

Non-equilibrium Thermodynamics and the Free Energy Principle in Biology

Matteo Colombo

Department of Philosophy, Tilburg University

Patricia Palacios

Department of Philosophy, University of Salzburg

March 2021

Abstract

According to the free energy principle, life is an “inevitable and emergent property of any (ergodic) random dynamical system at non-equilibrium steady state that possesses a Markov blanket” (Friston 2013). Formulating a principle for the life sciences in terms of concepts from statistical physics, such as *random dynamical system*, *non-equilibrium steady state* and *ergodicity*, places substantial constraints on the theoretical and empirical study of biological systems. Thus far, however, the physics foundations of the free energy principle have received hardly any attention. Here, we start to fill this gap and analyse some of the challenges raised by applications of statistical physics for modelling biological targets. Based on our analysis, we conclude that model-building grounded in the free energy principle exacerbates a trade-off between generality and realism, because of a fundamental mismatch between its physics assumptions and the properties of actual biological targets.

Keywords: Free Energy Principle; Dynamic equilibrium; Homeostasis; Phase space; Ergodicity; Attractor

1 Introduction

Life scientists use the term “equilibrium” in various ways. Sometimes they use it to refer to an inert state of *death*, where the flow of matter and energy through a biological system stops and the system reaches a life-less state of thermodynamic equilibrium (Schrödinger 1944/1992, 69-70). More often, they use it to mean *homeostasis*, which is the ability of keeping some variable in a system constant or within a specific range of values (Bernard 1865; Cannon 1929, 399-400). Some other time, the term “equilibrium” is associated with the concept of *robustness*, which refers to the capacity of systems to dynamically preserve their characteristic *structural and functional stability* amid perturbations due to environmental change, internal noise or genetic variation (Kitano 2004).

Over the past twenty years, theoretical neuroscientist Karl Friston and collaborators have developed an account of the conditions of possibility of a certain kind of *dynamic equilibrium* between a biological system and its environment. The core of this account is a “free energy principle”, according to which all biological systems actively maintain a dynamic equilibrium with their environment by minimizing their free energy, which enables them to avoid a rapid decay into an inert state of thermodynamic equilibrium (Friston 2012, 2013; Ramstead, Badcock, and Friston 2018; Parr and Friston 2019).

The free energy principle has received a lot of attention. But its foundations in statistical physics and dynamical systems theory have not been probed yet. This lack of attention is unfortunate, because the theoretical adequacy of the free energy principle, as well as its practical utility for the study of key biological properties such as homeostasis and robustness depend on the validity of those foundations. Here, we begin to fill this gap.

We start, in Section 2, by introducing Friston and collaborators’ free energy principle (readers familiar with the free energy principle may want to skip this section.) In Section 3, we put into better focus the phenomenon free energy theorists intend to account for, noting that it remains somewhat unclear what this phenomenon exactly is and whether it is a distinctively biological phenomenon. In Section 4, we critically examine the validity and role of the concepts of *phase space*, *ergodicity* and *attractor* in the free energy principle. On the basis of our discussion, we conclude, in Section 5, that, because of a fundamental mismatch between its physics assumptions and properties of its biological targets, model-building grounded in the free energy principle exacerbates a trade-off between generality and biological plausibility (or realism).

As we are going to argue, the foundations in physics concepts allow models built from the free energy principle to achieve maximal generality; but these assumptions also decrease the ability of these models to include enough factors to provide biologically plausible representations of the causal networks responsible for living systems’ dynamic equilibrium. To better appreciate why this trade-off arises and what its epistemic consequences are, we should pay closer attention to fundamental differences between physics and biology and how these differences interact with importing concepts from physics into the modelling of biological targets.¹

¹To forestall any confusion, our focus is on free-energy theorists’ assumptions that all biological systems at any scale can be modelled within the framework of ergodic random dynamical systems, and that homeostasis and/or robustness can be defined and studied in terms of dynamical attractors. While our discussion should suggest that concepts and mathematical representations from statistical physics and dynamical systems theory are sometimes abused in biology (May 2004), it should not suggest that those concepts and representations have no use in the life sciences. In fact, dynamical systems theory has been used to explain various biological phenomena, make predictions about cognitive behaviour, and suggest novel hypotheses and experiments (Brauer and Kribs 2016 Beer 2000; Izhikevich 2007).

2 Introducing the free energy principle

The free energy principle (FEP) says that unless a biological system minimizes its surprise, it will rapidly die (Friston 2012, 2013; Ramstead, Badcock, and Friston 2018).² More specifically, FEP presupposes a view of biological systems as “essentially persists” (Godfrey-Smith 2013; Smith 2017), and foregrounds the conditions of possibility for the persistence of biological systems (Colombo and Wright 2018).

Free energy theorists have formalized their principle by relying on concepts and mathematical representations from statistical physics and random dynamical systems. Parr and Friston, for example, write that “[t]he minimisation of free energy over time ensures that entropy does not increase, thereby enabling biological systems to resist the second law of thermodynamics and their implicit dissipation or decay” (Parr and Friston 2019, 498). Hohwy (2020) refers to a system’s *periodic, phase attractive dynamics* in state space to define what it is for a biological system to exist. He writes: “biological inexistence is marked by a tendency to disperse throughout all possible states in state space (e.g., the system ceases to exist as it decomposes, decays, dissolves, dissipates or dies). In contrast, to exist is to revisit the same states (or their close neighbourhood)” (Hohwy 2020, 3).

Given this definition of biological (in)existence, the problem free energy theorists set out to address is specify the conditions under which a system that is far from thermodynamic equilibrium attains a non-equilibrium steady state. In Parr and Friston’s (2019) words: “if a system is alive, then it must show a form of self-organised, non-equilibrium steady-state that maintains a low entropy probability distribution over its states” (498).³

In order to specify these conditions, free energy theorists relate the notion of *entropy* to the information-theoretic quantity of *surprise*,⁴ and suggest that a biological system’s attaining “homeostasis amounts to the task of keeping the organism within the bounds of its attracting set” (Corcoran, Pezzulo and Hohwy 2020; Friston 2012, 2107).

This suggestion implicates that life scientists can use the mathematics of random dynamical systems to build dynamic models of target biological systems that focus attention on some of the core factors responsible for homeostatic

²By “biological system”, free energy theorists refer to individual organisms, parts of organisms (e.g., their genome or brain) and ensembles of individuals (e.g., species). We use “biological system” in a similar, encompassing way.

³In relation to this idea, it is worth pointing out one of the historical threads of the FEP is W. Ross Ashby’s work in cybernetics (Seth 2014), where one leading hypothesis is that a biological system’s survival can be adequately explained in terms of the stability of a pertinent dynamical system (e.g., Ashby 1956, 19; see Froese and Stewart 2010 for a critical evaluation and refinement of this hypothesis).

⁴Informally, surprise is a measure of the uncertainty of an outcome. Different states of the environment – say, an external temperature of 30 degrees Celsius vs. two degrees Celsius – can generate different outcomes – say, rain vs. snow. Given the state of the environment, some outcomes are more uncertain than others, and so, more surprising. The outcome snow is more surprising than the outcome rain, for example, if the external temperature is 30 degrees Celsius.

processes. By finding a suitable attractor in the dynamic model, life scientists would be able to gain understanding of the workings of homeostasis in the system being modelled. The details of different random dynamical models will vary, depending on idiosyncrasies of particular organisms and the scale at which it is studied. But, in each of these different dynamical models, there should be a set of numerical values, an attractor, that denotes a homeostatic state in the target system; and finding that attractor would depend in all cases on minimizing a free energy functional, which is presumed to be one fundamental underlying constraint common to the dynamics of any biological system. To better understand what this means, let's examine how such idealised models are developed, and why free energy theorists opt to build their models starting from concepts from physics.

2.1 Self-organizing dissipative systems

The starting point of Friston and collaborators (e.g., Friston and Stephan 2007) is that biological systems are kinds of self-organizing dissipative systems. The concept of *dissipative system* can be traced back to another of the historical threads of the FEP, that is, the work by Ilya Prigogine and collaborators. Prigogine and Nicolis (1967) used the term “dissipative system” to refer to processes that take place under far-from-equilibrium conditions and display *order through fluctuations*. Dissipative systems are *self-organizing*, in the sense that local interactions between their micro-components can produce new spatial or temporal structures, or new functions. An example of self-organizing dissipative (non-biological) systems is convection instability, which occurs when a fluid – water in some container, for example – is heated from the bottom and kept at a fixed temperature at the top. If the temperature difference (temperature gradient) is small, then the fluid remains unchanged. However, if the temperature gradient reaches a critical value, then the fluid starts a macroscopic motion with beautiful and well-ordered patterns such as hexagons and rolls (Haken 1983).

According to Turner (1982, 57; see also Goldbeter 2008 and Janson 2012), there are three necessary conditions for self-organization in biological systems: i. the system is *open* to exchange energy and matter with its surroundings; ii. the interactions among the various components of the system are *nonlinear*, meaning that the response of a component receiving inputs from other components or from the environment is not equivalent to the sum of its responses to the individual inputs; and iii. the system operates in *far-from-equilibrium* conditions. Because there are several non-biological systems that satisfy these conditions and present self-organization – including turbulent flows, hurricanes and economies, those three conditions do not suffice to pick out some distinguishing feature of biological self-organization.

In fact, free energy theorists have emphasised that

biological systems are more than simply dissipative self-organising systems. They can negotiate a changing or non-stationary environment in a way that allows them to endure over substantial periods

of time. This means that they avoid phase-transitions that would otherwise change their physical structure.” (Friston and Stephan, 2007, 422)

To clarify the idea that biological systems “can negotiate a changing or non-stationary environment in a way that allows them to endure over substantial periods of time”, it will be useful to introduce a toy example discussed in Friston et al. (2006), namely the example of a winged snowflake. This toy example will help us put into better focus fundamental differences between the domain of physics and biology, and how these differences affect model-building grounded in the FEP.

2.2 The explanatory scope of the FEP. A winged snowflake

Snowflakes are crystals of ice that grow from droplets of water and water vapour. Given the right temperature and air pressure, an initial few water molecules in a cloud will freeze together in ringed unit cells. As the frozen droplets accumulate water vapour from the surrounding cold air, the molecules of water will get fixed in a crystal with increasingly distinct boundaries. Internal to their boundary, snowflakes have some degree of structure, meaning that their component molecules and dynamics have a distinct morphology, and spatial and temporal organization. Because ice growth is sensitive to the temperature, humidity, pressure and other states in the local environment, the growth behaviour and morphology of snowflakes will change over time as they are passively dragged around. After some time, snowflakes will encounter a phase-boundary, at which the temperature in the environment will cause them to lose their boundaries and internal structure, and melt.

Friston et al. (2006, 72; see also Friston and Stephan 2007, 422-3) ask us to imagine a snowflake endowed with wings. This winged snowflake could move, use its wings as solar reflectors or fans, and exchange energy with the environment for a much longer period than we would expect of an inanimate piece of ice to keep going under identical circumstances. Unlike familiar snowflakes, the winged snowflake can choose actions, which are the means by which different outcomes are brought about in different states of the environment. Its wings allow the snowflake to bring about outcomes such as lowering its core temperature in response to an increase in air temperature.

If the winged snowflake has the capacity to “regulate itself with respect to the boundaries of its own viability” (Di Paolo 2005, 430), that is, has the capacity to bring about outcomes that causally contribute to its own survival, then the winged snowflake can behave adaptively. But, the winged snowflake can also make choices that are detrimental to its existence; for example, if it decides to visit a sauna, this choice is likely to cause a rapid loss of its “thermodynamic homeostasis” (Friston et al. 2006, 72). But, if we do not distinguish between different kinds of perturbations, then it is plausible that visiting a sauna will produce not only a loss of “thermodynamic homeostasis” in the winged snowflake, but also a loss of structural and functional integrity,

that is, a loss of robustness: its constituent water molecules will degenerate into less orderly configurations. And the snowflake will eventually melt, losing its own constitutive organization.

In fact, free-energy theorists want to account for both biological homeostasis and biological robustness at the same time, as they often emphasise that to the extent an organism minimizes its expected free energy, it will maintain its structural and functional organization amid change in the environment (Allen and Friston 2018, 2473; Kirchoff 2018).

Both homeostasis and robustness contribute to the dynamic stability of a biological system. But the networks of causes responsible for robustness and homeostasis are different; and the types of interacting factors in such different networks vary, both between different types of biological systems and within the same type of system (Bich et al. 2016). If a model-based representation of stability in biological systems fails to distinguish between self-organizing processes aimed at maintaining approximately normal homeostasis and self-organizing processes aimed at maintaining robustness, and also fails to capture enough factors of the relevant causal structure, but can be applied to many systems in the world, then that representation sacrifices biological realism for increased generality.

Models grounded in the FEP seems to make exactly this trade-off. By relying on a mathematical formalism and assumptions from physics, these models are applicable to any target system that exists in some sense – whether the system is biological or not. But in gaining this generality, their degree of biological realism is minimal, which compromises their explanatory and/or predictive power with respect to actual biological phenomena. Ongoing discussions about the scope of the FEP are aware of this trade-off between generality and realism, and appeal to various properties that might distinguish biological systems from other kinds of systems in a principled way, so as to avoid that the FEP generalizes to all “existing” systems, risking triviality.

Kirchhoff and Froese (2017), for example, suggest that distinctive biological properties such as adaptivity and autopoiesis are not built into the mathematics of the FEP, but should be viewed as added, external constraints to better appreciate the biological significance of FEP. Kitchoff et al. (2018) make a different proposal. They introduce the distinction between “mere active inference” and “adaptive active inference,” where the “key difference... rests upon selecting among different actions based upon deep (temporal) generative models that minimize the free energy expected under different courses of action” (5). This distinction emphasises that, unlike purely physical systems, biological systems can coordinate their actions with their here-and-now sensorimotor state, anticipate possible future states and act to realize these states. Corcoran, Pezzulo, and Hohwy (2020) make a similar suggestion, arguing that biological cognition is distinctively grounded in the ability to make counterfactual (active) inferences.

This work has helped clarify the intended scope of the FEP. Yet, one problem is that the recursive, hierarchically organized behaviour of some non-biological, dissipative systems, such as whirlpools, tornadoes, hurricanes, Benard cells, economies and the Earth’s climate, have also been represented as engaged in

adaptive active inference (e.g., Rubin et al. 2020). And more importantly here, existing discussions of the scope of FEP do not explicitly address the reasons behind the trade-off between maximal generality and minimal realism. To address these reasons, we now turn our attention to the question of how biological systems can mathematically be represented as random dynamical systems.

3 Biological systems as random dynamical systems

In the previous sections, we pointed out that Friston and collaborators use the mathematics of random dynamical systems and statistical physics in order to build model-based representations of biological systems aiming to account for the conditions of possibility of biological persistence. We have also alluded to the active inference models grounded in the FEP (e.g., Friston et al. 2017; Friston et al. 2020).

Active inference models are phase-space representations of biological systems as forming expectations over observable external states and inferring policies (i.e., state-action mappings) that minimize the expected free energy of those states under a generative model in some pre-defined Markov decision process. By minimizing expected free energy, the modelled system would attain a non-equilibrium steady state, and so, it would “maintain”, in some sense, a low entropy probability distribution over its states.⁵

An assumption made by the active inference models developed by free energy theorists is that target biological systems can be represented as random dynamical systems. What does this assumption mean exactly?

In physics, random dynamical systems consist of two elements: i. a model of “noise” formalized by a base flow $\theta(t) : \omega \rightarrow \omega$, which comprises measure-preserving measurable functions for each time $t \in \mathbb{R}$ on a probability space (Ω, \mathcal{B}, P) , where Ω is the sample space, \mathcal{B} is a sigma algebra over Ω and $P : \mathcal{B} \rightarrow [0, 1]$ is a probability measure; ii. and a model of the dynamics on the phase space X affected by the noise formalized as a measurable flow $\varphi : \mathbb{R} \times \Omega \times X \rightarrow X$, which satisfies the co-cycle property: $\varphi(\tau, \theta(\omega)) \circ \varphi(t, \omega) = \varphi(t + \tau, \omega)$. This measurable flow can be interpreted as solutions to the stochastic differential equations governing the dynamics of the system and the noise can be interpreted

⁵Active inference modelling evolves rapidly. As Maxwell Ramstead helpfully reminded us in conversation, up to roughly 2012 the dynamics of internal and active states in active inference models were determined through gradient descent on variational free energy. From around 2012 up until recently, active inference models have been equipped with algorithms for policy selection, which evaluate the average free energy expected under each policy. In their latest work, the active inference community have endowed their models with a recursive expected free energy functional, which enables the models to represent target systems’ “higher-order, counterfactual beliefs,” which are “beliefs” a system has about the “beliefs” it would have as a consequence of action. Because these modelling advances all seem to make assumptions about the ergodicity of biological systems and the biological meaning of dynamical attractors, the points we make in what follows may help explain why all active inference models thus far tend to trade off realism for maximal generality.

as the environment in which the system is immersed.

In order to motivate and define the FEP, Friston (2012)⁶ allows for a partition of the state space $X = R \times S$, where $R \subset X$ precludes direct dependency on the base flow and, in this sense, constitutes an internal state space. On the other hand, $S \subset X$ constitutes an external state space. This means that there exists a Markov blanket that separates two other sets in a statistical sense.⁷ Internal states have a dynamics that depends on external states, on their relationships with other internal states and on internal noise, while external states have a dynamics that depends on internal “active” states, on their causal relationships with other environmental states, and on fluctuations in the environment.

According to Friston (2012), this partition applies to all biological systems, in which external states can be regarded as causal influences on the sensory receptors, while internal states as effects of sensory input. In a single cell, for instance, external states can be identified with the causal influences of, e.g., ambient temperature on the cell’s trans-membrane receptors. External states generate sensory samples (aka sensory inputs), which can be identified as energy arrays impinging on the cell’s sensory surfaces. These energy arrays influence the state of the cell’s transmembrane receptors, and have causal consequences on the internal state of the cell, for example on the concentration of intracellular metabolites. Crucially, internal states include *active states* denoted by $a(t) \in A \subset R$, which control how environmental fluctuations are sampled by sensory states. An example of an active state in a single cell is the motion of flagella, which can change the position of the cell in its surrounding environment, and so, how the cell’s environment influences the cell’s receptors.

Given this set-up, free energy theorists are interested in finding the active states of those systems that are confined to a bounded subset of states and remain there indefinitely. Such active internal states in an organism can allegedly be represented as random dynamical attractors in a dynamical model, which are defined as a random compact set $A(\omega) \in X$ that is invariant under the flow map such that $\phi(t, \omega)(A(\omega)) = A(\theta_t(\omega))$. If one assumes that the dynamics of the target organism is ergodic, then random dynamical attractors can be associated with the *ergodic density* $p(x|m)$, which is proportional to the amount of time

⁶In what follows, we concentrate on Friston’s (2012) and (2013) treatments of the FEP, because of the transparency of their exposition and explicit focus on all biological systems.

⁷More precisely, take a Bayesian network, which includes a set of random variables and their probabilistic conditional (in)dependencies. The Markov blanket of a target random variable X in that network is a subset of the network that contains all random variables that screen off X from all the other variables in the network. Fixing the values of the variables in the Markov blanket leaves X conditionally independent of all other random variables. This means the Markov Blanket of X suffices to infer the value of X . In philosophy, the notion of a Markov Blanket has generated some discussion. Is there a Markov Blanket that picks out the mind? Is the network of autopoietic processes in a biological system identical to the Markov Blanket of that system? What’s the metaphysical status of Markov blankets? These debates in the metaphysics of the FEP can be left on the side, since we are not interested in metaphysics here.

each state is occupied by the organism:

$$p(x|m) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \delta(x - x(t)) dt. \quad (1)$$

In other words, the ergodic density is an invariant probability measure that can be interpreted as the probability of finding the target system m in any state x when observed at a random point in time (Friston 2012). The assumption of ergodicity is important to get the FEP off the ground, since it ensures that organisms can be modelled as having invariant characteristics over time.

Next, Friston (2012) shows that the entropy H of the ergodic density of a random dynamical system m is upper bounded by the Lebesgue measure of its attractor (if it exists):

$$H(X|m) \leq \ln \lambda(A(\omega)) \quad (2)$$

This means that if the measure is small, then the entropy must be smaller.

Friston (2012) then uses this result to give a more precise definition of *active states*: Internal states are said to be *active*, if they “minimize the entropy of the ergodic density over external states” (2108).

The a priori rationale for this definition is that biological systems that do not minimize their informational entropy (and therefore their surprisal) would be unlikely to exist. After all, if the measure of the random dynamical attractor of a dynamical system representing an organism can be arbitrarily large, then the time required to achieve asymptotic stable behaviour can be arbitrarily long. And if the asymptotic stable behaviour of the dynamical model denotes homeostasis or robustness in the biological target, then the target could not feasibly attain homeostasis.

With this definition of active states in hand, Friston (2012) finally formulates the principle of least action as follows:

Principle of least action: The internal states of an active system minimize surprisal L , such that the variation δS of action S with respect to its internal states $r(t) \in R$ vanishes (2108).

The FEP would suffice to derive the principle of least action, and is defined by Friston (2012, 2109) as follows:

The free energy principle: Let $m = (\mathbb{R}^d, \phi)$ be an ergodic random dynamical system with state space $X = R \times S \in \mathbb{R}^d$. If the internal states $r(t) \in R$ minimize free energy, then the system conforms to the principle of least action and is an active system.

Bearing out the ambiguity we noted above about the intended explanatory target of the FEP (i.e., Is it homeostasis? Is it robustness? Is it biological persistence? Is it existence?) Friston (2013) explains that “active states [...] will appear to place a [free energy] upper bound on the dispersion [entropy] of biological states. This homeostasis is informed by internal states, which means

that active states will appear to maintain the structural and functional integrity of biological states.” (5)

When expressed in these terms, the FEP provides us with an account of biological persistence broadly understood, as a generic type of stability attained by targets modelled as random dynamical systems. To illustrate the FEP at work, Friston (2012, 2013) discusses a number of computer simulations. But, as he recognizes: “The examples presented above are provided as proof of principle and are as simple as possible. An interesting challenge now will be to simulate the emergence of multicellular structures using more realistic models with a greater (and empirically grounded) heterogeneity and formal structure” (Friston 2013, 11). Let’s then put into better focus some of the difficulties involved in this challenge, and ask: Is there any reason to believe that all biological systems, at any scale, can be modelled as random dynamical systems? And how does this maximal generality bear on the degree of realism of these models?

4 Modelling biological active states as random dynamical attractors

Free energy theorists’ account of biological systems’ persistence relies on three main modelling assumptions: i. ergodicity, ii. the existence of Markov blankets that imply a partition of states into internal and external, iii. the existence of random dynamical attractors. In this section, we concentrate on assumptions i. and iii., and on the more fundamental challenge of defining phase spaces for target biological systems in the life sciences.

4.1 Phase spaces in biology

Dynamical systems theory tracks the behaviour of one or more variables over time (occasionally other variables). This framework has proven to be useful in the construction of many mathematical models in biology. A popular example of the successful use of dynamical modelling in the life sciences is the Lotka-Volterra model usually applied to account for the dynamics of biological systems where two species interact: predator and prey (Matsuda et al. 1992).

In order to apply dynamical systems theory and assign probabilities to the different states of a target biological system, one first needs to define a phase space for that target. In dynamical system theory, the phase space of a system is a ‘giant’ space formed by the relevant degrees of freedom of the target system. A point in the phase space of a target system corresponds to a micro-state, which consists in a complete determination of the system in terms of the variables and parameters required for analyzing its dynamics. The dynamics of the system is determined by its equations of motion, which describe the evolution of points, or trajectories, in phase space.

In order to assign probabilities to the micro-states of a system, the phase space is usually coarse-grained into a set of macro-states of the target system that supervene on its fine-grained micro-states. The latter means that

any change in the macro-state must be accompanied by a change in the fine-grained micro-state, and correspondingly that to any given micro-state there corresponds *exactly one* macro-state (Frigg 2008). For example, macro-states of a system of water vapour include its pressure, volume and temperature. For each of these macro-states, there is some set of micro-states of the system, for example the velocity and position of its water molecules, that fix the macro-state.

In statistical physics, the formulation of a pertinent phase space for a given kind of system, the assignment of probabilities to the different micro-states and macro-states of the system, and the selection of the criteria for the coarse-graining of the phase space are non-trivial tasks. They require background knowledge about some features of the causal structure of the target system and its laws of motion; they also require to choose a relevant set of observables and symmetries, which determine what properties of the system being modelled are preserved (or remain invariant) under some transformation (Espanol 2004; Robertson 2020).

In the domain of biology, these challenges are compounded by at least three facts: first, usually a biological system has many more degrees of freedom than non-biological systems; second, the symmetries (or invariant preserving transformations) underlying observed regularities in biological phenomena are more unstable than in physics; third, the phase space for many real-world biological targets is much less stable than that of the kinds of target systems studied in physics.

The degrees of freedom of a target system are the number of independent factors that define its state at any point in time. Biological systems usually have more degrees of freedom than non-biological ones. In some cases, the number of their degrees of freedom can be so large that it can be unfeasible to construct a phase space that takes account of all those causal features of the system modelled that have a genuine influence on the behaviour of interest.⁸

Kauffman (1993) develops various computer simulations of dynamic, Boolean networks of molecules defined in an idealized “protein space”. While this idealized protein space trades off biological realism for computational tractability, it allows us to evaluate whether or not evolutionary processes of natural selection are required to produce auto-catalytic cycles of peptides exhibiting autopoietic, self-producing features characteristic of living systems. Kauffman (1993) identifies general molecular regimes under which an auto-catalytic system can emerge, showing that it is possible that a system of molecules spontaneously becomes to display ordered macroscopic behaviour, without the cumulative selection of

⁸One might try to construe the a phase space as a genotype space for a type of biological system. Behind this approach is the assumption that the phenotype of a biological system corresponds to its coarse-grained macroscopic states (e.g., cells, organs, tissues, agent-ecosystem interactions) and the genotype to fine-grained microstates. However, the size of the genotype space Ω for a protein consisting of N residues chosen from an alphabet of amino acids of size m is $\Omega = m^N$, which can be an enormous number. To give a better idea of the size of this phase space, in the case of a 100-residue bacterial protein, the population size of possible sequences with 20 amino acids to choose at each residue is $20^{100} \approx 10^{130}$ (McLeish 2015, 3; Kaufmann 1993, Ch. 2).

many, smaller increases in order. By constructing and analysing an idealized phase space for some generic biological system, and exploring the behaviour of this system with computer simulation, one can learn whether or not some causal factors are always necessary to give rise to the phenomenon of interest. This model-based representation sacrifices biological realism for tractability and generality. But, given Kauffman’s purpose, the generality of its models does not result in an oversimplification or distortion of target causal networks, and thus preserves some explanatory and predictive power.

A second fact that makes the construction of phase spaces in biology more challenging than in physics is that biological phenomena are characterised by a continuous breaking of symmetries (Longo, Montévil and Kauffman 2012). Typically in physics, the key observables, such as energy, momentum and charge, are the invariants of relevant mathematical representations. Such invariants allow modellers in the domain physics to construct phase spaces for “generic objects [that] will follow a specific trajectory determined by its invariants obtained by calculus” (Longo and Montévil 2014, 128). If the domain of biology does not present sufficiently stable symmetries or invariances, then one may not be able to rely on observable biological properties to construct and analyse the phase space of a given target system.

Furthermore, typically in statistical physics, symmetries or invariant preserving transformations enable modellers to assign probabilities to the different macro-states of a system. An important assumption in Boltzmann’s statistical mechanics is the so-called “combinatorial argument”, which says that in a partition of the phase space into cells, the macro-states are determined only by the number of particles in each cell and not by the specification of which individual particle is in which cell. This means that there is a symmetry or invariant transformation under permutation, and so, macro-states with the same number of particles in each cell will correspond to the very same macro-state.

In biology, symmetries or invariants are sometimes insufficiently stable to allow for an empirically informed attribution of probabilities to the macro-states. One reason is that different levels of biological organization affect each other both in an “upward” and “downward” manner at various temporal scales (e.g., Okasha 2012; Bechtel 2017). For instance, methylation, i.e. the addition of a methyl group on a substrate, can downwards modify the expression of genes (Longo and Montévil 2014, 198). Furthermore, small changes at the molecular scale may be amplified by cell to cell interaction and affect the phenotype, whose changes may downwards affect tissues and even entities at the molecular scale itself. This inter-level dependency compounds the challenge of establishing which changes at the genotype (or microscopic level) are causally irrelevant or relevant for the phenotype (or macroscopic level). In default of background empirical knowledge of lawful relationships between observables at different levels of biological organization, it can then become even more challenging to assign probabilities to the macro-states of large and complex target systems.⁹

⁹Parr et al. (2019) argue that any (weakly mixing) random dynamical system that possesses a Markov blanket will be equipped with an information geometry, which permits defining a

A third fact that makes the construction of phase spaces in biology harder than in physics is that phase spaces for biological systems can themselves change persistently if they are to capture the continual symmetry changes in these systems. This is particularly salient in biological processes studied in evolutionary biology. Longo, Montévil and Kauffman (2012) point out, for instance, that while in physics one can “pre-state” the phase space for target systems based on stable invariances and symmetries, historical processes studied principally in evolutionary and population biology (e.g., processes of mutation, migration, developmental plasticity and modification of gene regulatory processes, inclusive inheritance, niche construction, and so on) involve symmetry breaking, which makes phase spaces structurally unstable, ever-changing and unpredictable.

Because we cannot prestate the ever changing phase space of biological evolution, we have no settled relations by which we can write down the “equations of motion” of the ever new biologically “relevant observables and parameters”, but that we cannot prestate. More, we cannot prestate the adaptive “niche” as a boundary condition, so could not integrate the equations of motion even were we were to have them. (Longo et al. 2012, 1379).

The three challenges just discussed cast doubt on the possibility of modeling most biological systems in terms of random dynamical attractors, and therefore on the possibility of formulating a principle for the life sciences based on this assumption, that is at the same time biologically plausible and general. The generality of the FEP would come at the cost of not including enough biological factors to provide plausible representations of causal structures responsible for persistence in relevant biological targets.

But let us grant for the moment that this challenge is successfully met regarding the feasibility of constructing the phase space of any biological system of interest. Let us also assume that we know the equations of motion governing the dynamics of all biological systems, which would enable modellers to assign probabilities to different states of an organism of interest. There still remains the question of whether this phase space representation has the appropriate properties that allow one to define a random dynamical model; and there still remains the question of whether the attractors in the model meaningfully correspond to the homeostatic states of real-world biological systems. We discuss these two issues next.

4.2 Ergodicity

One of the assumptions of the FEP is that all biological systems are ergodic. Ergodicity (or more precisely metric transitivity) in random dynamical systems means that there exists a probability measure P on X (phase space), such that

phase space and assigning probabilities to the relevant states. However, in order to have an information geometry for all biological systems, one should be able to model them as a (random) dynamical systems, which is precisely what we are calling into doubt in this section.

for almost each $x \in X$, the flow $\varphi(t, \omega)$ converges weakly to P in the asymptotic limit $t \rightarrow \infty$ (Scheutzow 2007). This means that there is an ergodic density $p(x|m)$ representing the probability to find the system (m) in a particular state, which is proportional to the amount of time each state is occupied (Friston 2012, 13). Formally, the ergodic density can be expressed as follows:

$$p(x|m) = \lim_{T \rightarrow \infty} \int_0^T \delta(x - x(t)) dt$$

From the previous expression, it is clear that, in order to warrant ergodicity assumptions, one has to demonstrate that the infinite time limit ($T \rightarrow \infty$) exists, which means that, for long times, the values of the observables associated with the measurable flow φ will converge towards the asymptotic (stationary) values.¹⁰

Birkhoff (1931) demonstrated that if a dynamical system is ergodic, then the infinite time limit exists and coincides with the phase average for almost all initial conditions $x \in X$. This is called the “ergodic theorem”. Friston (2012, 2018) relies on this theorem to prove the “principle of least action” discussed in the previous section, which states that the internal states of an active system minimize surprisal. “The proof [of the principle of least action] is straightforward and rests on noting that action and the entropy of the ergodic density over external states are related via the ergodic theorem [Birkhoff 1931]” (Friston 2012, 2108).

However, the application of the ergodic theorem to real systems, whether in physics or biology, presents several challenges (Earman and Rédei 1998), and there are good reasons to think that this theorem cannot realistically be applied in the domain of biology. One reason is that, in order to make the theorem meaningful for a target system, it does not suffice to prove that the infinite limit exists. One has to show that the convergence rate, i.e. time needed to attain the time average, is also plausible for that target (Palacios 2018). In situations in which the values of the functions describing the target’s behaviour are constantly changing, this time can be of the order of the recurrence time, which even in simple models, such as a small sample of diluted hydrogen gas, can be much larger than the age of the universe (Gallavotti 1999).

Compared to a non-biological system, it is harder, if not unfeasible, to estimate a characteristic time needed to attain asymptotic values in most biological systems, because of non-uniformities in its dynamics. For instance, Mendell and Seltz (1990) show that the dynamics of a biological system associated with distinct phenotypes can a-periodically oscillate between a small number of states in time scales that range from minutes, to days, to years. This is referred to as the “production of distinct alternative phenotypes” (Mendell and Seltz 1990). The dynamics may also suddenly escape a parameter region of previous stability and gouge out an equilibrium with the emergence of new structural phenomena (e.g.

¹⁰The assumption of ergodicity also implies that eventually the system will visit almost every measurable region in X .

peramorphosis, or speciation). Furthermore, invasion, disturbance, evolution, species movement and other fluctuating resources can all destabilize the values of the observables, and prevent asymptotic behaviour. As Drake et al (2006) put it, “asymptotic behavior is seldom realized in the real world because nature happens” (168).

Another reason why ergodicity cannot plausibly be assumed in biology is that the ergodic theorem requires the *dynamics* of the system to be ergodic, which means that, eventually, almost every point $x \in X$ will visit every measurable region in X . Proving that ergodic systems exist has been a challenging enterprise in statistical physics, and even if there have been some positive results stating the existence of concrete examples of ergodic systems (e.g., Sinai 1970), it is widely recognized that most of the systems studied in statistical mechanics are most likely non ergodic (Wightman 1985; Earman and Rédei 1996; van Lith 2001).

Do we have any positive reason to believe some biological systems are actually ergodic? *Prima facie*, we don't, especially if we interpret ergodicity in biology as the assumption that any (complex) phenotype is the result of a random exploration of all possible molecular combinations along a path that will (eventually) explore all molecular possibilities. There are studies suggesting that no biological system can be ergodic in this sense. For instance, Kauffman (2002) argues that the universe will not make all possible proteins of length 200 amino acids in 10^{39} times its lifetime. So, proteins' organization into a characteristic phenotype cannot be the result of the ergodicity of physical dynamics. The same conclusion applies to cells, since the presence of a membrane in a cell canalizes the whole cellular activities along a very restricted form of possible dynamics. Furthermore, the same conclusion applies to any biological system, since any biological system is an historical entity, where processes of differential selection and retention, at either developmental or evolutionary time scales, exclude most of the paths the system might take, which means that the system prevents ergodic exploration.

Friston and collaborators recognize that adaptivity may restrict ergodic exploration in biological systems. However, instead of concluding that biological systems are not ergodic, they use this restriction to motivate a definition of *active systems* as ergodic random dynamical systems that minimize surprise:

The motivation for Definition 1 [active systems] is simple: systems that do not minimize their entropy [i.e., surprise] are unlikely to exist, in the sense that the measure of their random dynamical attractors $A(\omega)$ can be arbitrarily large and their recurrence times arbitrarily long. (Friston 2012, 2108)

In subsequent work, Ramstead et al. (2018) talk of “local ergodicity.” They define this notion informally, as the idea that “all living systems revisit a bounded set of states repeatedly (i.e., they are locally ergodic)” and appeal to intuitive examples to illustrate it. The problem here is that this informal notion of “local ergodicity” may not suffice to prove the Principle of Least Ac-

tion, and, in any case, it is controversial if it picks out a genuine property of biological systems being modelled.

A different strategy is to rely on an analogous theorem called “Multiplicative Ergodic Theorem,” which can be used in differentiable random dynamical systems to prove the existence of random dynamical attractors (Engel and Kuehn 2019). The problem here is that this notion of ergodicity implies that the base flow θ preserves the probability P , which means that the model of noise (environmental fluctuations) is fixed and the probability measure μ is invariant (Engel and Kuehn 2019). In other words, in ergodic random dynamical systems, there is a one-to-one correspondence between invariant random measures (noise) and stationary measures of the associated stochastic process (dynamics of the system). The problem is that the invariance of the probability measure μ and the existence of stationary measures in biological systems are very hard to prove, because, as we explained in Section 4.1, biological systems continuously change, species evolve and interact with other species and with different types of environments modifying the dynamics and sometimes the phase space itself (Wright 1932, Van Valen 1973, Drake et al. 2006).

The assumption of ergodicity to give a definition of equilibrium states is a controversial assumption even in the domain of physics, especially due to existence of physical systems in equilibrium that have proven to be non-ergodic (Palacios 2021). This is the reason why physicists and philosophers have offered alternative approaches to equilibrium that do not rely on ergodicity (e.g. Lebowitz 1993; Goldstein and Wallace 2016). Even philosophers that defend a “quasi-ergodic approach to equilibrium”, recognize that this definition of equilibrium may plausibly apply only to a restricted class of systems such as gases. Frigg and Werndl (2013) say, for instance:

Two scenarios seem possible. The first is that epsilon-ergodicity will turn out to be a special case of a (yet unidentified) more general dynamical property that all systems that behave TD-like possess. In this case epsilon-ergodicity turns out to be part of a general explanatory scheme. The other scenario is that there is no such property and the best we come up with is a (potentially long) list with different dynamical properties that explain TD-like behaviour in different cases. But nature turning out to be disunified in this way would be no reason to declare explanatory bankruptcy: ‘local’ explanations are explanations nonetheless!

In this section, we have argued that there is less reason to believe that ergodicity plausibly characterises any biological systems. This does not mean of course that ergodicity, or a weaker notion of “local ergodicity”, may sometimes be a reasonable assumption to make for some biological systems. Indeed, there are some attempts to demonstrate approximate ergodic behaviour in a certain class of biological systems and this can even serve to give local explanations of homeostatic states, analogously as in physics (McLeish 2015). However, the fact that one cannot plausibly extrapolate the assumption of ergodicity to the

behaviour of all relevant systems in biology casts doubt on the project of formulating a principle that depends on this assumption, and is both maximally general and biologically realistic.

4.3 Random dynamical attractors

We have just explained that ergodic assumptions are generally not easy to prove, and why they are implausible for biological systems. This is problematic, because free energy theorists use ergodicity to define active steady states as random dynamical attractors. What makes the application of the FEP to accounting for the behaviour of real-world biological systems even more challenging is that one also needs: (i) to prove the existence of an attractor in the relevant dynamical model of the target (Scheutzow 2007, 331), and (ii) justify why certain attractors but not others (if they exist) plausibly denote homeostatic states in the target.

Informally, attractors can be understood as solutions to the dynamics towards which the system evolves over time. For instance, the familiar carrying capacity (K) of a population growing logarithmically in continuous time is an attractor of the dynamics. The set of all initial conditions flowing towards an attractor is called the basin of attraction, which can lead exclusively to a fixed point (like a carrying capacity), or to a repeating sequence of states, which may exhibit complicated and itinerant behaviour (Ruelle and Takens 1971). Because attractors are invariant to the dynamics, they correspond to asymptotic solutions.¹¹

As we mentioned in Section 3, random dynamical attractors are specifically defined as a random compact set $A(\omega) \in X$ that is invariant under the flow map such that $\phi(t, \omega)(A(\omega)) = A(\theta_t(\omega))$.¹²

Finding biologically interesting attractors in a dynamic model of a target organism is distinctively challenging compared to finding attractors in models of non-biological systems. One reason is that various processes at different spatial and temporal scales – including, plasticity, development, invasion, evolution, species movement and fluctuating resources – eliminate, reshape, and replace the attractors associated with target biological systems altogether (Drake et al. 2006). For instance, in the model of population growth proposed by May (1976), if one assumes that two systems originally identical interact at different points during the evolution they may end up exhibiting different attractors representing phenotypic differences (Drake et al. 2006, 170). Differences among replicate source communities may develop suddenly, because of sensitive dependence upon

¹¹Kaufmann (2016, 48ff) distinguishes the main types of attractors of nonlinear dynamical systems into four types. One type includes stable steady states and unstable steady states. A second type is a “limit cycle.” A third type of attractor winds through and around a torus. The fourth main type of attractor Kaufmann discusses is “chaotic attractors.”

¹²In the ergodic theory, this means that there is a unique and stationary *ergodicity density* $p(x|m)$, where m represents the dynamical system, which is proportional to the amount of time each state is occupied. Free energy theorists interpret this density in either of two ways: as a description of the flow of a system’s state, or as a joint probability distribution over all the system’s variables and (implicitly) external states.

small variations in their interactions with other systems or with the environment. More generally, the interaction of a system with other systems may change the dynamics, and whether or not an interaction develops between systems is contingent on many factors. A small contingent event that permits or prohibits colonization by some species may act as a dynamical switch, which may prevent the evolution towards an attractor or change the value of the attractor. When the result of coupling is a new system, extinctions can occur. As Drake et al. (2006) put it: “[a]ttractors can dynamically break, after which may be no link, trajectory or solution from the original attractor or parent to the new or child attractor” (179).

Another reason is that a biological system can possess several attractors, but only a subset of them can plausibly be interpreted in terms of a homeostatic state, or some other biologically interesting property. In theoretical models of self-organization in biology (e.g., Nicolis and Prigogine 1977), the non-linearity of the equations governing the evolution of a system allows for the co-existence of multiple (kinds of) attractors. While the presence and functional significance of two stable steady states has been validated experimentally in various biological systems at molecular and cellular scales (Goldbeter 2018, Sec. 2), for most other biological systems at higher scales it remains theoretically and empirically unclear whether the co-existence of two or more attractors should be interpreted as a form of functional organization where an actual living system can operate in one or more homeostatic states.

For example, building on developmental biologist Conrad Waddington’s ideas of “epigenetic landscapes” (Goldberg, Allis and Bernstein 2007), gene regulatory networks have been represented as far-from-equilibrium dynamical systems (Kauffman 1971; Karlebach and Shamir 2008). Depending on their initial conditions, internal constraints, and local molecular relationships of inhibition or activation, these regulatory networks can evolve into different possible attractors, which are taken to represent distinct types of cell phenotypes individuated as arrays of mRNA transcripts (Huang et al. 2005). As noted by Huang, Ernerberg and Kauffman (2009, 871-3), some gene regulatory networks can produce complex landscapes with many attractors. Some of these attractors correspond to observed gene expression patterns (i.e., to actual cell types), including cancer cells. But, most of them do not correspond to any actual cell type; and it is currently unclear which attractors in a complex epigenetic landscape correspond to viable gene expression patterns, or more generally, which attractors in a dynamical model of a target have biological significance.

Increasing the biological plausibility of attractors in some dynamical model generally involves reducing the generality of the model. This increase in plausibility generally relies on knowledge of the historical and environmental contingencies that influence particular target systems and on experimental data about relevant causal mechanisms that constrain model-building (Brigandt 2013; Goldbeter 2018). For example, studies of the processes of canalization in embryos of the fruit fly *Drosophila melanogaster* have employed dynamical systems models of a network of genes known to be involved in generating the basic body plan of this organism. Validated and tested with high-precision empirical data about

gene expression in *Drosophila melanogaster*, these dynamical models indicate that coupled chemical reactions with multiple attractors can explain two properties of canalized developmental systems, namely: their discrete and buffered responses to external perturbations (Manu et al. 2009a, and Manu et al. 2009b).

A type of attractor that can potentially serve to plausibly understand the dynamic and homeostatic equilibrium of at least some biological systems is the Milnor attractor, which has the peculiar property of being asymptotically unstable (Drake et al. 2006). An advantage of Milnor attractors is that they can account for cases in which trajectories visit attractor after attractor exhibiting metastable “equilibrium” behaviour that is far from stable (thermodynamic) equilibrium. The latter is called “switching”, and can plausibly capture phenomena exhibited by biological systems, such as binocular rivalry (see, e.g., Friston, Breakspear and Deco 2012 for an application grounded in the FEP). Another advantage of Milnor attractors is that they allow modellers to take small fluctuations into account and they are compatible with situations in which the trajectory of the system visits regions of phase space that lay outside of the original attractor basin of the system. Finally, Milnor attractors have the advantage of being compatible with the itinerant (wandering) dynamics, which is an ubiquitous feature of self-organizing systems, whether in physics or biology. However, despite these attractions, in order to give a more general characterization of homeostatic equilibrium in biology in terms of Milnor attractors, considerable theoretical and experimental work remains to be done, which could bridge attractors in abstract dynamical systems models and measurable observables in real-world biological systems and it is not clear whether this generality can be achieved.

In summary, making a convincing case for the biological relevance of attractors found in dynamical models of target biological systems requires taking into account enough of the relevant causal factors of particular systems that can account for their homeostasis. The greater the number of the causal factors included in a dynamical model, the more realistic the model. But this increase in realism decreases the scope of the generalizations the model allows.

5 Discussion

In the previous sections, we have analysed three challenges involved in the justification of the FEP, namely: the problem of pre-stating phase spaces for most biological systems, the lack of theoretical and empirical warrant for making ergodicity assumptions in biology, and the challenge of identifying, in general, homeostatic states with attractors in the phase space.

Based on our analysis, one overall conclusion is that, because of a fundamental mismatch between its physics assumptions and properties of biological targets, model-building grounded in the FEP achieves maximal generality for minimal biological plausibility. Its foundations in concepts and mathematical representations from physics allow free energy theorists to build models that are applicable to theoretically any biological system. But, achieving this gen-

erality comes at the cost of minimal biological realism, as those models fail to accurately capture any real-world factor for most biological systems.

While this conclusion coheres with other analyses of trade-offs in model building in various fields of biology (Levins 1966; Matthewson 2011; Elliott-Graves 2018), one initial objection to our conclusion is that, at best, we have shown that the FEP makes false idealizing assumptions, such as ergodic assumptions. But idealization is pervasive throughout science, and allows free energy theorists to focus attention on only the core causal factors responsible for the phenomenon of interest (see, e.g., Batterman 2001; Weisberg 2013; Potochnik 2017). The FEP, and, more generally, any physics approach that models biological systems as ergodic far-from-equilibrium random dynamical systems possessing a non-equilibrium steady state are nothing special in these respects. Like for any other idealized scientific representation, even if the FEP involves simplifying distortions, it does not follow that it must be minimally realistic, and have limited explanatory or predictive power.

The problem with this initial objection is that it is actually controversial if models grounded in the FEP take account of *any* of the core causal factors that give rise to homeostasis (or robustness) in real-world biological targets. Consider the Ising model. This is a mathematical model originally created to explain ferromagnetism in statistical mechanics, which represents atoms or other physical particles as points along a linear lattice, and allows these points to be in one of two states (i.e., +1 or -1). Although the Ising model is very simple and includes almost no realistic feature of its target systems, it has been fruitfully used to explain phase transitions in a variety of systems, by capturing *some* of the core interactions that can give rise to them. For example, it has been used to generate explanations of how a large number of simple components (like neurons) can acquire abilities such as memory (Hopfield 1982), and even the occurrence of stock market crashes (Sornette 2003). It is in virtue of, and not despite, its simplified distortions that the Ising model plays an explanatory role in highlighting a minimal set of factors that suffice to explain how a system can undergo a phase transition between an ordered and a disordered phase in two dimensions. Furthermore, the explanations generated by the Ising model can be evaluated for their explanatory adequacy on the basis of relevant empirical data, for example on the basis of measurements of the activity of an ensemble of neurons on some relevant memory task or to associate the occurrence of stock market crashes with the collective interaction between traders (Jhun, Palacios and Weatherall 2018).

We mentioned the Lotka-Volterra model above. Physicist Vito Volterra used his dynamical model of predator-prey interactions to fit observations about the fish catches in the Adriatic Sea. Alan Hodgkin and Andrew Huxley developed and used their model of the action potential, which consists of a set of non-linear differential equations, to fit recordings of action potentials in the squid giant axon. Most relevant here, non-equilibrium models of cellular processes are developed and used to fit data about the detailed balance within a cell (e.g., Fakhri et al. 2014; Battle et al. 2016).

However, if the FEP or the active inference models it grounds make sim-

plifying, distorting assumptions about the phase space of a target system, the ergodicity of biological systems and the existence of an attracting set corresponding to homeostatic states for these systems, then these idealizations should earn their keep. These idealizations should allow life scientists to tractably draw some explanation and prediction about relevant biological observables in real-world systems, and assess those predictions against measurements relevant to understanding some aspect of homeostatic processes in actual biological systems (cf., Da Costa et al. 2020, Table 1 for a summary of some existing applications of active inference models). Our contention is that the idealizations made by free energy theorists do not play these practical and epistemic roles.

At this point, the free energy theorists can reply that we are unfair, or misinformed. In fact, they may argue that the biological plausibility of the FEP is much clearer than what we are suggesting. Maximal generality need not detract from realism. Specifically, they can point out two things to us: first, many researchers routinely define and manipulate free-energy functionals for estimating coupling parameters in models of actual biological systems; second, core assumptions of the FEP, such as the assumption that the phase space of any biological system has an underlying non-equilibrium steady state density, are not idealizations but are accurate representations warranted by empirical evidence from actual biological systems.

Let us consider the first part of this reply. This part alludes to dynamic causal modelling, which is one possible approach to causal inference from neuroimaging data (Marinescu, Lawlor, and Kording 2018). With dynamic causal modelling, researchers work with idealized, but well-defined models of the probabilistic dependencies between latent state variables in neural systems - like synaptic activities - and observed measurements - like blood-oxygen-level-dependent fMRI signals. The relative level of free energy (i.e., relative evidence) of different models of the target neural variables and their relationships is used to identify which one within a given set of models is the best supported given observed BOLD signals. This would demonstrate the epistemic and practical utility of defining and manipulating free energy functionals for inference and estimation.

However, this reply misses our point. The example of dynamic causal modelling in neuroscience does not demonstrate that the FEP provides life scientists with a realistic account of the behaviour of real-world biological systems. Successful applications of dynamic causal modelling do not give us any reason to believe that neural systems plausibly minimize free energy, or that they are (ergodic) active inference systems. After all, we would not conclude, from successful applications of structural equation modelling or Granger-causality to fMRI data, that it is plausible neural systems make causal inferences based on structural equation modelling or Granger-causality.

Let us now turn to the second part of the reply, which claims that core assumptions of the FEP are warranted by empirical evidence. Ramstead et al. (2018, 3), for example, say that a conception of a biological system’s “extended phenotype as the set of attracting states of a coupled dynamical system is supported by... studies of cancer genesis and progression... and by work on early myelopoiesis in real biological systems.” Specifically, they refer to two studies

– one by Yuan et al. (2017) and the other by Su et al. (2017) – which are aimed at understanding cancer as a dynamic attractor state in an epigenetic landscape, as providing support to the assumption that all biological systems have a non-equilibrium steady state density.

There are three things to say here. First, those two studies provide evidence that cancer can fruitfully be modelled and understood as an attractor in a dynamical system, as opposed to a breakdown in specific molecular pathways in a causal mechanism (Gross 2011; Green et al. 2018, Sec. 3.2). Those studies do not provide us with evidence that all biological systems occupy an attractor or that attractors can plausibly be associated with homeostatic properties, nor do those authors claim otherwise.

Second, this reply does not distinguish between attractors that plausibly correspond to homeostatic states and attractors that have no obvious biological significance because they are not observed in actual cells. The dynamical models of molecular endogenous networks described in Yuan et al. (2017) and Su et al. (2017) yield several “structurally robust states,” but, as the authors note, their biological significance can be interpreted only in relation to known cellular phenotypes. For example, those two studies interpret certain states with a relatively large basin of attraction as cancer, based on empirical evidence indicating that there is high inter-individual variance in the mutations in sequenced tumours for non-hereditary cancers.

Third and finally, claiming that certain stable cancer cell phenotypes are attractors, and that all attractors in dynamical models of biological systems can plausibly be interpreted as homeostatic states risks to conflate two scales of explanation. At the cellular scale, cancer cells might be considered homeostatic, functional states, because they contribute positively to the persistence of a tumor. At a higher scale, however, cancer cells cannot be homeostatic states, since they contribute negatively to the persistence of their host system, and tumors themselves consist in a homeostatic failure to regulate the balance between cell growth and cell death.

In summary, some dynamical systems analyses of biological networks play various epistemically and practically successful roles in systems biology (Green 2018 et al. 2018, Sec. 3). The success of these analyses depends on (i) the availability of large sets of data that could be used to constrain and empirically evaluate the dynamical models, (ii) background knowledge about the history of target biological systems, and (iii) the existence of warranted conceptual bridges linking abstract state representations and experimentally accessible observables. In particular, as systems biologist Hiroaki Kitano (2004, 835) put it, “[t]he greatest challenge [for dynamical systems approaches to biological robustness] will be to formulate theories that account for thermodynamics in heterogeneous and structured systems... [and that at such a level of abstraction they can] be practically applied to biological systems.” The FEP theorists have not handled that challenge yet and have not given us good reasons to believe that the FEP serves to explain and predict the behaviour of most biological systems.

6 Conclusion

In this paper, we have argued that free energy theorists have pursued maximally general models, by relying on foundational concepts and mathematical representations from statistical physics and dynamical systems theory. This pursuit means a sacrifice in biological realism with the risk of minimizing the explanatory (and predictive) power of free energy theorists' account of biological, homeostatic, far-from-equilibrium persistence.

The FEP can perhaps be better understood as a maximally general *definition* of any system that persists, viz. “any system that exists must, on average, have a stationary (non-changing) free energy, which (by the ergodic assumption) corresponds to the time integral” (Friston 2019, 184). But, this definition does not seem to provide us with any new insight into biological systems. To the extent free energy theorists treat all biological systems, at any scale, as pre-specified generic objects with fixed (currently unknown) equations of motions, their account risks missing all features that make biological systems interesting kinds of thermodynamic systems.

References

- Allen, M., and Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 195(6), 2459-2482.
- Ashby, W. R. (1956). *An Introduction to cybernetics*. London: Chapman Hall.
- Batterman, R. W. (2001). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press.
- Battle, C., Broedersz, C. P., Fakhri, N., Geyer, V. F., Howard, J., Schmidt, C. F., and MacKintosh, F. C. (2016). Broken detailed balance at mesoscopic scales in active biological systems. *Science*, 352(6285), 604-607.
- Bechtel, W. (2017). Explicating top-down causation using networks and dynamics. *Philosophy of Science*, 84(2), 253-274.
- Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in cognitive sciences*, 4(3), 91-99.
- Bernard C. (1865). *Introduction a l'Etude de la Médecine Expérimentale*. JB Baillière, Paris.
- Bich, L., Mossio, M., Ruiz-Mirazo, K., and Moreno, A. (2016). Biological regulation: controlling the system from within. *Biology Philosophy*, 31(2), 237-265.

- Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12), 656-660.
- Brauer, F., and Kribs, C. (2015). *Dynamical systems for biological modeling: An introduction*. CRC Press
- Brigandt, I. (2013). Systems biology and the integration of mechanistic explanation and mathematical explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 44(4), 477-492.
- Cannon, W. B. (1929). Organization for physiological homeostasis. *Physiological Reviews*, 9, 399-431.
- Chan, H. S., and Dill, K. A. (1991). Polymer principles in protein structure and stability. *Annual review of biophysics and biophysical chemistry*, 20(1), 447-490.
- Elliott-Graves, A. (2018). Generality and causal interdependence in ecology. *Philosophy of Science*, 85(5), 1102-1114.
- Colombo, M., and Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*, 1-26.
- Corcoran, A. W., Pezzulo, G., and Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology Philosophy*, 35(3), 1-45.
- Da Costa L, Parr T, Sajid N, Veselic S, Neacsu V, and Friston K. (2020). Active inference on discrete state-spaces: a synthesis. *arXiv preprint*. arXiv:2001.07203.
- Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the cognitive sciences*, 4(4), 429-452.
- Drake, J. A., Fuller, M., Zimmerman, C. R., and Gamarra, J. G. (2007). Emergence in ecological systems. In *From energetics to ecosystems: The dynamics and structure of ecological systems* (pp. 157-183). Springer, Dordrecht.
- Earman, J., and Rédei, M. (1996). Why ergodic theory does not explain the success of equilibrium statistical mechanics. *The British Journal for the Philosophy of Science*, 47(1), 63-78.
- Español P. (2004). Statistical Mechanics of Coarse-Graining. In: Karttunen M., Lukkarinen A., Vattulainen I. (eds) *Novel Methods in Soft Matter Simulations*. Lecture Notes in Physics, vol 640. Springer, Berlin, Heidelberg.

- Evans, D. J., Cohen, E. G. D., and Morriss, G. P. (1993). Probability of second law violations in shearing steady states. *Physical review letters*, 71(15), 2401-4.
- Frigg, R. (2008). A field guide to recent work on the foundations of statistical mechanics. In D. Rickles (Ed.), *The ashgate companion to contemporary philosophy of physics*, 99–196. London: Ashgate.
- Fakhri, N., Wessel, A. D., Willms, C., Pasquali, M., Klopfenstein, D. R., MacKintosh, F. C., and Schmidt, C. F. (2014). High-resolution mapping of intracellular fluctuations using carbon nanotubes. *Science*, 344(6187), 1031-1035.
- Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86), 20130475.
- Friston, K. (2012). A free energy principle for biological systems. *Entropy*, 14(11), 2100-2121.
- Friston, K., Da Costa, L., Hafner, D., Hesp, C., and Parr, T. (2020). Sophisticated Inference. arXiv preprint arXiv:2006.04120.
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural computation*, 29(1), 1-49.
- Friston, K., Breakspear, M., and Deco, G. (2012). Perception and self-organized instability. *Frontiers in computational neuroscience*, 6, 44.
- Friston, K., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3), 417-458.
- Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3), 70-87.
- Froese, T., and Stewart, J. (2010). Life after Ashby: Ultrastability and the autopoietic foundations of biological individuality. *Cybernetics Human Knowing*, 17 (4), 83-106.
- Gallavotti, G. (1999). *Statistical mechanics: A short treatise*. Springer-Verlag Berlin Heidelberg.
- Godfrey-Smith, P. (2013). Darwinian Individuals. In Frédéric Bouchard and Philippe Huneman (Eds.). *From Groups to Individuals: Evolution and Emerging Individuality*, 17–36. Cambridge, MA: MIT Press.
- Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, 128(4), 635-638.

- Goldbeter, A. (2018). Dissipative structures in biological systems: bistability, oscillations, spatial patterns and waves. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2124), 20170376
- Green, S., Şerban, M., Scholl, R., Jones, N., Brigandt, I., and Bechtel, W. (2018). Network analyses in systems biology: new strategies for dealing with biological complexity. *Synthese*, 195(4), 1751-1777.
- Gross, F. (2011). What systems biology can tell us about disease. *History and Philosophy of the Life Sciences*, 33, 477–496.
- Haken, H. (1983). *Synergetics. An Introduction*. Berlin Heidelberg: Springer-Verlag.
- Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational properties. *Proceedings of the National Academy of Science (USA)* 79, 2554-2558.
- Hohwy, J. (2020). Self-supervision, normativity and the free energy principle. *Synthese*, 1-25.
- Huang, S., Ernberg, I., and Kauffman, S. (2009). Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. In *Seminars in cell developmental biology*, 20(7), 869-876.
- Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. E. (2005). Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical Review Letters*, 94, 128701.
- Izhikevich, E. M. (2007). *Dynamical systems in neuroscience*. MIT press.
- Janson, N. B. (2012). Non-linear dynamics of biological systems. *Contemporary Physics*, 53(2), 137-168.
- Jhun, J., Palacios, P., Weatherall, J. O. (2018). Market crashes as critical phenomena? Explanation, idealization, and universality in econophysics. *Synthese*, 195(10), 4477-4505.
- Karlebach, G., Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10), 770-780.
- Kauffman, S. A. (1993). *The Origins of Order: Self-Organisation and Selection in Evolution*. New York: Oxford University Press.

- Kauffman, S. (1971). Differentiation of malignant to benign cells. *Journal of theoretical biology*, 31(3), 429-451.
- Keller, E. F. (2009). Organisms, machines, and thunderstorms: a history of self-organization, part two. Complexity, emergence, and stable attractors. *Historical Studies in the Natural Sciences*, 39(1), 1-31.
- Keller, E. F. (2008). Organisms, machines, and thunderstorms: A history of self-organization, part one. *Historical Studies in the Natural Sciences*, 38(1), 45-75.
- Kirchhoff, M.D. Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese* 195, 2519–2540 (2018). <https://doi.org/10.1007/s11229-016-1100-6>
- Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface*, 15(138), 20170792.
- Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11), 826-837.
- Levins, R. (1966). The Strategy of Model Building in Population Biology. *American Scientist* 54 (4): 421–31.
- Longo, G., Montévil, M. (2014). From physics to biology by extending criticality and symmetry breakings. In *Perspectives on Organisms* (pp. 161-185). Springer, Berlin, Heidelberg.
- Longo, G., Montévil, M., and Kauffman, S. (2012). “No entailing laws, but enablement in the evolution of the biosphere,” in *Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computation Conference Companion* (New York, NY, USA: ACM), 1379–1392.
- Mandell, A. J., Selz, K. A. (1990). Heterochrony as a generalizable principle in biological dynamics. In *Correlations and Connectivity* (pp. 281-294). Springer, Dordrecht.
- Manu, S. Surkova, A.V. Spirov, V.V. Gursky, H. Janssens, A.-R. Kim, et al. (2009a). Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors. *PLoS Computational Biology*, 5 (2009), p. e1000303
- Manu, S. Surkova, A.V. Spirov, V.V. Gursky, H. Janssens, A.-R. Kim, et al. (2009b). Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation. *PLoS Biology*, 7 (2009), p. e1000049

- Marinescu, I. E., Lawlor, P. N., and Kording, K. P. (2018). Quasi-experimental causality in neuroscience and behavioural research. *Nature human behaviour*, 2(12), 891-898.
- Matsuda, H., Ogita, N., Sasaki, A., & Satō, K. (1992). Statistical mechanics of population: the lattice Lotka-Volterra model. *Progress of theoretical Physics*, 88(6), 1035-1049.
- Matthewson, J. (2011). Trade-offs in model-building: A more target-oriented approach. *Studies in History and Philosophy of Science Part A*, 42(2), 324-333.
- May, R. (2004). Uses and abuses of mathematics in biology. *Science*, 303, 790-794.
- McLeish, T. C. (2015). Are there ergodic limits to evolution? Ergodic exploration of genome space and convergence. *Interface Focus*, 5(6), 20150041.
- Nicolis, G. and Prigogine, I. (1977). *Self-Organization in Nonequilibrium Systems*. Wiley-Interscience, New York.
- Nikolov, S., Yankulova, E., Wolkenhauer, O., Petrov, V. (2007). Principal difference between stability and structural stability (robustness) as used in systems biology. *Nonlinear Dynamics, Psychology, and Life Sciences*, 11(4), 413-33.
- Okasha, S. (2012). Emergence, hierarchy and top-down causation in evolutionary biology. *Interface focus*, 2(1), 49-54.
- Palacios, P. (2018). Had we but world enough, and time... but we don't!: Justifying the thermodynamic and infinite-time limits in statistical mechanics. *Foundations of Physics*, 48(5), 526-541.
- Parr, T., and Friston, K. J. (2019). Generalised free energy and active inference. *Biological cybernetics*, 113(5-6), 495-513.
- Potochnik, A. (2017). *Idealization and the Aims of Science*, Chicago, IL: University of Chicago Press.
- Prigogine, I. and Nicolis, I. (1967). On Symmetry-Breaking Instabilities in Dissipative Systems. *Journal of Chemical Physics* 46, 9, 3542-50.
- Ramstead, M. J. D., Badcock, P. B., and Friston, K. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of life reviews*, 24, 1-16.
- Robertson, K. (2020). Asymmetry, abstraction, and autonomy: justifying coarse-graining in statistical mechanics. *The British Journal for the Philosophy of Science*, 71(2), 547-579.

Rubin, S., Parr, T., Da Costa, L., Friston, K. (2020). Future climates: Markov blankets and active inference in the biosphere. *Journal of the Royal Society Interface*, 17(172), 20200503.

Scheutzw M. (2007) Attractors for Ergodic and Monotone Random Dynamical Systems. In: Dalang R.C., Russo F., Dozzi M. (eds) *Seminar on Stochastic Analysis, Random Fields and Applications V*. Progress in Probability, vol 59, pp. 331-344. Birkhäuser Basel.

Schrödinger, E. (1944/1992). *What is life?: With mind and matter and autobiographical sketches*. Cambridge University Press.

Seth, A. K. (2014). The cybernetic Bayesian brain. In *Open Mind* (Ed.), Open MIND. Frankfurt: MIND Group.

Smith, S. (2017). Organisms as persisters. *Philosophy, Theory and Practice in Biology*, 9: 14. <http://dx.doi.org/10.3998/ptb.6959004.0009.014>

Sornette, D. (2017). *Why stock markets crash: critical events in complex financial systems*. Princeton University Press.

Su, H., Wang, G., Yuan, R., Wang, J., Tang, Y., Ao, P., Zhu, X. (2017). Decoding early myelopoiesis from dynamics of core endogenous network. *Science China Life Sciences*, 60(6), 627-646.

Turner, J S (1982). Self-organization in non-equilibrium chemistry and biology. In W. Schieve and P. M. Allen (Eds). *Self-organization and Dissipative Structures*, 40-63. Austin, TX: Texas University Press.

van Lith, J. (2001). Ergodic theory, interpretations of probability and the foundations of statistical mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 32(4), 581-594.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. New York: Oxford University Press.

Wightman, A. S. (1985). Regular and Chaotic Motions in Dynamical Systems Introduction to the Problems. In *Regular and chaotic motions in dynamic systems* (pp. 1-26). Springer, Boston, MA.

Yuan, R., Zhu, X., Wang, G., Li, S., and Ao, P. (2017). Cancer as robust intrinsic state shaped by evolution: a key issues review. *Reports on Progress in Physics*, 80(4), 042701.

.
....