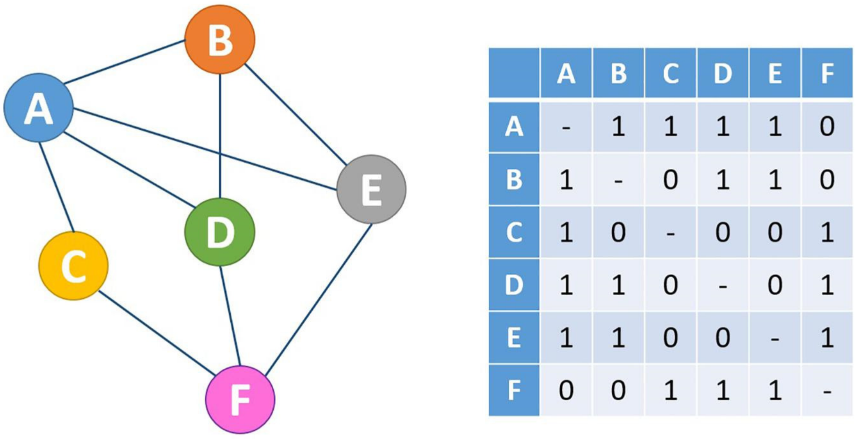
4. The Coding/Decoding Circle: A Case Study

Relational thinking has its most fundamental mathematical counterpart in the concept of a network. A ´network´ is not necessarily a ´real´ thing; rather, it must be considered a ´cognitive schema´ [29], which is an abstract collection of concepts used to make sense of the unknown world of life. In other words, a network is a formal system arising from the correlation structure of the studied system that can eventually be decoded in terms of the original natural system interaction pattern. A network (graph) structure is fully described by its adjacency matrix, isomorphic to the usual node-and-edge representation (Figure 2).

Biological systems are complex entities that both adapt to their environment and interact with other systems while, at the same time, being naturally amenable to network formalization. Therefore, the peculiarities of information transfer across networks are crucial to understanding the basic principles of biological organization.

Protein molecules are the most microscopic objects displaying adaptation and interaction properties, being a perfect playground for the study of complex systems [31].

Proteins are biopolymers made of linear series of N (with N ranging from 30 to more than 5000) monomers (amino acid residues) held together by covalent chemical bonds between subsequent monomers. Proteins are composed of 20 different amino acid species, and their linear ordering (primary structure) corresponds to a string in which the 20 different elements juxtapose with no evident periodicity. Protein molecules are the molecular agents responsible for the unique properties of life, such as metabolism, signaling, and the immune response, thanks to their ability to adapt their three-dimensional structures to carry out specific biologically relevant functions in which information transfer is of the utmost importance.



	A	B	C	D	E	F
A	-	1	1	1	1	0
B	1	-	0	1	1	0
C	1	0	-	0	0	1
D	1	1	0	-	0	1
E	1	1	0	0	-	1
F	0	0	1	1	1	-

Figure 2. Mathematically, every network (**left**) can take the form of an adjacency matrix (**right**). In this case, a network with undirected, unweighted edges corresponds to a symmetric adjacency matrix with 0/1 values for the absence/presence of connections. These binary values can be derived by any pairwise correlation index between system features (nodes) by the agency of a threshold on the actual correlation values (e.g., (r > ∣0.7∣ with r = Pearson's correlation coefficient between two i,j variables) [30].

The most straightforward paradigm of information transfer through a network in proteins is the allosteric effect. Allostery is a neologism coming from the Greek language, which is related to the ability of proteins to transmit a signal from one site on a molecule to another in response to environmental stimuli. This ability stems from the transmission of information across the protein molecule from a sensor (allosteric) site to the effector (binding or active) site [32].

When in solution, proteins fold, acquiring a three-dimensional (native) structure corresponding to an energy minimum. Figure 3 reports a sketchy example of a protein molecule in solution.

Weak noncovalent forces between amino acid residues are responsible for information transduction across the network, with the consequent re-arrangement of the protein structure in response to external signals driving the allosteric process. It is worth noting the order of magnitude of these intermolecular forces approximately corresponds to the thermal noise the molecules experience in physiological conditions.

The molecule perceives ligand binding (or any other micro-environmental perturbation) happening at a distance from the active site (where the specific reaction catalyzed by protein takes place) and adapts its configuration accordingly. For example, a hemoglobin molecule senses the partial pressure of oxygen (p[O_2]) at the allosteric site, whereby when the p[O_2] is high, the affinity of hemoglobin for oxygen increases, and the protein binds to oxygen molecules at the active site. On the contrary, when the p[O_2] is low, the affinity decreases, and the bound oxygen is released into the cells. This process is crucial for life. In the lungs, there is a very high oxygen pressure, and the red blood cells containing hemoglobin must catch the oxygen molecules, which, in turn, must be released into peripheral tissues (with a low p[O_2]) to make oxidative metabolism possible. How can the protein molecule discriminate such a semantically (but not energetically) relevant signal from the continuous motions coming from thermal noise and transmit the information at a distance to reach the active site?

To answer this question, it is useful to consider a protein molecule as a network, with the amino acid residues as nodes and the noncovalent interactions between them as edges. The adjacency matrix of a protein contact network (PCN) has amino acid residues as nodes and the scoring of efficient non-trivial (where neighboring amino acid residues along the sequence have obliged and, thus, trivial contacts) contacts between residues as edges. When applied to hemoglobin, the network formalism gives rise to the pattern reported in Figure 4.

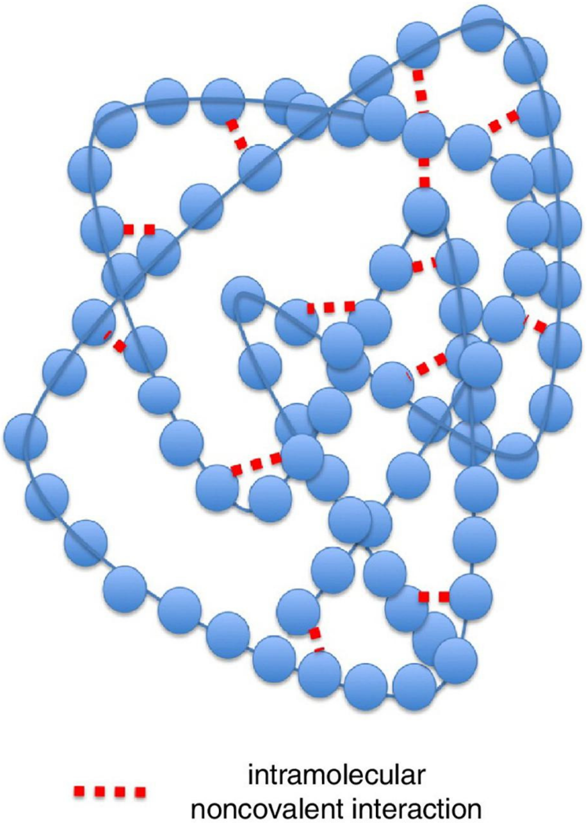


Figure 3. Spheres = amino acid residues; continuous blue line = covalent bond series responsible for the primary structure; dashed lines = noncovalent (weak) interactions between residues distant in the sequence put in contact by folding process (modified from [30]).

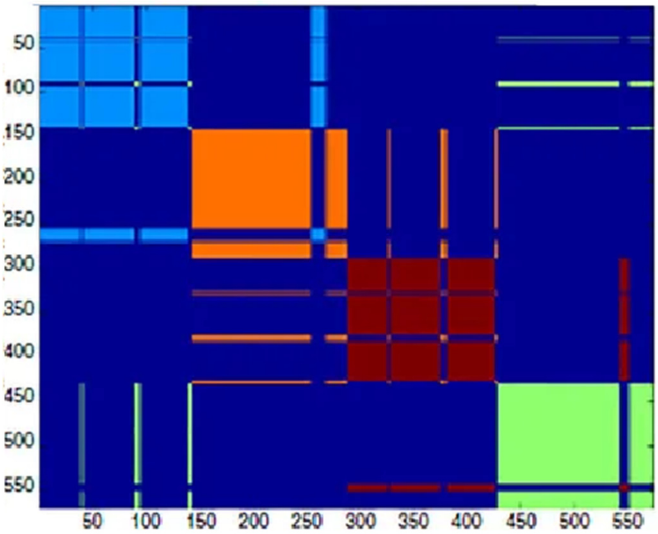


Figure 4. The numbering of the axes corresponds to the sequential order of residues along the primary structure. The adjacency matrix is shown as a clustering color map that reports the cluster partition along the sequence. The spectral-clustering technique decomposes the space through the adjacency matrix eigenvalues so that the partition relies on the topological role of the residues in the interaction network rather than on their spatial positioning (modified from [30]).

The adjacency matrix was colored by a spectral-clustering procedure [33], with specific color points corresponding to residues having a much higher number of contacts between them than with other nodes of the network and blue areas being devoid of intermolecular contacts. As expected, the clusters approximately correspond to residues located nearby along the sequence, pointing to the highly modular structure of the hemoglobin PCN. The long 'whiskers' evident in the figure as 'displaced contacts' (e.g., the 250–270-residue patch having the majority of contacts with the 'blue' cluster instead of the orange one 'nearer' in the sequence) correspond to the so-called 'fast lanes' of communication responsible for allosteric signaling. The 'whiskers' can be quantified in terms of the 'participation coefficient' (P) corresponding to the proportion of edges starting from node i and ending up in node j pertaining to a different cluster. A perturbation affecting specifically a ´high-P´ node travels a long distance across the network, passing by subsequent ´high-P´ nodes and arriving at the destination (the active site), thereby supporting allosteric effects. On the contrary, generic (noisy) thermal motion rapidly dissipates, distributing across non-directional cycles through intra-module motions. High-P nodes create a ´fast lane´ for relevant information neatly separated by noise. This finding, initially coming from hemoglobin analysis by a purely data-driven procedure (the only input data being the pairwise distances between residues coming from X-ray spectrographs), was experimentally corroborated by the simple inspection of the role of high-P nodes not present in other protein systems [34,35]. Moreover, in networks having completely different constitutive laws (e.g., genetic regulation networks where nodes correspond to the expression levels of different genetic elements), high-P elements play the same role of fast and reliable fast lanes for information transfer [36].

Let us now imagine an ML procedure to predict amino acid residues endowed with allosteric properties, i.e., in charge of transmitting a relevant environmental signal (in the case of hemoglobin, the oxygen partial pressure p[O_2]) to the active site and driving the molecular configuration change. The structural changes driving the allosteric behavior of hemoglobin have been known for many years [37]; thus, the training data set contains the three-dimensional coordinates of the 574 hemoglobin amino acid residues labeled as 'allosteric' and 'non-allosteric'. An ML algorithm exploiting the mutual spatial relations between the protein residues allows us to correctly predict the class label, while, at the same time, the network formalization permits us to 'decode' the obtained solution (the recognition of residues in charge of signal transmission) in topological terms. This relational (network-based) approach generates the hypothesis that 'the allosteric residues correspond to those having a high participation coefficient in the protein contact network', which can be immediately tested with other proteins (test sets) in which the allostery dynamics are unknown (see [35]). The obtained model of allosteric signaling is expressed in terms immediately understandable to a biochemist, maximizing the explainability of such an ML procedure [15].

5. Conclusions

Relational systems theory, exploiting the power of network thermodynamics [19], allows for the reconciliation of data- and theory-driven styles of reasoning. This reconciliation passes through the progressive blurring of the boundary between methodological and theoretical work, allowing for the discovery of organizational principles largely independent from specific scientific fields [38], as we observed in the previous chapter.

It is worth noting that the notion of a purely 'data-driven' research style relative to the network-based analysis of information spreading across PCNs can be misleading. As stressed in Chapter 2, 'theory-free' perfectly objective approaches do not exist. In the PCN case, scientists rely on theoretical principles coming from biochemistry, like the a priori definition of a distance threshold between residues for the establishment of an effective interaction between two amino acid residues or the choice of α-carbons as reference points for distance computations. The same implicit use of theoretical principles emerges from methodological aspects of the described example, like the adoption of topological spectral clustering instead of more common Euclidean metrics.

In classical systems science, based on the relative controllability of different systems, the notion of a 'grey box' is adopted [39]. This notion points to the fact that any modeling approach is a hybrid integrating both empirical data and theoretical constructs to enhance model accuracy and interpretability [15,39]. In synthesis, we can affirm that the 'tension' between data and theory is by no means an opposition but a fruitful cooperation.

The relative balance between theoretical and empirical aspects strictly depends upon the studied system. Here, I stressed the possibility offered by a relational approach to enlarge the reach of the systems approach to complex entities, such as biological ones, almost totally out of reach of mechanistic laws [40]. Moreover, the adoption of a style of reasoning tailored to thermodynamics greatly simplifies the transfer of both solutions and hypotheses across different fields of inquiry.

It is worth stressing that the adoption of a network-based relational approach is not a panacea and, in some cases, can be counterproductive, like in situations in which the definition of the atomic elements (the nodes of the network) is problematic or when in the presence of continuous variables.

From a more general perspective, it is worth noting that relational systems thinking asks for a re-shaping of scientific education; it is useless to put together an interdisciplinary group made of different specialists if they do not share a common cultural basis. On the contrary, we urgently need to raise a new generation of scientists able to encode (and then decode in the terms proper to the different scientific fields) the studied phenomena in systems science terms.

It is not an easy task to offer specific recommendations for changing scientific curricula in order to raise such a 'new generation', but, in my opinion, we should take inspiration from the artisan nature of modeling activity. Like in all artisan activities, in the case of data analysis (in the broad sense of systems science, statistics, and operation research), we deal with a relationship between a client (the expert in the field) and an artisan (the data analysis person). A good relationship is reached when the client can go inside the logic of the proposed solution, and the artisan understands what the client really needs. This asks for the hybridization of the specific knowledge of both the artisan and the client, which can be reached only by a drastic simplification of the data analysis methods [11,24], reducing the need for formalism to the minimum to make possible an immediate translation of the suggested analytic procedure in the language of the client. Meanwhile, the client should be trained to acquire an intelligent use of perspective via an educational style that makes clear the separation between key concepts and specialist details. The exaggerated emphasis on the 'last breakthrough' blurred the above perspective, increasing the fragmentation of scientific culture. In terms of scientific curricula, I do not think that we should eliminate or add some specific course but instead promote a global change in style to place much more emphasis on the 'what is it?' than on 'how can I do it?' questions.

In this respect, I am thankful to an anonymous reviewer who drove my attention to an illuminating website (https://norvig.com/chomsky.html accessed on 10 October 2024) in which the prejudices arising from the fragmentation of scientific fields are very clearly exposed and criticized. The most perturbing prejudice (arising from the prevalence of 'how can I do it?' over 'what is it?' questions) is the contempt for the statistical approach that (even by prominent scientists) is put in evidence by the derogatory statement, 'it is only statistics'.

The contempt for statistics does not only come from the side of 'experts in the field' but also from 'pure computer scientists', who often equate the actual physical entity (e.g., a specific i-th patient) to the corresponding feature vector within a data set. This acritical superposition between real and formal systems is related to overlooking the context in which the selection of statistical units (e.g., patients) happens. In their work in [41], Beaulieu-Jones and colleagues, by the analysis of a huge data set containing millions of patient records, offer an illuminating view of the problem. The authors of [41] observed that two i,j statistical units (patients) having exactly the same feature vector (and, consequently, not being distinguishable by a data-driven ML approach) have a widely different average life expectancy in the case of the feature vector derived from routine examinations or following a specific suggestion by a clinician. This is both vivid proof of the inconsistency of any 'theory-free' scientific ideal and of the urgency of fostering a fruitful exchange between different scientific fields. I hope this work can represent a (very small indeed) contribution to this goal.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Huang, S. The tension between big data and theory in the "omics" era of biomedical research. *Perspect. Biol. Med.* **2018**, *61*, 472–488. [CrossRef]
2. Dai, X.; Shen, L. Advances and trends in omics technology development. *Front. Med.* **2022**, *9*, 911861. [CrossRef]
3. Topliss, J.G.; Costello, R.J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1066–1068. [CrossRef]
4. Young, S.S.; Karr, A. Deming, data and observational studies: A process out of control and needing fixing. *Significance* **2011**, *8*, 116–120. [CrossRef]
5. Ioannidis, J. Why most published research findings are false. *PLoS Med.* **2005**, *2*, e124. [CrossRef]
6. Yuan, J.; Ran, X.; Liu, K.; Yao, C.; Yao, Y.; Wu, H.; Liu, Q. Machine learning applications on neuroimaging for diagnosis and prognosis of epilepsy: A review. *J. Neurosci. Methods* **2022**, *368*, 109441. [CrossRef]
7. David, A.; Islam, S.; Tankhilevich, E.; Sternberg, M.J. The AlphaFold database of protein structures: A biologist's guide. *J. Mol. Biol.* **2022**, *434*, 167336. [CrossRef]
8. Gorban, A.N.; Tyukin, I.Y. Blessing of dimensionality: Mathematical foundations of the statistical physics of data. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2018**, *376*, 20170237. [CrossRef]
9. Pascual, M.; Levin, S.A. From individuals to population densities: Searching for the intermediate scale of nontrivial determinism. *Ecology* **1999**, *80*, 2225–2236. [CrossRef]
10. Zimatore, G.; Tsuchiya, M.; Hashimoto, M.; Kasperski, A.; Giuliani, A. Self-organization of whole-gene expression through coordinated chromatin structural transition. *Biophys. Rev.* **2021**, *2*, 031303. [CrossRef]
11. Webber, C.L., Jr.; Marwan, N.; Facchini, A.; Giuliani, A. Simpler methods do it better: Success of Recurrence Quantification Analysis as a general-purpose data analysis tool. *Phys. Lett. A* **2009**, *373*, 3753–3756. [CrossRef]
12. Dorogovtsev, S.N.; Goltsev, A.V.; Mendes, J.F. Critical phenomena in complex networks. *Rev. Mod. Phys.* **2008**, *80*, 1275–1335. [CrossRef]
13. The Royal Society. *Science in the Age of AI: How Artificial Intelligence Is Changing the Nature and Method of Scientific Research*; The Royal Society: London, UK, 2024; ISBN 978-1-78252-712-1.
14. Andrews, M. The Immortal Science of ML: Machine Learning & the Theory-Free Ideal Preprint. 2024. Available online: https://www.researchgate.net/publication/371982028 (accessed on 10 October 2024).
15. Ho, S.Y.; Wong, L.; Goh, W.W.B. Avoid oversimplifications in machine learning: Going beyond the class-prediction accuracy. *Patterns* **2020**, *1*, 100025. [CrossRef]
16. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
17. Lu, Y.; Lu, J. A universal approximation theorem of deep neural networks for expressing probability distributions. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3094–3105.
18. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired Mag.* **2008**, *16*, 16.07.
19. Mikulecky, D.C. Network thermodynamics and complexity: A transition to relational systems theory. *Comput. Chem.* **2001**, *25*, 369–391. [CrossRef]
20. Longo, G.; Montévil, M.; Pocheville, A. From bottom-up approaches to levels of organization and extended critical transitions. *Front. Physiol.* **2012**, *3*, 232. [CrossRef]
21. Minati, G.; Licata, I. Emergence as mesoscopic coherence. *Systems* **2013**, *1*, 50–65. [CrossRef]
22. Mikulecky, D.C. Robert Rosen (1934–1998): A snapshot of biology's Newton. *Comput. Chem.* **2001**, *25*, 317–327. [CrossRef]
23. Minati, G.; Abram, M.; Pessa, E. (Eds.) *Towards a Post-Bertalanffy Systemics*; Springer: New York, NY, USA, 2016; pp. 211–218.
24. Giuliani, A. The application of principal component analysis to drug discovery and biomedical data. *Drug Discov. Today* **2017**, *22*, 1069–1076. [CrossRef]
25. Cai, S.; Mao, Z.; Wang, Z.; Yin, M.; Karniadakis, G.E. Physics-informed neural networks (PINNs) for fluid mechanics: A review. *Acta Mech. Sin.* **2021**, *37*, 1727–1738. [CrossRef]

26. Krotov, D. A new frontier for Hopfield networks. *Nat. Rev. Phys.* **2023**, *5*, 366–367. [CrossRef]
27. Smart, M.; Zilman, A. Emergent properties of collective gene-expression patterns in multicellular systems. *Cell Rep. Phys. Sci.* **2023**, *4*, 101247. [CrossRef]
28. Gigante, G.; Giuliani, A.; Mattia, M. A novel network approach to multiscale biological regulation. *Cell Syst.* **2023**, *14*, 177–179. [CrossRef]
29. Palumbo, M.C.; Farina, L.; Colosimo, A.; Tun, K.; Dhar, P.K.; Giuliani, A. Networks everywhere? Some general implications of an emergent metaphor. *Curr. Bioinform.* **2006**, *1*, 219–234. [CrossRef]
30. Uversky, V.N.; Giuliani, A. Networks of networks: An essay on multi-level biological organization. *Front. Genet.* **2021**, *12*, 706260. [CrossRef]
31. Frauenfelder, H.; Wolynes, P.G. Biomolecules: Where the physics of complexity and simplicity meet. *Phys. Today* **1994**, *47*, 58–64. [CrossRef]
32. Hilser, V.J.; Wrabl, J.O.; Motlagh, H.N. Structural and energetic basis of allostery. *Annu. Rev. Biophys.* **2012**, *41*, 585–609. [CrossRef]
33. Cumbo, F.; Paci, P.; Santoni, D.; Di Paola, L.; Giuliani, A. GIANT: A Cytoscape plugin for modular networks. *PLoS ONE* **2014**, *9*, e105001. [CrossRef]
34. Tasdighian, S.; Di Paola, L.; De Ruvo, M.; Paci, P.; Santoni, D.; Palumbo, P.; Mei, G.; Di Venere, A.; Giuliani, A. Modules identification in protein structures: The topological and geometrical solutions. *J. Chem. Inf. Model.* **2014**, *54*, 159–168. [CrossRef]
35. Di Paola, L. The discovery of a putative allosteric site in the SARS-CoV-2 spike protein using an integrated structural/dynamic approach. *J. Proteome Res.* **2020**, *19*, 4576–4586. [CrossRef]
36. Fasoli, M.; Dal Santo, S.; Zenoni, S.; Tornielli, G.B.; Farina, L.; Zamboni, A.; Porceddu, A.; Venturini, L.; Bicego, M.; Murino, V.; et al. The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *Plant Cell* **2012**, *24*, 3489–3505. [CrossRef]
37. Baldwin, J.; Chothia, C. Haemoglobin: The structural changes related to ligand binding and its allosteric mechanism. *J. Mol. Biol.* **1979**, *129*, 175–220. [CrossRef]
38. Laughlin, R.B.; Pines, D.; Schmalian, J.; Stojković, B.P.; Wolynes, P. The middle way. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 32–37. [CrossRef]
39. Deng, J.-L. Control problems of grey systems. *Syst. Control Lett.* **1982**, *1*, 288–294.
40. Dhar, P.K.; Giuliani, A. Laws of biology: Why so few? *Syst. Synth. Biol.* **2010**, *4*, 7–13. [CrossRef]
41. Beaulieu-Jones, B.K.; Yuan, W.; Brat, G.A.; Beam, A.L.; Weber, G.; Ruffin, M.; Kohane, I.S. Machine learning for patient risk stratification: Standing on, or looking over, the shoulders of clinicians? *NPJ Digit. Med.* **2021**, *4*, 1–6. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.