New Mechanism for complexity:
How to enable understanding of emergent phenomena through the lens of Machine-Learning

Abstract

In this thesis, I deal with a topic in epistemology of Machine Learning (ML). With an outstanding predictive accuracy and its ability to handle large amounts of data, ML is increasingly applied to complex systems science. However, ML models are often opaque and sometimes described as ”ruthless correlation extractors”, which makes them ineffective for understanding on a process-level. I seek to improve upon the concept called ”link-uncertainty”, introduced by Emily Sullivan, who addressed the question of how we could gain understanding through ML. In her drawn picture, mechanistic knowledge is just a passive precondition for an abstract level of understanding that is not further specified. Instead, I focus on mechanisms as a desired target of understanding, while grounding my analytical terminology within the recent movement of ”New Mechanism”. On the backdrop of a symbiotic (statistical/mechanistic) modelling framework, I first use case studies that apply ML in the field of climate science, to further centre my ideas around a ML model, called *AgentNet*, which deals with agent-based complex systems in a physically transparent way. Based on my analysis, I introduce a novel concept that I labelled ”Correspondence Principle for Mechanistic Interpretability”, or short **CPMint**. It features a threefold correspondence-scheme between a ML model and the target system - First, on the ontological, second on the functional, and third on the predictive, phenomenological level, thus serving as a recipe to establish ”mechanistic interpretability”. In contrast to Sullivan’s ”link uncertainty”, **CPMint** capitalises on introducing physical transparency into the ML model, which makes it a guide to setting up ML models that aim at contributing to procedural knowledge within complex systems.

Contents

Introduction

In the last decades, there has been progress towards more powerful computational tools. Regarding data-driven technology, Machine-learning (ML) is a particularly outstanding example. High predictive accuracy and an ability to extract patterns in vast amounts of data are two of the virtues justifying why ML entered almost every branch of science; It established itself especially in disciplines that are known to amass huge piles of data, like microbiology, medical sciences, neuroscience or atmospheric and climate research. However, ML models, especially deep neural networks (DNNs) are notoriously opaque, meaning that the working mechanisms of these systems are very unintelligible. In contrast to traditional deterministic systems, for which we can derive a degree of certainty that a system is working correctly, for example through end-user testing, design documentation, or code reviews, ML approaches are inherently non-deterministic in nature. They are perfect correlation-extractors, trained to find a functional relationship between sometimes high-dimensional vector spaces. Another way to express this opacity is that DNNs do not provide information about *how* the weights have been constructed to approximate an arbitrary function, but merely that such a construction is possible. All these aspects are summarised with the term ”black-box”, and as a consequence, the epistemic role of ML within the broader scientific modelling context has been - and still is subject of philosophical debate.

In part fuelled by these realisations, research has gone some way in recent years towards making ML models interpretable and explainable.[1] Some attempts at making these systems more transparent utilised physical constraints within the algorithmic architecture, which led to the creation of hybrid models. This is a very plausible strategy, since ML models can easily deal with real-world uncertainty that normally requires elaborate statistical treatment, thus filling the gaps where conservative physical models are failing. It is a discernible trend that moves into the direction of ”symbiotic modelling”; Physical models benefit from ML, which is statistical and data-driven, while physically constrained ML models benefit through increased accuracy. For example in climate science, in which this hybrid modelling is known as ”Medium AI”, there is now a stronger emphasis that ML models (especially deep neural networks) ”should not only have a high predictive performance (e.g., high accuracy) but further provide accessible predictive reasoning consistent with existing theory.”(Bommer et al. 2023) Yet, we might interpret such statements as expressions of dissatisfaction with algorithmic transparency or post-hoc explainability alone; And rightly so, because explainability methods cannot sufficiently guarantee that the model is *explanatory about a specific phenomenon.*

Often times we hear that technological advancements like ML will help us understand nature and the environment around us. But ”understanding”, especially on a process level, cannot readily derived from any application of ML, which is a mainly data-driven technology.

There is an interesting tension between prediction and understanding: On the one hand, existing theories were for the most part developed through empirical research and were highly informed by mechanistic modelling efforts. Understanding *how* a phenomenon arises, has been - and still is the purpose in many areas of science, for example physics, chemistry, biology, or sociology, to name a few. Although there is a varying degrees of detail, all these disciplines have benefited from this modelling strategy. because mechanistic descriptions aim at characterising the relevant causal dependencies among parts of a system, facilitating understanding on a process-level, and hence, control. On the other hand, the adoption of ML technology across many scientific disciplines fortified the statistical modelling paradigm, which has become very strong amongst scientists who seek to predict the dynamical behaviour of systems, instead of understanding *why* certain phenomena arise within them. ”Predicitivism” does plausibly characterise such modelling commitments.”[2] It becomes apparent in the rising prominence of ’complex systems science’, which is perhaps best described as a set of methodologies that are increasingly applied across many scientific disciplines. ML technology, following its spread across the

[1] There is often drawn a distinction between ML-model-interpretability, which is more directed towards the internal functionality of the algorithm, and ML-explainability, which is concerned with understanding a decision made by the algorithm and usually requires less detail.

[2] As mentioned by Kaplan and Craver 2011, predictivists hold that a sufficiently accurate prediction of a target system qualifies as ”understanding”, which does not require a representation of the underlying causal structure.

sciences, has secured a safe place within this methodological toolbox; On the backdrop of an enormous increase in predictive power, which these tools provide, it is understandable that every strenuous search for procedural knowledge not only slows down scientific ”progress”, but also is of lower priority compared to a faithfully ML prediction of any system evolution, regardless of its epistemic opacity.

So why should we try to understand the mechanisms responsible for complex macroscopic phenomena when we can use ML algorithms to directly predict them instead? It is true that, considering the study of our climate system, we are interested in *understanding* which role we as humans play in our ecosystem, to then extrapolate risk-factors and mediate our possible damaging influence. Put in the words of the ecologist Simon Levin: *Understanding patterns in terms of the processes that produce them [...] is the key to the development of principles for management. Without an understanding of mechanisms, one must evaluate each new stress on each new system de novo.*(Levin 1992)[3]

I am interested in the science of complex systems - in particular because the diverse modelling methods presents to us in its purest form the tension between prediction and understanding. Given that in climate science we already see the emergence of hybrid models that make use of ML together with physical constraints, begs the question if such a recipe could be of benefit for complex systems science more generally. Understanding the micro-dynamics that give rise to certain macroscopic phenomena could in consequence enable the level of control we are interested in, especially in scientific domains where direct experiments are almost impossible. *Agent-based modelling* for instance, as it is applied in complex systems science, seeks to gain explanatory insight into the collective behaviour of agents obeying simple rules. In other words, these models are supposed to probe deeper into the mechanism of a given phenomenon. Interestingly, ML is not unfamiliar with treating agent-based systems from a data perspective, leading to the development of interesting methods that allow for physical transparent interpretations.(S. Ha 2021) Having motivated my commitment to a genuine symbiotic approach that respects both, mechanistic and ML modelling efforts, I would now like to give a brief overview of how I am going to proceed from here on.

The path of this project - aims and methods

The broadest framework of my Thesis is provided by an underlying idea that it is possible, and even useful, to reconcile the statistical data-driven modelling paradigm with the mechanistic one. In the context of molecular biology, it has already been argued for the potential of such hybrid modelling.(Baker et al. 2018) Yet, I wish to see how it can actually play out at the intersection of Machine-Learning and complex systems, in particular emergent collective phenomena within such systems. What will be of particular importance, is a clear focus on how the ML model framework is ”linked” to the (complex) target system, and how this link bares upon the level of understanding that can result. My central question for this Thesis takes the following form:

> **How can we link a ML algorithm to a complex target system, such that the resulting model provides mechanistic insight into emergent phenomena within that system?**

In the first part, I will motivate my chosen setting of complexity, which is characterised as systems that exhibit emergent phenomenology. I wish to clarify what it means to aim at ”mechanistic understanding” of those systems, which hinges on the notion of what constitutes a *mechanism*; For a conceptual grounding of ”emergence” and ”mechanism”, I am relying on literature within the philosophical branch called ”New Mechanism”. Then I will go on to lay out the analytical idea that I used as a starting point to approach the central question above: The concept of ”link-uncertainty”, originally introduced by Emily Sullivan.(Sullivan 2022) It implies the notion that there is a link connecting the ML model to a target system of interest. How much understanding a ML model can provide, according to Sullivan, is not obstructed by

[3] I interpret the term ”processes” to be synonymous to ”mechanisms”, and I believe it is justified to use them interchangeably for this thesis topic.

the ”black-boxing” of the algorithmic implementation. Instead, understanding is affected by the extent of empirical evidence supporting the claims that any ML model makes about the target phenomenon - or put differently, the link between a ML model and the target phenomenon. Sullivan explains that mechanisms are part of the background knowledge, in which the ML has to operate in order to have low ”link-uncertainty”.

An analysis of her chosen ML examples reveals that they are all instances of statistical correlation extraction, in virtue of predictivist modelling, and the emerging picture will provide reasons why it does not qualify as ”mechanistic”, in the sense of New Mechanism. In the remainder of my thesis, my task is therefore to qualify a stronger link between the ML model and the target system, which can satisfy the mechanistic level of understanding within the chosen complexity setting. It requires me to dive deeper into some technical aspects of Machine-learning like deep neural networks (DNN), in search for a philosophically viable formulation of mechanistic understanding that applies to the ML-aided study of emergent phenomena.

In part two, my focus shifts towards benefits of a symbiosis between the statistical and the mechanistic modelling paradigm, which is already apparent in ”physics-aware” ML applications for the study of our climate system (also called *Medium AI*)[4] A kind of interpretability will gain prominence, which lets us spectate the ML predictions through a physical lens. One realisation that will be of great importance for the formulation of a ”mechanistically effective” link is that Graph Neural Networks (GNNs) can provide a first mechanistic sketch, because they characterise a system’s components and their structural relations. Yet, as I aim to demonstrate, applications of physics-aware GNNs confront us with an epistemic limit regarding the mechanistic understanding of dynamical interaction rules.

Methodologically, my research takes into account literature from different scientific disciplines: First, case studies in the area of Machine-learning for climate modelling, showing the hybrid-modelling commitments of ”Medium-AI”. Second, biological research about the collective phenomenon of bird-flocking that are necessary to anchor what ”empirical evidence” means in the context of ML. Third, I will consider philosophical literature on complex systems and mechanistic modelling, for which I build on work from philosophy of ”New Mechanism” - in particular, because I aim to derive key components of my extended link-framework from insights that have been expressed in detail by Stuart Glennan, a prominent New Mechanist.(Glennan 2017, Glennan, Illari, and Weber 2021) These different branches of literature help situate my arguments in a broader context, while endowing the conceptual steps taken with meaning. Next to Emily Sullivan’s work (Sullivan 2022), another case study will be of central relevance, in which the authors leverage a deep learning algorithm (named *AgentNet*) to ”unravel hidden interactions in complex systems”.(S. Ha 2021)

In part three I will provide an in-depth analysis of the methodology behind AgentNet. Specific attention is payed to the establishment of physical interpretability, which the authors of this study motivate through *correspondence* between the ML model and an abstract mathematical formulation of agent-based complex systems. I wish to highlight how their ML model realises an implicit commitment to the mechanistic modelling paradigm; Although originally introduced as a ”model free” framework, it exemplifies, as I will claim, some of the relevant parts for gaining ”mechanistic” understanding of emergent phenomenology. The further path of this section will follow two different test systems, to which AgentNet has been applied. Subsequently moving from simpler to more sophisticated systems, the point is to deliver justification for correspondence at the functional level of the ML evolution operator, making the whole ML modelling framework fit for posing a philosophically sound mechanistic hypothesis. This, as I argue, is achieved with a combination of two components: First, *prediction* of the macroscopic time evolution of the phenomenon, and second, *Relevance Realisation* through targeted visualisation of ML variables.[5]

[4] The concept of ”symbiosis” stems from biology and refers to situations in which distinct organisms mutually benefit from one another. Here, I am applying it in a metaphorical sense, referring to two different modelling paradigms.

[5] The term ”Relevance realisation” has been originally employed in the philosophy of cognitive sciences. In particular, John Vervaeke argued for a refined concept of how human intelligence manages to zero in on the *relevant* information in the context of a specific task.(Vervaeke, Lillicrap, and Richards 2012) In part III, I will provide More information about why I chose this particular term.

In the final part of my thesis, I will illustrate and summarise my previous analysis by putting everything together in one conceptual frame, which I call ”Correspondence Principle for Mechanistic Interpretability”, or short **CPMint**. I aim to make explicit how the ”extended link” for mechanistic understanding improves upon the initial link-concept as employed by Emily Sullivan, moving through the development of the key illustrations that are supposed to run in parallel to the analytical parts. Whenever ML technology is leveraged to *understand* complex emergent phenomenology in terms of the micro-processes producing it, questions of system-decomposition, constitution and causation need to be addressed. **CPMint** is an attempt at formalising *how* these questions can be addressed. The idea of correspondence down to the functional level of a physical system, as developed on the basis of New Mechanism and exemplified through AgentNet, is the essential piece that establishes an episitemically effective ”symbiotic modelling framework”. Afterwards, I will discuss some limitations of this framework, which again requires to look back at the ways in which ML technology is applied in practice, especially regarding the detail of mechanistic knowledge. I try to exemplify these limitations in part through AgentNet, while motivating possible strategies for alleviating some of these limitations. Within the concluding section of my thesis, I will argue that **CPMint** is presumably of relevance beyond the grounds of agent-based complex systems, for which it has been developed. I wish to justify this claim by touching on some recent modelling efforts in the field of cognitive science, whose scope is to mimic the functional properties of neuron in the human brain with ML technology. This might open up a pathway towards ”biologically plausible ML” in the sense of enabling ontological (and maybe even functional) correspondence to biological organisms.

Part I: Philosophical basis - Complexity, Machine-learning and link-uncertainty

1.1 Understanding Emergence through the lens of New Mechanism

Much like the concept of ”time”, we seem to be very familiar with *complexity*, unless our task is to properly define it. Approached from a colloquial point of view, we often refer to something as ”complex” when there is (subjective) opaqueness that shuts us off from understanding how a certain phenomenon arises. In this sense, it just means that a thorough investigation into the (presumably) many interacting parts that caused it might be complicated.

The study of various intertwined and often non-linear dynamics in natural systems does surprisingly well fit with this colloquial interpretation, because it can be seen as a very complicated task in itself. There are huge uncertainties associated with the predictions of such systems, resulting from deficiencies of our modelling tools and possible inadequacies in our knowledge about the underlying physical processes. This mismatch between the actual physical dynamics and our imperfect descriptions of the same is captured within the epistemic meaning of *complexity.* But there is another meaning, which treats ”complexity” as an (intrinsic) *system property* and which is presumably responsible for huge difficulties to extract *hidden micro-dynamics from the observed data of an unknown complex system.*[6] This notion is thus directed to the ontology of the system. Aside from debates on whether to locate complexity at the epistemic or ontological level, views are diverging among scientists when prompted with the task to define complexity or complex systems.(Ladyman and Wiesner 2020) Numerous attempts have been made to characterise such systems as exhibiting some similar features across many domains of research, but none has been successful so far in defining a coherent cross-disciplinary set of basic characteristics, which is sparse *and* exhaustive. There is an empty spot where a suitable definition should rest, and I take this as an opportunity to settle on a concept that sufficiently captures at least one of the main features associated with complexity. After all, I need to define a starting point; without it, the quest for mechanistic understanding would be meaningless.[7]

Ladyman and Wiesner give us a short list of ”truisms in complexity science”, which consists of nine statements that are all concerned with characterising ”complexity”. The first one is identical to the title of a highly influential paper by Philip Anderson, named *More is different*, and it reflects ”one of the most fundamental ideas in complexity science.”[8] Expressed in the words of Anderson himself, ”the constructionist hypothesis breaks down when confronted with the difficulties of scale and complexity[...]” such that at any scale, ”entirely new properties appear.” These properties would require research that is ”as fundamental in its nature as any other.”(Anderson 1972) When trying to understand complex systems, we usually think of a very high number of interacting parts.[9] As a result, the phenomena of interest that these systems produce are necessarily studied on a different level of organisation. The hierarchy of sciences reflects that need for ontological reduction; Considering for example biology or sociology, which make extensive use of ”mechanisms”, rather than physical ”laws”. Thus we can see that Anderson’s formulation does sit well within the broader setting of this paper.

It is worth pointing out that ”more is different” does by no means exclude the possibility of epistemological reductionism; Not least because of an increasing development of more powerful computational tools that can help to simulate and thus *understand* macroscopic phenomena in terms of microscopic interaction rules. There is an increasing awareness that the emergence of some macroscopic phenomena is in fact dominated by only few organising principles, intelligible

[6] See for Example S. Ha 2021. In their introduction and characterisation of complex systems, the authors talk about some basic properties, but also refer to this mysterious ”intrinsic complexity”, which is not further illustrated.

[7] It is worth keeping in mind that any well-motivated characterisation of complexity given here will still be insufficient to properly capture the whole breadth of complex systems phenomenology.

[8] ibid., p.3

[9] In many cases, these are assumed to be identical in their nature. Yet, often times there are distinct features and rules governing their individual behaviour, complicating the process of modelling their collective behaviour even further.

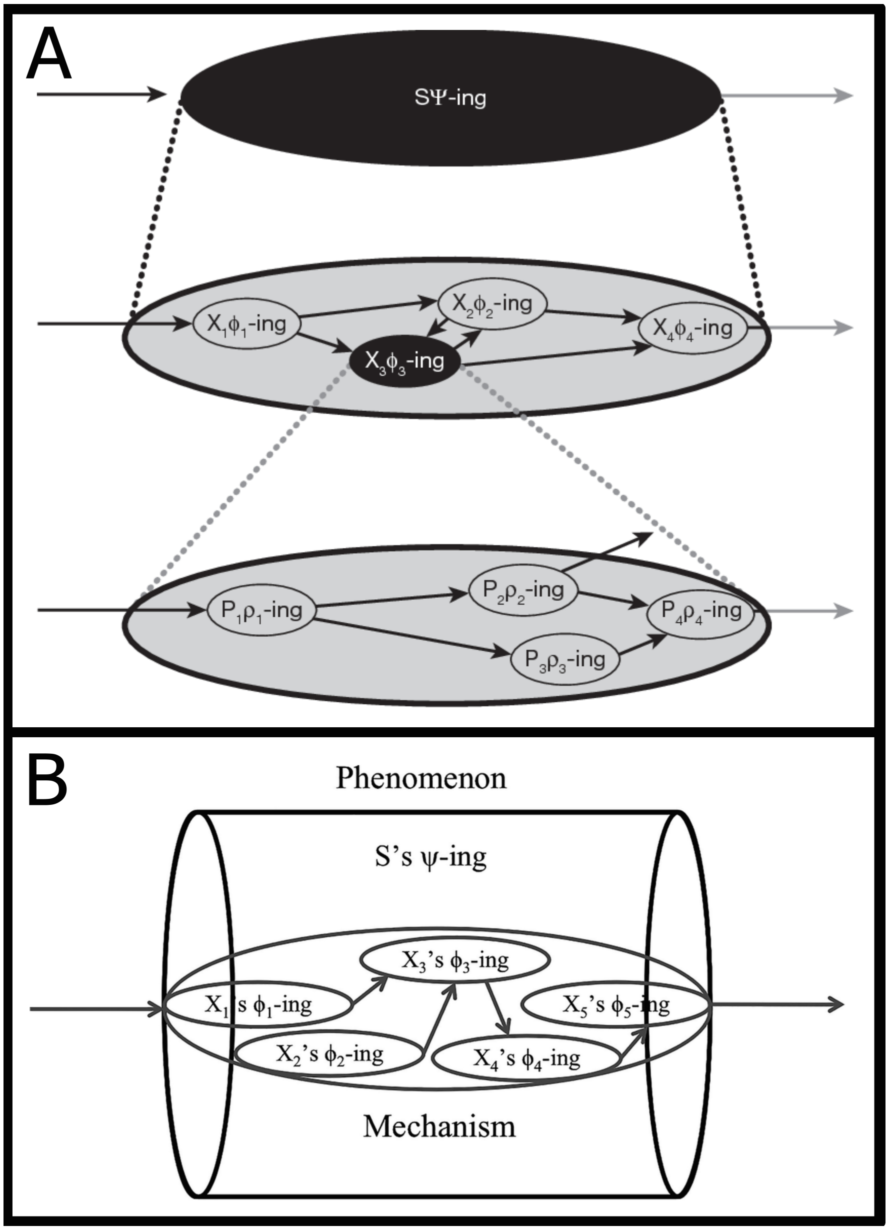


Figure 1: **A** Does show a ”Craver-diagram”, as it was reproduced in (Glennan 2017). It illustrates the hierarchical constitution of mechanisms through the vertical levels, while indicating causal, spatio-temporal dimension with horizontal arrows; **B** refers to the same relation between a mechanism and a phenomenon in constitutive mechanistic explanations, according to the ”behaving entity view”. (Krickel 2023). In both panels, the system S is composed of lower-level entities, which, by acting and interacting, influence each other and produce a macroscopic phenomenon. This phenomenon is observed as the ”activity” of system S, shown as ψ in both panels, despite slight differences in notation.

through simpler models. It is thus plausible that, rather than getting every microscopic detail of the system right, we can go about those crucial organising principles and still end up with mechanistic understanding. However, we first need to define what ”mechanism” means in the context of complex phenomena. Grounded within the philosophical movement called ”New Mechanism”, it is thought to be most fruitful to consider a mechanism for a phenomenon as consisting of ”entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon.” Put differently, ”mechanisms are identified and individuated by what they do (the ”phenomenon”), what they are made of (“components”/“entities”/”parts”), what their parts do (“activities”/“interactions”/“operations”), and how they are structured (“organization”).” (Glennan, Illari, and Weber 2021)

There is value in distinguishing between two dimensions of mechanisms - a horizontal and a vertical constitutive one. A point of consensus among New Mechanists is that ”according to minimal mechanism, all mechanisms are mechanisms for some phenomenon”. The relationship between a mechanism and a phenomenon brings into contact both of these dimensions - the causal dimension, ”where the output of a mechanism is caused by its inputs”, and the constitutive dimension, relating the mechanism as a whole to the activities of its parts.[10] It is the vertical dimension that is of key relevance for our colloquial understanding of *emergence*, and I decided to capitalise on phenomena that are in virtue of this definition ”emergent”, because they span more than one level of hierarchical organisation. Figure 1 illustrates two different versions of these dimensions, while both panels (**A** and **B**) endorse the principle that a system S, which performs a certain action (abstractly termed ψ-ing), is made up of individual component entities (X_i) that act in their own way (ϕ_i-ing) and are causally connected with each other through interaction (arrows). This relation can go down more than just one level, depending on the detail of mechanistic resolution. Yet, there is a slight difference between those two characterisations. According to the ”behaving entity view” (panel **B**), the system S *contains the mechanism* (thus the big tube around the lower-level processes), while the system S in the Craver-diagram (panel **A**) is made up by the mechanism, isolating it from the environment. Both views will be important throughout this thesis, because mechanisms are usually only one part of a macroscopic system-behaviour, while there are many other components of the system that do not take part in the mechanism. The behaving entity view is thus the more realistic picture. Yet, as it is the desired working definition of ”mechanistic understanding”, we are interested in the *relevant* and mechanistically effective parts of the system, which are central drivers for the emergence of a collective phenomenon, allowing us to drop all the non-relevant parts (that would be contained in the large cylinder of panel **B**); The Craver-diagram does best capture this situation.

It makes evident, why the conceptual grounds given by New Mechanism does provide a viable working definition of mechanistic understanding in the context of emergent phenomenology, which I am centring my Thesis around. It should be noted, though, that there is usually a problem of demarcation - where does a mechanism start, and where does it end? Because I will employ the term ”target system” very often throughout this thesis, I need to put emphasis on the importance of Spatio-temporal, as well as constitutive boundaries of mechanisms. (A more detailed account of the demarcation problem will follow in section 2.2 below.)

[10] See (Glennan 2017, p.22-23).

1.2 Emily Sullivan's Link-Uncertainty concept

Having the philosophical introduction in place, let me now turn to ML technology. Operating on the basis of integrating simpler computational units into higher-level systems, we have managed to endow our modelling tools with the ability to ”learn” an internal representation from data, as for example Machine-learning algorithms do. But they are themselves hierarchically organised systems of activation, reflecting both dimensions (horizontal and vertical) of a mechanism mentioned before. The algorithmic implementation on the ”micro-level” and the resulting emergent functional approximation power makes the phrase *more is different* apply to them as well. Hence, when using ML to model natural complex systems, complications can arise, especially if the goal is to *understand* these systems. For this reason, there is continuing effort in making them computational intelligible (for example with methods of ”Explainable AI”) or even partly interpretable. In the following, I would like to introduce some philosophical work that deals with the question how we could gain understanding from ML models.

Emily Sullivan proposed the concept of *link-uncertainty* in the context of models that employ Machine-learning (ML) technology. According to her, it is not the black-box nature, nor the complexity of those models that is associated with the biggest hindrance to understanding. Instead, ”it is a lack of scientific and empirical evidence supporting the link that connects a model to the target phenomenon that primarily prohibits understanding.”(Sullivan 2022) ML algorithms are especially powerful in dealing with big datasets and do well in pattern recognition, up to the extent that they are providing us with higher accuracy in prediction. Yet, only when the ML model is operating within the context of existing scientific knowledge, or, as put by Sullivan, ”is informed by empirical evidence”, it can provide us with understanding of the phenomenon. However, Sullivan does not qualify the nature of this *link* directly; Further, regarding the relationship between explaining and understanding, she chose to stay on a more superficial level as well, where it is solely important if, given a lack of information, ”it is still possible to gain insight about a phenomenon.”[11] Thus, the setting is left broad, as she omits any specification of the desired *kind* of understanding. After introducing this concept on the basis of Schelling's segregation model[12], Sullivan gives three different ML examples, each of them characterised by a distinct degree of link-uncertainty, demonstrating how the concept plays out in practice. All three case studies are situated in the field of medical practice or in a social science context. Social phenomena would definitely qualify as very complex in their nature, since there are many layers of organisation that can lead to the emergent behaviour.[13] In the following, I would like to give an account of her chosen case studies with increased attention on the epistemic role of the ML algorithm; Furthermore, I want to analyse the way Sullivan indirectly motivates the ”link” connecting ML model and target phenomenon, and which level of understanding follows as a consequence.

The first example Sullivan starts off with is a deep-learning model that was trained on electronic medical records of over 700,000 patients. The main purpose was to build a generalisable abstract patient representation for ”improved clinical predictions”, that could ”provide a machine learning framework for augmenting clinical decision systems.”(Miotto et al. 2016) Interestingly, certain medical problems were kept out of their set of predictions, for example HIV, since the behaviour of patients has a large effect on the development of this disease. Sullivan further qualifies the claim that the model is ”informed by existing empirical evidence concerning diseases”, which then serves ”to explain and enable understanding for patients about their risk factors.”

At this stage, it becomes very clear that *link-uncertainty* is nothing intrinsic to the ML model itself, but can only exist in a composite system of ML model and established background knowledge about the target phenomenon. Significant for the analysis, however, is that ”prior

[11] See Sullivan 2022; In her work, Sullivan starts off with the notion that explanations aims at understanding, while the term ”explaining” is applied in a broader sense, in contrast to its meaning within the context of ML.

[12] See Schelling 1971

[13] One might argue that purely data-driven models appear methodologically more adequate, since deep-learning algorithms that potentially neglect many micro-details and only look for regularities or faint data correlations are able to predict a macroscopic phenomenon more effectively. With a lack of detail, some of these models are also much easier to handle. However, gaining a deeper understanding of the macro-phenomenon seems more difficult, which is one of the aspects I would like to address within this thesis.

knowledge about which records are salient for medical diagnosis helped lead to the success of the model."(Sullivan 2022) This might seem trivial at first glance, but it shows how background knowledge plays an essential part in the process of model-creation;[14] It means that the establishment of such a "link" has a direction. In the deep patient case, the model has been tailored to the existing background knowledge. Yet, results showed that it had trouble predicting for example "diabetes mellitus without complications". One reason is that it does not seek to clarify the underlying biological mechanisms *causing* these medical conditions to correlate with certain preconditions; In the words of Sullivan, it could only track "proxies of disease development, such as previous physicians' decisions to carry out a diagnostic test."[15] It requires still a bit more work to disentangle the various processes promoting understanding, which are all mixed into the "link" which Sullivan refers to.

A second example is given with a ML model that is trained to detect cases of Melanoma by classifying images of skin-lesions.(Esteva et al. 2017) It works through finding correlations in graphical information of healthy and diseased skin to then classify out-of-sample data upon the internal learned representation. Sullivan qualifies this model as having reduced link-uncertainty, because there is extensive scientific evidence that can link appearance of moles to stages of melanoma. Hence, as Sullivan explains, it does not matter that we are dealing with a computational black-box, because "the model is operating within a background of existing scientific understanding."[16] Again, it is not clear what kind of understanding is meant here, while it can almost certainly be said that *mechanistic* understanding of the biological processes leading to skin cancer is one part of this background knowledge. It further justifies why the melanoma model can guide clinicians to explain certain medical interventions to patients, even without an additional method of explainability accompanying the ML model. However, there is a limit to understanding, determined by the type and diversity of the data that was used to train the model. For example, as Sullivan explains, the Melanoma model was mainly trained on white skin colour, leading to its failure when faced with images of darker skin colours. In other words, the model fails to extrapolate to data, for which the training data set has not been representative enough. Consequently, this increases the amount of "link-uncertainty", and therefore puts a limit on the level of understanding possible - an understanding that is much more constrained than mechanistic understanding, which is part of the background knowledge. Consider, for example, that mechanistic characterisations would likely apply to a much bigger group of people, and not just to those with brighter skin tones.

As a last example that shows very high degree of link-uncertainty, Sullivan introduces a model that employs facial recognition to predict sexual orientation of people.(Wang and Kosinski 2018) The story here is plain an simple, because so far, there is no solid empirical evidence supporting a causal connection between facial features and the sexual orientation of a person. Consequently, it severely limits any notion of understanding, because it removes the necessary condition for the establishment of a link to empirical knowledge. All the model can do (and which it does do with surprising accuracy) is find correlations within the type of datasets on which it was trained. It should not be taken to mean more than that.

With this elaboration on all given ML examples, I would like to summarise some essential components within Sullivan's concept of "link-uncertainty". What she introduces as a "link" connecting ML model and target phenomenon, can be qualified as a set of assumptions about which features are causally responsible for a specific phenomenon. This is visualised with the blue arrow above **B** in Fig. 2. Furthermore, it is implied the representational content of the data used to train the ML model. The amount of empirical evidence supporting those causal connections (denoted by **A**) forms the "background knowledge", in which the ML model has to operate to enable any kind of understanding at all.[17] This background knowledge can vary in its extent; Consequently, the level of link-uncertainty differs as well. Although there is no strict measure to which we could attribute a numerical value, it is possible to get a grasp of

[14] Here, I am referring again to the fact that the authors of the deep-patient model did not seek to predict HIV in order to make it work.

[15] Sullivan here refers to the reasoning of the authors, explaining that the frequency tests made at routine checkups was not valid discriminant factor.

[16] Sullivan 2022, p.24

[17] In Sullivan's discussion, it appears as if she uses "empirical evidence" interchangeably with "background knowledge", while they mean slightly different things in the context of analysis.

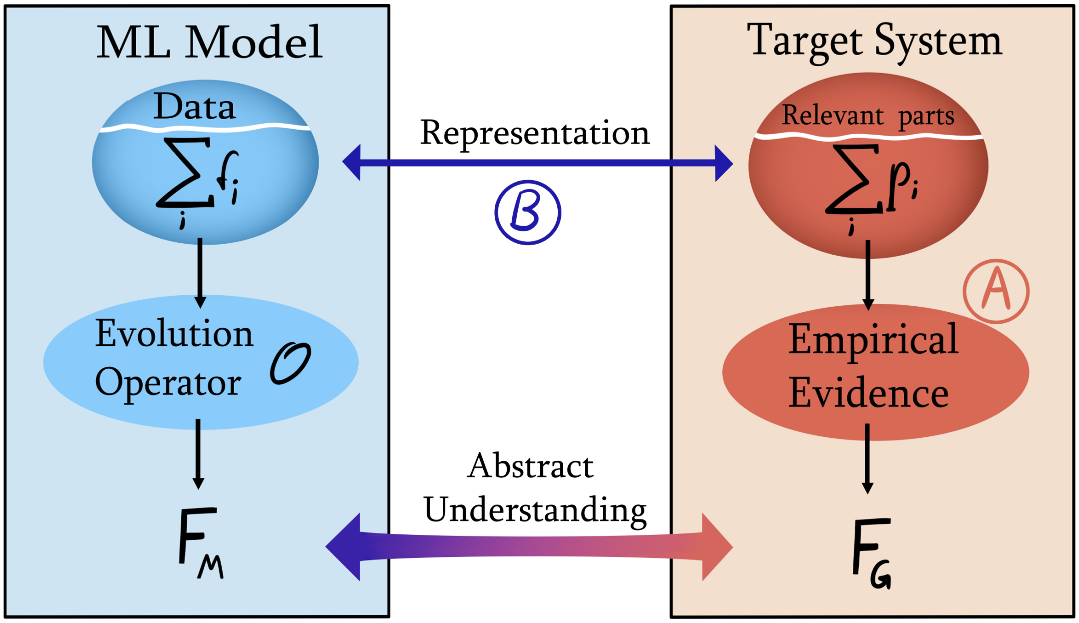


Figure 2: Visualised here, the core elements within Emily Sullivan's "link-uncertainty" concept. The ML model is depicted on the left, while the target system is shown on the right. Both components, a representational one (**B**) and the enabling empirical evidence (**A**), are illustrated, as well as the evolution operator $\mathcal{O}$, which upon training leads to the ML output, shown as F_M. The resulting *kind* of understanding is confined to the phenomenological level of F_G, and highly constrained by the already existing background knowledge (for example mechanistic representations). The graphic has been created by myself.

the two core elements that determine the strength of the link between ML model and target phenomenon:

- The primary one is the amount of empirical evidence that some parts of a system (p_i) are causally connected to the target Phenomenon.(**A**).
- The secondary element is a representative connection, shown as **B**. First, it depends on the extend to which the ML features (f_i) actually represent those aspects of the target system that are known to be phenomenologically relevant. Second, representation is also limited through data that is used in the training phase.

In Fig. 2, these two components are shown. On the left is the ML model, operating with defined features f_i. It is trained on data that relates to those features, which are usually numerical values, but can also be strings or higher-order units like graphs. The final functional relationship of the ML algorithm is denoted as "evolution operator" $\mathcal{O}$, and with it the ML model is able to describe the development of a phenomenon in time and/or space.[18] For the concept of "link-uncertainty", it is irrelevant whether the evolution operator is black-boxed or not. The resulting ML output in the form of a prediction or classification is denoted as F_M. Visible on the right is the theoretical description of the target system, in which a number of causally relevant parts p_i are thought to produce the observable "ground truth"-phenomenon that is denoted as F_G. In between, shown with **A**, sits the empirical evidence, linking the individual parts of the system to the emergence of this phenomenon; It is the most essential part here, independent of the ML algorithm, and, according to Sullivan, "determined by the standards of the domain in question."(Sullivan 2022) The arrow above **B** visualises the second requirement that the ML

[18] This is facilitated through a training procedure, usually leading to an adjustment of the weights within the artificial neural network, such that its output optimally fits the training data samples.

features represent the parts that are causally relevant for the phenomenon.[19] The bottom arrow shows the resulting ”abstract understanding”. Both components **A** *and* **B** are necessary for this understanding, deliberately indicated by the choice of colour. Through Sullivan’s discussion of these case examples, it becomes clear that the resulting understanding is necessarily confined by the background knowledge.

Since my central question is how to enable mechanistic understanding of emergent phenomena through ML, the choice of Emily Sullivan’s work seems odd. After all, none of her case-studies uses a genuine complex-systems approach. Furthermore, she did not trade on any positive notion what constitutes ”understanding”. However, there are two compelling reasons why I nevertheless decided to include Sullivan’s work as a suitable starting point: First, in order to deal with the question above, it provides the framing of a link-relation between a ML model and a target system. In this sense, it establishes the central analytical ideas for this work. Second, building on this framing, Sullivan already points out two key components (illustrated in Fig. 2 **A** and **B**). Representativeness of data, as well as empirical evidence are both likely to stay relevant for a mechanistic link-connection. In the next section, I will give reasons why her framework is not sufficient if we want to explain a causal relationship that could allow a mechanistic understanding of complex systems. Put differently, the quest from here on is to try treading the distinguishing line between what is *real* and what is *useful*.

1.3 The missing link

In the last section, I have analysed the ML case studies, chosen by Emily Sullivan to demonstrate the merits of her ”link-uncertainty”. I arrived at an illustration that concludes the main components of her concept, piecing out the foundation of this ”link” that connects ML model and target system. However, there appear to be limits to understanding provided by this link, (or by low link-uncertainty respectively). As a first example, the authors of the deep patient model argued that clinicians should use an additional ”feature selection tool” that would help them understand what drove the ML model decisions.(Miotto et al. 2016) Why has this point been raised? Presumably, because different areas of application hold different standards of understanding. In a clinical setting, where doctors decide upon medical interventions, some causal understanding is required to effectively treat a patient’s disease. The broad and abstract level of understanding, as provided with Sullivan’s concept, cannot be sufficient in these settings.[20]

Let me make this point more clear. Each of Sullivan’s ML case examples stays more or less on a higher level ontology; None of them does seek to be explanatory about the underlying *mechanisms* producing the phenomenon. Hence, they are perfect paradigm cases of statistical modelling, in virtue of what Kaplan an Craver called ”predicitivism”. (Kaplan and Craver 2011) There is a very dangerous pitfall looming nearby: When sticking to the predictivist modelling paradigm, we are left with the unsettling realisation that ”Deep Neural Networks (DNNs) are ruthless correlation extractors unconstrained by the information that is actually consumed by downstream processes.” (Kieval 2022) Put simply, they cannot distinguish causation from correlation. To account for this pitfall, Emily Sullivan included the requirement of additional empirical background knowledge; Serving as a necessary condition for low ”link-uncertainty”, it forms the basis for any link to be established between a ML model and the target system. Consequently, the strength of this link is constrained by the existing empirical background knowledge. Mechanism representations can be part of this background, and with them the embedded knowledge about causal and constitutive dependencies among parts of the target system.

[19] I did not specify these features in more detail, since it is meant to show the abstract nature of this link. The most important aspect of arrow **B**, is the representational content. Imagine the case-scenario of a well established empirical background knowledge that some parts of a system are causally relevant for a given phenomenon. Even then, the ML model can still fail to incorporate a link to those parts in its feature-space - maybe due to insufficiency of data or computationally over-idealised implementations.

[20] At best, a ML model with low link-uncertainty can provide indication where future empirical investigation could lead to new discovery.

In an attempt to move beyond the qualified ”link” in Sullivan’s concept, I am interested in exactly the kind of knowledge we would call ”mechanistic”. Rather than just being a *precondition* for understanding, I will make mechanistic representations a desired *target* of understanding. Asking in the spirit of New Mechanism: It it possible that a ML algorithm could represent a mechanism driving a phenomenon? Arguably, this requires me to introduce a causal dimension into the link framework, without which the resulting level of understanding would remain superficial and abstract, incapable of distinguishing causation from correlation. Why do I think this is a fruitful topic to investigate? In my chosen setting of complex systems exhibiting collective emergent phenomena, we could easily imagine situations where there is no coherent empirical knowledge as to which parts (entities) and processes (interactions between them) are relevant for a given phenomenon. For this reason, there has been been progress on the front of predicting such phenomena on the macro-level; a field that is increasingly dominated by ML technology. But as it is common knowledge, this kind of statistical input-output mapping does little to illuminate the phenomenon on a micro-level. It distorts mechanistic transparency.

The ML algorithm’s representation of the target system (introduced as ”evolution operator), does play an important role for understanding its output in two ways: First, it is the target of ”explainable AI” methods that aim at making the ML computation process logically intelligible. And second, if constrained in certain ways, the evolution operator is sometimes made ”interpretable” in a specific context of application.[21] I am purposefully leaving aside all additional considerations that need to go into the ML training process. Although it is crucial for the algorithm to ”learn” a working representation of the target at all, there are huge sensitivities to different training processes.[22] It is thus highly unrealistic to pose, without any degree of algorithmic transparency, that the evolution operator learned by the ML model sufficiently captures the relevant causal connections between the input features; and because ”link-uncertainty” is acclaimed to work completely independent of algorithmic transparency, I see this as the main limitation of Emiliy Sullivan’s concept, with respect to mechanistic insight. But I suspect that it is possible to leverage ML technology for the purpose of understanding emergent phenomena in a mechanistic sense.

My aim from here on is to formulate the *missing link* that is able to epistemically connect the ”evolution operator” of the ML algorithm to the internal organising principles, which are relevant for a given phenomenon. I want to make my work relevant for the field of ML epistemology in a descriptive and a normative way. Descriptive in the sense of illuminating how ”physical interpretability” is established in practice; and normative, because I try to develop a conceptual basis which makes interpretability contribute to mechanistic understanding. Right from the outset, I filled my analytical toolbox with the terminology and insights from New Mechanist philosophy, for example how to decompose a mechanism into a horizontal, causal dimension and a vertical, constitutive one. At the same time, I would like to justify the value of applying this terminology for the crossroads of mechanistic understanding of complex systems with ML.

But before diving deeper into the technical part of how such a link could be established, there is an extra step in between, which concerns the different modelling-paradigms at play. Juxtaposing the statistical ML modelling paradigm with the mechanistic one will clarify their complementary strengths and weaknesses, which is why I want to focus on a symbiotic modelling approach, as it has been already argued for in the context of biology. This discussion will properly situate several case-studies, while taking momentum from already present (often tacit) mechanistic modelling commitments at the crossroads between ML and complex systems.

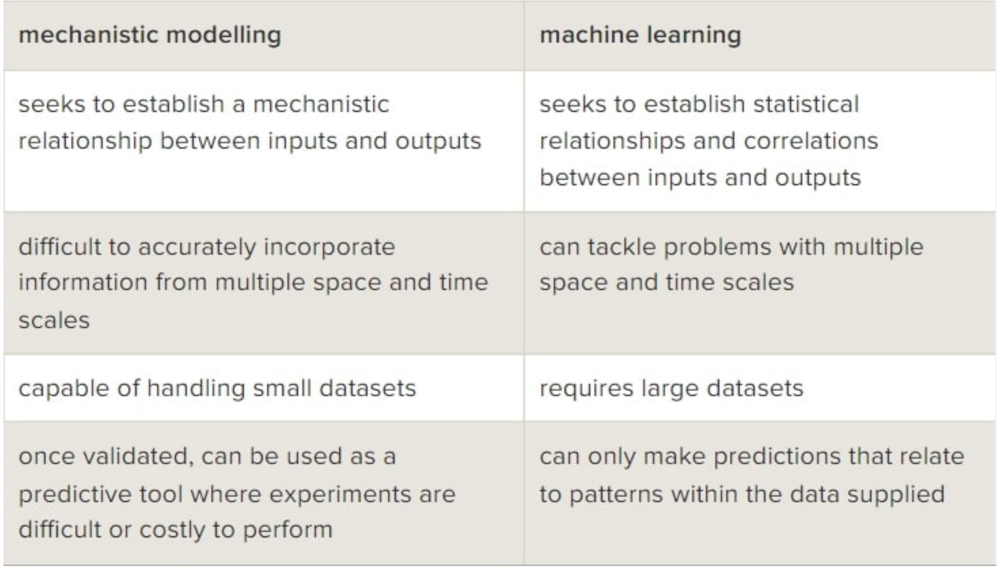
[21] In climate physics, it is often referred to as ”Medium AI”, which I will discuss in more detail in Part II.

[22] Those various methods have to be tested carefully, and since the ML hyperparameters need to be finetuned, often times through a lot of experimentation, ”scientific ML” is regarded as a highly empirical enterprise. Unfortunately, there is no coherent rule book for adjusting such parameters successfully in every situation. Furthermore, the fact there has been so much effort in bringing algorithmic transparency into the ML applications through ”Explainable AI”, does also point to our limited understanding of this learning process. In light of my chosen complexity setting, it is thus possible to view the predictive power of ML technology as an ”emergent phenomenon” in itself, which is dependent upon a number of (only partly well known) factors.

Part II: Ways towards physical interpretability

2.1 The symbiotic modelling framework

One of the greatest pitfalls of ML modelling has been touched on by Thomas Carlson and his colleges, who stress that ”precisely because it [ML] offers up simple patterns, it can be easy to read too much into data – to see phenomena that are not really there.”(Carlson et al. 2018) Thus, the application of ML as a discovery tool has continuously been subject of discussion among philosophers of science. Following the adoption of ML technology in many scientific disciplines, it is worth looking at the underlying modelling commitments, which I decided to demonstrate with a brief detour into the field of neuroscience and neuroimaging. In those areas, the problem is one of representation to understand what is going on in the brain. On the one hand, the model framework has to deal with messy data from neural activation patterns, while on the other hand, it is sought to represent the underlying causal principles responsible for these patterns. The latter condition requires continuous emphasis on *mechanistic* modelling. As argued by Carlson, there is a ”gap between our ability to measure information and our capacity to determine the underlying neural source.” Importantly, he later explains that (only) ”successfully locating neuronal sources serves as a foundation of mechanistic understanding.(Carlson et al. 2018) So we see that there is an interesting modelling-intersection between the statistical tradition and the mechanistic one. However, these two modelling approaches seem to be at odds, because, ”they rely on different types of data and provide access to different types of information; In short, they are two different paradigms.”(Baker et al. 2018) I chose to include the tabular in Fig. 3 that emphasises their complementary role. As argued by Ruth Baker and colleagues, there is reason to try embedding both modelling structures into a symbiotic framework. Although originally justified in the context of biological research, it parallels the broader movement towards interdisciplinary approaches that has led to many cross-applications of data-driven technology. Especially in complex systems science, we need to look beyond the borders of individual disciplines, which is why I will now try to give an account of how this symbiotic paradigm popped up in settings of physical complex systems.



mechanistic modelling	machine learning
seeks to establish a mechanistic relationship between inputs and outputs	seeks to establish statistical relationships and correlations between inputs and outputs
difficult to accurately incorporate information from multiple space and time scales	can tackle problems with multiple space and time scales
capable of handling small datasets	requires large datasets
once validated, can be used as a predictive tool where experiments are difficult or costly to perform	can only make predictions that relate to patterns within the data supplied

Figure 3: Complementary character of predictive modelling by means of ML and mechanistic modelling. Reproduced from Baker et al. 2018.

To prevent confusion, let me clarify the term ”symbiotic”. First, it is meant in a metaphorical way, in that there is a mutual relationship between two different modelling strategies, which normally aim at different epistemic values - statistical ML modelling, aiming at prediction of a system’s behaviour, and mechanistic modelling, aiming at ”understanding” in the sense of gaining causal knowledge of that system. Second, I seek to demonstrate this particular ”symbiosis” in the context of physical complex systems. Being physically constrained lets ML models benefit with higher predictive accuracy, while the physical descriptions benefit from the ability of ML tools to capture unresolved processes. Although *physical description* is not synonymous to *mechanistic description* of a system, physical models usually aim at a micro-level formulations, employing an analytical framework.[23] In other words, physical modelling in huge parts encapsulates mechanistic commitments.

In recent years, some ML applications in the context of climate science have already gone towards symbiotic modelling approaches, and it is sometimes referred to as *Medium AI*. One of the interesting aspects is that these models aim at representing physical knowledge inside the ML architecture. In reviewing some examples that employ different strategies, I hope to illuminate the common theme running through all of them.

The first example is a study that brings together a Graph Neural Network with a discrete representation of the diffusion equation to create a predictive climate model. Hwang and colleagues developed as a hybrid model, which they call ”Neural Diffusion Equation” (NDE).(Hwang et al. 2021) Their essential improvement upon other works is that the model learns the heat capacity of the diffusion equation, and thus is able to consider the uncertain nature of noisy real-word climate data. A representation of physical knowledge enters through the diffusion equation, which nevertheless has to be modified to fit the discrete dynamics on a graph. The overall aim they managed to achieve is higher predictive accuracy. Interpretability, in turn, is very limited, and therefore we cannot readily extrapolate to ”new” physical understanding beyond what the statistical model assumes, since heat diffusion is essentially treated as ”stochastic noise” within the model. This point becomes more clear when considering additional examples of ”physics aware deep learning”.(Zanna and Bolton 2021) Zanna and Bolton review some studies in the field of turbulent ocean mixing, which include constraints to ”ensure that the parameterizations remain faithful to the physics of the underlying process which we are trying to capture.” They touch on the issue of interpretability, ”since not all parameterizations included in ocean models are correct or encompass all the missing processes.” The NDE-framework arguably suffers form the same issue, as we do not know to which degree the ”noise” that the ML part deals with is resulting from misrepresented physical dynamics.

This is an example why we need to ask how well these approaches provide us with the methods to study complex systems mechanistically; Even though ”Medium AI” models do incorporate physical knowledge within their architecture and thus go some way towards Baker’s vision of a ”symbiotic modelling paradigm”, it is still the case that many of those models are deployed for the purpose of more accurate predictions. It is only plausible that they are in consequence unfit for extrapolating to new physical knowledge. An interesting observation, however, is that we see the role of ML in a symbiotic framework more clearly reflected. It is supposed to take on real-world uncertainty, which is not captured in the physical equations.[24] In other words, the ML part is supposed to ”correspond” to something that the physical expressions so far miss. At this point, one could argue that a full-blown application of explainable AI methods (also called XAI) could lead to increased interpretability, and thus to an understanding of the physics that is reflected in the ML algorithm. It might seem plausible at first, but I would like to briefly digressing along this line of reasoning, explaining why I rule out that possibility for mechanistic transparency. It is true that in some cases XAI can provide us with post-hoc prediction-plausibility. Yet, at the same time, it would complicate the process of model setup, since we have to take into account a vast number of different XAI methods for various ML models.(Bommer et al. 2023) Considering every possible combination with respect to their epistemic effectiveness would be unfeasible and almost impossible practically. Furthermore, it would completely divert our attention at how the ML algorithm itself works, which ”takes our focus away from the phenomenon it bears on.” Emily Sullivan points out that ”there is a distinction between understanding and explaining

[23] I am fully aware of the fact that many physical models employ statistical analysis within their methodological framework. But in contrast to deep neural networks, these statistical methods are arguably much better understood. In other words, we can reconstruct ”how” the statistics work, while this is not as easy in the context of many ML algorithms.

[24] They observe in their experiments that diffusion processes around large cities have non-trivial uncertainties that cannot be solely described by the diffusion equation.”(Hwang et al. 2021)

how the model works and using that model to understand a phenomenon of interest." With a reference to the deep-patient model she further explains that "learning more about the exact fine-grained weights of different data points does not improve our understanding of how it is *possible* schizophrenia can be predicted and correlated with features found in medical records." (Sullivan 2022) Yet, applications of XAI are more directed towards an understanding of how the model itself operates, which does little to enable understanding *about a phenomenon.* searching the vast zoo of XAI methods would thus be of limited use for my aim to formulate a conceptual basis of mechanistic understanding through ML.[25] If I decided to go down the path of XAI methodology further, I would run the risk of not seeing the forest for the trees, resulting in all sorts of contradictions and philosophical incoherent arguments.

In the next section, I will therefore motivate another way of securing a basis for interpretability, by appeal to Graph Neural Networks (GNNs). First, I aim to demonstrate how GNNs are realising the conceptual idea of interaction networks in the computational space of ML algorithms, clarifying how this offers the possibility of structurally decompose the desired target system. In order to do this, I decided to capitalise on other "Medium-AI" case-studies, which are methodologically different from the first one mentioned above. In a second step, I will then illuminate the advantages of GNNs through the philosophical lens of New Mechanism, arguing that they provide a first "mechanistic sketch". This step will be crucial along the way of formulating an extended link-concept for mechanistic understanding, when applying ML to complex emergent phenomenology.

2.2 Graph Neural Networks: A step towards mechanistically transparent ML

Back in 2016, there has been work on *interaction networks*, based on a deep learning architecture, which can deal with graph structured data. The first introductory part of a paper by P.W. Battaglia and colleagues bares high significance for my further discussion. Without paraphrasing, it says:

> *Many everyday problems, such as predicting what will happen next in physical environments or inferring underlying properties of complex scenes, are challenging because their elements can be composed in combinatorially many possible arrangements. People can nevertheless solve such problems by decomposing the scenario into distinct objects and relations, and reasoning about the consequences of their interactions and dynamics. Here we introduce the interaction network – a model that can perform an analogous form of reasoning about objects and relations in complex systems.*(Battaglia et al. 2016)

Importantly, the inference of underlying properties in "complex scenes" has already been a desired purpose, justifying an extra minute of attention. As explained by Battaglia, their model behaves like a physical simulation engine, "which generates sequences of states by repeatedly applying rules that approximate the effects of physical interactions and dynamics on objects over time."(Battaglia et al. 2016) Although Graph Neural Networks (GNNs) were already around at that time, their value for physically modelling complex systems was realised.[26] The central idea of GNNs is therefore already reflected within "interaction network". Starting from a graph consisting of objects (nodes) and relations (edges), it is able to conceptually separate the object-centric part of the model from the interaction-centric part. This setup makes interaction networks flexibly applicable and lets them "automatically generalize their learning across variable numbers of arbitrarily ordered objects and relations[...]."(Battaglia et al. 2016) The resulting broad applicability is a desired feature, especially in complex system settings. In addition to the structural information already present in the graph, GNNs are also able to incorporate *dynamical* update rules. A combination of structured modelling, which is more tailored

[25] This is of course simplifying the problem a little, because different modelling setups sometimes require a suitable method to make the working process internally "transparent". Later on, I will highlight the importance of "visualisation" in a broader sense.

[26] Broadly speaking, GNNs are ML algorithms that can process graph-structured data.

to the mechanistic paradigm, and deep learning that provides the tools to extract patterns in large datasets, makes GNNs effectively symbiotic.

A study by Sungyong Seo and Yan Liu, dealing with ”differentiable physics informed Graph Networks”, is a suitable example to demonstrate how GNNs manage to represent physical quantities inside a ML framework. Figure 4 shows essential elements within their framework and illustrates the fundamental computational update steps of the employed GNN.(Seo and Liu 2019) Equation 1 tells us how the edge e_{ij} between node i and j is updated based on its previous state, the individual values of both neighbouring nodes (v_i and v_j) and the global attributes $\mathbf{u}$. Equation 2 shows how the vertices are updated based the previous value of the vertex, on every edge value connected to the vertex, and the global attributes. 3 tells us how the global attributes are updated.

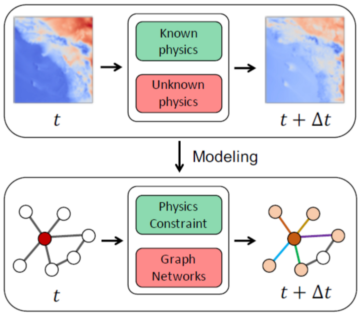


Figure 1. Concept of the proposed DPGN. The behaviors of sequential observations (Temperature) are governed by physics rules. Some of the physics rules are known and we inject them into a model explicitly. The remained unknown patterns will be extracted from data.

Given a set of nodes ($\boldsymbol{v}$), edges ($\boldsymbol{e}$), and global ($\boldsymbol{u}$) attributes, the steps of computation in a graph networks block are as follow:

1. $\boldsymbol{e}'_{ij} \leftarrow \phi^e(\boldsymbol{e}_{ij}, \boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{u})$ for all $\{i, j\} \in \mathcal{E}$ pairs.
2. $\boldsymbol{v}'_i \leftarrow \phi^v(\boldsymbol{v}_i, \bar{\boldsymbol{e}}'_i, \boldsymbol{u})$ for all $i \in \mathcal{V}$.
   $\bar{\boldsymbol{e}}'_i$ is an aggregated edge attribute related to the node i.
3. $\boldsymbol{u}' \leftarrow \phi^u(\boldsymbol{u}, \bar{\boldsymbol{e}}', \bar{\boldsymbol{v}}')$
   $\bar{\boldsymbol{e}}'$ and $\bar{\boldsymbol{v}}'$ are aggregated attributes of all edges and all nodes in a graph, respectively.

where ϕ^e, ϕ^v, ϕ^u are edge, node, and global update functions, respectively, and they can be implemented by learnable feed forward neural networks. Note that the computation order is flexible. The aggregators can be chosen freely once it is invariant to permutations of their inputs. Furthermore, a set of input for each mapping function can be also customized.

Figure 4: On the left hand side, an illustration of how the GNN works in a physics-informed environment to make predictions for the next time step $t + \Delta t$; On the right hand side, a formulation of the internal functional dependencies. The relevant symbols are explained in the subtext. The figure is a composite of parts and images in Seo and Liu 2019.

However, as mentioned in the same paper, ”it is hardly practical for modeling complicated real-world problems with the operators solely because it is only possible when all physics equations governing the observed phenomena are explicitly known.”(Seo and Liu 2019) Thus, we are clearly facing an epistemic limit regarding the extrapolation to unknown physical dynamics, which makes the term ”unknown physics” in Fig. 4 not optimally suited for the context. What essentially happens, is that through the GNN, possibly unknown patterns can be extracted from data; Yet, this does by no means imply that we gain a viable and physically interpretable representation of those patterns with regards to their origin. Consequently, there is still the need to represent non-resolved physics in some intelligible way to enable understanding, which the authors of this study do not cover. In the cases of uncertain ground-truth, as it is also apparent in the earlier mentioned NDE-study (Hwang et al. 2021), physical interpretability can still be minimal, although their GNN works with respect to physical knowledge. The translation of such knowledge into the latent space of the ML algorithm is one part of the ”link”, but without any kind of ”backwards-translation” into the space of physical knowledge, new data-representations are mechanistically opaque and seem very ineffective to gain new understanding. Undoubtedly, GNNs are of great importance for my main goal of qualifying the conditions for an effective link between ML model and target phenomenon, but for the reasons mentioned above, they can only provide one building block.

Because GNNs can incorporate information about the *structure* of the target system and its underlying parts within the ML architecture, this is the optimal place for a first refinement of the extended link-concept. Considering again Fig. 2, the most reasonable place to put this structural information would be at the level of representation, indicated with the arrow above **B**. Alongside the link between the ML features f_i and the causally relevant parts of the target system p_i, there is now also information about the organisation and interaction of those features. The very important point here becomes apparent through Fig. 1, showing how interacting parts (X_i) explicitly produce and underlie a given phenomenon (ϕ) within a system S. Borrowing a term used by Kakplan and Craver, we are essentially dealing with a ”mechanism sketch” of the target system.(Kaplan and Craver 2011) In the context of neuroscience, Kaplan and Craver refer to the fact that models of mechanisms can be more or less complete, where the less complete ones ”might be more of a *mechanism sketch*, identifying one or two significant internal variables from which the vast majority of the variance in the phenomenon can be accounted[...].”(Kaplan and Craver 2011) Concerning my chosen complexity setting of emergent phenomena, knowledge about the significant internal variables can also be limited. However, GNNs constrain the ML architecture itself to the ”significant internal structure” of the target system (that is, the composite entities and their organisation), which provides a stronger link in terms of representation between the ML features and the relevant parts of this system. The exact kinds of interactions are not yet fully transparent, which is apparent in that they are modelled by straight lines/edges connecting nodes.[27]

The next step is then to ask how a given target system is meaningfully decomposed into parts and their interactions, ”nodes” and ”edges” respectively. It comes down to a mechanism boundary-problem; Graphically speaking, how big should we draw the circle around the X_i's in Fig. 1 to confine the horizontal, causal dimension? And how many vertical levels should be included downwards? For my overall research aim to formulate an extended link for the mechanistic study of emergent phenomena with ML, this problem of boundary-setting is a secondary one, since I am mainly interested in the broader outlook of such a link. Nevertheless, the particularities have to be specified within the confines of an application to a certain test system, which is worth acknowledging for the further path of my project. In New Mechanist literature, this is usually termed the ”phenomenon-dependence of decomposition”. Coming back to what has been mentioned in the introduction, mechanisms are always mechanisms for some phenomenon; ”but depending upon the phenomenon one seeks to explain, one will get different decompositions of a system into parts.”(Glennan 2017) Not all parts of a system need to be working parts of the mechanism that is responsible for a given phenomenon. For a collective of birds (anticipating one of the target systems which will be of interest soon), behaving in a coherent manner when flocking, it does not matter which wingspan they have or if their feathers are coloured differently - even the constituents of individual birds might be irrelevant, if one tries to understand what mechanism drives the emergence of such flocking phenomena. Getting at what exactly the relevant parts and their interactions are, is the key challenge for understanding an emergent phenomenon mechanistically, in virtue of New Mechanism.

With the flexibility of a GNN to provide the template for decomposing the target system based on available data, we arguably have the conceptual ground prepared that will allow us to fill in the details. But since GNNs by themselves do not pose any boundary-setting criteria for mechanisms, the term ”mechanism sketch” is perhaps best fitting. In this sense, they can provide mechanistic guidance and thus seem to fulfil a necessary condition for mechanistic understanding in complex systems.

2.3 Inspecting the Evolution Operator

In the context of a symbiotic ML-mechanistic modelling approach, there is another study worth mentioning, because it proceeds one step further in the direction of understanding through ”physical interpretability”. Miles Cranmer and colleagues introduced a framework, which makes it possible to extract physical relations from the data representations of various many-particle systems. Their strategy consists of first modelling these systems with a GNN, because it makes use of three internally separate and interpretable functions (an edge-model ϕ^e, node-model

[27] Although the context in which the concept of a ”mechanism sketch” was developed, does not match my chosen setting, I see value in adopting the terminology for the reason just explained. Furthermore, it has to be mentioned that this terminology can be traced back to two normative distinctions made by Carl Craver: ”sketchiness”, according to which models form a continuum from mechanism sketches to mechanism schema to ”complete mechanistic models”; And ”plausibility”, also referring to a continuum between ho-possibly to how-plausibly to how-actually models. Crucially, ”progress in building mechanistic explanations involves movement along both the possibly-plausibly-actually axis and along the sketch-schema-mechanism axis”(Craver 2007)

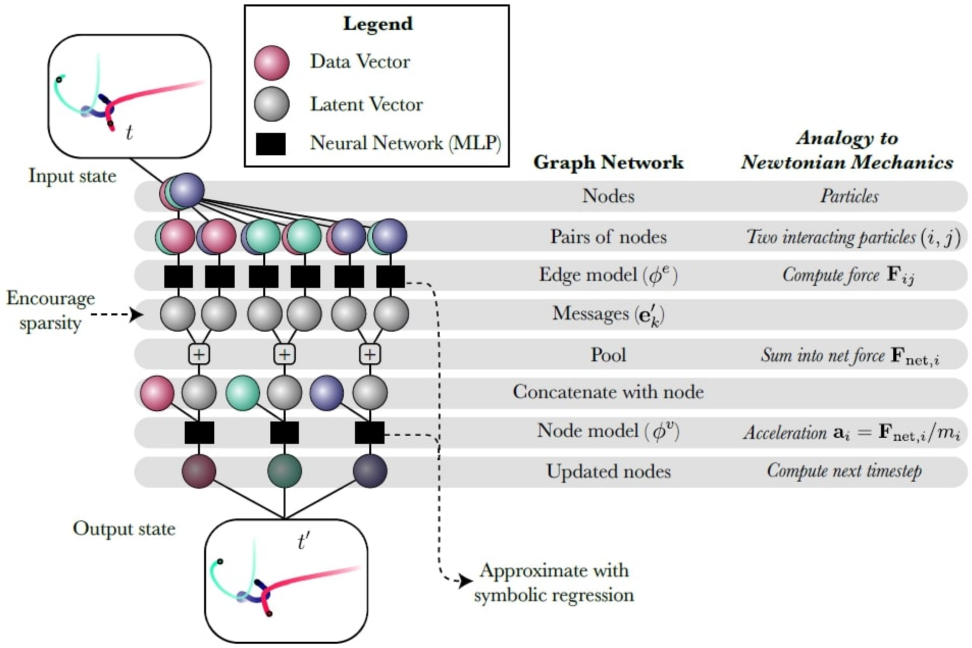

Legend
Data Vector
Latent Vector
Neural Network (MLP)
t
Input state
Graph Network
Analogy to Newtonian Mechanics
Nodes
Particles
Pairs of nodes
Two interacting particles (i, j)
Edge model (ϕ^e)
Compute force $\mathbf{F}_{ij}$
Encourage sparsity
Messages $(\mathbf{e}'_k)$
Pool
Sum into net force $\mathbf{F}_{\mathrm{net},i}$
Concatenate with node
Node model (ϕ^v)
Acceleration $\mathbf{a}_i = \mathbf{F}_{\mathrm{net},i}/m_i$
Updated nodes
Compute next timestep
Output state
t'
Approximate with symbolic regression

Figure 5: Here is a schematic demonstration of how the model setup works, which was chosen by Miles Cranmer and colleagues. On the left, a visualisation of the internal algorithmic GNN architecture with a compared Graph-terminology in the middle. On the right hand side, they decided to include a rough analogy with Newtonian mechanics, which, as they put it, ”is purely for explanatory purposes”. Figure is reproduced from Cranmer et al. 2020

ϕ^v and a global model ϕ^u).(Cranmer et al. 2020) The central element, however, is a separate symbolic regression algorithm, which then approximates the learned edge-model ϕ^e and the node-model ϕ^v by finding algebraic expressions for those latent representations. (Fig. 5). The ML model is first tested on various known force laws, proceeding with an application to a detailed dark mater simulation, about which exact physical relations are not yet formulated. What arguably lies underneath this methodological approach, is a realisation that they needed to ”back-translate” form the ML space where the learned latent representation lives, to the physically intelligible operator-space.[28] For the purpose of formulating an extended link, the general idea to make the *evolution operator* physically transparent, is a very fruitful one; With the use of this terminology, I am explicitly referring to Fig. 2, which was developed based on Emily Sullivan's link-concept and does not further acknowledge the role played by the evolution operator.

However, there are some points worth mentioning, because they bare philosophical relevance for my further discussion. First, the physical structure of the studied systems is constrained through the underlying GNN, while the additionally applied symbolic regression algorithm that provides an approximate fit with mathematical expressions, does mainly serve as a ”transparency tool”. Yet, it is unclear how these algebraic representations can be understood with (or derived from) existing physical principles, which means that the ”link” to existing physical knowledge is obscured. In other words, we are left with a high degree of epistemic opacity, but now camouflaged through mathematical intelligibility. The second point, arguably reducing opacity, is the fact that they included a ”rough equivalency between this architecture and physical frameworks”, which presumably helps ”to *interpret* learned formulas in terms of existing physics.”[29](Cranmer et al. 2020) It has to be stressed, though, that analogy can sometimes be

[28] By ”intelligible” I mean the fact that a practising researcher in the relevant field of study can understand the output with regards to the discipline's concepts and terminology. The evolution operator of the fully trained ML algorithm could be regarded as very unintelligible; Although based on mathematical rules, the pure amount of variables and weights is simply too high.

[29] See (Seo and Liu 2019) for a similar reference to physical analogy.

misleading; It might help to ”interpret” the internal dynamical update-rules in terms of physical principles, but it does by no means imply identity. Following along that lines, I see an important connection to the concept of *representation*, much emphasised in New Mechanist literature, especially on its role in the process building of mechanistic models. Cranmer and his colleagues, with their reference to physical analogy, implicitly pose a hypothesis that the model resembles the target in some respects.[30] But how exactly this representation bares on the resulting level of mechanistic understanding, is still not entirely clear. The analytical tool needed here is perhaps best illustrated through an account of mechanistic models by Stuart Glennan:

> *We may think of a mechanistic model as having two parts - a model of the phenomenon, and a model of the mechanism that is responsible for that phenomenon. We can call these two parts of the model the phenomenal description and the mechanism description. While the phenomenal description [...] is useful for describing and predicting the mechanism's behaviour, it is the mechanistic model, which characterises the entities or parts and their activities and operations that actually explains.*(Glennan 2017)

Put in those terms, the framework as seen in Fig. 5 would qualify as an intelligible phenomenal model of the target system, since it predicts the behaviour of the same and finds algebraic expressions that fit those predictions. Yet, the mechanistic part of the model does very little to extent beyond the ”mechanism sketch” through a GNN, as I have already discussed in the previous section. On the other hand, what has been termed ”rough analogy” within their study (Cranmer et al. 2020), together with their aim to make the evolution operator physically intelligible, provides the guiding intuition of what the ”link” between ML model and target system needs to be mechanistically effective. It seems to be very important *how* the evolution operator is made transparent, so that strong connection between the ontology of the ML features and the target system is maintained.

In summary, those points are hinting towards a missing component, which could epistemically link the evolution operator to the *actual* physical processes - like the GNN is able to link the ontology of the target system to the ML architecture. On a procedural level, we require a stronger condition of *correspondence* to enable comprehensive mechanistic understanding. In what comes next, I want to use the idea of a GNN as ”mechanistic sketch”, to proceed and bridge over to my seminal case-study, dealing with ML for agent-based complex systems. My task is now to apply these conceptual tools to analyse the ML model framework in this case study. It will clarify the relevance of this discussion for the development of my broader link-concept, ultimately leading towards an argument for the adoption of mechanistic terminology in ML for the study of complex systems.

[30] This has been termed *theoretical hypothesis* in philosophical literature (Giere 1988), and in this case it is based on the the representational content of the GNN, in both, its structural features and internal functions.

Part III: Leveraging ML to understand hidden interactions in complex systems

Seungwoong Ha and Hawoong Jeong, who are the authors of a study titled ”Unravelling hidden interactions in complex systems with deep learning”, attempt to leverage ML technology to tackle what they identify as one the main complications in complex systems science: Investigating and extrapolating to ”the micro-level processes that govern emergent behavior.” As indicated in Part II, they were clearly influenced by the previous work of P.W. Battaglia on ”interaction networks”, and their broader epistemological aim is to ”explain collective behaviors of active matter like a bird flock with minimal microscopic description.”(S. Ha 2021) They chose to include three complex simulated test systems in their study: Cellular automata, active Ornstein Uhlenbeck particles (AOUP) and the Vicsek model that aims to capture bird-flocking phenomena. Each of those has a distinct characteristic, as for example the Vicsek model is continuous in space, while cellular automata are discrete. Furthermore, the AOUP is described as non-Markovian.[31] All three models can be considered representative cases of systems that show emergent collective behaviour, thus exhibiting both dimensions of a mechanism - the horizontal (causal), and a vertical (constitutive) dimension. The central task for gaining mechanistic understanding, in the sense of New Mechanism, is to build a model which can faithfully grasp the whole system-evolution with respect to a phenomenon, while representing the organisation and interactions of the lower-level entities that are relevant for producing this phenomenon.

I decided to include this work as the seminal case study in my Thesis, because of the following two reasons: First, the authors are conscious about the importance of leveraging powerful ML tools for the study of complex systems, and emergent phenomena in particular. Second, they are very explicit about how the ML framework enables *direct* physical interpretability without the need for additional XAI methods. What makes this study especially outstanding is the way they setup and maintain the strong correspondence between the ML architecture and the theoretically motivated description of agent-based complex systems.[32] I now want to follow along their methodology, which fits right into the context of a symbiotic modelling paradigm. I would like to argue that it has already set in place one main supporting pillar to enable mechanistic understanding within the chosen setting of complexity, which makes this study valuable for the formulation of an extended link-concept.

3.1 AgentNet and the formulation of correspondence

After the introduction into the theoretical background of complex systems, the Ha and Jeong (2021) start off with two subsequent sections, in which they establish what they call *spontaneous correspondence* between the complex system and the structure of AgentNet by providing formulations of both systems. The first part deals with the ”system formulation”, stating that they focus on a general agent-based system of n agents, whereby their aim is to identify the transition function of its constituents up to some time T.[33] The observed states of all agents at time t are denoted as a set of vectors $\mathbf{S}^t = \{\mathbf{s}_1^t, \mathbf{s}_2^t, ..., \mathbf{s}_n^t\}$, where each state variable vector of a single agent consists of k different variables itself. In addition, there are j global external and time dependent variables $\mathbf{u}^t = \{u_1^t, u_2^t, ..., u_j^t\}$. The composite state of all agents $\mathbf{S}^t$, and the global external variables $\mathbf{u}^t$ basically define the system's state at time t, and both quantities are used to compute the next time step $t+1$. The transition function $\Delta\mathbf{S}^{t+1}$ that entails all the information about how the system evolves between those time steps can be expressed as follows:

$$\mathbf{S}^{t+1} = \mathbf{S}^t + \Delta\mathbf{S}^{t+1} = \mathbf{S}^t + F(\mathbf{S}^{t,m}, \mathbf{u}^{t,m}) \tag{1}$$

[31] A Markov process assumes that the next state of a process only depends on the present state and not the past states; Consequently, non-Markovian means that the system exhibit memory-effects, where the actual state depends on a sequence of past states. For a more detailed account of the different characteristics, see S. Ha 2021, table 2.

[32] Their proposed novel variable-wise attention mechanism, which they call ”neural attention”, introduces an intrinsic transparency into the system; They further claim that their framework would be, to the best of their knowledge, the first ML algorithm to deal with dynamic nodes on a spatiotemporal graph.

[33] The following formal mathematical part closely follows the author's formulation in (S. Ha 2021), p.2

where $\mathbf{S}^{t,m} = [\mathbf{S}^t, \mathbf{S}^{t-1}, ..., \mathbf{S}^{t-m}]$ is a time sequence of all the prior states diwb to $t - m$. The same notation applies to $\mathbf{u}^{t,m}$. Depending on the characteristic of the system, the transition function F can be deterministic or stochastic, for which the authors chose one example each. The macroscopic model, which is here formulated with capital letters, is a mathematical shorthand and can be thought of as an aggregation of all the basic agent-agent interactions to obtain global dynamics. Correspondence, however, can only happen at the level of local interactions between agents. Therefore, the focus is shifted to the individual agent's state differences $\Delta \mathbf{s}_i^{t+1}$, which is like "zooming in" on the complex system. It can be expressed as follows:

$$\Delta \mathbf{s}_i^{t+1} = f(\mathbf{S}^{t,m}, \mathbf{u}^{t,m}) = f\left(h_{self}(\mathbf{s}_i^{t,m}, \mathbf{u}^{t,m}), \sum_{i \neq j} h_{pair}(\mathbf{s}_i^{t,m}, \mathbf{s}_j^{t,m}, \mathbf{u}^{t,m}) \right) \tag{2}$$

Because every agent could potentially interact with every other agent, including itself, the individual transition function f for a single agent can be decomposed in a self-interaction part h_{self} and a pairwise interaction term with the jth agent h_{pair}.

Following what Kaplan and Craver noted in the context of neuroscience, the modelling of a bird collective as a graph affords us with a rough *mechanistic sketch* agent-based complex systems and can be regarded as an abstraction. It is worth pointing out the terminology of "systems" employed by the authors, which is coherent with the New Mechanist account of a "higher-level entity".(Glennan 2017) To clarify this point further, the central element of their "system-formulation" runs directly parallel to how an abstract entity is composed of lower-level entities, interacting with each other, as captured within Fig. 1.[34] The abstract mathematical decomposition carries with it assumptions about the relevant internal structure, but does not identify the "significant internal variables" that account for most of the variance in a given phenomenon. In the same way as I argued for GNNs in section 2.2, it constrains the interpretation to facilitate mechanistic understanding, while at the same time leaves enough room for the details to be filled in.[35] Yet, when facing emergent phenomenology, these "details" can make a huge difference, which is why modelling complex systems is such a challenging task; Finding exact analytical expressions of h_{pair} and h_{self}, which would have to be formulated with respect to a specific system, is very close to fiction.[36] Instead, I would like to focus on *how* the authors chose to idealise the general agent-based complex system in such a way that, instead of finding analytical expressions, the representations of the same are learned from data by the ML-algorithm. A graphical illustration of this process (Fig. 6) helps to see the central points. First, it is useful to focus on the representational content: The initial modelling process is an abstraction step, moving from the macroscopic observation of the target system to the microscopic formulation. the whole flock of birds (on the right) is modelled as a graph G consisting of nodes V and edges $\mathcal{E}$ (on the left); Nodes correspond to individual birds and edges to their interactions.[37] Put differently, the circle on the left "zooms in" on the whole flock, like a magnifying glass, revealing individual components and their structural dependence. With that in mind, it is now time to focus on the ML-model part.

[34] The only difference is that the author's of AgentNet move from the macroscopic system to the microscopic interactions, by decomposing equation (1) into (2), while the emphasis in New Mechanist literature also takes into account the reverse - how the micro-level entities are organised and interact in such a way to produce the macro-phenomenon within system S. Both views are valid and necessary, if one sets out to gain a coherent mechanistic understanding.

[35] There is a gradual distinction, to which degree a mechanistic model is able to fill in these details, as touched on by Stuart Glennan. (Glennan 2017; Later in my work, I will provide a more complete discussion about these degrees of mechanistic understanding.

[36] Not only would we lose generality, but as well face the problem of how to test for accuracy in a space of virtually infinite possible solutions. Once more, Kaplan and Craver seem right when asserting that mechanistic modelling without abstraction and idealisation is a "philosopher's fiction".(Kaplan and Craver 2011)

[37] This might give us a false sense of ease when modelling a flock of birds or other agent-based complex systems in this way. However, using ML technology for the purpose of investigating different scales of organisation, requires in some way a symbiotic modelling framework. Especially if the macroscopic behaviour is to be understood on the basis of microscopic processes, a purely ML-driven setup is likely to muddle correlation with causation.

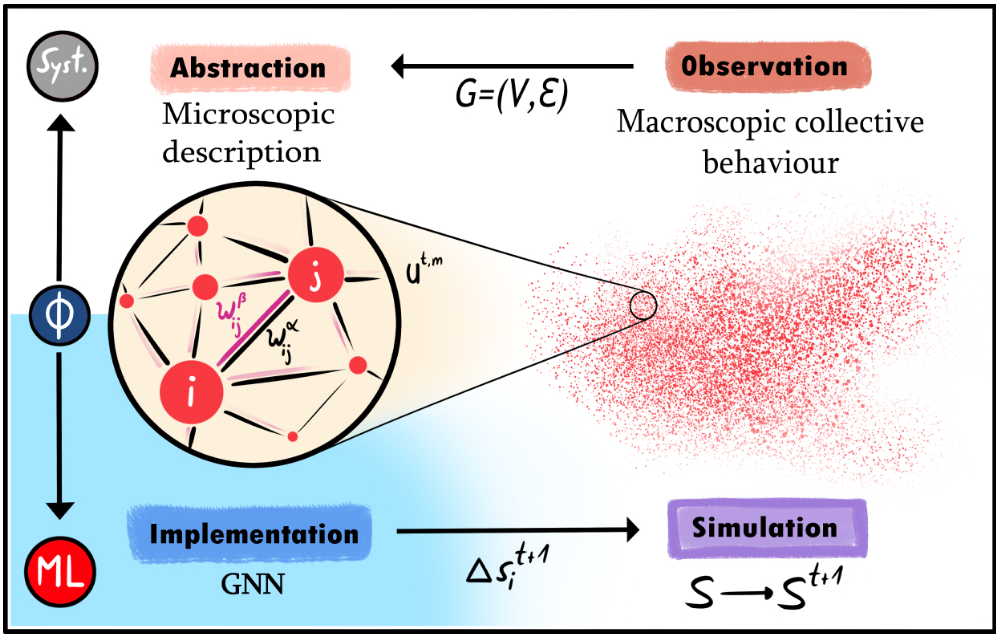


Figure 6: This graphic highlights important aspects of how the authors of AgentNet established *correspondence* between the ML framework and the complex system formulation, exemplified with an arbitrary agent-based system. It illustrates the central modelling steps taken to illuminate microscopic interaction rules in a "physically interpretable" way. Further, the mathematical underpinning of their corresponding formulations are depicted accordingly. The graphic has been created by myself.

The whole architecture, named *AgentNet*, works on the basis of a GNN, providing a "generalised framework" for data-driven modelling of "agent-based complex systems". With the introduction of a "mechanism sketch" in part II, I have demonstrated that a GNN already realises one central condition for a mechanism representation, since it establishes a link between the ontology of the target system and the ML framework. The authors of AgentNet seem to be well aware of this, as they explain that "one may depict a complex system as a dynamically changing graph in which each vertex is an agent, with links between agents indicating interactions."(S. Ha 2021) Taken along, however, the ontological constraint does not secure interpretability, which is why *AgentNet* incorporates a novel "attention mechanism". Usually, attention is performed over a GNN, specifying individual importance to neighbouring nodes in the network.[38] In other words, it enables each node to attend over the features of it's neighbouring nodes, assigning them different "attention values" that roughly correspond to the importance of each of the associated node connections (edges). There are various desirable properties, as for example efficient computation, applicability to graphs with varying node-degrees, and the ability of graph attention networks (GATs) to generalise to unseen graphs, making them effective for inductive learning problems.(Veličković et al. 2017) Further, as mentioned by Veličković, is "that they allow for dealing with variable sized inputs, focusing on the most relevant parts of the input to make decisions." But before jumping to conclusions about the interpretability of GATs, it is important to consider a deciding difference between the attention mechanism employed by Veličković and the novel "variable-wise attention", which Seungwoong Ha and Hawoong Jeong used for AgentNet. In contrast to conventional GATs that assign only one attention value per agent, AgentNet does learn an attention value for every state variable with a separate individual decoder.[39]

The clever move that tightens the link between the ML model and the system formulation is to further split up the pair-wise interaction functions h_{pair} into an interaction-strength component, and a directional component. Thereby, they introduce the variable-wise interaction strength functions w_{ij}^{q} as follows:

[38] For the computational basis of GNNs and the mathematical underpinnings, see Veličković et al. 2017.

[39] In the introduction of S. Ha 2021, the authors introduce *neural attention* that improves upon other GATs architectures, thus demonstrating a path to interpretable GNNs, as for example (Heras et al. 2019).

$$h_{pair}(\mathbf{s}_i^{t,m}, \mathbf{s}_j^{t,m}, \mathbf{u}^{t,m}) = \sum_q w_{ij}^q(\mathbf{s}_i^{t,m}, \mathbf{s}_j^{t,m}, \mathbf{u}^{t,m})\mathbf{v}_j^t(\mathbf{s}_j^{t,m}) \tag{3}$$

Here, the w_{ij}^q are scalar values that refer to the interaction strength of the qth variable between the ith and jth agent. Borrowing an analogy from the paper, w_{ij}^q can be identified with the magnitude of a force vector, while $\mathbf{v}_j$ does only convey the "directional" information; In this case, it is information about which node is interacting with the ith one. In Fig. 6 on the left, two attention values for separate variables with differently coloured edges are shown. As already mentioned above, instead of formulating analytical expressions of these interaction functions, the ML algorithm can now learn the unknown transition rules for individual agents $\Delta\mathbf{s}_i^{t+1}$ from data, which is, as I would argue, the equivalent of "filling in the details" into the mechanistic sketch.[40] The authors designed AgentNet to initially operate on a fully connected graph, in order to let it narrow down the number of interaction partners to the most relevant ones trough training. Hence, it realises the fact that not every agent of the whole collective is a potential interaction partner, but rather only a limited number of neighbours within a certain interaction range of the agent (and only at time t).

Already laid out by Veličković, an analysis of the learned attention weights "may lead to benefits in interpretability [...]."(Veličković et al. 2017) Yet, the kind of interpretability meant there is a more conservative one, tailored to the ML modelling paradigm. As such, it can only help visualise the varying importance that each node has given to its neighbours, but it cannot clarify physical relevance of those importance values for the purpose of *understanding* a certain system. It obscures the "back-translation" from latent variable-space to the output space (consisting of observable quantities relevant for understanding a prediction of a system's future state). Like in the case of applying a symbolic regression algorithm to the latent representations of a GNN.[41], this poses a hindrance for extrapolating *new* physical understanding. The authors of AgentNet, however, seemed aware of this issue, since they opted for a variable-wise attention mechanism and noted that it would enable "more extensive physical interpretation for the first time, which was impossible for conventional GATs such as identifying directional forces separately."(S. Ha 2021)

Again, Fig. 6 provides a graphical intuition for how this correspondence is approached from the ML model direction. In addition to the representational mode, a second way of reading it captures the sequence of modelling steps taken, and is thus more practically focused. It starts with the observed macroscopic collective behaviour of the bird flock on the right, traversing to an abstract microscopic formulation (system formulation) on the left, to arrive at the implementation of a ML algorithm (bottom left). On the basis of a GNN with an additional attention-mechanism (indicated with w_{ij}^α and w_{ij}^β), the latter is then trained with data to learn a functional representation of the system evolution, which is used to simulate the behaviour of a collective swarm of birds.[42] This predictive step is shown in the bottom part of the picture, moving from left to right - from microscopic to macroscopic.

There is a third mode of reading this graphic, which takes into account the formal mathematical background. It defines the system $\mathbf{S}$ at time t, for which AgentNet predicts (simulates) the next time-step $t+1$ as an aggregate of all the update-functions $\Delta\mathbf{s}_i^{t+1}$ of individual birds. $\mathbf{u}^{t,m}$ denotes the external variables present in the background, and both ML quantities w_{ij}^q refer to the separate attention-values for different variables (q) between node i and j within the GNN.[43] On the basis of this mathematical formulation, it is established what the authors termed "spontaneous correspondence" between system formulation and the ML model, which is shown as Φ (on the left) and supposed to be the very core element for "physical interpretability".[44] It clarifies the position of this study within a symbiotic modelling paradigm, because there is some implicit mechanistic modelling effort in the way the system is formulated, even though the

[40] Specifically, the model framework enables the ML system to learn a separate transition rule for every variable of each agent.

[41] Here, I am referring to Cranmer et al. 2020

[42] I will refer to this functional representation, also called "latent" representation, with the term "evolution operator", as it has been foreshadowed in figure 2 already.

[43] These variables are shown as α and β in Fig. 6.

[44] I've chosen to exemplify this picture through birds and the phenomenon of flocking, but Fig. 6 can be seen independent of any particular (agent-based) target system.

authors claim in their abstract that the proposed framework would be ”model-free” and entirely ”data-driven”.(S. Ha 2021) In the subsequent sections, I will try to argue for the benefits of adopting this mechanistic terminology, because it is reflective of this symbiotic modelling strategy, and therefore suitable for ML applications that aim to capture the micro-processes behind a macroscopic phenomenon.

Before moving on, I would like to return to the initial starting point. In Part I of my thesis, I chose to include a brief introduction into the kind of complexity I am concerned with, characterised by P.W. Anderson’s phrase ”more is different”. The lack of a shared complexity-definition across different disciplines arguably requires to first pick a specific class of systems. Now comprehensible through the previous analysis, the framework of AgentNet, dealing with agent based complex systems, provides a suitable starting point that does not clash with my chosen setting. What makes this study even more relevant, is that the authors claim to realise a kind of universality with AgentNet; The many diverse characteristics of those systems, which are ”difficult to incorporate into a single modelling framework”, are resolved by modifying encoder and decoder, (which is the way how data is translated into latent space) and setting a proper optimisation function. The core module is maintained throughout the whole process.[45] This way, they are able to address a variety of different system characteristics. I would like to juxtapose this with an account by Ladyman and Wiesner, who pointed out that ”there are various kinds of universality and forms of universal behaviour in complex systems”, which are ”often captured by a common mathematical description of the different systems in question.”(Ladyman and Wiesner 2020) Exactly that is the way, the authors of AgentNet established the *correspondence* between the (abstract) and broad formulation of agent-based complex systems and the ML model, where the latter one is fixing the particularities.

In this section, I tried to demonstrate the correspondence principle with the Vicsek model in more detail, highlighting the important aspects that link the ML model parameters to the ontology of the target system. Throughout this discussion, I kept track of the conceptual underpinning, pointing towards tacit mechanistic modelling commitments and a fruitful link to New Mechanist literature. It puts this case study in a position of seminal relevance for establishing mechanistic understanding of complex systems through hybrid ML modelling. However, there are some limitations of the correspondence as it has been analysed up until now and to which I would like to turn in the following section. I will also provide the means to build the bridge between Emily Sullivan’s concept and my developing extended link framework.

3.2 Limitations and the need for ”external” evidence

As with every other modelling framework, there are limitations and guiding assumptions within the setup of AgentNet. For example, agents with identical state variables are assumed to behave the same way, which is a very plausible assumption, and arguably a necessary one to make the work at all.[46] The second assumption, building on the first one, is that those state variables capture the *full* interaction strengths with different agents. The emphasis here is on ”full”, because it puts a limit on how much AgentNet can extrapolate to *unknown* ”interaction channels” or different ordering mechanisms, beyond what is already captured within the state-variables. Furthermore, it is assumed that the system is dominated by pairwise interactions at the expense of higher-order ones; Here, the authors claim that this would be the only inductive bias introduced by the ML algorithm.(S. Ha 2021)[47] However, the Vicsek model is meant to describe the flocking behaviour of ”macroscopic” individuals like real birds, which exhibit a higher behavioural diversity than what is captured within the state variables of AgentNet. Thus, the framework might only enable low degrees of mechanistic understanding about what features of individual birds drive the organisation in flocks. The ”link” that their setup of *correspondence* can provide, even though it is a much stronger one than any GNN alone could

[45] Here, ”core module” refers to the variable-wise attention mechanism, which they call ”neural attention”, enabling a very transparent physical interpretability. See S. Ha 2021, p.3

[46] AgentNet, like any other ML algorithm, has a mathematical underpinning. Let’s imagine now that two agents would have the same state-variable, but behave differently. This would introduce analytic ambiguities, quite in the same way as if a function $f(x)$ takes two or more values at one point x.

[47] In my view, the assumption that the state variables cover the full interaction strength could also be described as an inductive bias - and perhaps an even more radical one.

achieve, requires some additional plausibility criteria, not obtainable through the same ML algorithm. In other words, there is some leftover *link-uncertainty* within the correspondence principle itself. Accounting for such uncertainty requires the model to respect already existing empirical evidence, as captured by Sullivan.

It is usually standard of practice in ML applications to make those models applicable in situations where they have not been trained on - also referred to as ”generalisability”. On the one hand, the authors of AgentNet posed minimum prior assumptions of the unknown nature of the target agents. Probably with this condition in mind, it is claimed that the generality of their framework makes AgentNet flexible and applicable to many agent-based complex systems. On the other hand, it bares the danger of misrepresenting systems, missing crucial organisational features or important internal variables in complex settings. Consider for example that the mathematical system formulation in the paper does not clarify any relevant internal parameters contributing to most of the variability of a phenomenon. Yet, when applied to a specific target system, every ML algorithm needs some set of input variables, and so does AgentNet. I would argue that this is the right moment to consider ”external” empirical evidence. Optimally, there is already some evidence which features could be relevant for the phenomenon. Emily Sullivan qualified ”link-uncertainty” through the absence of such ”empirical evidence”; In her concept, the question if the ML features are actually casually related, and whose dependency the ML model could then pick up, is a matter of background knowledge. This evidence is thus ”external”, because in Sullivan’s characterisation, the ML algorithm itself cannot deliver any proof about the actual causal relations among the input variables. Yet, as I set out from the beginning, such causal relations form the basis of mechanistic knowledge, which is my desired target of understanding. What did the authors of AgentNet do with regards to such external evidence?

In the context of flocking birds, it was chosen to feed AgentNet with the bird’s positions and their velocities.[48] Besides the obviously important feature of every bird’s location, the plausibility of their choice to include the velocity is backed up by biological studies that investigated bird-flocks through means of statistical correlation analysis. (Ballerini et al. 2008; Bialek et al. 2012; Evangelista et al. 2017) In general, velocity correlation is very high among coherently moving parts of a flock. It is not yet completely clear, whether bird-flock behaviour is actually governed by interaction rules based on physical or topological distance.[49] It could be argued that this point of conflict could be resolved by taking into account empirical evidence from studies of comparably similar target systems. But without posing any conditions on the kind of empirical evidence needed, the first plausibility of ML features has to be grounded in such ”external evidence” drawing on an already existing body of knowledge about the target system in question; One might think of this as an enabling criterion for the ”link” to work at all - This condition for understanding, as anticipated by Emily Sullivan (visible in fig. 2 as **A**), is one I would like to keep in the extended link-concept.[50]

But if we decide to make mechanistic representations of a system the desired target of understanding, in order to gain insight into emergent phenomenology, the ML model takes on a more active part. Entering into a symbiotic modelling framework, which I’ve been motivating throughout this Thesis, it essentially means that we should consider the complementary kind of evidence - one, which is delivered *by* the ML model. If the ML setup includes the structural information to constrain the ML architecture to the ontology of the complex target system (as in the case of AgentNet), then the first logical conclusion is that this ML model, when trained, is able to create a ”mechanism hypothesis”, or at least could provide some mechanistic insight by pointing towards central organising features. One could think of this as *internal* evidence. However, this is not as straightforward, as it seems. Even if we make the ML model correspond on an ontological level (i.e. with a GNN that establishes a first mechanism sketch), it does not follow that this model is able to deliver physically plausible mechanisms. We need to remind ourselves that we are still dealing with a correlation extraction tool.

I would like to go back once more to how Sullivan qualified the level of understanding that

[48] First, they let AgentNet deal with a simulated test case of the vicsek model by generating data in 2D space. Later, they apply their model to trajectories of real birds, in particular Chimney Swifts.

[49] I will refer to this in more detail in the following section (3.3)

[50] The kind of understanding, which is resulting from a tight link in Sullivan’s concept, can only be abstract, since there is no way the modeller has access to the opaque evolution operator of the ML algorithm and how the internal model features are used to predict the data-correlations.

her link-concept could provide. According to her, even in the presence of link-uncertainty, a ML model can still deliver ”how-possible” knowledge; ”It is the patterns that the [...] model indicates that gives researches hypotheses to test and gain additional evidence for the strength of these patterns in real-world cases.” Exploring such hypotheses could further reduce link-uncertainty, because ”it points to possible correlations that are worthy of future scientific and empirical research.” [51] But, as philosophers of New Mechanism would put it, extracting correlations is not enough to gain mechanistic insight; When facing the struggle to mechanistically understand emergent phenomena, in which both dimensions of a mechanism appear, an indication of fundamental causal relations seems essential, pinning down at least the ”horizontal dimension”. (Glennan, Illari, and Weber 2021) So what exactly do we need to formulate an effective link between the ML model and the target system, in order for the former to offer mechanistic insight into the latter? In fact, the Authors of AgentNet claim that, different from correlations, their model ”provides a *causal* interaction strength since the attention value is strongly connected to the predictability of future dynamics [...].”(S. Ha 2021) This claim is particularly bold. What justification does AgentNet provide that it is able to models causal relations?

In the following section, I wish to deal with these questions in more depth, since it provides the necessary means for making the ML framework mechanistically effective for the purpose of understanding collective phenomena. I will first give an account of how AgentNet has been applied to the other (simulated) agent-based complex systems, namely a cellular automaton (CA) and ”Active Ornstein Uhlenbeck Particles” (AOUP). In comparison to the Vicsek model, these are arguably simpler cases with less degrees of internal variability.[52] Thus, they are good grounds to explore to which extend AgentNet provides correspondence on a functional level.

3.3 Establishing functional correspondence

Through training, AgentNet learns an internal representation of the target system, based on which it then predicts the time evolution of the system, which should make visible the emergent phenomenon. In this context, what I have called ”evolution operator” is basically the solidified version of the learned functional dependency between the latent variables of the ML model.[53] But as such, the ML evolution operator is a static and epistemically inaccessible unit, which has to be probed in certain ways to become functionally transparent. How this probing has to happen does strongly depend on the context; But in the case of AgentNet that aims at *physical interpretability*, the authors employed a very insightful strategy, which I would like to unravel in the following sections. My aim is to shed light on how AgentNet delivers validation that its evolution operator captures the central organising features for a specific macro-phenomenon, thus securing what I would call ”functional correspondence”. An analysis of AgentNet’s test cases will help qualify what ”functional correspondence” means and how it is integrated within my developing extended link concept for mechanistic understanding.

3.3.1 AgentNet applied to CA and AOUP

A strong link between the ML algorithm and the system’s ontology is likely to promote understanding on a process-level. To see this, I would like to consider how AgentNet has been applied to a cellular automaton (CA), which consists of a (14 x 14) cell-grid, where each cell can be in two states: alive or dead. The decision rules of the system govern how each cell evolves over discrete time steps; In the paper(S. Ha 2021), they are equivalent to those in Conway’s famous ”Game of life”.[54] The authors initialised a random pattern on the grid and simulated the progression, whereby the pattern of each time step was then used to train AgentNet, up to 200 updates. The ML algorithm was given the positions of each cell on the grid $(x_i(t), y_i(t))$ and their respective states $c_i(t)$, for which it predicted probabilities of becoming alive. After the training procedure, the attention weights of a single target cell were high almost exclusively for the eight directly neighbouring cells that surround the target. All the other weights nearly

[51] See Sullivan 2022; p. 22; Sullivan does argue for such how-possible knowledge with the deep-patient case-study.

[52] By ”variability”, I am referring to the range of behaviour that individual agents can exhibit.

[53] For AgentNet, these are all internal weights of the Neural Network layes, as well as the attention weights of the graph-attention core as shown in figure 1 of S. Ha 2021.

[54] The details can be found in Gardner 1970.

vanished. Remember that these weights correspond to the interaction strength values between the cells, with direct reference to Fig. 6 as an illustration of the underlying graph structure.[55] The CA is arguably a very simple test case to see what the attention mechanism of AgentNet does. It is able to navigate through the space of possible interactions to focus on the ones that are *functionally relevant* for the system's evolution.

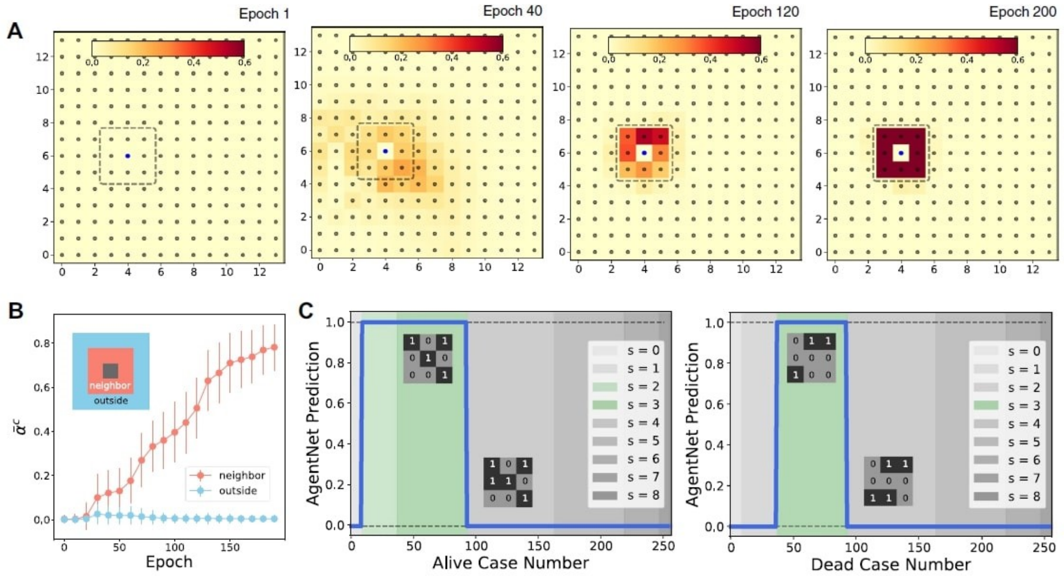


Figure 7: Shown here ar the result of AgentNet for cellular automata. In **A**, the attention weights of a single target cell (blue dot) across the whole grid are illustrated with intermediate steps throughout the training. It becomes visible, how the model learns to narrow down the initially equal-distributed attention weights to the eight nearest neighbours surrounding the target cell. The final pattern of concentrated attention (after 200 training steps) coincides perfectly with the theoretical interaction range upon which the CA decision rule is based. In panel **(B)**, it is demonstrated that the attention values $\bar{a}^c$ are approaching higher values for directly neighbouring cells. The data is averaged from 100 test samples. In **(C)**, the evaluation takes place with dead-cases and alive-cases of a given test cell. Importantly, "the total number of alive cells in the neighbourhood is denoted by s, which is the sole parameter of the CA decision rule". The figure was copied from (S. Ha 2021)

We can faithfully claim, that in this case, AgentNet captured the "true" interaction range. Put differently, it successfully retrieved the ground-truth of *relevant* interaction partners for a single cell. However, this is only the first half of the puzzle, because the whole CA evolution is determined by a *decision rule*, which is, of course, formulated through relevant neighbours. The question is: Did AgentNet also recapture the correct decision-rule? There is no additional output channel that could let the model communicate its decision rule in a semantic or algebraic sense. The only possible way to evaluate the direct evolution operator is through *prediction* and a targeted visualisation of ML variables. As visible in Fig. 7 panel **C**, the predicted system evolution can be evaluated with respect to the single relevant parameter for the decision rule, which is the total number of alive cells within the eight surrounding ones. It is in perfect alignment with the theoretical expectation. Thus, AgentNet is not only able to narrow the interaction range down to the relevant number of eight neighbouring cells, but also gets the whole system evolution right for this test system. Before integrating these results conceptually, I would like to look at another test system, to which AgentNet has been applied.

[55] Although I opted for the vicsek model to illustrate the correspondence principle, the graphic is representative of other agent-based complex systems as well; In this case, the collective "macroscopic behaviour" of the system is a discretely changing pattern on the grid, whereby the the ML algorithm does not know anything about the underlying decision rule governing the system's evolution.

The authors decided to further test the capabilities of AgentNet on grounds of an active particle system. Some of the most studied ones among those systems are bacteria, cells, and the so-called motor-proteins.[56]

The authors of AgentNet simulated a system confining many particles in a harmonic external potential ($V_{ext}(x) = \frac{1}{2}kx^2$), while the interaction between them is modelled through a soft-core potential, which reads:

$$V_{int}(x_j) = \sum_{i<j} \phi(r_{ij}), \qquad \phi(r_{ij}) = exp\left(-\frac{|r_{ij}|^3}{R^3}\right) \tag{4}$$

where r_{ij} is the relative distance between the particles i and j, and R being the typical interaction length. The whole system can then be described with an over-damped Langevin equation on the position $\mathbf{x}_i$ of each particle:

$$\gamma \dot{\mathbf{x}}_i = -\nabla(V_{ext} + V_{int}) + \sqrt{2\gamma T}\eta_{\mathbf{i}} + \gamma \mathbf{f_i} \tag{5}$$

Here, γ is the drag coefficient, $\eta_{\mathbf{i}} = (\eta_x, \eta_y)$ describes a 2D Gaussian white noise with zero average, and the whole term $\sqrt{2\gamma T}\eta_{\mathbf{i}}$ is essentially the effect of the thermal bath at temperature T, in which the particles are immersed. $\mathbf{f_i}$ captures the intrinsic propulsion force of the particles, in this case driven by a stochastic process, called ”Ornstein Uhlenbeck process”. This process is modelled as follows:

$$\tau \dot{\mathbf{f}}_{\mathbf{i}} = -\mathbf{f_i} + \sqrt{2D_a}\mathbf{w_i} \tag{6}$$

where τ refers to the correlation time, D_a is the diffusion constant, and $\mathbf{w_i}$ denotes a standard Gaussian white noise. As input, AgentNet did receive four different variables: The position and velocity of each particle in two dimensions. I would like to draw the attention again to Fig. 6 for one moment. The AOUP system does fit within the picture, as different nodes of the graph correspond to the individual particles, while their interactions, which are modelled with a soft-core potential V_{int}, can be identified as edges in between. The global variable u that refers to an overall system property is in this case the typical interaction distance R.

AgentNet is perfectly capable of predicting many time steps beyond the eight sampling steps used for training (as visible in Fig. 8, panel **A** and **B**). More interesting, however, are the other visualisations of the learned velocity attention weights in both directions (x-direction in panel **C** and y-direction in panel **D**). Both weights are plotted against the respective x-directional and y-directional force components of $F_{int} = -\nabla V_{int}$ and show a linear relationship, while a single attention value from another graph attention network (as comparison baseline) does not show a correlation with those force components at all. In panel **E**, there is even made a stronger case for the success of direct correspondence between the velocity attention weights and the soft-core interaction force(components) between the particles, demonstrated with the x-directional component.[57]

Considering that the ML algorithm did not have any information about the independent variable $\mathbf{f_i}$, which has its own Ornstein-Uhlenbeck dynamics, it had to infer this variable from just eight steps of trajectory-observation. Yet, one detail has only been briefly mentioned in the paper: Apparently, the proposed particle system in its ”passive” form (i.e. without the active self-propulsion $\mathbf{f_i} = 0$ and no external potential $V_{ext} = 0$) does already show an equilibrium phase-aggregation: ”Particles form stable clusters, which arrange into a periodic configuration. In two dimensions the triangular or hexagonal lattice is the only stable pattern, occurring for small enough temperature T at a large density.”[58] In other words, the collective phenomenon of clustering is already observable in a system without any stochastic process that drives the

[56] As claimed in a comparative study between two models of active cluster crystals, understanding this self-propulsion could lead to important applications, for example in drug delivery: ”A propulsion mechanism could enhance and direct the transport process with a consequent increase of its efficiency.”(Caprini et al. 2019)

[57] This is true for a constant scaling factor of c=0.28; The authors further remark that there is a slight mismatch for small r_{ij}.

[58] See Caprini et al. 2019; It is further mentioned that ”the cluster-formation phenomenon despite the repulsive interaction between all the particles can be physically interpreted by considering the force balance between the intra-cluster repulsion (the force felt by particles inside clusters) and the inter-cluster effective interaction, that is, the force exerted by neighboring clusters.”

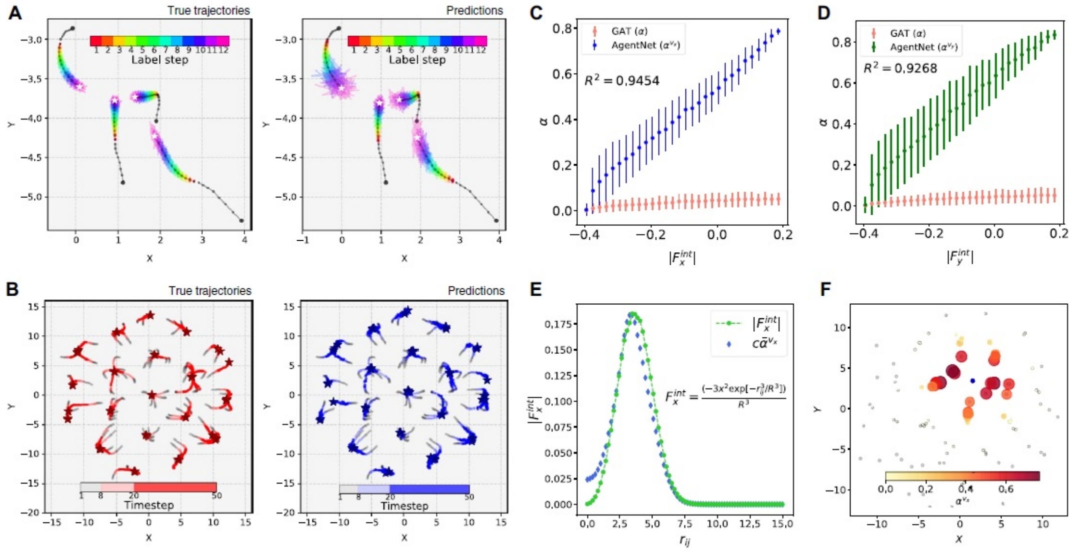


Figure 8: Panel **A** and **B** show the true trajectories of individual particles (top) and a collective of particles (bottom), compared with the predictions of AgentNet. Panel **C** and **D** shows a linear relationship of the velocity attention values, where the x-directional value α^{v_x} is plotted against the x-directional force component. (The total force on the particles, encapsulated in V_{int}, has to be considered as $\mathbf{F}_{int} = F_x\mathbf{e}_x + F_y\mathbf{e}_y$) The linear relationship between these values, in contrast to a conventional single-head attention value, indicates a strong correspondence. Panel **E** makes the same point even more explicit, where the velocity attention value is plotted together with the the x-directional force component F_x against the relative particle distance r_{ij}. The figure was copied from (S. Ha 2021)

particle's self-propulsion. Because there is no additional transparency module that would let us observe the evolution operator of AgentNet directly, we cannot recapture the exact dynamics of the self-propulsion process. But, concluding from the evidence about the passive systems, the self propulsion of the particles cannot be the central locus of collective organisation that drives the clustering phenomenon. Hence, the presence of such an additional stochastic dynamics is perhaps better characterised as additional "noise" within the system, possibly having an effect on the final spatial distribution of the cluster. The authors of AgentNet state that their model is "one the most complex systems possessing internal potential, external potential, memory effects, and stochastic noises."(S. Ha 2021) All these processes challenge AgentNet in its performance to still get the macroscopic system evolution right. As shown in Fig. 8, it is in fact able to capture the relevant aspect - the internal electromagnetic potential between the particles, which has been shown to drive the clustering phenomenon.(Caprini et al. 2019) At the same time, it faithfully predicts the whole system evolution beyond the eight training steps. Two things are worth mentioning at this point: First, "external" evidence from outside the ML framework still needs to play a valuable role for the extended link framework, because, as demonstrated within this short discussion, it takes part in evaluating the ML performance (in this context to see its robustness to noise.) Second, the desired kind of mechanistic understanding does not seek to make every detail of the complex system fully transparent. As New Mechanism holds, we can still have mechanistic understanding of a system, if the central organisational principles, upon which a certain phenomenon arises, are represented. Once settled on a phenomenon, it is then necessary to discard some irrelevant aspects, which might complicate the process of modelling, but have (almost) no bearing on the emergence of this macroscopic phenomenon. AgentNet seems to enable mechanistic insight into the complex agent-based system, which is in alignment with this New Mechanist characterisation of *relevance*.

Following this point, let me return to what I called "targeted visualisation of ML variables", which is necessary to evaluate the predictive capabilities of AgentNet. At the same time, visualisation is part of a strategy for justifying *physical interpretability*. Because it is a common thread running through the entire case study, I would like to use the term *relevance realisation*, which is supposed to capture the whole strategy as a combination of establishing and evaluating physical interpretability. The term ”relevance realisation” was originally introduced in the context of cognitive science, where it refers to the ability of ”intelligent” agents like humans to ignore most of the available options in the process of decision-making, while acting only on the basis of *relevant* ones. Which ones are relevant cannot be defined in itself, because relevance is highly situation-dependent. However, as argued by John Vervaeke and his colleagues, the process by which relevance is *realised* can be characterised in more general terms.(Vervaeke, Lillicrap, and Richards 2012). I borrowed the term to indicate that something similar is happening here within the ML model, which is able to realise the relevance of different agents for the emergence of a defined phenomenon. Importantly, which variables are relevant is also highly phenomenon-dependent. This is in agreement with the New Mechanist characterisation of relevance, where emphasis is also placed on the context-sensitivity of mechanistic details. Thus, I chose to evoke a process that manages to ”realise” relevance in a given situation. What I mean by situation-dependent relevance of ML variables is, that *they take part in the system evolution, while pointing towards central causal drivers for the emergence of the macroscopic phenomenon.* Part of this strategy, therefore is the targeted visualisation of those ML variables, in order to understand *if* and *how* they correspond to physical quantities within the target system.

In the test case with the Cellular Automaton, AgentNet does successfully demonstrate that the attention weights $\bar{a}^c$ are indicative of the ”true” interaction range for individual cells - which, in turn, is a causally relevant factor for the system evolution. In the AOUP case, AgentNet establishes a tight correspondence between the attention weights and individual components of the force that particles experience - a physical quantity within the target system.

Both aspects taken together, *prediction* and *relevance realisation* give rise to what I already indicated with *functional correspondence.* This conceptual integration is supposed to be an extension to the ontological correspondence between the ML architecture and the structure of the target system. Furthermore, both dimensions of a mechanism are reflected therein: The horizontal causal dimension is apparent in the *relevance realisation*, while the *prediction* of the whole phenomenon secures the vertical dimension, because it ensures that the micro-details are causally linked to the whole system’s evolution. Without a decent prediction of the macroscopic system evolution, purported insights into micro-details would be meaningless; Without relevance realisation, the insights gained would be epistemically hollow and could not provide mechanistic understanding of a phenomenon. This allows for the claim that AgentNet, through establishing *functional correspondence*, becomes interpretable in a mechanistic way.[59]

3.4 Functional correspondence employed - a step towards ”internal evidence”

In the last section I outlined the general line of reasoning how AgentNet enables *physical interpretability* and manages to establish *functional correspondence* by means of prediction and relevance realisation. Yet, the focus has so far been on applications to specific test systems. But in order to be a valuable contributor to the body of background knowledge, it has to stand grounds in applications to a real-world scenario as well - preferably even to a whole class of complex systems, which is often captured within the term ”generalisability”. In this section, I will therefore focus on how AgentNet has been applied to model the phenomenon of flocking birds, based on data samples of Chimney Swifts. My aim is to employ the conceptual developments that came clear throughout the previous analysis of both simulated test cases. I will argue that *functional correspondence* makes AgentNet contribute to the (partly incoherent) body of empirical background knowledge, by providing ”internal evidence” for central drivers of the emergent phenomenon. This section is supposed to prepare the grounds for the final integration of my results into a conceptual framework for mechanistic understanding.

[59] I need to emphasise that the CA is in itself a model, which abstracts away from real-life processes. Much like a graph model, cells can represent individual agents. Thus, the link between a Cellular Automaton and a certain target system is arguably of the same character as a ”mechanism sketch”. Consequently, there would be two links to consider: One between he ML model and the target model (In this case the CA) and another between the target model and the ”real” target system. AgentNet, however, does not make any claims about the second link-connection between the CA and other more behaviourally diverse target systems. Taking the CA as an idealised ”target system”, I deemed it worth including, because it demonstrates how AgentNet establishes correspondence on a functional level.

3.4.1 Shaky grounds of background knowledge

As I have mentioned already, empirical evidence has to accompany the ML model ”externally”, like Emily Sullivan argued for in her link-concept.[60] But similar to her staying vague about the desired kind of understanding, she abstained from qualifying empirical evidence further. Rather the problem is outsourced to the specific scientific discipline of interest, which, according to her, sets the standards for the nature of this evidence.(Sullivan 2022) What if there is only a partly coherent or ambiguous body of empirical background knowledge? Is there any way for AgentNet to contribute to more understanding of the subject matter? Before I can try to answer these questions, I would like to first demonstrate the issue with an example in the context of bird flocking behaviour:

In a study by William Bialek and his colleagues, methods from statistical mechanics were used to model the macroscopic phenomenon of bird-flocking based on micro-principles. The authors derived statistical quantities like perpendicular and longitudinal velocity correlations between neighbouring birds from those microscopic descriptions, and then evaluated their predictions on the basis of empirical data. (Bialek et al. 2012) Their general theoretical commitment is that fundamental correlations are between birds and their directly interacting neighbours. It is further mentioned that their results provide ”strong support for the claim [...] that birds interact with a fixed number of neighbors, rather than with all the birds within a fixed metric distance.” Another empirical study finds evidence along the same lines, in favour for topological interaction distance.(Ballerini et al. 2008) I would like to contrast this with an analysis that ”supports the presence of physical distance based interaction rules”, as briefly touched on in section 3.2.(Evangelista et al. 2017)

The bottom line is the following: We are here confronted with a disjoint body of empirical and theoretical work on the flocking behaviour of birds, in particular chimney-swifts and starlings. It is important to note that two of the studies concerned with starlings (Ballerini et al. 2008 and Bialek et al. 2012) find similar evidence for topological interaction distance, while the single study about chimney-swifts by Evangelista and colleagues comes to the opposite conclusion. One simple explanation could be that different bird species organise differently when it comes to collective behaviour. In general, it seems like the literature on flocking behaviour of European starlings is mostly agreeing on the presence of topological distance-based interaction rules, while for Chimney-swifts it is hypothesised to be otherwise. However, none of the studies mentioned above performs a time-simulated prediction of the whole bird-flock. Rather, statistical quantities are computed based on the model choice, which are then compared to the same quantities obtained from empirical data.[61] Regarding the prediction of a flocking phenomenon, this empty space is ready to be filled in by ML models. But we have to ask the question if and to what extent their learned functional relations are representative of the real organising principles upon which the phenomenon arises inside the target system. Exactly at this point we can see the advantages of having *functional correspondence* available, as AgentNet does, although so far only justified for specific simulated systems. To understand how a ML model like AgentNet could contribute to the disparate ”external” evidence of bird-flocking behaviour, it is necessary to apply the results of my previous analysis to the example of AgentNet dealing with real world data. To this I would like to turn now.

[60] By ”external”, I am referring once again to all the evidence obtained independently of any ML application.

[61] In (Bialek et al. 2012), the correlation functions predicted by the maximum entropy model are plotted against the values obtained experimental data. In (Evangelista et al. 2017), the similarity in three-dimensional heading is computed and stored within a matrix W, for which a dendrogram is constructed, representing hierarchical clustering; In (Ballerini et al. 2008) it used a reconstruction of the birds positions in 3D space, based on calculated statistical quantities that were obtained through empirical data. Later in the paper, the time-evolution of a starling flock under predatory attack is simulated, one time with physical distance interactions and under topological distance-based interactions. While this comes closest to a prediction of the bird-flocking behaviour, it does study a different setting, hence a different ”mechanism” that maintains organisation in cases of predatory attack.

3.4.2 AgentNet meets empirical data

As demonstrated with the test case applications, we are now equipped with more confidence that AgentNet is actually delivering in terms of its ability to model a causal interaction strength. On this basis, the authors of AgentNet further argue for its usefulness to infer and model the microdynamics of individual birds.[62] For this, they performed an application of AgentNet to predict the trajectories of flocking Chimney Swifts; In particular, they used parts of the data from Evangelista and colleagues (Evangelista et al. 2017), and thus complement the methodology of statistical cluster analysis, which in itself was lacking a predictive component of the collective phenomenon.

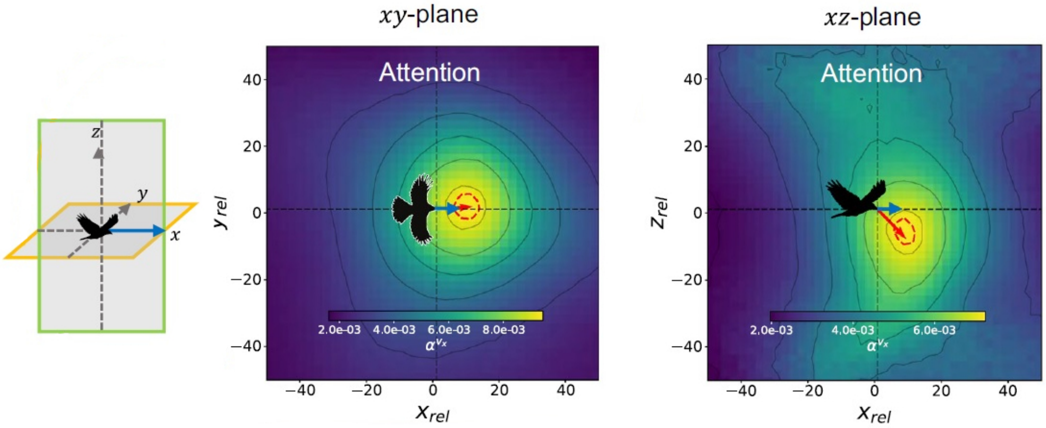


Figure 9: This picture shows two-dimensional heatmaps of averaged attention values α_v for individual birds in the xy-plane and xz-plane. Both axis refer to the relative coordinates between birds. The direction of movement is visible as a blue arrow, the red arrow indicates the direction towards maximum attention values, which is oriented downward to the front. The figure was reproduced from S. Ha 2021

To deal with data of real birds, however, requires a lot more preparatory considerations, which the authors of AgentNet provide in their supplementary notes. One of the main differences, precluding a hasty application of their ML model, is that due to the limited capabilities of capturing every bird's full trajectory in a given time interval, the system can no longer be modelled as a static graph. Instead, it needs to be accounted for cases, where a bird newly enters the scene, or another one escapes the camera's focus for a while to re-enter later. Thus, we are dealing with a dynamically changing graph; A naive application of a GNN to this situation, as the authors argue, would be doomed to failure.[63] For this reason, they decided to additionally supervise the chosen data with an *inspection algorithm scheme* that manually checks the hidden state at each time step and updates each node's status to the appropriate form.[64] The central point here: Without such considerations of the structure and incompleteness of data from a real bird-flock, the correspondence between the ML agents and the target agents would be obstructed and the resulting mechanistic insight would be at best limited, and at worst impossible. Undoubtedly, the attention mechanism of AgentNet does the heavy lifting in terms of physical interpretability; But the "Inspection algorithm" is the crucial part that needs to be there in order to preserve this transparency. Put differently, functional correspondence is very dependent on an active maintenance and careful thought about how the ML evolution operator can stay connected to the target system. In real life settings, this is highly dependent on the structure of data available.

[62] In the case of the AOUPs, this would refer to the movement of the individual particles, based on the effect of the electromagnetic potential.(S. Ha 2021)

[63] See S. Ha 2021, supplementary notes.

[64] More precisely, the inspection method "creates a mask for the attention matrix to ensure that the attention between a valid agent and non-existent agent should be strictly zero at any time." (See supplementary notes in S. Ha 2021)

AgentNet performs a faithful *prediction* of the flocking phenomenon at the macro-level with sufficient accuracy.[65] A visualisation of averaged attention weights is shown in Fig. 9, and AgentNet's ability of *relevance realisation* ensures that these weights functionally correspond to some quantity at the micro-level of individual birds. Thus, it is now possible to pose plausible hypothesis about quantities, to which these weights could correspond. At this point, external evidence has been considered by the authors. Drawing on other research about individual agent's properties, as for example visual field studies of birds, they argue that the resulting attention focus (Fig. 9) as oriented to the front downwards is "physically plausible".[66] Regarding European Starlings, there is also some speculation about an underlying "mechanism" consistent with their modelled interaction rules. As put by Ballerini and colleagues, "[...] we believe that the only mechanism compatible with our result is vision."(Ballerini et al. 2008) But this is only one possibility. Since AgnentNet dealt with Chimney Swifts, it is not perfectly clear how precise the attention weights do actually corresponding to the "visual fields" of individual birds. Perhaps, it is an intrinsic behavioural property of individual birds that better matches the heatmap of averaged velocity-attention (Fig. 9). For instance, it could also mean that during flocking, Chimney Swifts (on average) pay most attention to other birds that fly in a region indicated by the highest attention values.

Let me return to the fact that there has not been secure knowledge as to which kinds of interactions at the micro-level, are governing the collective behaviour. It is this uncertain background through which AgentNet becomes such a valuable extension to the statistical modelling paradigm. As demonstrated at length, it is arguably fit to contribute "internally" to the body of background knowledge by posing "mechanistic hypotheses" in the sense of New Mechanism.[67] I do not try to argue that AgentNet is able to deliver hard evidence in favour for any particular organising mechanism. Viewed in the context of New Mechanism, it is perfectly fine to point towards central *relevant* aspects within the system that are causally connected to the emergence of the macroscopic phenomenon. AgentNet established a strong enough link that makes it possible for the ML model to pose hypotheses about cognitive capacities of individual birds while flocking. Although it reaches over into biological mechanisms, it is not quite obvious why we should exclude such explanatory accounts from the framework of mechanistic understanding; After all, it is the cognitive capacity of birds that allows them to make sense their surrounding and make decisions in the first place. It needs to be emphasised that these considerations are only applicable to the tightly constrained situation when birds aggregate collectively. The phenomenal target does not include any other individual habitual decision strategies that may have led to this collective behaviour in the first place.

What I have tried to demonstrate is that AgentNet is able to viably connect to that body of empirical background knowledge for an emergent phenomenon - collectively flocking birds. It constrains the abstract ML agents to the ontology of the target system, as well as secures a level of *functional correspondence* through *prediction* and *relevance realisation*. In this sense, AgentNet points out features at the micro-level, which are of central relevance for the emergence of the macroscopic phenomenon. It works purely on the basis of available data, but manages to contribute to mechanistic understanding in the spirit of New Mechanism. Correspondence on a functional level, however, has to be carefully maintained throughout the process of modelling, because of the sometimes incoherent structure of data available. In the following part, which is also the final one, I will wrap up by putting every analytical step together into a full-fledged conceptual framework. Afterwards, there will be a discussion of its limitations, and the resulting implications for future investigation.

[65] This is usually evaluated against other benchmarks. See S. Ha 2021, Fig. 5.

[66] Mentioned studies are for example Martin and Shaw 2010 and Martin 2010

[67] A side-note: Because ML is in huge parts a very empirical process that includes a lot of experimentation, the output could be considered "empirical knowledge" as well.

Part IV: ”Correspondence Principle for Mechanistic Interpretability” - Discussion and outlook

It is time to summarise the most important points that came clear through my analysis. I now conceptually integrate them into a full-fledged extended link-framework, which I call *Correspondence Principle for Mechanistic Interpretability*, or short, ”CPMint”. Figure 10 illustrates the main features of the concept. I will now provide a thorough description of it, which should guide the intuition for the following summary. I hope to shed light on how the concept manages to improve upon Emily Sullivan’s link-uncertainty framework, as I have promised in the beginning of my Thesis.

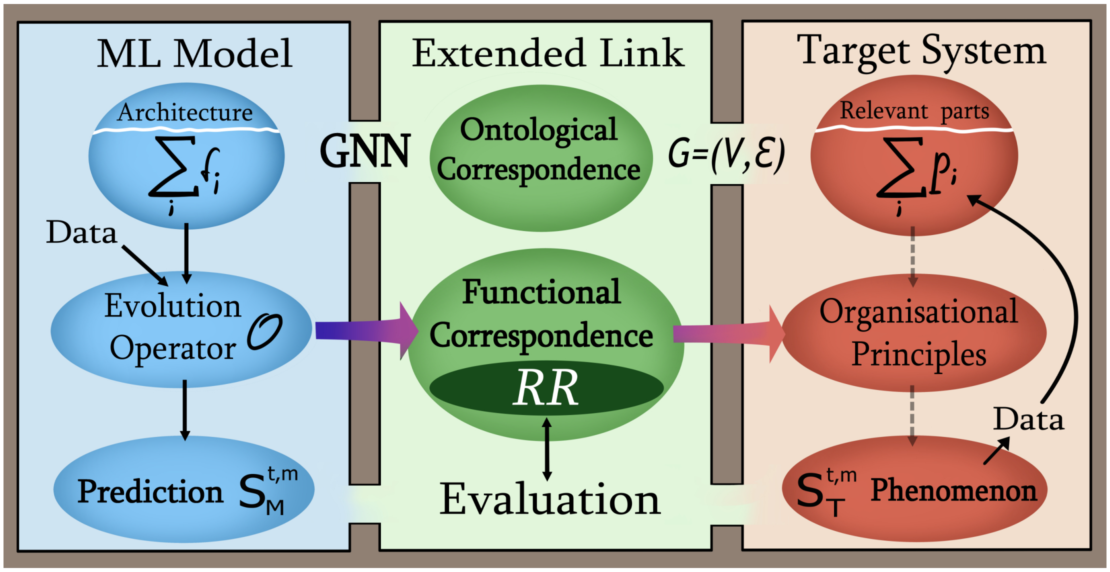


Figure 10: Illustration of **CPMint**, emphasising correspondence at three levels: At the ontological, the functional and the evaluation level. The upper layer reflects Part II, in which I argued for the ontological constrained ML architecture (or features), realising a ”Mechanism sketch”. The middle layer encapsulates how the evolution operator is made physically interpretable. The big colourful arrows indicate the direction of epistemic inference to the relevant organising principles, once the ML model established *functional correspondence.* This is essentially the horizontal, causal dimension of a mechanism. The bottom layer shows the evaluation of the predictive ML model output with the time-evolution of the (macroscopic) target phenomenon, thus capturing the vertical, constitutive dimension.

4.1 Pictures and words - two intertwining tracks of analysis

First of all, the final concept emphasises once more a link connection between the ML model (on the left) and the target system (on the right), and thus it picks up the framing that has been central to this whole work. It could be claimed that the images employed tell their own story; Yet, this is only true in part, because the analytical and the graphical track have been co-created simultaneously - so both are supposed to tell the same story through different media. By constantly moving back and forth between them, I attempted to make my conceptual framework as concise, plausible and intelligible as possible. But why did I chose the name ”**CPMint**”? And does the concept deliver in terms of mechanistic understanding? To answer both questions in one sweep, I will now trace my thought process behind this co-creational process, which should be just as much a reflection on the graphical content as it is a summary of the main analytical points.

Let me revisit the graphical summary of Emily Sullivan’s link-uncertainty concept again, shown in Fig. 2. There, neither the evolution-operator of the ML model, nor the relevant mechanism producing the target phenomenon is addressed. The kind of understanding is not specified further, therefore staying quite abstract. It is not *confined to* mechanism representations, but instead *confined by* the empirical background knowledge about the studied phenomenon; Mechanistic representations are just a part of this body of empirical knowledge. From there on, I essentially switch perspectives. Rather than just being a *precondition* for understanding, I set out to create a way to make mechanistic representations the desired *target* of understanding.

In part II, I motivate why there is benefit in adopting a symbiotic modelling framework, which openly acknowledges the complementary strengths and weaknesses of data-driven modelling (ML) and mechanistic modelling. By reviewing some case studies in the field of ”Medium AI”, I come to the intermediary result that GNNs can provide a mechanism sketch for complex (agent-based) systems exhibiting emergent collective phenomena. Here, the establishment of ontological correspondence between the ML architecture and the structural organisation of the target system takes shape, which is reflected in the top ”extended-link-layer” of Fig. 10. I end this part with the realisation that this alone does not yet make the ML evolution operator physically transparent. Hence, the level of understanding would be confined to those of phenomenological models, still lacking the mechanistic component.[68]

With the final version of **CPMint**, illustrated in figure 10, I take this idea all the way through, such that there are three levels of correspondence between the ML model (again on the left hand side) and the target system (on the right hand side). Both are connected with the *Extended Link* in the middle, which I set out to formulate. The illustration is designed to mirror the structure of my thesis. Beginning with insights from Part II, the top layer shows the ontological component of the extended link, featuring the GNN, which is able to process data that is represented as a graph ($G = (V, \mathcal{E})$). This correspondence between ML features f_i and parts within the system p_i confines the ML architecture to the theoretical system decomposition, while putting a structural foundation underneath it. It already shows crucial components of a mechanism - thus the term *mechanism sketch*.[69]

Moving on to Part III, the focus is now pointed towards agent-based complex systems. In the case of a collective, macroscopic phenomenon, as sometimes occurs in these systems, the central goal is to find out which parts and interactions at the micro level are *relevant* for the emergence of that particular phenomenon. As visible in Fig. 6, I deliberately placed the theoretical system-decomposition on the left hand side, emphasising its abstract nature, while putting the target system of interest on the right hand side. The ”emergent” phenomenon can be regarded as the time-evolution of states with respect to that system, as seen with ($S^t \to S^{t+1}$) on the bottom right of the same image. One central element is the establishment of correspondence between the theoretical system formulation and the ML model description, indicated with ϕ on the left of Fig. 6, which points towards some correspondence beyond the ontological level. At the top we see how a GNN delivers boundaries to decompose the system into parts, while the same is done mathematically in parallel. The authors of AgentNet show which quantities in the system the ML features correspond to. For example, the attention weights, illustrated with w_{ij}^q, are plausibly linked to the interaction strength between different nodes (or ”agents”) in the system. These weights are the central elements, through which the ML algorithm becomes ”physically interpretable”. I then used two test-case applications of AgentNet (CA and AOUP) to analyse how it delivers justifications for the proposed ”physical interpretability”, making possible some insight into the dynamical interaction rules that govern the collective phenomenon. The central point of analysis could be summarised in the form of a guiding hypothesis:

(FCH) Functional Correspondence Hypothesis: *Correspondence between the ML model and the target system, initiated on the structural level of system-ontology and ML architecture, needs to be preserved down to the functional level of the ML evolution operator, in order to be epistemically effective for studying dynamical interaction rules that govern an emergent phenomenon within that system.*

[68] This distinction has been made in New Mechanist literature. See Glennan 2017, p. 66-67)

[69] The theoretical content of this decomposition by means of mathematical relations and a graph representation should be evident. It has to be mentioned, however, that this decomposition is not only theoretical, but also practically confined by the kind of data that is collected. If we are observing a flock of birds, for example, then it follows logically to decompose the system individual birds, because only their trajectories are observed. Emily Sullivan's already noted that biased data is limiting representativity of the ML model. In **CPMint**, I tried to account for this problem, making data a consideration from the outset, which impinges upon the choice of ML architectures.

Working on the backdrop of the already established ontological correspondence, I qualify functional correspondence through a combination of two aspects: First, the *prediction* of the whole system evolution on a macroscopic scale, and second, a successful *relevance realisation* by strategically visualising ML variables.[70] This conceptual move is the main pillar underneath **CPMint**, therefore occupying the centre within Fig. 10. These qualities nicely fit together with the horizontal and the vertical dimension of a mechanism, anchored in the philosophical insights from New Mechanists: The vertical one being prediction of the whole system-evolution; The horizontal one referring to the transparency for causal understanding of organisation principles. I argue that a successful realisation of functional correspondence can make a ML model deliver ”internal evidence” for some relevant organising principles, upon which the collective phenomenon emerges. This way, the symbiotic, mechanistically interpretable ML model would be a valuable contributor to the body of empirical knowledge. I wrap up part III by reviewing how such contributions look like in the example of AgentNet applied to data from a real flock of Chimney Swifts.

When the ML model is trained on data obtained from a target phenomenon of interest (visible as a time evolution of observable states $S_T^{t,m}$, or ”trajectory”), this operator emerges on the basis of ML architecture and input features f_i; It then should be able to predict the time-sequence of states $S_M^{t,m}$ for the whole system. At the bottom layer of **CPMint** (Fig. 10), correspondence is illustrated as an evaluation step. The accuracy of the ML model prediction can be assessed on various levels, but for modeling the dynamical interaction rules, comparing how well $S_M^{t,m}$ matches with $S_T^{t,m}$ beyond the training data is arguably the most important way.[71] Visualisation takes again part in this evaluation process, although only in a passive way, because it does make the predictive accuracy intelligible, which is normally computed in the background. The main power of **CPMint** rests within its emphasis on correspondence that needs to be preserved down to the functional level, as formulated with the *FCH* above. Yet, every claim of correspondence has to be justified.[72] This essentially demands a second dimension of evaluation, which is directed towards a ML model's ability to realise the relevance of ML features for the emergence of the macroscopic phenomenon. However, there are other dimensions of evaluation, not covered in this illustration. These could also include the ML model's ability to extrapolate to unseen dynamics of different systems, or its ”robustness” to variation of initial conditions.[73] I will come back to these aspects once more in the final subsection, where I talk about the fine line between the general and the particular.

The concept illustration (Fig. 10) suggests a flow of reading with different arrow styles. The small dashed arrows on the right denote our normal way of causal thinking, meaning the microscopic structures produce a certain macroscopic phenomenon through organisation and dynamical interaction. The passage of information throughout the modelling process is indicated through black arrows, which starts by collecting data from the observable phenomenon (at the bottom right), progressing to a theoretical description of the individual interacting parts (top right). From there, the first correspondence level is established, and the data is used for training the ML model. Finally, the ML prediction is evaluated against ground truth

[70] Although there is no tight threshold for when a ML is sufficiently accurate in predicting the phenomenon, a plausible way to define it would be through benchmarks. Comparison with the performance of other ML models gives a good sense of a higher bound for this threshold. Arguably, there is no lower bound, since the better accuracy is always preferable - in the predictive context. In the context of combining predictive accuracy with mechanistically viable visualisations, the picture slightly changes, maybe in favour for ML models, which are slightly worse at prediction but much more physically transparent. I do not want to argue for any particular stance here, since it would fall outside the scope of this thesis.

[71] A prominent technique that is widely used is ”cross-validation”, where the whole data is split into a larger training part and the loss of the fully trained model is then computed with a baseline comparison

[72] At the ontological level, this justification is provided by theoretical considerations, at best with reference to already existing empirical knowledge about relevant features of the target phenomenon or the target system. Of course, the ontological link could be justified by the available type of data, which requires certain architectural features.

[73] Another mode to evaluate the performance of a ML model is its ”robustness to intervention”. In a recent paper, it is emphasised that robustness notions in ML point in different conceptual directions, voicing the need for a unified framework. Further, the authors claim that ”robustness is an independent epistemic notion, presupposing the generalization capacities of the model under training conditions, but going beyond that by shifting the focus to the reliable real-world deployment of ML models.” (Freiesleben and Grote 2023) Since it goes beyond the generalisation, I do not wish to go much deeper into this discussion for now.

by means of visualisation, while the same visualisation of epistemically relevant ML features justifies functional correspondence. Note that the "organisational principles" are not connected by these arrows, because we have to infer the dynamical interaction rules based on the established correspondence and ML model transparency. Hence, I put in big colourful arrows that indicate the direction in which this inference step is taken.

In this summary, It should have come clear, how *prediction* and *relevance realisation* work together in order to establish functional correspondence. Furthermore, they can be associated with the two dimensions that appear in mechanisms, thus justifying that the ML model becomes "interpretable in a mechanistic way". The "Extended Link" *between* ML model and target system (Fig. 10), illustrates this strategy to gain insight into the most relevant dynamical interaction rules, governing the emergence of a collective phenomenon. By doing so, **CPMint** provides a strategy for making a ML model contribute to the body of empirical evidence with mechanistic guidance. Hence, it manages to improve upon Emily Sullivan' framework of link-uncertainty, that I used to justify why we need a stronger link connection for mechanistic understanding in the first place. Finally, I would like to highlight the symbiotic nature of this strategy, since key elements of mechanistic modelling are combined with deep neural network (ML) technology to gain insights into causal interactions at the micro level (which physical models are usually good at) while capturing the macroscopic system evolution (which is usually a high performance area of ML algorithms). In the following section, I will look at some limitations of **CPMint** and in addition provide a few perspectives along the same lines.

4.2 Between limitations and perspectives - Three challenges for correspondence

The way I ended up at the final version of **CPMint** makes apparent my commitment to mechanistic modelling. I conceptually decomposed the whole modelling process into individual subunits, while framing each of them in terms of *correspondence.* It highlights crucial aspects of the link between a ML algorithm and a complex target phenomenon, which make this connection epistemically viable for mechanistic investigation. But this recipe is in itself an abstraction; Its limitations need to be addressed and discussed, which, as I believe, is best done by putting the concept back into the context of application. Let me therefore revisit the philosophical foundation that I started off with, since they consequently have a bearing on any possible practical deployment. New Mechanists have acknowledged that there is a "phenomenon-dependence of decomposition", meaning some parts of the system can be relevant for producing one phenomenon, while the same parts could be negligible in the context of another phenomenon. The "system" is often much larger and comprises many more parts than just those that are phenomenologically relevant. Returning to Fig. 1, panel **B** therefore is the more realistic one, although panel **A** is reminding us of the fact that mechanisms can sometimes encompass many layers of organisation. Considering the top layer of **CPMint**, which features correspondence at the level of system ontology, there are some practical and theoretical issues, related to the purpose of application.

4.2.1 Structure and resolution of data

The first and perhaps least severe problem is concerned with the structure of data available. If one wishes to gain mechanistic understanding of how lower-level entities aggregate to produce a higher level phenomenon, the ML model requires some data of those individual entities. If the resolution of data is not high enough, then the ML model needs to provide additional justification that it actually establishes correspondence to non-resolved scales and/or processes. AgentNet delivers an example with its application to interacting AOUP. While being mechanistically interpretable for the phenomenon of clustering - even in the presence of a stochastic noise - the exact dynamics of the *intrinsic* propulsion force of individual particles is left inaccessible. AgentNet cannot deliver any insight into the propulsion mechanism (of the particles), because it essentially treats these "agents" as devoid of internal structure. This arguably becomes a problem, if the framework would be confronted with other kinds of agents that exhibit internal structure, like self-propagating cells. Without doubt, AgentNet can still be a good model for predictive reasons, if their propulsion mechanism is also assumed to be stochastic. Yet, mechanistic interpretability would be lost, since the ML part does not establish correspondence to that scale.[74] To account for such limitations, I put a black arrow into the concept illustration (on the right) that moves from the phenomenon of interest to "data", which then impinges upon the decomposition of the system into parts. I doubt that the authors of AgentNet worked out the AOUP test case in this way; but because **CPMint** is meant to guide ML that purposefully aims at mechanistic understanding, approaching the target system like this could be a game-changer. In perspective, it is plausible to first define the phenomenon of interest, and afterwards consider the kind of data available, leading to an appropriate theoretical decomposition. There is also room for potentially including "external" empirical evidence, which, following any theoretical considerations, initiates a connection of the ML model to the already existing body of empirical knowledge. However, the role of external evidence within **CPMint** is something I did not manage to include into Fig. 10, because it is - as the terminology suggests - external and thus serving as a corrective and/or enrichment for the whole ML model process.

4.2.2 Assumed uniformity

A second issue, probably more severe, has to do with the assumed uniformity of both, the individual agents and the determined target phenomenon as a whole. I developed **CPMint** based on the notion of finite and non-changing parts of a system. Yet, when it comes to more behaviourally diverse agents, whose collective interactions also change over time, this assumption does not hold anymore. Consider birds that do not only flock their whole life, but behave in many different ways. Even within the bounds of the collective flocking, Chimney Swifts for example have been observed to behave differently in different areas of the flock, moving from cooperative to more competitive upon entering the chimney zone.(Evangelista et al. 2017) This does take into account that the whole flocking phenomenon as such is non-stable and changes over time, while the flock itself might also exhibit spatial substructures, in which birds might behave slightly different. As a consequence, establishing ontological and functional correspondence becomes much more difficult. One way to circumvent this problem is to tackle intermediary stages within the time-evolution of a complex system, where the entities can be assumed to behave uniformly, leading to a temporally stable phenomenon. Yet, if one is interested in the collective reaction to an outside influence, the uniformity condition breaks down at the phenomenological level. Ballerini and colleagues give an example by modelling the reaction of a bird flock in the face of a predator attack, comparing interaction rules based on the topological distance to ones based on metric distance.(Ballerini et al. 2008) It is evident that many phenomena feature different *kinds* of agents, working together in a much more diverse way than captured with a simple graph model. Can **CPMint** help alleviate these limitations through uniformity assumptions?

By invoking ontological correspondence by means of a "mechanism sketch", I do not suggest that ML model and target system should correspond to the tiniest detail (which would be rather impractical). Rather, central structural features of agents and their interactions should be represented. **CPMint** grants the possibility to first have a very rough structural correspondence, which treats every agent as uniform, but then allows for differentiation of agents at the functional level. For example, in a study by Charlotte Hemelrijk it is made the claim for situation-dependent interaction rules, which are more biologically plausible, because they take into account behavioural changes along the formation and stabilisation of bird-flock phenomena.(Hemelrijk and Hildenbrandt 2015) One possible outlook for future investigation that would form a natural extension of **CPMint**, is to allow for behavioural change of target agents by working with different architectural blocks within the ML model. Perhaps, endowing them with more freedom to deal with certain environmental disruption through individual strategies might deliver a more realistic picture of an organising mechanism. Another approach to took for in this context is adaptive machine learning, because it is good at dealing with rapidly changing data and thus applicable to more realistic real-world scenarios. Yet, establishing and keeping functional correspondence in such cases might be a bigger challenge, arguably requiring extensive

[74] For a graphical intuition, imagine that the lowest process layer at the bottom of panel **A** in Fig. 1 would be black-boxed, and there is no data about the entities interacting at this level available, it becomes very hard and virtually impossible to gain mechanistic insight into that layer. Arguably, this point is valid independently of a reference to ML.

justification in the face of such changing data. However, the guiding strategy that the extended link-framework can provide is its emphasis on a strong coupling of *prediction* and *relevance realisation*, - two aspects that are likely to play a central role, regardless of higher functional diversity. Furthermore, disentangling these complex adaptive problems into two separate ones, which are centrally linked through the ML evolution operator, can help to see and fulfil the requirements for mechanistic understanding.

4.2.3 Epistemic blindness to feedback and training

The third issue, possibly the most severe, is the fact that many complex systems exhibit feedback effects, as they are well known in the climate system. Especially if there are not only many kinds of agents, but also different scales of organisation involved at once, the analytical concept of a ”mechanism sketch” is no longer tightly connected with ontological correspondence. An informed speculation suggests quite the opposite, since it is unclear how many layers of organisation to include for the ontological correspondence, striving to establish it one or the other way could actually impede the interpretability of the ML model. I highly doubt that **CPMint** in its present form could be employed to guide ML research into complex systems with feedback effects. Perhaps, we would even need to rethink the category of phenomena, because feedback effects often depend on multiple (sub)systems, which might dampen or amplify certain processes, as a consequence of their interconnections. However, the notion of a ”mechanism sketch” could arguably still be of value here, but much more abstract and removed from any correspondence at the ontological level of individual agents; since it is possible that structures at macroscopic scales exert a forcing ”backwards” on micro-level processes. Maybe future investigation can tackle the question on which grounds to establish functional correspondence to investigate such feedback effects. Especially in real life settings, where so many components co-depend in their activity, ”mechanisms often have too many constituents and their behaviours are too dynamically complex for us to even reasonably approximate a complete explanation.”(Glennan 2017) Hence, it is clear that the mechanistic understanding gained through **CPMint** has limits, because it only delivers a partially complete picture of all the requirements that presumably need to be met for such understanding. The whole training process, as one of the practical steps in every ML model setup, has been left completely untouched. based on the assumption that it has purely practical implications, I left the final concept epistemically blind to training. After all, ML is often very experimental, and as to my knowledge, the adjustment of hyperparameters can sometimes be a test of endurance for data scientists, until the model actually learns sufficiently well. Regarding the influence of training procedures on the mechanistic modelling aspects is a topic for future investigation, especially with a focus on how **CPMint** could incorporate such considerations. More insights might be gained by closer analysing Neural Differential Equations (NDEs), because they provide uncertainty modelling through ML that already works on a physically constrained architecture. Going from here could open up more strategies to make a NDE framework mechansitically transparent, for example by trying to connect it with symbolic regression strategies, whereby the latter finds algebraic expressions for the ML operator that is supposed to represent uncertain or non-resolved processes.

4.3 Walking the ridge along the particular and the general

I would like to end this thesis with a philosophical reflection on the balance between the particular and the general, which is at the heart of the underlying symbiotic modelling approach. Once more, the philosophical background of New Mechanism provides a comprehensive account of this balance. On the one hand, for mechanistic understanding we need to specify the phenomenon of interest, which in consequence directs the decomposition of the system into parts; On the other hand, a significant hallmark of a good ML performance is the algorithm's ability to "generalise" (in both ways - beyond the training data, and preferably even beyond the test system). It appears that the requirement for a ML model to be sufficiently "general" is in tension with the demarcation of a clear phenomenon to study, which (viewed through New Mechanism) is more likely to be a "particularity". Therefore, it is essential to consider how common or "general" certain phenomena are across different kinds of systems. Admittedly, this fine balance is very hard to get right. But it also gave me reason to unpack the idea of correspondence around the ML model AgentNet that targets the broader class of agent-based complex systems. As far as AgentNet's ability to generalise goes, I cannot add more than what has already been mentioned by S. Ha and colleagues.[75] Arguably, the ML model itself does not admit of making broader claims about different classes of systems. Thus, I essentially tried to contribute with a recipe that is sufficiently general, while still paying attention to the particular. This philosophical commitment gave me a sense of meaning behind the development of an extended link concept for mechanistic understanding. Analysing a system with respect to mechanisms can only fully unfold its power once the results are put back into a wider context, and thus it requires a constant moving back and forth between the abstract and the actual world. In the previous section, I aimed to demonstrate how **CPMint** relates to this wider context by confronting it with three challenges, which opened up new perspectives along the way.

Throughout this Thesis, the philosophy of the New Mechanism has underpinned the central ideas, and so I would like to be guided by this analytical basis for my final outlook. Let me extend the perspective into the field of cognitive sciences, which is concerned with systems of much higher connectivity, like our brain. Even though a highly integrated biological neural network is not the typical type of "agent-based" system, it does nevertheless relate to them in some ways. It is true that individual neurons could be regarded as "agents"; but one notable difference is the bandwidth of interaction between neurons, compared to the internal information processing of each neuron individually. In this work, I went from systems consisting low-level entities like "cells" in a CA to much more complex agents like birds, who exhibit a very high degree of autonomy, compared to neurons that are confined inside the brain. However, the amount and diversity of connections between a large number of neurons is in no way comparable to that between birds.[76] Although neurons can be regarded as "agents" in a very abstract and non-autonomous way, the main feature of biological neural networks is their high functional integration. In his book "The New Mechanical Philosophy", Stuart Glennan reviews a cognitive model of emotional decision-making, the GAGE model.(Wagar and Thagard 2004) It is supposed to provide hypotheses about how various regions in the brain contribute to emotional cognition - in this sense, it provides "mechanistic hypothesis", as I employed the term. Furthermore, the GAGE model links its entities, as well as their activities and organisation to the same qualities within the target system, thus showing the character of a mechanistic model. The authors further justify their work with empirical studies, which give evidence that these parts of the brain take part in the process of decision-making.[77] Crucial for the GAGE model is the employment of so-called "artificial spiking neurons". As Glennan puts it, "in contrast to the artificial neurons in [...] basic connectionist networks, spiking neurons simulate variations in

[75] More precisely, the authors stress that the generalised system formulation is compatible with the transition functions of various different systems "such as the Monte Carlo simulation of the Ising model, the voter model, systems governed by Newtonian dynamics, and phase space dynamics driven by the Liouville equation." (S. Ha 2021), p.3

[76] It has been referred to this as the "bandwidth criterion", according to which the boundaries between components of a system should be drawn by looking at the kinds of interaction, and not just the overall strength. Only limited kinds of interaction will occur at interfaces between components, or "agents". See Glennan 2017, p.39

[77] On a side-note: Decision-making itself could be regarded as the "emergent" phenomenon that we are observing at the macro-level. Relating it to the picture I have drawn with CPMint, it is the desired target phenomenon, for which we would like to gain mechanistic understanding.

rate and duration of firing that are characteristic of biological neurons”.(Glennan 2017)

Although GAGE does not include any reference to ML, it shows how mechanistic understanding does not require a full-blown account of every micro-detail and therefore it is worth mentioning in the given context. The model highlights once more the importance of correspondence, although not on the ontological level; It is rather focused on functionally relevant features in biological neural networks, for which the ”spiking neurons” are supposed to establish correspondence. If I want to extend **CPMint** to take such highly integrated biological systems into account, ontological correspondence might no longer be of central relevance for mechanistic interpretability at all.[78] What is left to say about the generality of my proposed concept for mechanistic interpretability? A recent PhD thesis, written by Bojian Yin, takes the ideas of ”spiking neural networks” over into the field of Machine Learning, where their mathematical underpinning has been notoriously hard to handle. Yin developed methods to overcome these mathematical difficulties for modelling energy efficient neural networks like those in our brain.(Yin 2022) It makes evident that in the near future, biologically plausible ML methods could be advanced in cognitive sciences - in perspective even ones that grant mechanistic insight into our cognitive capacities. It might even be the case that, analogous to ”physical interpretability”, we see the emergence of *”biological interpretability”*. This requires careful investigation into the different levels of correspondence between the ML model and the target system, and it would be naive to apply **CPMint** in its current form. After all, we are dealing with a completely different system composition consisting of minimally autonomous ”agents” for which it is difficult to find suitable system boundaries. If the concept should continue to be relevant for mechanistic considerations in this territory, it would requires a change of focus and a reorientation of central analytical components within it. For example, it would be necessary to investigate what ”biological interpretability” would mean in the context of ML, further leading to a different prioritisation of ontological and functional correspondence. Again, the mechanistic understanding requires us to walk a fine line between the particular and the general, which is usually a trade-off relation - even more apparent in the cognitive sciences. For example, our brains are very different from monkey brains, and thus more detailed brain models will apply to fewer targets. It might also be the case that we need many models, depending on the phenomenon we want to study. As Catherine Stinson has argued, cognitive models cannot always be continuously filled with mechanistic detail at the neurological level.(Stinson 2016) Stuart Glennan, in my opinion, put it perfectly:

> *Ultimately, the need for many models of mechanisms derives from the variety of purposes we have in modeling, the plurality of phenomena with a complex hierarchy of mechanisms, and the inevitable abstractions and idealisations that help us find generality in a world of mechanisms that are ultimately particular, localised and heterogeneous.*(Glennan 2017)

With this work, I hope to have delivered some good reasons that **CPMint** provides a sufficiently general recipe, which, up to some necessary adjustments, applies even beyond the bounds of agent-based systems, arguably providing mechanistically guided navigation through this zoo of many models.

One aspect, however, has intrigued me since the first time I encountered the ML model AgentNet. Why *attention*? It seems trivial and quite obsolete to inquire into mere linguistic similarity, since in the context of ML, this term, as many would claim, has no deeper connection to what is associated with ”attention” in cognitive psychology. But looking closer, the term implies that there is something that *attends to* something else, which means that ”paying attention” is thus an act of prioritisation. This is true in the psychological sense, as well as in the sense of this particular ML setup of AgentNet, because in a Graph Attention Network, nodes give different priorities to neighbouring nodes - they ”attend to” them, which is an information channel on top of their normal connections (edges). However, it is essentially treated as another function, built into the ML system, whereas in cognitive psychology, ”attention” is treated not just as another function alongside other cognitive functions, but as ”something prior to functions

[78] It might be that it is more of an ”artefact”, due to my concerns with agent-based systems. A rough mechanism sketch by means of feature-correspondence to phenomenologically relevant parts of the system, and the subsequent establishment of functional correspondence, might be the more generally applicable recipe.

and even to things." As Iain McGilchrist unpacks in his book *The Master and his Emissary*, "attention changes *what kind of* a thing comes into being for us: in that way it changes the world."(McGilchrist 2019) He does not refer to an objective reality, but to our experience of it, which lends the term attention some very tight connection to how we perceive - and in consequence how we *understand* the world. I do not seek to invoke all these notions to retain their meaning when applied to ML algorithms; Yet, it is even more fascinating that an "attention mechanism" seems to have so much potential to enable an *understanding* that goes deeper than correlating different features through lots of data. Perhaps, it is "attention" that helps the ML algorithm "see" - not only *what* is most relevant for a phenomenon, but also *in a way* that bears some similarities to how we perceive the world.

In the end, we should keep in mind that we do not perceive the world as "mechanistic". It is just that science has in huge parts come to investigate our cosmos in this particular way. I dealt with the topic of how we could enable mechanistic understanding of a target phenomenon through ML - and I wish to emphasise that such understanding is in every case meant to be a stepping stone along the path to action. What action - what we want to *do* with this knowledge, remains to be seen. We should pay close *attention*, so that hopefully it will be in service of what is meaningful and *relevant* for the well-being of humankind.

References

Anderson, P. W. 1972. “More Is Different.” *Science* 177, no. 4047 (August): 393–396.

Baker, Ruth E, Jose-Maria Pena, Jayaratnam Jayamohan, and Antoine Jérusalem. 2018. “Mechanistic models versus machine learning, a fight worth fighting for the biological community?” *Biology letters* 14 (5): 20170660.

Ballerini, Michele, Nicola Cabibbo, Raphael Candelier, Andrea Cavagna, Evaristo Cisbani, Irene Giardina, Vivien Lecomte, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, et al. 2008. “Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study.” *Proceedings of the national academy of sciences* 105 (4): 1232–1237.

Battaglia, Peter W., Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. 2016. “Interaction Networks for Learning about Objects, Relations and Physics,” 4502–4510.

Bialek, William, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak. 2012. “Statistical mechanics for natural flocks of birds.” *Proceedings of the National Academy of Sciences* 109 (13): 4786–4791.

Bommer, Philine, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M.-C. Höhne. 2023. “Finding the Right XAI Method: A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science,” eprint: arXiv:2303.00652.

Caprini, Lorenzo, Emilio Hernández-García, Cristóbal López, and Umberto Marini Bettolo Marconi. 2019. “A comparative study between two models of active cluster crystals.” *Scientific reports* 9 (1): 16687.

Carlson, Thomas, Erin Goddard, David M Kaplan, Colin Klein, and J Brendan Ritchie. 2018. “Ghosts in machine learning for cognitive neuroscience: Moving from data to theory.” *NeuroImage* 180:88–100.

Cranmer, Miles, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. 2020. “Discovering symbolic models from deep learning with inductive biases.” *Advances in Neural Information Processing Systems* 33:17429–17442.

Craver, Carl F. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience.* Clarendon Press.

Esteva, A., B. Kuprel, R. Nova, J. Ko, S. Swetter, H. Blau, and S. Thrun. 2017. “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.” *Nature* 542:115–118.

Evangelista, Dennis J, Dylan D Ray, Sathish K Raja, and Tyson L Hedrick. 2017. “Three-dimensional trajectories and network analyses of group behaviour within chimney swift flocks during approaches to the roost.” *Proceedings of the Royal Society B: Biological Sciences* 284 (1849): 20162602.

Freiesleben, Timo, and Thomas Grote. 2023. “Beyond generalization: a theory of robustness in machine learning.” *Synthese* 202 (4): 109.

Gardner, Martin. 1970. “MATHEMATICAL GAMES.” *Scientific American* 223 (4): 120–123. ISSN: 00368733, 19467087, accessed February 16, 2024. http://www.jstor.org/stable/24927642.

Giere, Ronald N. 1988. *Explaining science: A cognitive approach.* University of Chicago Press.

Glennan, Stuart. 2017. *The new mechanical philosophy.* Oxford University Press.

Glennan, Stuart, Phyllis Illari, and Erik Weber. 2021. “Six theses on mechanisms and mechanistic science.” *Journal for General Philosophy of Science,* 1–19.

Hemelrijk, Charlotte K, and Hanno Hildenbrandt. 2015. “Diffusion and topological neighbours in flocks of starlings: relating a model to empirical data.” *PLoS One* 10 (5): e0126913.

Heras, Francisco JH, Francisco Romero-Ferrero, Robert C Hinz, and Gonzalo G de Polavieja. 2019. "Deep attention networks reveal the rules of collective motion in zebrafish." *PLoS computational biology* 15 (9): e1007354.

Hwang, Jeehyun, Jeongwhan Choi, Hwangyong Choi, Kookjin Lee, Dongeun Lee, and Noseong Park. 2021. "Climate modeling with neural diffusion equations." In *2021 IEEE International Conference on Data Mining (ICDM),* 230–239. IEEE.

Kaplan, David Michael, and Carl F. Craver. 2011. "The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective." *Philosophy of Science* 78 (4): 601–627.

Kieval, Phillip Hintikka. 2022. "Mapping representational mechanisms with deep neural networks." *Synthese* 200 (3): 196.

Krickel, Beate. 2023. "Different Types of Mechanistic Explanation and Their Ontological Implications." *New Mechanism,* 9.

Ladyman, J., and K. Wiesner. 2020. *What Is a Complex System?* New Haven & London: Yale University Press.

Levin, Simon A. 1992. "The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture." *Ecology* 73 (6): 1943–1967.

Martin, GR. 2010. "Bird collisions: a visual or a perceptual problem." *BOU Proceedings–Climate Change and Birds. http://www. bou. org. uk/bouproc-net/ccb/martin. pdf.*

Martin, GR, and JM Shaw. 2010. "Bird collisions with power lines: failing to see the way ahead?" *Biological Conservation* 143 (11): 2695–2702.

McGilchrist, Iain. 2019. *The master and his emissary: The divided brain and the making of the western world.* Yale University Press.

Miotto, R., L. Li, B. A. Kidd, and J. T. Dudley. 2016. "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records." *Scientific Reports* 6 (26094): 1–10.

S. Ha, H. Jeong. 2021. "Unraveling Hidden Interactions in Complex Systems with Deep Learning." *Scientific Reports* 11, no. 1 (June): 12804. https://doi.org/10.1038/s41598-021-91878-w.

Schelling, T. C. 1971. "Dynamic Models of Segregation." *The Journal of Mathematical Sociology* 1 (2): 143–186.

Seo, Sungyong, and Yan Liu. 2019. "Differentiable physics-informed graph networks." *arXiv preprint arXiv:1902.02950.*

Stinson, Catherine. 2016. "Mechanisms in psychology: Ripping nature at its seams." *Synthese* 193:1585–1614.

Sullivan, Emily. 2022. "Understanding from Machine Learning Models." *British Journal for the Philosophy of Science* 73 (1): 109–133.

Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. "Graph attention networks." *arXiv preprint arXiv:1710.10903.*

Vervaeke, John, Timothy P Lillicrap, and Blake A Richards. 2012. "Relevance realization and the emerging framework in cognitive science." *Journal of Logic and Computation* 22 (1): 79–99.

Wagar, Brandon M, and Paul Thagard. 2004. "Spiking Phineas Gage: a neurocomputational theory of cognitive-affective integration in decision making." *Psychological review* 111 (1): 67.

Wang, Yilun, and Michal Kosinski. 2018. “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.” *Journal of personality and social psychology* 114 (2): 246.

Yin, Bojian. 2022. “Efficient and Accurate Spiking Neural Networks.”

Zanna, Laure, and Thomas Bolton. 2021. “Deep learning of unresolved turbulent ocean processes in climate models.” *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences,* 298–306.