

Information Width: A Way for the Second Law to Increase Complexity

David H. Wolpert

SFI WORKING PAPER: 2012-09-015

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu



SANTA FE INSTITUTE

Information Width: a way for the second law to increase complexity

David H. Wolpert
The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501;
Information Sciences Group
Los Alamos National Lab, MS B-256, Los Alamos, NM 87545
david.h.wolpert@gmail.com

September 19, 2012

Abstract

Many systems appear to increase their “complexity” in time and then robustly maintain a high complexity once achieved [15, 5, 22, 35]. To investigate this phenomenon it is necessary to formalize “complexity”. Here I build on recent work arguing that complexity of a system should be formalized as how much the patterns exhibited on different scales and/or at different locations of that system differ from one another. I quantify this variation in patterns — this type of complexity — as the Jensen Shannon (JS) divergence among the patterns.

Next I construct a highly stylized model of off-equilibrium, steady-state, network systems whose structure is maintained by depletion forces. Such networks can be viewed as highly abstracted models of living systems (organisms, ecosystems, or entire biospheres), bypassing considerations of reproduction and natural selection to focus on the underlying physics and information theory.

Finally, I show how the second law can drive the growth of these depletion force network systems. I also show that this growth causes such networks to have high JS divergence. In this way the second law can actually drive the increase of complexity in time.

1 Introduction

1.1 Rising complexity, natural selection, and thermodynamics

It seems that many system both increase their complexity if initialized in a low complexity state, and then reliably and robustly maintain high (but finite) complexity once it is attained. Some of the most prominent examples are the many biological systems undergoing natural selection that seem to start with low complexity and then increase their complexity [15, 5, 22, 35]. Such systems are typically modeled as localized individuals, that reproduce in an error-prone process, with their offspring weeded out in competitions with other individuals that select for higher complexity. In this natural selection process the individuals in a line of biological descent increase their complexity in time.

Some have argued from these examples that natural selection is a necessary condition for complexity to increase. The idea is that for a particular lineage to have large fitness advantage over its competitors, it must become increasingly “complex”. However we should not confuse the properties of an example of a phenomenon with the phenomenon itself: complexity increase is not synonymous with adaptionist natural selection. Indeed, one can engineer by hand models of systems undergoing natural selection where the competition selects for *small* complexity of the individuals in a line of descent, not high complexity.¹ One can even engineer models where the competition ends up weeding out *all* the individuals, so that the natural selection causes all the lines of descent die — in which case the complexity of the system has been driven to its minimal value. So natural selection, by itself, need not cause complexity to increase.

Conversely, there are biological systems that appear to increase their complexity with time but that do not involve the adaptionist process discussed above. Examples are constructive neutral evolution [10], and arguably auto-catalytic systems [14], in which natural selection plays a different (and less central) role in the increase of complexity than it does in adaptionist processes.

Another example is embryogenesis of a single embryo developing in a womb; the increase in complexity of the embryo is not due to its “competing” with other embryos in any sense. At best, one might argue that the embryo’s increase in complexity arose via competition occurring in the past, between its ancestors and their antagonists. This is a rather tortuous connection between a current rise in complexity of a system and the process of natural selection. It also doesn’t address the possibility of hand-crafting an artificial embryo so that its complexity will rise in an artificial womb, without that womb having any ancestors.

¹For example, the model might have typical individuals of low complexity be more robust against external shocks than typical individuals with high complexity, which confers a selective advantage to those low complexity individuals.

A particularly simple kind of process that increases complexity without any natural selection is depletion forces [2, 8, 1, 20]. As an example of such forces, consider a hollow sphere, filled with some balls that share a large sizes, and very many balls that share a small size. Assume all interactions are simply elastic hard-sphere collisions. Start the system with the large balls uniformly distributed throughout the interior of the sphere, and the small balls uniformly distributed throughout that part of the sphere's interior where there are no large balls. Then the second law will generate "depletion forces" that drive the large balls to hug the interior wall of the enclosing sphere.² So the large balls will be driven from uniformly filling the interior of a sphere to uniformly filling the shell of the sphere's boundary, i.e., they will be driven to assume a more complex configuration than the one they started with. With shapes slightly more complicated than balls, far more elaborate structures can be driven to arise, e.g., sheets, helices, etc. Recent work has suggested that some of the apparent complexity of biological systems (e.g., in the internal structure of cells) arises from such processes [20].

For perhaps the most striking example of a biological system that increases in complexity without natural selection, note that at least over certain periods and timescales, it seems that entire terrestrial biosphere has increased its complexity, e.g., immediately following mass extinction events, or in the major life transitions. However the "individual" of the biosphere does not undergo reproduction, and certainly is not engaged in competition with other biospheres. So external natural selection pressures cannot be the underlying cause of the increase of its complexity.³

Evidently then complexity in biological systems can be driven to increase by many different processes in addition to natural selection. Since both natural selection and many of these other processes are emergent phenomena, it is natural to wonder whether there is a more fundamental and broadly applicable process that underlies and unites many of them. In particular, given the example of depletion forces, one might wonder whether the second law of thermodynamics is an underlying driver of complexity rise in many biological systems.⁴ Might it somehow be that the second law, which increases disorder in a full (closed) system, not only

²When the large balls hug the interior wall, they exclude less of the volume of the sphere's interior to the small balls, i.e., increase the available volume to the small balls. Since there are so many more small balls, this increase in available volume to the small balls more than compensates for the decrease in available volume to the large balls, and therefore increases total Boltzmann entropy.

³Of course, there is natural selection occurring *within* the biosphere. But that is quite different from the way that natural selection arises in the biological examples of increasing complexity given above, where a lineage grows in complexity due to *external* natural selection pressures, not internal ones.

⁴Of course thermodynamics is deeply involved in biochemical systems [32, 33, 34]. The concern here is with thermodynamic processes that are more broad-ranging.

allows (open) subsystems to increase their order, but actually *drives* them to do so, under certain circumstances?

1.2 Dynamical systems theory perspective

The goal of this paper is to investigate this question. Ultimately, one would like to answer it in terms of dynamic systems theory. To be more precise, let $\rho(\gamma)$ refer to the phase space density of a system over phase space position γ , and let H refer to the Hamiltonian governing its dynamics. Then our goal is to understand what characteristics of H determine whether the resultant dynamics of ρ has an attractor throughout which ρ has high complexity, and understand how the dynamics of the complexity of ρ is determined by H .

To illustrate this dynamic systems perspective in a biological context, say our system is initially described by a ρ within a basin of attraction of a high-complexity attractor. Say that the system experiences an external shock knocking it to another point in the basin that has lower complexity. After the shock, the system would start increasing its complexity back to the value it had before the shock. Examples of this arguably include asteroid impacts, volcanic eruptions, etc., that cause a mass extinction, thereby reducing the complexity of the terrestrial biosphere, after which the biosphere's complexity grows back. Note though that if the shock were big enough to knock the systems completely out of the basin of attraction, then the system would "die", and not increase its complexity back to what it was.

Our goal then is to investigate how the Hamiltonian of a system determines its high-complexity attractors, and in particular how its thermodynamic properties do so. Some of the particular questions pursuant to this goal are:

1. How many separate high-complexity attractors are there for a given H ?
2. What fraction of the space of all $\rho(\gamma)$ lie in the basins of attraction of those high-complexity attractors?
3. Are those attractors fixed points, limit cycles, strange attractors, etc.?
4. How narrow are those basins of attraction? In other words, within a single such basin, what is the shape of the function taking complexity level χ to the volume of all $\rho(\gamma)$ within a given radius of the basin's attractor that have complexity at most equal to χ ?
5. How does the rate of increase of complexity of a particular $\rho(\gamma)$ depend on how close its complexity is to the complexity value of the attractor it is approaching?

6. Is complexity a (negative of a) Lyapunov function of the dynamics within such a basin of attraction? In other words, is the attractor the local peak of complexity in the space of $\rho(\gamma)$? Or are there $\rho(\gamma)$ in the basin of attraction that have higher complexity than the attractor does?
7. As an associated question, what is the maximal value of complexity per unit volume in space-time that a system can have?
8. As an empirical issue, are almost all examples of systems that reliably increase their complexity in time biological?
9. How big is the average high-complexity attractor of a given Hamiltonian? In other words, how much of a disruption must there be to $\rho(\gamma)$ to knock it out of the basin of attraction? (In the context of biological systems, this question amounts to asking how much of an external shock it takes to “kill” such systems.)
10. What fraction of densities $\rho(\gamma)$ with high complexity lie in basins of a high-complexity attractor? In other words, how likely is it that a randomly chosen density with high complexity only has that high complexity temporarily?

Presumably these questions can be investigated without recourse to thermodynamics. Indeed, one might expect there to be situations where the dynamics resulting in high-complexity attractors has nothing to do with the second law. However, in this paper I focus on situations where much about the dynamics can be understood in terms of the second law.

1.3 Paper roadmap

The very first step in analyzing “high-complexity attractors” is to fix a definition of “complexity”. To that end, I begin in Sec. 2 by arguing in general terms that a crucial feature of complex systems is that they contain patterns that vary greatly across locations and/or scales within themselves [38, 39, 40]. Here I quantify that variation with the Jensen-Shannon (JS) divergence among the patterns across the locations and/or scales of the system [12, 6, 17].

In the next section, I consider a particularly simple model of steady state, off-equilibrium open systems, as an archetype for systems that maintain high complexity. This model can be viewed as an abstraction of a photosynthetic organism, open to an environment of sunlight, reducing the interaction between the organism and the sunlight into what is essentially a generalization of depletion forces [37, 7, 19]. I show that the second law will cause systems described by this model to stochastically grow networks, assuming certain conditions are met.

(Loosely speaking, these networks can be viewed as abstractions of food webs.) I also show that such networks can be expected to have high Jensen-Shannon divergence. In this way, I show how the second law not only *allows* an open system to have high complexity, but can actually *drive* it to have high complexity.

2 Complexity and Jensen Shannon divergence

2.1 Self-dissimilarity

In almost all large systems commonly characterized as complex, the spatio-temporal patterns exhibited on different scales and/or in different regions differ markedly from one another. Conversely, for systems commonly characterized as simple the patterns are quite similar.

The human body is a familiar illustration of this; as one changes the scale of the spatio-temporal microscope with which one observes the body, the distribution of patterns that one sees varies tremendously. Similarly, as one changes from one region in the body to another, the distribution of patterns varies tremendously. The (out of equilibrium) terrestrial climate system is another good illustration, having very different dynamic processes operating at all spatiotemporal scales and in different regions, and typically being viewed as quite complex. Complex human artifacts also share this property, as anyone familiar with large-scale engineering projects will attest.

The following examples describes such variations in patterns across scales and/or regions in more detail:

Example 1. *Consider a spatially extended physical system involving many particles. The phase space position of such a system, γ , is a mass distribution $w(r)$ across $r \in \mathbb{R}^3$, together with an associated momentum field, π . As an example, if the system is a human body, $w(r)$ would be the mass distribution of all elementary particles in that human body at a particular time.⁵ In light of this meaning of γ , a phase space density $\rho(\gamma)$ evaluated at a particular time t fixes a probability density function $p(w, \pi)$ over mass distributions and associated momentum fields.*

Use i to specify the pair of a scale and location of a three-dimensional sphere, i.e., $i = (i_{scale}, i_{location})$. Define x_i for a particular i to be the possible contents of a mass distribution over r when that distribution is given by masking $w(r)$ with the sphere specified by i . So x_i is a mass density function over a Euclidean variable.

⁵As discussed in [40], we may be interested not in the distribution $w(r)$ of all particles in the system, but only in the distribution of carbon atoms, or of water atoms, or of electrons, or some other subset of the particles of the system. For current purposes though, we don't need to specify what types of particles $w(r)$ describes.

To be precise, given a particular mass distribution $w(r)$, and a scale / location i , the associated x_i is defined as

$$x_i(s) = \begin{cases} w(i_{scale}s + i_{location}) & \text{if } |s| \leq d \\ 0 & \text{otherwise.} \end{cases}$$

for some fixed window width d . Note that for any set of i 's, a given $w(r)$ fixes an associated set of values $x_i(s)$, which we can represent as a vector of functions, \vec{x} . Therefore a distribution $p(\vec{x})$ is induced by any distribution $p(w)$ (which in turn is induced by a distribution $\rho(\gamma)$).

As an illustration, we could have $w(r)$ be the mass distribution of the human body. We could also take $d = 1$ mm, and have a set of i 's sharing the value $i_{scale} = 1$, where $i_{location}$ varies across the entire human body. So the contents $x_i(s)$ would be the mass distribution $w(r)$ masked in various ways, as described above. We could approximate $p(\vec{x})$ as a product of the distributions at the different locations, assuming those locations are sufficiently far apart, on the scale of 1mm, so that the mass distributions (inside 1mm-wide spheres centered) at those locations are approximately statistically independent. The mass distributions at those locations differ quite a lot from one another (assuming $i_{location}$ ranges across the many organs in the human body), so self-dissimilarity is large.

Example 2. As a variation on Ex. 1, for the same human body $w(r)$, we could have $i_{location}$ be fixed, but have i_{scale} vary across many orders of magnitude. In this case, we are interested in how much the mass distributions at a single location, but a wide range of magnifications, differ from one another. Again, for a human body, in which the distributions at the scales of organs differs drastically from that at the scale of cells which in turn differs drastically from that at the scale of organelles, this variation is quite high.

Example 3. Another example is to use $w(r)$ of the entire terrestrial biosphere. As with a human body, at any fixed moment in time, the mass distributions in the biosphere differ drastically from one location (i.e., within one organism) to another. Also like with the human body $w(r)$, mass distributions in the biosphere $w(r)$ differ drastically from one scale to another (i.e., as one moves among levels of magnification with which a single organism is examined). So again, the variation in patterns is quite high.

In contrast to such complex systems, in many “simple” systems the distributions of the patterns at different scales and/or in different regions do not vary significantly from one another. For example, at scales sufficiently large compared to the size of individual molecules, the patterns at different scales and/or in different regions in a fully equilibrated ideal gas do not vary at all. Similarly, at sufficiently large scales, the patterns do not vary in a crystal.

Based on such examples, one could even argue that it is the self-*similar* aspects of simple systems, as revealed by allometric scaling, scaling analysis of networks, etc. [36], that reflects their inherently simple nature. After all, such self-similarity means that the pattern across all scales can be encoded in a short description (e.g., have low algorithmic information complexity). Such small code lengths is often taken to mean *ipso facto* that the system is not complex.

More generally, even if one could find a system commonly viewed as complex that was clearly self-similar in all important regards, it is hard to see how the same system wouldn't be considered even more "complex" if it were self-dissimilar. Indeed, it is hard to imagine a system that is highly self-dissimilar in both space and time that would not be considered complex.

Evidently then the self-*dissimilarity* among the patterns within many systems is an important component of their complexity [38, 39, 40, 28, 30]. Intuitively, the self-dissimilarity of a system reflects how much the information stored at one region or scale, and / or its processing there, differs from that at the other regions or scales. Indeed, viewing a physical system as a computational device, different regions in the system are simply different addresses, using an "addressing system" based on location. Different scales are also different addresses in the system, just using a different addressing system.

For an engineering perspective on this, note that if a system has very similar information stored at different addresses, then as an information storage device, the system is quite inefficient. It has lots of redundancy among the contents of its different addresses. Conversely, if the system has quite *different* information stored at its different addresses, then the system is efficient at encoding as much information into its state as possible. It has distributed its different parts of the information it stores into different addresses, rather than just duplicating all its information among all those addresses. This (in)efficiency is captured in the notion of self-dissimilarity.

The importance of how much the patterns in a system vary across addresses extends beyond addressing schemes based on either spatial location or spatial scale. As an illustration, complex systems often exhibit great variation in their patterns at different times, whereas simple systems typically exhibit less. For example, once it has fossilized a dead organism is static across time, i.e., completely self-similar along the time axis. What relatively little spatio-temporal complexity it still possesses is purely spatial, a relic of its complex past.

It is not being claimed that the definitive answer to the old question of what it means to say a system is "complex" is that the information at different addresses in the system varies greatly. (Ultimately, what scientists mean by the word "complex" is an issue perhaps best addressed by linguists and anthropologists.) Rather the claim is that how distributed the information is is an important component of complexity. Furthermore, as elaborated below, this aspect of complexity turns

out to be particularly well-suited to an analysis of the relation between rise in complexity and the second law.

2.2 Quantifying self-dissimilarity as JS divergence

How should we formalize a system’s self-dissimilarity, the amount that the “information at different addresses in the system varies”? To answer this, first we must fix some notation. Consider a vector-valued random variable, \mathbf{X}^n , taking values $\vec{x} \in \mathbf{X}^n$. Write a component of \mathbf{X}^n as \mathbf{X}_i with elements x_i . The indices i are the “addresses” referred to above, and the x_i ’s are the contents of those addresses. Write the probability of any \vec{x} as $p(\vec{x})$. Write the associated marginalizations as $p_i(x_i)$. When I want to refer to a generic $p_i(x_i)$, I will often write $p(x)$. (So the argument of p determines whether it refers to the full distribution defining \mathbf{X}^n or to a generic marginalization of that full distribution.) For much of what’s below, I will take each \mathbf{X}_i to be the same finite space X , with $m \equiv |X|$ elements.

Finally, write the *Shannon entropy* of $p(\vec{x})$ as

$$S_\mu(p) \equiv - \sum_{\vec{x} \in \text{supp}[\mu]} p(\vec{x}) \ln \left[\frac{p(\vec{x})}{\mu(\vec{x})} \right], \quad (1)$$

where conventionally, $\mu(\vec{x})$ is viewed as a “prior probability” of \vec{x} . Since μ is a proper probability distribution, and so normalized, S_μ is non-positive, being maximized as 0 iff $p(x) = \mu(x)$.⁶ Up to an irrelevant overall additive constant, the “information” in a distribution p is taken to be $-S_\mu(p)$. It is non-negative (due to the additive constant), being minimized when p is exactly what was expected *a priori*, namely μ .

An important class of scenarios is where $p(\vec{x})$ is well-approximated as a product distribution, $p(\vec{x}) = \prod_i p_i(x_i)$. This means that the random variables at the different addresses are physically separated enough so that we can treat them as decoupled. For example, in the case of a human body, with addresses being locations and the random variables being spatial configurations of amino acids, this approximation means that we consider the distribution of of such configurations located in the hippocampus as statistically independent of the distribution of such configurations located in the pancreas. Note that when p is a product distribution we can write $S_\mu(p) = \sum_i \sum_{x_i \in \text{supp}[\mu_i]} p_i(x_i) \ln [p_i(x_i)/\mu_i(x_i)]$.

In this paper I restrict attention to the case where p is a product distribution. In addition, for simplicity (and to agree with much of the literature), I will take μ

⁶To recover the usual formulation where entropy is non-negative, with a *minimal* value of 0, one adds the p -independent constant $-\min_{\vec{x} \in \text{supp}[\mu]} (\ln(\mu(\vec{x})))$ to the definition of S_μ .

to be uniform. For convenience I will define

$$S(p) \equiv - \sum_{\vec{x}} p(\vec{x}) \ln[p(\vec{x})] \quad (2)$$

which equals $S_\mu(p) + n \ln[m]$ evaluated under the uniform μ .

2.3 Formalizing Self-dissimilarity

In information theory, the conventional way to quantify how much a set of distributions $p_i(x)$ vary is as their *Jensen-Shannon divergence* (JS divergence [12]). Taking those p_i to be marginals of a product distribution p , JS divergence is defined as

$$JS(p) \equiv S\left(\frac{\sum_i p_i}{n}\right) - \frac{\sum_i S(p_i)}{n}. \quad (3)$$

$JS(p)$ can be viewed as the average “information-theoretic distance” from a randomly chosen distribution p_i to the “center of masses” of the set of distributions $\{p_i\}$. In this sense the JS divergence can be viewed as a symmetrization and extension of Kullback-Leibler distance [6]. In particular, the JS divergence equals 0 iff all of the distributions p_i are identical.

As an illustration, consider the case of a window-based addressing scheme (as in Ex. 1), where each address specifies a location, and say that we approximate p as a product distribution. Then on scales substantially larger than molecules there is *no* variability among the p_i for a perfectly equilibrated ideal gas. The same is true (at least to first order) for a perfect crystal. For such systems JS divergence equals 0. This is stated a bit more formally in the following example.

Example 4. *Return again to the setting for Ex. 1. For simplicity, assume the densities within each window are statistically independent, so that p is a product distribution.*

Like any other density function, $p(w)$ contains information. In particular, we are interested in how much information in $p(w)$ is distributed across different addresses. Physically, a large amount of information distributed across the system would mean that the patterns stored in the different addresses i typically vary a lot from one another, assuming those distributions are formed by sampled from $p(w)$.

For any fixed address i , $p(w)$ induces a distribution over x , given by

$$p_i(x) = \int dw p(w) \delta(x_i(s) - w(i_{scale}s + i_{location})).$$

(where both the x and w arguments of the Dirac delta function are being viewed as vectors). $JS(p)$ is very small (or zero) for some systems typically viewed as

simple, like fractals, gases, and (at large enough scales) crystals. On the other hand, $JS(p)$ is quite high for systems typically viewed as complex. For example, in a human, the distribution at any single region tends to have lots of structure, i.e., high information. Furthermore, the structure in different addresses tend to be quite distinct. The result is a high value of the difference in entropies that defines $JS(p)$.

For another perspective on JS divergence, view the index i as a random variable with uniform prior, and identify $p(x | i) \equiv p_i(x)$. So the marginalization of $p(x, i)$ over i is just $p(x) \equiv \sum_i p_i(x)/n = \mu(x)$, the average $p_i(x)$, and the marginalization of $p(x, i)$ over x is a uniform distribution. Then the JS divergence equals the mutual information between x and i :

$$\begin{aligned} JS(p) &= S(p(x)) - \sum_i p(i)S(p(x | i)) \\ &= S(p(x)) + S(p(i)) - S(p(x, i)). \end{aligned} \quad (4)$$

In other words, for a product distribution p , $JS(p)$ quantifies how much knowing the value x of a sample of one of the distributions p_i tells you about what i is, and vice-versa. As an illustration, in Ex. 1, JS divergence is high if being provided an image in a window, where all you know about it is that it was sampled from one of the distributions of images, tells you a lot about what distribution of images it was likely sampled from. In contrast, self-dissimilarity is low if being provided an image that you know was sampled from one of the distributions of images tells you little about what distribution of images it was likely sampled from.

In [41] I consider the more general scenario where p need not be a product distribution. I then solve for the optimal product distribution μ . I do this using four separate arguments based on Bayesian statistics, the maximal entropy principle, minimizing coding length, and minimizing algorithmic information complexity, all of which result in the same (non-uniform) answer for the optimal μ . Using this μ rather than the uniform μ changes the quantification of self-dissimilarity. This new quantification is the sum of JS divergence and the multi-information among the marginals of p . It reduces to $JS(p)$ when p is a product distribution however, and so does not affect the analysis of this paper, which assumes a product distribution p .

3 Analyzing systems that rise in JS divergence

Given the choice of JS divergence to measure complexity, the next step is to investigate how the dynamics of a system's JS divergence might be related to the second law. This section presents a preliminary investigation of this issue by analyzing a simple model involving a generalized version of depletion forces.

Throughout this analysis coarse-graining will be assumed, so Liouville’s theorem does not force phase space volume to be conserved. In addition, no external heat baths or any source of energy uncertainty will be assumed. So (Gibbs) physical entropy reduces to the number of distinct states available to the system.

Since the systems considered below will often be far from equilibrium, with widely varying temperatures within them, the analysis will be cast in terms of entropy gaps (between the actual entropy of a system and its maximal entropy) rather in terms of free energy. In keeping with convention though, rather than refer to such an “entropy gap”, I will use the term “negative entropy” (negentropy) [31, 4, 3, 18].

3.1 Examples of systems that rise in JS divergence

To begin, note that many systems that appear to have an attractor with high complexity are thermodynamically open. In such cases there is actually a composite system $\mathcal{A} \times \mathcal{B}$ undergoing Hamiltonian dynamics. It is the dynamics of \mathcal{B} that, evolving under the external influence of \mathcal{A} , has an attractor with high complexity. In such systems, loosely speaking, \mathcal{A} is a negative-entropy (negentropy) flux flowing from some low-entropy third system \mathcal{T} . \mathcal{B} “harvests” that flux, and by doing so \mathcal{B} increases its complexity. Some examples of this are:

- I) The terrestrial biosphere is \mathcal{B} , robustly maintaining a high complexity by harvesting the negentropy flux of sunlight (which plays the role of \mathcal{A}) flowing from the sun (which plays the role of \mathcal{T});
- II) An ecosystem at a hydrothermal vent is \mathcal{B} , robustly maintaining a high complexity by harvesting the negentropy flux of the flow through the vent (which plays the role of \mathcal{A}) from within the earth (which plays the role of \mathcal{T});
- III) Lineages of a bird that regularly rides atmospheric drafts, and of plants that regularly release fertilized seeds on the wind, are examples of systems \mathcal{B} that harvest the negentropy flux of the wind (which plays the role of \mathcal{A}) to maintain a high complexity (i.e., to live and reproduce), with sunlight and the earth’s rotation playing the role of \mathcal{T} (since they are what drives the wind);
- IV) Organisms resident in a fixed region in a stream that use that stream’s current to flush waste are examples of systems \mathcal{B} that harvest the negentropy flux of the stream \mathcal{A} to maintain a high complexity (i.e., to live).

Note that in all these examples the system \mathcal{B} has relatively high JS divergence in its attractor state.

In addition to the broader questions raised in Sec. 1.2, these examples raise many questions of their own. For example, in all these examples, much of the negentropy flux is not harvested. Is that always the case? Is there an upper bound to what fraction of the flux can be harvested for an extended period of time by a system with an attractor of high complexity? Is such harvesting of negentropy flux the only way that a system can have an attractor with high complexity? Or can thermodynamically closed systems have such attractors as well? A final example of such questions is whether the existence of such attractors of high JS divergence in these examples is a result of the second law somehow. This question forms the basis for the rest of the analysis of this section.⁷

3.2 JS divergence and off-equilibrium steady state systems

In all of the examples of Sec. 3.1, the attractor state is

1. Stationary (steady state);
2. Stable against small perturbations;
3. Out of equilibrium (and therefore given (1), in contact with an external environment);
4. Would quickly move to an equilibrium if the system were isolated from its environment.

For example, the terrestrial biosphere is (over certain timescales) in a steady state, with many of its large-scale physical characteristics stable against small perturbations. It is also out of equilibrium, and if it were isolated from its environment (i.e., were deprived of sunlight and hydrothermal vent effluent) would quickly move to equilibrium (i.e., all life would cease).

I call states with properties (1-4) **vital** states.⁸ Intuitively, systems in a vital state are exploiting their environment to stably maintain a far from equilibrium configuration, despite strong thermodynamic “forces” pushing them to equilibrium. The analysis of such systems and their (non-equilibrium) thermodynamics is a deep and mature field with roots going back decades [25, 26, 29, 9, 16]. Here I will only be interested in sketching some relevant properties of such systems, to allow them to be related to the second law and JS divergence.

⁷This is similar to the question of whether an external flux of negentropy drives the emergence of life. A seminal analysis of this issue can be found in [23].

⁸Note that each of properties (1), (2) and (4) have an implicit timescale; I will take this timescale to be the same for all of them.

Since a vital state is stable against small perturbations, it is an attractor. By hypothesis it is also off-equilibrium, when considered separately from its environment. I assume that to be such an off-equilibrium attractor, a vital state cannot dissipate entropy (and thereby approach equilibrium) via internal thermalization faster than a rate σ . (In particular, it cannot dissipate entropy faster than it acquires negentropy from the environment.) In addition to holding in many biological systems, this **thermalization rate limit** seems to hold in many artificial complex systems. For example, it is very difficult to make large computers dissipate heat beyond a certain rate. (This is why it is extremely important to design large computers so that they do not generate heat too quickly.)

3.3 Vital states formed by coupling two subsystems

I now consider vital systems that comprise one or more pairs of coupled subsystems. Each of those subsystems will correspond to a separate “address” of the system, in the sense of Sec. 2.3. So I will assume that probability distributions over possible states of the joint system is well-approximated by a product distribution over states of the constituent subsystems (at least at appropriate moments in time).

Vital systems that comprise more than one pair of coupled subsystem often have an elaborate network structure linking the pairs. There has been a lot of analysis done on these kinds of networks (e.g., see [24, 13], and other work analyzing free energy flows and free energy networks in ecosystems.) However for current purposes I only need to consider very simple aspects of such systems.

I start with an analysis for systems comprising a single pair of subsystems. Let A be the set of states a of a “background” subsystem \mathcal{A} , and let B be the set of states b of a “foreground” subsystem \mathcal{B} , where for simplicity I restrict attention to the case where both A and B are finite. Assume that during a time interval $[\tau, \tau + \delta t]$ the subsystems \mathcal{A} and \mathcal{B} are coupled, and that there are no other systems interacting with either \mathcal{A} and/or \mathcal{B} during that interval. To ground intuition, \mathcal{B} can be a (highly abstracted) model of a population of photosynthesizing organisms, or even of a photovoltaic cell, and \mathcal{A} can be a (highly abstracted) model of some sunlight incident upon that population during the interval $[\tau, \tau + \delta t]$.

Let $N_{AB}(t)$ be the number of joint states $\{ab : a \in A, b \in B\}$ that can occur at time t .⁹ Assume that those states are indistinguishable to an external observer. Let $N_A(t)$ be the number of states in A that can occur at t , again assuming that they are indistinguishable to an external observer. Define $N_B(t)$ similarly. Note that $N_{AB}(t) \leq N_A(t)N_B(t)$.

⁹These can be the number distinct of eigenstates if \mathcal{A} and \mathcal{B} are quantum mechanical systems, or for classical systems, they can be the number of coarse-grained bins of phase space.

Assume that due to the Hamiltonian governing the interaction of the two subsystems, the following two conditions hold:

- i) $N_B(\tau) > N_B(\tau + \delta t)$
- ii) $N_A(\tau)N_B(\tau) < N_{AB}(\tau + \delta t)$

Assumption (i) means that the coupled system ends the interval with lower Boltzmann entropy for \mathcal{B} . However assumption (ii) means that the coupling increases total Boltzmann entropy. (Recall we are assuming coarse-graining, so that available phase space volume need not be conserved.)

As a simple example, both assumptions (i) and (ii) hold with systems being subject to depletion forces, where \mathcal{B} refers to the larger balls, \mathcal{A} to the far more numerous tiny balls, τ is the initial time when the balls in \mathcal{B} are uniformly distributed throughout the interior of the sphere, and $\tau + \delta t$ is after the system has reached equilibrium with the balls \mathcal{B} hugging the inner wall of the sphere. Note that the fine-grained states a and b are highly correlated at $\tau + \delta t$. But due to coarse-graining, we do not see any of that correlation. This justifies the approximation that we describe the distribution over the joint system as a product distribution.

Note that assumption (i) would be violated if \mathcal{A} were a heat bath for \mathcal{B} and we used fine-grained states (rather coarse-grained). This is because in such a case the ‘‘coupling’’ between the two subsystems is simply energy transfer between them. More complicated coupling is needed for assumption (i) to hold, coupling that relates the configuration of \mathcal{A} to that of \mathcal{B} . (In the case of depletion forces coupling, where (i) holds, no particle ever changes its energy, in contrast to the case of heat bath coupling.)

It is always true that

$$\min\{N_A(\tau + \delta t), N_B(\tau + \delta t)\} \leq N_{AB}(\tau + \delta t) \leq N_A(\tau + \delta t)N_B(\tau + \delta t) \quad (5)$$

(As a simple example, in depletion forces the second bound in Eq. (5) holds exactly: $N_{AB}(\tau + \delta t) = N_A(\tau + \delta t)N_B(\tau + \delta t)$.) When assumptions (i) and (ii) hold though, by plugging (i) into (ii) we see that

$$N_B(\tau + \delta t)N_A(\tau) < N_{AB}(\tau + \delta t). \quad (6)$$

If we now plug the second inequality in Eq. (5) into Eq. (6), we see that when (i) and (ii) hold, $N_A(\tau) < N_A(\tau + \delta t)$. Indeed, plugging assumption (ii) into the second inequality in Eq. (5) gives

$$\ln\left[\frac{N_A(\tau)}{N_A(\tau + \delta t)}\right] < \ln\left[\frac{N_B(\tau + \delta t)}{N_B(\tau)}\right]. \quad (7)$$

So the shrinkage in the entropy of \mathcal{B} is more than offset by the growth in the entropy of \mathcal{A} .

Whenever assumptions (i) and (ii) hold I say that \mathcal{A} is **harvested** by \mathcal{B} . I define the **harvest rate** of the coupling during $[T, \tau + \delta t]$ as

$$\begin{aligned}\epsilon(\tau) &= \frac{\ln\left[\frac{N_{AB}(\tau+\delta t)}{N_A(\tau)N_B(\tau)}\right]}{\delta t} \\ &= \frac{\ln[N_{AB}(\tau + \delta t)] - \ln[N_A(\tau)] - \ln[N_B(\tau)]}{\delta t}\end{aligned}\quad (8)$$

where δt is implicit. The **instantaneous harvest rate** is defined as $\lim_{\delta t \rightarrow 0} \epsilon(\tau)$. By assumption (ii) $\epsilon(\tau)$ is positive, reflecting the fact that entropy increases during the coupling.

We can construct a long-lasting, stationary off-equilibrium version of this harvesting process by chaining instances of it one after the other. In the first step of such a process, the harvest described above occurs, in a time interval $[\tau, \tau + \delta t_1]$. Next during $[\tau + \delta t_1, \tau + \delta t_1 + \delta t_2]$, two things happen. First, \mathcal{A} and \mathcal{B} decouple from each other, without increasing N_A or N_B . For simplicity I restrict attention to cases where once the decoupling has occurred, the number of possible states of \mathcal{A} is $N_A(\tau + \delta t_1)$ and the number of possible states of \mathcal{B} is $N_B(\tau + \delta t_1)$. This means that the subsystems are made statistically independent in the decoupling. This is in keeping with our presumption that the probability distribution over the joint system is a product distribution at appropriate moments in time.

Next one of two essentially equivalent processes occurs:

1. \mathcal{B} and \mathcal{A} get reset to the states they were in at τ . One way for this to occur is for the entropy of \mathcal{B} to increase due to the second law while the entropy of \mathcal{A} decreases due to its coupling to an external system.
2. \mathcal{B} gets reset this way. However \mathcal{A} leaves the picture, and is replaced by a new subsystem that is identical to \mathcal{A} as it was at time τ .

As an example, the second process is what happens when \mathcal{B} is a photovoltaic cell “harvesting negentropy flux” in the form of sets of photons \mathcal{A} streaming by.

Whichever two-step process we use to return the joint system to its initial state, the total amount of entropy gained is the amount of negative entropy that had to be introduced at the end of the process to return the joint system to its starting state. This equals

$$\left(\ln[N_A(\tau + \delta t_1)] - \ln[N_A(\tau)]\right) + \left(\ln[N_B(\tau + \delta t_1)] - \ln[N_B(\tau)]\right). \quad (9)$$

Note that the first term in big parentheses is positive while the second is negative. However by Eq. 7 the magnitude of the first term is larger. So this total gain of entropy — the total negentropy harvested — is positive.

Note also that the entropy dissipated by thermalization in the second step of the 2-step process is $\ln[N_B(\tau)] - \ln[N_B(\tau + \delta t_1)]$. So by the thermalization rate limit,

$$\ln[N_B(\tau)] - \ln[N_B(\tau + \delta t_1)] \leq \sigma \delta t_2. \quad (10)$$

This gives a lower bound on how quick the second step can be, and therefore an upper bound on the harvest rate:

$$\begin{aligned} \epsilon(\tau) &= \frac{\left(\ln[N_A(\tau + \delta t_1)] - \ln[N_A(\tau)]\right) + \left(\ln[N_B(\tau + \delta t_1)] - \ln[N_B(\tau)]\right)}{\delta t_1 + \delta t_2} \\ &\leq \frac{\left(\ln[N_A(\tau + \delta t_1)] - \ln[N_A(\tau)]\right) + \left(\ln[N_B(\tau + \delta t_1)] - \ln[N_B(\tau)]\right)}{\delta t_1 - \left(\ln[N_B(\tau + \delta t_1)] - \ln[N_B(\tau)]\right)/\sigma}. \end{aligned} \quad (11)$$

Consider the case where \mathcal{A} can either go through this two-step process, or can never couple to \mathcal{B} , simply thermalizing by itself. Of the two options, the two-step process will be thermodynamically preferred by the second law if the harvest rate is greater than the thermalization limit of \mathcal{A} , since under those conditions entropy will be higher at the end of $\tau + \delta t_1 + \delta t_2$ if the coupling occurs than if it does not. This can be illustrated by the example mentioned above where “ \mathcal{A} is sunlight and \mathcal{B} is a vastly simplified version of photosynthetic organisms”. Say the photosynthetic organisms, with their thermalization rates, can harvest the negentropy of \mathcal{A} faster (as determined by Eq. 11) than \mathcal{A} can lose that negentropy by itself. Then \mathcal{A} will couple to \mathcal{B} . Otherwise the light will not couple to \mathcal{B} this way, and will simply pass through or reflect off of \mathcal{B} .

3.4 Vital states formed by coupling more than two subsystems

We can construct a long-lasting, stationary off-equilibrium version of this process that involves more than two steps. As an example, say that at time $\tau + \delta t_1$, \mathcal{B} gets decoupled from \mathcal{A} , as before. However rather than have the two-step process repeat, \mathcal{B} gets coupled to a third subsystem \mathcal{C} . Then during $[\tau + \delta t_1, \tau + \delta t_1 + \delta t_2]$, rather than get reset, \mathcal{B} itself gets harvested, by \mathcal{C} , in a process that obeys assumptions (i) and (ii). Presume that this causes the entropy of \mathcal{B} to increase back to the value it had at τ . Furthermore, assume that just as in the two-step

process, in the three-step process, by the time $\tau + \delta_1 + \delta t_2$, the subsystem \mathcal{A} gets reset to its time- τ state (or is replaced by a subsystem whose state at $\tau + \delta t_1 + \delta t_2$ is identical to what the state of \mathcal{A} was at time τ). Finally, assume that in an interval $[\tau + \delta_1 + \delta t_2, \tau + \delta_1 + \delta t_2 + \delta_3]$, \mathcal{C} has its entropy increase to the value it had at time $\tau + \delta_1$. So by time $\tau + \delta_1 + \delta t_2 + \delta_3$ the entire three-step process has returned the joint system to the original joint state it had at τ .

In this three-step process the total entropy gained is

$$\left(\ln[N_A(\tau + \delta t_1)] - \ln[N_A(\tau)] \right) + \left(\ln[N_C(\tau + \delta t_1 + \delta t_2)] - \ln[N_C(\tau + \delta t_1)] \right). \quad (12)$$

The second term in big parentheses in Eq. (12) has smaller magnitude than the second term in big parentheses in Eq. (9). Accordingly, more entropy is gained in this three-step process than in the two-step process it starts with. However the three-step process also takes more time than the two-step process. So to see if it is thermodynamically preferred to attach \mathcal{C} to $\mathcal{A}\mathcal{B}$ — to compare the highest possible harvest rates of the two-step and three-step processes — we have to consider the time lengths of those two processes.

To do this, first note that in the three-step process we never need to thermalize \mathcal{B} . So the thermalization rate limit of \mathcal{B} does not provide bounds on δt_2 . Instead we need to thermalize \mathcal{C} , so that the thermalization rate limit of \mathcal{C} applies, to δt_3 :

$$\ln[N_C(\tau + \delta t_1)] - \ln[N_C(\tau + \delta t_1 + \delta t_2)] \leq \sigma_C \delta t_3. \quad (13)$$

Combining, we see that the harvest rate for the three-step process is

$$\begin{aligned} \epsilon(\tau) &= \frac{\left(\ln[N_A(\tau + \delta t_1)] - \ln[N_A(\tau)] \right) + \left(\ln[N_C(\tau + \delta t_1 + \delta t_2)] - \ln[N_C(\tau + \delta t_1)] \right)}{\delta t_1 + \delta t_2 + \delta t_3} \\ &\leq \frac{\left(\ln[N_A(\tau + \delta t_1)] - \ln[N_A(\tau)] \right) + \left(\ln[N_C(\tau + \delta t_1 + \delta t_2)] - \ln[N_C(\tau + \delta t_1)] \right)}{\delta t_1 + \delta t_2 - \left(\ln[N_C(\tau + \delta t_1 + \delta t_2)] - \ln[N_C(\tau + \delta t_1)] \right) / \sigma_C}. \end{aligned} \quad (14)$$

The δ_2 in Eq. (14) does not need to be long enough for \mathcal{B} to be thermalized in the three-step process, and therefore is not limited by the thermalization rate limit of \mathcal{B} . So it can be shorter than the δt_2 occurring in the two-step process. The term $\ln[N_C(\tau + \delta t_1 + \delta t_2)] - \ln[N_C(\tau + \delta t_1)]$ in Eq. (14) is negative, just like the analogous term $\ln[N_B(\tau + \delta t_1)] - \ln[N_B(\tau)]$ from Eq. (11). However it has smaller magnitude. Therefore the numerator in Eq. (14) is larger than the numerator in Eq. (11). In addition, up to the term δt_2 , the denominator in Eq. (14) is smaller than

the numerator in Eq. (11). Therefore whether the two-step or three-step process has a higher harvest rate — and is therefore thermodynamically preferred — is determined by the size of the δt_2 's in the two-step and three-step processes.

This means that for an appropriately small three-step δt_2 compared to the two-step δ_2 , if a composite subsystem $\mathcal{A}\mathcal{B}$ joined in a two-step process encounters a subsystem \mathcal{C} , it is thermodynamically preferable for $\mathcal{A}\mathcal{B}$ to couple to \mathcal{C} at its “tail” \mathcal{B} , rather than stay isolated from \mathcal{C} . So the second law will induce such coupling into a three-step process whenever the opportunity arises.

This phenomenon provides a general way for the second law to create chains of many steps linking subsystems with vital states. To illustrate this, recall the example discussed above where \mathcal{A} is sunlight and \mathcal{B} is a population of photosynthesizing organisms. In this case an example of a system \mathcal{C} is a (highly abstracted) population of herbivores. They increase the entropy of the photosynthesizing organisms (by digesting them) faster than those organisms would be able to by themselves if they existed in a steady state without any herbivores.

3.5 Vital states formed when more than two subsystems are coupled simultaneously

There is no *a priori* restriction that negentropy harvests can only involve coupling two subsystems at once. There are many extensions of assumptions (i) and (ii) to concern tuples of more than two subsystems, in which a “foreground” system loses entropy even though the total system gains entropy.

Since these kinds of harvest involve more than two subsystems being coupled, in principle multiple instances of them can be joined to build joint systems with vital states that are more elaborate than chains. Most generally, sets of harvests involving such tuples of subsystems can build systems that are graphs. In particular, whereas harvests involving two subsystems at a time cannot result in cycles, harvests involving more than two subsystems at once can result in an arbitrary number of cycles.

This is what often happens in a real biological system. To give an intuitive example, there are many organs in a human that need to all be functioning for the body as a whole to maintain homeostasis (i.e., to maintain a vital state). Those organs do not interact only pairwise. Typically, any single organ needs to be interacting with multiple other organs, in an elaborate graph, for the body as a whole to maintain homeostasis.

4 Vital states and JS divergence

In all negentropy harvesting systems, the second law “glues” the subsystems together. In general it can only glue such subsystems together if

1. Their interaction Hamiltonian obeys assumptions (i) and (ii) (or more generally, the interaction Hamiltonian of whatever tuples of subsystems are available obeys the generalizations of those assumptions).
2. The thermalization rate limits and associated harvest times are related by the equations discussed in Sec. 3.3 and 3.4.

Only those subsystems that are “available” in that they obey these conditions can be glued together by the second law. In general, a stochastic process creates the subsystems that are available in this sense and so can be glued together. (In the example of the biosphere, depending on the timescale at which one models the biosphere, the stochastic process can involve anything from mutation and cross-over to the major transitions in life.) *A priori*, one would not expect such a stochastic process to join subsystems in a way that makes them have similar configurations. Therefore one would expect it to create overall systems with vital states have relatively high JS divergence among their subsystems.

This relationship between the second law and complexity is a subtle one. The second law, operating through a stochastic process, “glues together” subsystems, to make a full system that has a vital state. In turn, that full system is likely to have high JS divergence, simply because there is no *a priori* reason for the constituent subsystems to have similar configurations in the vital state. In this sense, it is almost accidental that systems with vital states — attractors of the underlying dynamics that arise due to the second law — are complex.

There is a long-standing debate in the literature about whether increasing complexity in the biosphere has been driven or is a simple drift process starting from an initial condition of low complexity [5, 15]. In the process elucidated in this paper, the second law is “driving” the process of gluing each new edge in a harvest network. However the introduction of nodes to be glued together with an edge is instead determined by a “drift process”. In this sense, both sides of the debate are correct.

5 Future work

The analysis in this paper shows how to derive the information theoretic way to formulate “complexity” through its realization that the prior μ in the definition of Shannon information plays a crucial role. And it shows how the second law

can drive the construction of dynamical attractors that have high values of such complexity.

While necessary, this leaves all of the questions in Sec.'s 1.2 and 3.1 yet to be analyzed. In addition, even the analysis that *is* done in this paper is far from complete. To give one of the most obvious examples, how should we choose the coordinate system with which to decompose x ? Given that we want $\mu(x)$ to be a product distribution over that coordinate system, a natural set of ways to choose the coordinate system is to apply principle components analysis (when x is a Euclidean vector), or a more sophisticated technique like independent components analysis. It remains to be seen though whether such approaches result in reasonable measures of self-dissimilarity.

As another example of a question that needs to be analyzed, as mentioned above, assumptions (i) and (ii) need to be extended to allow negentropy harvests that involve more than two interacting subsystems at once. What such extensions are best suited to modeling the kinds of harvests that seem to occur in real-world biological systems? As another example of an important question not yet analyzed, can we bound the total harvest efficiency in a sequence of harvest links in terms of the efficiencies of each individual link? If so, does this provide an experimentally testable claim? (For example, it might mean that entropy increases faster when a plant is coupled to light, water, CO₂, etc., then when it is decoupled from them (and therefore starts to die).)

Another set of issues concern the vaguely described “stochastic processes” under which subsystems arise that jointly meet assumptions (i) and (ii) (and their extensions), and therefore can be glued together by the second law. What are the characteristics of such processes (biological or otherwise) in the real world? For example, are random growth networks appropriate models? Whatever form they take, do the stochastic process models allow us to relate harvest rates, the size of the harvest network, and JS divergence? (Note that until something precise is said about the stochastic process, little quantitative can be said concerning the probability distribution over possible values of JS divergence.) More generally, does the form of such processes provide an explanation of the many power laws that govern the biosphere, food webs, etc. (e.g., power laws of complexity vs. total biomass)?

There are several important questions to ask concerning any particular stochastic growth model of harvest nets. For example, given such a model, what is the associated probability of arising at a net with overall rate ϵ , and how does that compare to the ϵ distribution induced by uniformly sampling all nets? A related question is whether under some such growth model, any network G_1 with overall harvest rate $\epsilon_1 > \epsilon_2$, is *always* thermodynamically preferred to a network G_2 with overall rate ϵ_2 (and therefore has higher probability). One of the factors that goes into answering this question is whether there can be a subnet of G_1 that has

a harvest rate smaller than any of the subnets of G_2 , even though G_1 has a higher overall harvest rate.¹⁰

If under some given growth models nets do get more probable as their harvest rate increases, independent of the harvest rates of their subnets, then even something like a human city, which is a huge network, is driven to have a high harvest rate. This would be rather astonishing, given that one normally models how complexity forms in cities in terms of human-human cognitive interactions, not in terms of anything so physical as negentropy.

A related (empirical) question is whether the rise of human civilization, including all the complexity of human cultures, resulted in a growth or loss of self-dissimilarity of the entire biosphere.¹¹

All of these are issues for future research.

ACKNOWLEDGEMENTS: I would like to thank Giovanni Bellesia for very stimulating discussions.

References

- [1] M. Adams, Z. Dogic, S. Keller, and S. Fraden, *Entropically driven microphase transitions in mixtures of colloidal rods and spheres*, *Nature* **393** (1998), 349–352.
- [2] S. Asakura and F. Oosawa, *J. Polym. Sci.* **33** (1958), 183.
- [3] L. Brillouin, *Negentropy principle of information*, *Journal of Applied Physics* **24** (1953), 1152–1163.
- [4] ———, *Science and information theory*, Academic Press, 1962.
- [5] S. Carroll, *Chance and necessity: the evolution of morphological complexity and diversity*, *Nature* **409** (2001), 1102–1109.
- [6] T. Cover and J. Thomas, *Elements of information theory*, Wiley-Interscience, New York, 1991.
- [7] J. C. Crocker, J. A. Matteo, A. D. Dinsmore, and A. G. Yodh, *Entropic attraction and repulsion in binary colloids probed with a line optical tweezer*, *Physical Review Letters* **82** (1999), 4352–4355.

¹⁰A literature tangentially related to such modeling concerns the “maximum entropy production principle” [27, 21, 11]. This rather controversial literature traces its lineage back to [25].

¹¹One would likely want to analyze this question using an extension of this paper’s quantification of self-dissimilarity, since the analysis would concern p ’s that are not product distributions. For example, one might want to use the extension of self-dissimilarity that is discussed in [41].

- [8] B. Götzelmann, R. Evans, and S. Dietrich, *Depletion forces in fluids*, Phys. Rev. E **57** (1998), 6785–6800.
- [9] Jr Grandy, W.T., *Entropy and the time evolution of macroscopic systems*, Oxford University Press, 2008.
- [10] M. W. Gray, J. Luke, J. M. Archibald, P. J. Keeling, and W. F. Doolittle, *Irremediable complexity?*, Science **330** (2010), no. 6006, 920–921.
- [11] G. Grinstein and R. Linsker, *Comments on a derivation and application of the 'maximum entropy production' principle*, J. Phys. A: Math. Theor. **40** (2007), 971720.
- [12] I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. Oliver, and H. E. Stanley, *Analysis of symbolic sequences using the jensen-shannon divergence measure*, Phys. Rev. E **65** (2002), 041905.
- [13] M. Ho, *What is negentropy?*, Modern Trends in BioThermoKinetics **3** (1994), 50–61.
- [14] S. A. Kauffman, *At home in the universe: The search for the laws of self-organization and complexity*, Oxford University Press, 1995.
- [15] D. Krakauer, *Darwinian demons, evolutionary complexity, and information maximization*, Chaos **21** (2011).
- [16] G. Lebon, D. Jou, and J. Casas-Vazquez, *Understanding non-equilibrium thermodynamics: Foundations, applications, frontiers*, Springer-Verlag, 2008.
- [17] D.J.C. Mackay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2003.
- [18] S.P. Mahulikar and H Herwig, *Exact thermodynamic principles for dynamic order existence and evolution in chaos*, Chaos, Solitons and Fractals **41** (2009), 1939–1948.
- [19] D. Maranduzzo, K. Finan, and P. Cook, *The depletion attraction: an under-appreciated force driving cellular organization*, Journal of Cell Biology **175** (2006), 681–686.
- [20] D. Marenduzzo, K. Finan, and P. Cook, *The depletion attraction: an under-appreciated force driving cellular organization*, Journal of Cell Biology **5** (2006), 681–686.

- [21] L. Martyushev and V. Seleznev, *Maximum entropy production principle in physics, chemistry and biology*, Physics Reports **426** (2006).
- [22] D. McShea, *Complexity and evolution: What everybody knows*, Biology and Philosophy **6** (1991), 303–324.
- [23] H. Morowitz and E. Smith, *Energy flow and the organization of life*, Complexity **13** (2007), 51–59.
- [24] H.J. Morowitz, *Energy flow in biology*, Academic Press, 1968.
- [25] L. Onsager, *Reciprocal relations in irreversible processes: 1*, Physical Review **37** (1931), 405.
- [26] ———, *Reciprocal relations in irreversible processes: 2*, Physical Review **38** (1931), 2265.
- [27] G. Paltridge, *Climate and thermodynamic systems of maximum dissipation*, Nature **279** (1979), 5714.
- [28] L. Parrott, *Measuring ecological complexity*, Ecological Indicators (2010), 1069–1076.
- [29] I. Prigogine, *Moderation et transformations irreversibles des systemes ouverts*, Bulletin de la Classe des Sciences, Academie Royale de Belgique **31** (1945), 600–606.
- [30] R. Proulx and L. Parrott, *Measures of structural complexity in digital images for monitoring the ecological signature of an old-growth forest ecosystem*, Ecological Indicators (2008), 270–284.
- [31] E. Schrodinger, *What is life?*, Cambridge University Press, 1944.
- [32] E.D. Smith, *Thermodynamics of natural selection i: Energy flow and the limits on organization*, Journal of Theoretical Biology (2008).
- [33] ———, *Thermodynamics of natural selection ii: Chemical carnot cycles*, Journal of Theoretical Biology (2008).
- [34] ———, *Thermodynamics of natural selection iii: Landauer’s principle in computation and chemistry*, Journal of Theoretical Biology (2008).
- [35] J. Maynard Smith, *Time in the evolutionary process*, Studium Generale **23** (1970), 266–272.

- [36] M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Les chhorn, P. Maass, M. A. Salinger, and H. E. Stanley, *Scaling behaviour in the growth of companies*, Nature **379** (1996), 804–806.
- [37] Ritu Verma, J. C. Crocker, T. C. Lubensky, and A. G. Yodh, *Entropic colloidal interactions in concentrated dna solutions*, Physics Review Letters **81** (1998), 4004–4007.
- [38] D. H. Wolpert and W. Macready, *Self-dissimilarity: An empirically observable complexity measure*, Unifying Themes in Complex Systems, New England Complex Systems Institute, 2000, pp. 626–643.
- [39] ———, *Self-dissimilarity as a high dimensional complexity measure*, Proceedings of the Fifth International Conference on Complex Systems, 2004, Boston, Mass.
- [40] ———, *Using self-dissimilarity to quantify complexity*, Complexity **12** (2007), 77–85.
- [41] D.H. Wolpert, *How to salvage shannon entropy as a complexity measure*, Submitted, 2012.