MAGNETIZED MEMORIES:
ANALOGIES AND TEMPLATES IN MODEL TRANSFER

Tarja Knuuttila & Andrea Loettgers

Introduction

The title of the Bakerian 2001 lecture by David Sherrington, a renowned physicist, and the other author of the Sherrington–Kirkpatrick model, was "On magnets, microchips, memories and markets: the statistical physics of complex systems."[1] That the Sherrington–Kirkpatrick model of spin glasses (i.e. disordered magnets) should find applications in as distant fields as statistical physics, computer science, neural network theory, and financial markets is both outstanding and commonplace. Indeed, it serves as an epitome of the contemporary modeling practice, where the same function forms and equations, and mathematical and computational methods are being transferred and recycled across the disciplinary boundaries. While philosophers of science have only recently started to address the interdisciplinary dynamics of such a modeling practice, it is far from a new phenomenon. The transfer of theoretical and formal tools from one area of physics to another, and from physics to other disciplines such as economics and biology have marked many scientific breakthroughs in the 19th and early 20th century. More lately, engineering has had an increasing interdisciplinary influence on many fields, attested by, for example, the emergence of synthetic biology.

Two bodies of philosophical discussion, in particular, have addressed this distinctively interdisciplinary character of modeling; one has studied analogical reasoning and the other one has focused on the role of various kinds of templates in model-building and theoretical transfer. While the interest in analogical reasoning dates back at least to the 20th century discussion of physical and mathematical analogies (e.g. Hertz 1893/1962), the discussion of templates is of a much more recent origin (e.g. Humphreys 2004, 2018; Knuuttila and Loettgers 2011, 2016; Houkes and Zwart 2018). The two discussions have proceeded in parallel, largely separately from each other[2]. Philosophers and cognitive scientist studying analogical reasoning have addressed the material and formal analogies between different objects/systems, in different domains, licensing inferences from the source objects/systems to the target objects/systems. In contrast, the emphasis of the discussion of templates has targeted cross-disciplinary formal and computational templates that are detached from any particular objects, systems or domains.

At the outset, then, the analogy-based and template-based approaches seem to be addressing different kinds of things that is also manifest in their different takes on representation. In introducing the notion of a computational template Humphreys explicitly turns away from representation to computation, while analogical inference seems wedded to similarities and thus to "representation as". But appearances can be deceptive. Our claim is that any more fully-blown account of model transfer, or that of modeling more generally, needs both perspectives, although, of course, actual cases of model transfer may proceed without making use of both template-based and analogical reasoning. Moreover, the analogy-based and template-based approaches also lend more credence to each other. Without the perspective of analogical reasoning it seems difficult to explain what drives the transfer of formal and computational templates from one domain into another, especially in interdisciplinary contexts – given that various kinds of structures may exhibit the appearances of the target domain of interest. On the other hand, the template-based approach provides a more encompassing vision of modeling that focuses on the application of tractable tools, a phenomenon driving the model-based science.

In examining how the analogy-based and template-based approaches contribute to each other, we focus on the conceptual dimension of model transfer. It has so far been inadequately addressed by the template-based view in its emphasis on tractability. Inattention to the conceptual side of model transfer is also characteristic of those accounts of analogical reasoning that concentrate on structure mapping between two domains. Moreover, analogy-based approaches have typically addressed the local level of particular source and target domains. The focus on the conceptual dimension of model transfer bridges the local and the global in focusing on how general conceptual ideas embedded in formal templates facilitate the application of those templates across different domains. This is the representing-as dimension of template transfer that draws it close to analogical reasoning, anticipating and directing actual model construction.

We will examine analogical reasoning and template transfer through a study of the application of a spin glass model to neural networks in neurosciences. The Ising model of ferromagnetism (Ising 1925) provided the basic template for the Sherrington–Kirkpatrick model of spin glasses that in turn contributed to the Hopfield model of the associative memory (1982). The formal similarities between these three models are striking, underlining the insights of the template-based approach – yet the conceptual side of these model transfers is equally important. The reason for the transferability of the formal template underlying the Ising model, and those of its variants such as the Sherrington-Kirkpatrick model is due to these models providing modelers hypothetical systems with interesting theoretical properties. These properties have been conceptually rendered as phase transitions, critical points, and most importantly, cooperative phenomena, and they can be ascribed to various kinds of systems that appear to display similar kind of behaviour.

Such model transfers presume flexibility of the formal template, involve novel theoretical interpretations, and even development of new methods.

In the following, we will first present a brief overview of the philosophical discussion of analogies and templates (Section 2), and then go over to the transfers between the Ising model, the Sherrington- Kirkpatrick model, and the Hopfield model (Section 3). The concluding chapter discusses the lessons to be learned concerning the use of analogies and templates in model-based theoretical practice.

Analogies and templates

Analogies

In philosophy of science, analogical reasoning has often been discussed in the context of knowledge generation, being related to topics such as scientific discovery, theory development and hypothesis formulation. While the heuristic role of analogies in the aforementioned activities has generally been recognized, philosophers have disagreed whether or not any general account of analogical reasoning could be formulated that would warrant analogical reasoning in general (see Bartha 2016, 12-18; Norton 2011). In this regard, the use of analogical reasoning in modeling seems instructive. Analogical reasoning provides scientific modelers a frequently utilized powerful cognitive strategy for transferring concepts, formal structures, and methods from one field to another. That analogical transfer is so common practice in scientific modeling seems to request an analysis going beyond the conventional philosophical divide between discovery and justification. We will suggest that combining the insights concerning the modelers' use of cross-disciplinary templates with those of analogical reasoning will importantly contribute to such an analysis.

A classic treatment of analogical inference in the context of scientific modeling was provided by Mary Hesse (e.g. 1966). Hesse's account is two-dimensional: she distinguishes between *horizontal* and *vertical* relations. "Horizontal relations" refer to similar or dissimilar properties of two domains. Hesse approaches them in terms of positive, negative and neutral analogies. Positive analogies refer to those properties that the two analogs have in common, whereas negative analogies refer to known differences between them. Neutral analogies refer to the properties whose commonality or difference has yet to be established providing thus epistemic potential for further inferences and theoretical development. "Vertical relations" are relations between objects and properties *within* a domain. Two domains are formally analogous if the relations between certain elements within one domain are identical or at least comparable to the relations of the corresponding elements in another domain. Hesse contrasts such formal analogies to material ones that are shared,

frequently observable and/or pretheoretic, similarities between two domains. In her view, acceptable analogies need to be grounded in both horizontal and vertical relations; in her examples, material analogies seem to provide a basis for constructing formal analogies.

Cognitive scientist Dedre Gentner has made formal analogies the centerpiece of her influential theory of analogy (e.g. 1983). She distinguishes between attributes and relations claiming that an analogy does not necessarily become stronger if the two analogs only share more attributes. According to Gentner the key similarities are those that lie in the relations that hold within domains, and it is preferably those *relations* within a domain that are being transferred to another. These kinds of formal analogies display systematicity being governed by 'higher order relations' such as causal, mathematical, or functional relationships. The fact that relations like these are being sought after in scientific reasoning underlines, in Gentner's view, the importance of formal analogies. She seems thus not to agree with Hesse, who stresses the importance of causal relations, but does not contrast them to horizontal relations between observable properties. Given the overriding focus on formal analogies of Gentner's account, analogical transfer becomes that of structure-mapping between the source and target domains (Gentner 1983; see also Gentner & Markman 1997). With structure-mapping Gentner refers to mapping knowledge about the base domain into the target domain such that the mapping rules "depend only on the syntactic properties of the knowledge representation, and not on the specific content of the domains" (Gentner 1983, 155).

Yet, analogical inference cannot boil down to a mere projection of a syntactic structure. As any system can be conceptualized in different ways resulting in different kinds of structures, and so the structure a system is supposed to exhibit is intimately related to how the system is described (see Frigg 2016). Indeed, in his discussion of analogical reasoning, Bartha (2016, 32-33) argues that there is "no short-cut via syntax". He underlines the importance of focusing on the *relevant* features of both domains, and how they relate to the analogical inference in question.

Bartha's "articulation model" addresses relevance by paying particular attention to "prior association" and "potential for generalizability" (Bartha 2010). Prior association holds in the source domain between the known similarities and a further property that is then projected on the target domain. It picks up the features of the source that are deemed relevant for the analogical inference. Bartha also underlines the need to make the prior association explicit. The potential for generalization, in turn, stipulates that there must be good grounds to believe that the same kind of connection would hold in the target domain. In particular, there should be no critical disanalogies between the domains. For instance, in analogical transfer between the Sherrington–Kirkpatrick model and the Hopfield model, the prior association between the interactions between the magnetic moments and

cooperative phenomena, i.e. ferromagnetism, is generalized to hold also in the case of pattern matching performed by the nervous system.

What we want to focus our attention on, then, is that even in the case of formal analogies, the projection of a structure is never purely syntactic. The point is that the templates used in model construction come with associated concepts that aim to capture some theoretically interesting properties of the structures in question. These concepts suggest how to theorize about the phenomenon of interest, and function as springboards for further theoretical development in the target domains. Accordingly, it is possible to extend Bartha's requirement of a potential for generalizability from the non-existence of critical disanalogies to the utilization of some general principles. These principles are embedded in well-understood formal and computational templates that are used to study multiple phenomena. We will examine this insight more in detail in the next section by discussing templates of various kinds.

Templates

The interest in templates as distinctively cross-disciplinary vehicles of modeling has its origin in the work of Paul Humphreys (e.g. 2004, 2008). He called attention to one of the most conspicuous characteristics of the contemporary modeling practice: its reliance on "the relatively small number of computational templates in the quantitatively oriented sciences" (2004, 68). "Science," he suggested, "would be vastly more difficult if each distinct phenomenon had a different mathematical representation" (Ibid.) – and this observation, we suggest, is even more true of model-based science.[3]

Humphreys zooms in on something so "simple and well-known" (2004, 60) that it has escaped the explicit attention of philosophers of science: computational templates. Humphreys' computational templates are genuinely cross-disciplinary mathematical or computational forms and methods that can be applied to different problems in various disciplines. In using the word 'template' Humphreys invokes a pattern for developing a product that can simultaneously be configured in view of the aims of the modeler (2008, 5).

Computational templates may have their origin in formal disciplines such as the Poisson distribution in probability theory. While computational templates are genuinely subject-independent in their application, many of them have originally been introduced as theoretical models of a certain system being only subsequently applied to different domains. A theoretical model functions as a template first when it is separated from the original theoretical context and used to model other, usually very different types of phenomena. The Lotka-Volterra model[4] and the Ising model provide illustrative examples

of such templates that are so general that they have been used virtually in all areas in science, where scientists are engaged in mathematical model-building. But apart from generality, computational templates need to be tractable, to allow for computation. Humphreys (2004) considers tractability to be the distinguishing feature of computational templates. Interestingly, many models such as the Lotka-Volterra model and the Ising model have gained such tractability due to the subsequent development of mathematical and computational methods. Generic technologies such as digital computing made the Lotka-Volterra model a suitable case for studying the dynamics of nonlinear systems (Knuuttila and Loettgers 2011). On the other hand, mathematical methods, developed within physics, such as the renormalization group theory, played a crucial role in making the Ising model tractable.

Humphreys (2018) seeks to distinguish the template-based approach from analogical reasoning. His main argument is that while analogical reasoning relies on similarities typically left at least partially implicit, the template-based approach does not need to rely on any "vague" resemblances. Instead, theoretical templates that are candidates for developing into transdomain templates are typically results of construction processes[5], whose assumptions can be made explicit. As a consequence, Humphreys claims, there are cases where "there is no need to use analogical reasoning in applying a template – we can check directly whether the assumptions are satisfied for the system at hand." (Humphreys 2018, 4). Moreover, Humphreys (2018) notes, in line with his earlier writings on templates, that any transfer of a template usually involves refinement and adaptation of the template to a new domain, except for the off-the-shelf representations "that can be opportunistically justified at the system level by analogical reasoning from their previous successful applications that are recognized as similar" (ibid.). According to Humphreys' examples, statistical distributions as well many general equation forms belong to his group. However, when it comes to the latter group, their off-the-shelf nature is questionable, as our case shows.

Another related question concerns the application of formal templates with only mathematical interpretation, and whose "construction assumptions have only mathematical content" (Humphreys 2018, 3).[6] Humphreys seeks to explain how it is possible to apply formal templates given their purely mathematical interpretation. He approaches such an application as a mapping from a formal template to a target system. In such a mapping, all empirical content is contained within the mapping, and not within the template that explains why formal templates can be applied across a multitude of domains. Humphreys views the construction and use of formal templates as superior to analogical inference since in using them scientists do not need to invoke the language of the domain from where the template originally comes from: "There is therefore no need for vocabulary translations or for interdisciplinary knowledge" (Humphreys 2018, 7). He uses as an example Barabási-

Albert model (Barabási and Réka 1999) that is a random scale-free network making use of a preferential attachment mechanism. It has been applied to various kinds of natural and social networks that contain nodes ("hubs"), whose number of links within a network greatly exceeds the average.

The Barabási-Albert model fits Humphreys' views on formal template transfer as its origin is in mathematical theory, and so it is devoid of previous subject-specific empirical and theoretical content. But does such a case of formal template transfer also suit other kinds of template transfers characteristic of contemporary modeling practice? And what explains the seemingly unreasonable success of a relatively small number of templates? We do not believe that their success can be explained by tractability and generality alone, unless these two features are linked in a particular way. Namely, successful templates embody something more: a vision of the phenomenon exhibiting a particular kind of *general* pattern for the study of which the template offers *tractable*, or at least already well-studied tools. And seeing various kinds of systems as instances of some already familiar general patterns amounts – to use Kuhnian language – to a "gestalt switch" that enables scientists to approach various kinds of systems as being like each other at least in one important dimension.[7] In order to study this analogous moment in template transfer, we turn to our case study on the model transfer between the Sherrington-Kirkpatrick model and the Hopfield model. The idea of cooperative phenomena forms the conceptual core of this template transfer, already introduced by the Ising model that provided some basic formal templates and associated theoretical ideas for the Sherrington-Kirkpatrick and Hopfield models.

Modeling cooperative phenomena

The Ising model, originally presented as a mathematical model of ferromagnetism by Ernst Ising (1925) is nowadays used to study an amazing variety of phenomena in different disciplines ranging from physics to biology and social sciences. Although what appear to be transmitted between different fields in the case of the Ising model are mathematical structures, the conceptual side of these model transfers has been equally important. Physicist Daniel Amit has described the conceptual fruitfulness of the Ising model in physics the following way:

> "[The Ising model] has been a birthplace and the testing ground for a treasure of new concepts in essentially all fields of physics. Such fundamental ideas as symmetry breaking, cooperative phenomena, order parameters, disorder parameters, critical exponents, symmetry restoration etc., have had their first explicit, precise articulation in the framework of this apparently simple, naïve model." (Amit 1989, 105)

While most of these theoretical ideas have been developed and applied in the domain of physics, especially the concept of *cooperative phenomena* has proved more globally applicable, being applied beyond physics in biology, economics, and sociology. How did cooperative phenomena gain this global, cross-disciplinary nature? In the following, we will trace the journey of this concept from the context of modeling properties of magnetic systems in physics into neuroscience.

The Ising model

Cooperative phenomena are general in character, resulting from interactions between the constituents of a system. These interactions can be of various kinds. Ferromagnetism provides a standard example that is also of a historical importance. On the microscopic level, a piece of magnetic iron consists of magnetic moments, which below a certain temperature T_C align and result in a macroscopic net magnetization. Above the temperature T_C the thermal motion of the magnetic moments counterfeits this tendency and as a result the net magnetization vanishes. The piece of iron becomes paramagnetic. The transition from the ferromagnetic phase into the paramagnetic phase (and the other way around) is called *phase transition*, and T_C the *critical temperature*. This kind of transition can be observed in experiments on the macro level, but the interactions between magnetic moments on the micro level that give rise to the transition, are not experimentally accessible. The Ising model provides a conceptual and methodological framework by which these processes on the micro level can be approached.

At first glance, the structure of the Ising model seems astonishingly simple for such a consequential model. It consists of N magnetic moments, so-called spins S_i, which can take only two values, $S_i = +1 \text{ or } -1$, corresponding to their two possible discrete orientations up and down. In the two-dimensional case, the spins are located on the sides of a lattice. The interaction, which is central for the occurrence of cooperative phenomena such as ferromagnetism, is given by the interaction energy J_{ij} describing the interaction strength between the *nearest* neighbor spins S_i and S_j. In the original Ising model J_{ij} is constant and in the case of ferromagnetism, where all the spins are aligned, the interaction energy is positive ($J > 0$). To sum up, the interaction energy depends on the configurations of neighboring spins and, furthermore, tends to align them.

With each magnetic moment S_i comes an internal magnetic field h_i that is created by the interaction between the magnetic moments S_i:

$$h_i = \sum_{j, j \neq i}^{N} J_{ij} S_j \qquad (1)$$

with $J_{ij} = J_{ji}$.

For each of the 2^N configurations of spins $\{S\}$, where 2 stands for the two possible orientations of the spins and N for the total number of spins, an energy function is given for each of these microstates by:

$$E\{S\} = -\frac{1}{2}\sum_{i,j\neq i}^{N} J_{ij} S_i S_j. \qquad (2)$$

The overall energy of the system decreases if S_i and S_j point in the same direction. In this case the interaction energy J_{ij} is making a positive contribution, which, together with the "–" sign in front of the sum, leads to the decrease of the overall energy E. If S_i and S_j point in different directions, the overall energy of the system increases because the interaction energy J_{ij} is making a negative contribution.

The formal structure of the magnetic moments S_i, the internal magnetic field h_i and the energy $E\{S\}$ provide conceptual and methodological resources by which phenomena of cooperative can be explored. In Humphreys' terms, they can be approached as templates. One of the main challenges consists in calculating from the microscopic behaviour the macroscopic properties such as the magnetization M. In general, calculating some macroscopic property from the microstates of the system is one of the main subjects of statistical mechanics. The Ising model provided one successful framework for such a task, developed in the confined context of ferromagnetic systems.[8]

In statistical mechanics the magnetization M is calculated from the microstates $\{S\}$ in the following way:

$$\langle M \rangle = \frac{\sum_{\{S\}} M\{S\} \cdot e^{-E/kT}}{\sum_{\{S\}} e^{-E/kT}} \qquad (3)$$

In the equation, the sum $\sum_{\{S\}} e^{-E/kT}$ is the so-called partition function. The functional form of the energy E given by equation (2) is specific to the Ising model due to its form of interaction. The exponential function describes the probability of the realization of a

microstate at a given energy and temperature. Because of the minus sign in the exponential, microstates at a high energy are less probable than states at a lower energy. The probability that a microstate is realized can be calculated by:

$$P_{\{S\}} = \frac{e^{-E/kT}}{\sum_{\{S\}} e^{-E/kT}}$$

Accordingly, the magnetization is the sum over all 2^N possible states weighted by the probability that a microstate is taken. The partition function and energy function can be considered as general theoretical templates in the context of physics. The partition function has its origin in statistical mechanics, whereas the energy function is fundamental in all parts of physics. The probability distribution provides a computational template. Those templates are flexible enough for modeling different forms of interaction and therefore cooperative behaviour. The package consisting of the concept of cooperative phenomena and the associated theoretical and computational templates served as a framework within which further concepts such as phase transitions, critical exponents, and symmetry breaking got developed, together with further methods for their calculation. Within physics, these concepts and methods have not subsequently been confined to the case of ferromagnetism. They have been applied also to other physical systems exhibiting cooperative phenomena. The so-called spin glasses provide one particularly fruitful application. Spin glasses differ from ordinary glasses by containing a small number of magnetic moments interacting with each other. These interactions lead to interesting cooperative behaviour due to the fact that both ferromagnetic and antiferromagnetic couplings are present in the system. The simultaneous presence of ferromagnetic and antiferromagnetic couplings, in general, does not allow for the establishment of a conventional long-range order (of ferromagnetic or antiferromagnetic[9] type).

The Sherrington–Kirkpatrick model

In 1978 David Sherrington and Scott Kirkpatrick, introduced a model of a spin glass by drawing an analogy to the case of ferromagnetism (Sherrington and Kirkpatrick 1978). They hypothesized that the observed behaviour of spin glasses is caused by the interaction between their magnetic moments as is the case in the Ising model. The subsequent development of the Sherrington-Kirkpatrick model (hereafter the SK model) made available further concepts, templates and methods.[10] The flexibility of the theoretical and formal templates underlying the Ising model allowed for the construction of the SK model, which tries to capture spin glass specific observations, such as a transition into a disordered

state at low temperatures. However, at first glance the SK model cannot be distinguished from the Ising model.

Like in the Ising model the magnetic moments are represented by binary variables S_i in the SK model. Again the magnetic moments S_i can take either the value +1 or -1. The coupling between two spins, S_i and S_j, is, as before in the case of the Ising model, represented by the coefficient J_{ij} and the overall energy of the system is also of the form:

$$E = -\sum_{i,j \neq i} J_{ij} S_i S_j.$$

The main difference between the two models lies in the form of interaction. In the SK model the couplings are modeled as a function of the distance between the magnetic moments $J_{ij} = J(R_i - R_j)$, with R_i and R_j as the positions of the magnetic moments on, for example, a lattice. This kind of interaction leads to cooperative behaviour, which is different than in the case of the Ising model. Positive values of J_{ij} correspond to ferromagnetic and negative values to antiferromagnetic couplings. The spins in this model cannot at the same time satisfy both ferromagnetic and antiferromagnetic couplings that means that the couplings are of a competitive nature. The consequences of these competing interactions between the ferromagnetic and antiferromagnetic couplings become apparent at low temperatures when the system undergoes something like a phase transition. The system exhibits a "freezing transition" to a state with a new kind of order in which the magnetic moments are aligned in random directions (Binder and Young 1986). The topology of the energy landscape of spin glasses after undergoing this freezing transition is varied, consisting of a large number of valleys, representing metastable or stable spin configurations.
The flexibility of the partition and energy functions allowed scientists to explore what kinds of macroscopic properties such as the freezing transition caused by the microscopic behaviour. These calculations turned out to be very difficult and required the development of further mathematical tools such as the replica method, which is used in statistical mechanics in the calculation of the partition function (see footnote 12).

The Ising and SK models display how the concept of cooperative phenomena coupled with theoretical and formal templates and computational methods accounts for macrolevel phenomena, such as phase transitions, in terms of microlevel interactions. Such cooperative phenomena is not limited to physical sciences. The question is how template transfer from physics to other sciences, licensed by the notion of cooperative phenomena,

is bound to succeed. An example of such a transfer is provided by the Hopfield model, which transferred the SK model of spin glasses to neuroscience. While the SK model successfully applied the Ising model to spin glasses, explaining some characteristics of spin glasses via their ferromagnetic and antiferromagnetic moments, any such straightforward interpretation of microlevel phenomena was not possible in neuroscience. As a result, analogical reasoning played a much more substantive role in the case of the Hopfield model.

The Hopfield Model

Theoretical physicist John Hopfield introduced in 1982 a model that was to become one of the milestones in the study of artificial neural networks. (Hopfield 1982) He was interested in how the brain could fulfill the task of completing information, such as recognizing a friend's face, from an incomplete input (i.e. partial picture provided by our visual perception). Hopfield made use of both positive and negative analogies in his reasoning. First, he drew a negative analogy to small circuits such as electric circuits and computers, arguing that evolution does not proceed in the same way as an engineer. Second, he drew a positive analogy to many body systems suggesting that auto-associative memory could be understood as collective[11] phenomena.

> "Given the dynamical electrochemical properties of neurons and their interconnections (synapses), we readily understand schemes that use a few neurons to obtain elementary useful biological behavior. Our understanding of such simple circuits in electronics allows us to plan larger and more complex circuits, which are essential to large computers. Because evolution has no such plan, it becomes relevant to ask whether the ability of large collections of neurons to perform "computational" tasks may in part be a spontaneous collective consequence of having a large number of interacting simple neurons." (Hopfield 1982, 2554)

The turn towards collective phenomena does not, by itself, give too much understanding of how the brain possibly recognizes a pattern from an incomplete input. In the construction of the actual model, Hopfield began from visualizing the process of pattern recognition by water flowing from different directions into a valley. Take again the example of the friend's face; in addition to her we memorize a lot of other different people or objects. Translated into the visual analogy of Hopfield, such phenomena amount to a landscape consisting of many valleys, in which each valley stands for one memorized person and object. This valley analogy renders the problem of pattern recognition as that of finding how patterns (valleys) are stored and the dynamic by which an incomplete pattern is recognized and by doing so completed.

The complex structure of the landscape of the SK model, consisting of many energy minima, provided a suitable template for Hopfield's idea of how pattern recognition should be approached.[12] The notion of energy minima, which was one of the central theoretical elements transferred from the SK model, was interpreted by Hopfield in terms of stored patterns. In his analogical reasoning Hopfield modified and integrated different templates such as the energy function, rendered the neural components as binary variables, and introduced dynamic and storage rules from statistical mechanics and neuroscience. The construction of the Hopfield model was far from any straightforward application of the SK model.

One challenge was due to the randomness of the energy minima in the SK-model. This cannot be the case if energy minima stand for a stored pattern to be recovered by the recognition system. In order to accommodate this feature, Hopfield made use of the Hebb learning rule (Hebb 1949). According to this rule the synaptic efficiencies between neurons are described by the set of parameters J_{ij} in which the information is stored. The simultaneous activation of two connected neurons results in a strengthening of the synaptic coupling between the two neurons. This rule is formalized in the Hopfield model as follows:

$$J_{ij} = \sum_{\mu=1}^{p} \xi_i^{\mu} \xi_j^{\mu}$$

The ξ_i^{μ} are variables that describe a pattern, i.e., a given configuration of active and inactive neurons. The number of patterns stored into the network is given by p, and in each pattern the number of neurons is equal to the total number of neurons in the network, N. Each of the patterns is associated with an energy minimum. The topology of the energy landscape shows a similar complexity as in the case of the SK model. By implementing the Hebb rule, which even has some neurophysiological grounding[13], the patterns are not random anymore. On the formal level, the structure of the Hopfield model is akin to the Ising and SK models. The main difference between them – as was also the case between the Ising and SK models – lies in the choice of the coupling between the components of the network.

As in the case of the Ising and SK models, the neurons σ_i in the Hopfield model are binary variables. The neuron σ_i takes the value 1 in case it is active, and the value 0 if it is inactive. The state of each of the neurons is determined by its post-synaptic potential (PSP) h_i, produced by the activating signals arriving from all the other neurons to which it is connected. It is given by:

$$\boldsymbol{h_i = \sum_{j,j\neq i}^{N} J_{ij}\sigma_j}$$

The post-synaptic potential is of the same form as the internal magnetic field caused by the magnetic moments, although it has different interpretation in the case of the Hopfield model. The internal magnetic field in the Ising model is the magnetic field that the magnetic moment S_i experiences. It either aligns the two magnetic moments or let them point into different directions. In the case of the Hopfield model the magnetic field is replaced by a biochemical interaction, which either changes the state of the neuron e.g. from active to inactive or leaves it in its actual state. Moreover, the energy function is also of the same form as in the case of the Ising and SK models:

$$E = -\frac{1}{2}\sum_{i\neq j} J_{ij} S_i S_j$$

It assigns an energy value E to each system configuration $\sigma = \{\sigma_1,....,\sigma_N\}$. The energy function can be considered as a computational template that is adjustable to the respective system to be modeled. A further important difference between the Ising, SK model and the Hopfield models is given by their dynamics. The Ising and SK models are not dynamic; they calculate the properties of the system such as the magnetization of the microstates through the probability of their occurrence. In contrast, dynamics is essential for the Hopfield model; starting from an incomplete input the neural network develops into an energy minimum, associated with one of the stored patterns.

To conclude, the Hopfield model provides an example of how the notion of collective phenomena enabled Hopfield to draw an analogy between auto-associative memory and the phenomena modeled by the Ising model and the SK model. This analogy enabled Hopfield to transfer theoretical and computational templates from the study of magnetic phenomena into the field of artificial neural networks. These templates functioned as conceptual and methodological resources, which were used to construct an artificial neural network that was able to recognize patterns.

Analogies and Model Templates in Model Transfer

Above we have shown that what made the Ising model, and its off-springs, like the Sherrington-Kirkpatrick model, attractive candidates for model transfer is the conceptual and methodological framework they embody. It renders certain kinds of patterns as instances of cooperative phenomena coupled with associated mathematical forms and tools that enable the study of such phenomena. The Sherrington-Kirkpatrick model examines a situation where the behaviour of magnetic spins is disordered due to competing ferromagnetic and antiferromagnetic couplings between the magnetic moments. This situation leads to behaviour that cannot be anticipated from any single elements of the system but only from the competing interactions between a large number of individual elements, leading to a large number of local minima. In the course of cooling down, the spin glass gets trapped into one of the many local minima of the complex energy landscape. Hopfield was able to use this property in modeling auto-associative memory.
The notions of either a computational template or a formal template do not adequately recognize this intertwinement of the conceptual, mathematical and computational sides of model transfer.

To better capture the holistic aspect of modeling Knuuttila and Loettgers introduced the notion of a *model template* that is a mathematical structure or a computational method that is "coupled with a general conceptual idea that is capable of taking on various kinds of interpretations in view of empirically observed patterns in materially different systems" (2016, 396). As such, a model template provides "a formal platform for minimal model construction coupled with very general conceptualization without yet any subject-specific interpretation or adjustment" (ibid., 382). The Ising model provided such a model template for the SK model, and the SK model, in turn, provided a model template for the Hopfield model. This model template can be understood as a formally defined framework for modeling particular kinds of cooperative systems that instantiates the concept of cooperativity through the interaction energy J_{ij} ,[14] i.e. the coupling strength between the magnetic moments of the system. The interaction energy is central for the cooperative behaviour of the system. It defines the form of the energy landscape by being embedded into the interlocking theoretical templates of energy function $E\{S\}$, magnetization $< M >$, and the partition function $P_{\{S\}}$. Transferred to the Hopfield model these templates forgo their original interpretation becoming thus computational templates.[15]

Thinking about the SK model as a model template for the Hopfield model emphasizes the importance of the analogical dimension of template transfer. The notion of a cooperative mechanism provided the central conceptual idea shared by the Ising, SK and Hopfield models. That Hopfield was able to conceive of pattern recognition in terms of the energy landscape resulting from competing ferromagnetic and antiferromagnetic couplings

between the magnetic moments was a result of analogical reasoning, and not of any unequivocal structure mapping. Rather, the analogy enabled him to make use of the theoretical and computational templates provided by the Ising model and the SK model, leading to intricate model construction process in which Hopfield also drew resources from statistical mechanics.

According to Humphreys one can often dispose of analogical reasoning, because the model construction assumptions can be stated explicitly and checked empirically. This may hold in some cases in physics, but not even in the case of the SK model that does not lend itself to any straightforward empirical interpretation. And checking empirically the assumptions of the model in the case of transdomain transfer may even be more difficult.[16] In the case of the Hopfield model, it is difficult to see how the construction assumptions could be checked, as the concepts adopted from physics such as temperature or phase transitions do not map onto any empirical properties of neural networks.

In our view, analogy-based and template-based approaches can fruitfully be used to augment each other. In analogy-based approaches the formal and mathematical representations are often considered to be derived by abstraction from target and source domains. Analogy enables the mathematization of the target domain in terms of relational generalizations that may yield abstract schemas. For instance, Nersessian (2002) details how Maxwell formulated the mathematical representation of the electromagnetic field concept by making use of imaginary models of fluid medium, drawing, moreover, from continuum mechanics and machine mechanics. As he progressed in this theorizing, his conception of the aetherial medium became more abstract. Nersessian's discussion captures the conceptual and intertheoretical dimension of analogical exchange, but displays also the tendency of the analogy-based approaches to disregard the genuinely cross-disciplinary nature of many formal tools.

In contrast, the template-based approach focuses on the generalized methods for modeling various kinds of systems. It also addresses the question of why some tractable templates have proven so nearly universally applicable. Apart from mentioning their generality and tractability, Humphreys underlines the importance of the local construction and adjustment of templates. We have suggested yet another reason for this universality, highlighting the importance of the analogy-based approach: crucial for template transfer is the general conceptual core of the model template. This conceptual core is global in character, motivating local and domain-specific template construction and adjustment processes. While the templates themselves appear to be merely syntactic structures in transdomain exchange – given that their underlying ontologies change with the different material systems they are applied to – they do have an important conceptual dimension.[17] It is animated by analogies between various kinds of systems that are used to mobilize template

transfers across a wide spectrum of domains and disciplines, a practice that is particularly visible in contemporary complex systems theory and network science.

Last, and related to the global character of model templates, we wish to briefly consider how Hopfield himself understood model transfer from physics to neuroscience. In a symposium on the work of John Hopfield's former student Sir David McKay, John Hopfield referred to Niels Bohr and Max Delbrück, who both thought that in order to describe and explain biological phenomena, a new kind of physics would be necessary. Their question was: "How the diverse seemingly purposeful complex phenomena described by the word 'life' could emerge from lifeless physics?"[18].

Hopfield argued that from the present-day vantage point the idea of a new kind of physics may have become obsolete with the realization that the laws of neural-based behaviour in higher animals are macroscopic. Biological and large physical systems are alike in that both have robust emergent properties arising from the interaction between the components of the system. Herein lies the justification for Hopfield for drawing an analogy between magnetic systems and neural networks. The analogy is based on the shared structural and dynamical properties of systems giving rise to specific properties such as ferromagnetism, or pattern recognition. It is not justified on the basis of any observed fact of analogy between some particular systems (cf. Norton 2011). The justification is more general and theoretical in nature, related to a new grouping of systems and phenomena under the headings of many body systems and emergent behaviour that enables the analogical transfer of theoretical and model templates within physics, but also beyond physics to biology.[19]

References

Amit, D. (1989). *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge, MA: Cambridge University Press.

Barabási, A-L and A. Réka. (1999). Emergence of Scaling in Random Networks. *Science*, 286, 509-512.

Bartha, P. (2016). Analogy and Analogical Reasoning. Winter 2016 Edition. *In Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information Stanford University, Stanford, CA.

Bartha, P. (2010). *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. New York: Oxford University Press.

Binder, K., & Young, A.P. (1986). Spin Glasses: Experimental Facts, Theoretical Concepts, and Open Questions. *Reviews of Modern Physics*, 58(4), 801-976.

Caianiello, E.R. (1960). Outline of a Theory of Thought-Processes and Thinking Machines. *Journal of Theoretical Biology*, 1(2), 204-235.

Cragg, B.G., & Temperley, H.H. (1954). The Organization of Neurons: A Cooperative Analogy. *Electroencephalography and Clinical Neurophysiology*, 6(1), 85-92.

Frigg, R. (2006). Scientific Representation and the Semantic View of Theories. *Theoria*, 55, 49-65.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7, 155-170.

Gentner, D., & Markman, A. B. (1997). Structure Mapping in Analogy and Similarity. *American Psychologist*, 52(1), 45-56.

Goodwin, R. M. (1967). A growth cycle. In C. H. Feinstein (Ed.), *Socialism, Capitalism and Economic Growth* (Essays presented to Maurice Dobb). Cambridge: Cambridge University Press.

Hertz, H. ([1893] 1962). *Electric waves: Being researches on the propagation of electric action with finite velocity through space*. New York: Dover Publications.

Hesse, M. (1966). *Models and Analogies in Science*. Notre Dame: Notre Dame University Press.

Hebb, D. (1949). *The Organization of Behavior: A Neurophysiological Theory*, Mahwah: Erlbaum Books.

Hopfield, J. (1982). Neural Networks and Physical System with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences of the USA*, 79(8), 2554-2558.

Houkes, W., & Zwart, S. D. (2018). Transfer and templates in scientific modelling. *Studies in History and Philosophy of Science*.

Humphreys, P. (2004). *Extending Ourselves. Computational Science, Empiricism and Scientific Method*. Oxford: Oxford University Press.

Humphreys, P. (2018). Knowledge Transfer Across Scientific Disciplines. *Studies in History and Philosophy of Science.*

Ising, E. (1925). A Contribution to the Theory of Ferromagnetism. *Zeitschrift für Physik*, 31(1), 253-258.

Knuuttila, T., & Loettgers, A. (2011). The Productive Tension: Mechanisms vs. Templates in Modeling the Phenomena. In P. Humphreys, and C. Imbert (Eds.), *Representations, Models, and Simulations*, 3-24. New York: Routledge.

Knuuttila, T., & Loettgers, A. (2016). Model Templates Within and Between Disciplines; From Magnets to Gases—and Socio-economic Systems. *European Journal for Philosophy of Science*, 6(3), 377-400.

Mezard, M., Parisi, G., & Virasoro, M. A. (1987). Spin glass theory and beyond: An introduction to the replica method and its applications. *World Scientific Lecture Notes in Physics* (Vol. 9). Singapore: World Scientific Publishing.

Nersessian, N. (2002). Maxwell and 'the Method of Physical Analogy: Model-based reasoning, generic abstraction, and conceptual change. In D. Malament (Ed.), *Essays in the history and philosophy of science and mathematics*, 129-166. Lasalle, Il: Open Court.

Norton, J. (2011). Analogy. unpublished draft, University of Pittsburg.

Sherrington, D., & Kirkpatrick S. (1978). Infinite-Ranged Models of Spin-Glasses. *Physical Review B* 17, 11, 4384-4403.

Yang, G., Lai, C.S.W, Cichon, J., Ma L., Li W., & Gan., W. (2014). Sleep promotes branch-specific formation of dendritic spines after learning. *Science* 344(6188), 1173-1178.

[1] The Bakerian Medal and Lecture is awarded annually by The Royal Society, and it is one of the most prestigious lectures in physical sciences.

[2] See however Humphreys (2018).

[3] Paul Humphreys originally introduced the notion of a computational template in his study of computer simulations and their relation to traditional modeling techniques (see also Humphreys 2018, 3).

[4] See Houkes and Zwart (2018) for a study of the application of the Lotka-Volterra model to technology transfer. The model was originally used by Volterra to study population dynamics and Lotka to study biological and chemical systems more generally (Knuuttila and Loettgers 2011)

[5] Humphreys contrasts constructed transdomain templates to those theoretical templates that are part of the fundamental principles of a theory, such as Newton's Second Law, or Schrödinger equation.

[6] Humphreys (2004) does not specifically address formal templates.

[7] There is an analogous moment even in the transfer of what seem as a purely formal template. Such an analogous dimension of model transfer depends, we suggest, on the conceptualization of the phenomenon as being of a particular kind, and giving thus rise to some distinctive patterns.

[8] Another example is provided by the kinetic gas model.

[9] In the antiferromagnet neighboring spins point in different directions. In the paramagnetic phase the spins point due to the temperature into random directions.
[10] Probably the most important method developed in this context is the replica method, which allows for the calculation of the sum over the 2^N microstates that easily becomes a very large number. Also the different possible realizations of disorder pose a serious problem: The form of phase transitions varies depending on the distribution of the interactions between the magnetic moments. This means that there exists a correlation between disorder and the form of the phase transition. In order to get more representative results, an average of a large number of different realizations of interactions —replicas— are made use of (see Mezard, Parisi, and Virasoro 1987).
[11] Hopfield used the word 'collective' synonymously to what we call 'cooperative'.
[12] Hopfield was not the first one to draw an analogy between the Ising model and the organization of neurons (see Cragg and Temperley 1954, Caianiello 1960)
[13] Direct experiments on neurons have shown that changes in the signaling transfer is part of learning in the brain (e.g. Yang et al. 2014).
[14] Interaction energy is a general concept that can be found in physics, chemistry, as well as in biology.
[15] In the Hopfield model, the equations for modeling the dynamics of pattern recognition, i.e. Glauber dynamics, is more generally used to model the stochastic dynamics in the Ising model. Glauber dynamics can be compared to computational templates such as the Poisson distribution. A question, which remains unanswered by Humphreys is the relationship between computational and formal templates. The notion of a computational template has receded in the background in Humphreys (2018). Poisson distributions are in Humphreys (2004) examples of computational templates, but they are discussed as formal templates in Humphreys (2018, 4). Does this mean that the notion of a formal template covers computational templates? And how are then computational templates related to formal templates such as Barabási-Albert preferential attachment templates? How these lines are drawn does not seem to be of a consequence for our argument, since in addition to the formal and/or computational side of template transfer, the notion of a model template also encompasses a conceptual dimension.
[16] Humphreys does not deny that analogical reasoning may play some role, as shown by his discussion of the application of the Volterra's predator-prey model to the dynamical contradictions of capitalism (Goodwin 1967). By referring to the idea of a symbiosis of two populations that are partly complementary and partly hostile, he grants that "Kuhnian analogies can assist in the transfer of a representation from one domain to another" (2008, 5). Yet, he insists that "this analogical transfer can be made explicit by means of a formal set of assumptions" (ibid); in the case of the Goodwin model, the Lotka-Volterra equations can be arrived at via explicit economic assumptions. Moreover, in Humphreys' view, formal templates, such as Barabási networks, are instantiated by the mapping of a formal template on a target system, and the success of this mapping process is evaluated in terms of whether the formal construction assumptions are empirically justified. Thus, Humphreys appears to argue that analogies can either be dispensed with, or they are not needed in the first place.
[17] Although Humphreys (2004) argues that templates are endowed with intended interpretation, he also mentions that changing the ontology of the system comes "very close to starting a new [template] construction process" (Humphreys 2004, 80). If the template is used to model a new system, the original justification goes with the intended interpretation (ibid.). What we are arguing is that the general conceptual content of a model template bestows it with still some justification, along with tractability, and the interesting philosophical question is what kind of justification this is.
[18] Inference, Information, and Energy: A Symposium to celebrate the work of Professor Sir David McKay. University of Cambridge 15.3.2016.

[19] This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 818772).