Qeios

Thermodynamics of scientific ontology

Artem Chumachenko[1*] and Brett Buttliere[2†]

[1,2*]Science Studies Lab, University of Warsaw, Krakowskie Przedmieście 30, Warsaw, 00-927, Poland.

*Corresponding author(s). E-mail(s): a.chumachenko@uw.edu.pl;
Contributing authors: b.buttliere@uw.edu.pl;
†These authors contributed equally to this work.

Abstract

In this work, we propose a thermodynamic framework to analyze the creative potential of scientific fields by examining statistical data on the usage frequencies of scientific concepts within a corpus of publications from ArXiv. Using statistical mechanics and thermodynamics, we model the system of physical concepts that form the ontology of a scientific field. We explore the relationship between Clausius entropy and Shannon entropy in this context, assuming the interaction of concepts through their pairwise mutual information. Our approach enables us to leverage methods from statistical physics to analyze information systems during knowledge production and transfer. We demonstrate that the coarse-grained frequencies of scientific concepts follow a generalized Boltzmann distribution, allowing for a thermodynamic description. This study calculates internal energy, Helmholtz free energy, temperature, and heat capacity for scientific concepts as closed thermodynamic systems, and maps the state space of the concepts-based knowledge network using data-driven thermodynamic diagrams. This framework advances the methods of the computational theory of discovery by providing insights into the dynamics of scientific knowledge and the emergence of innovation.

Keywords: complex information network, statistical mechanics, thermodynamics, statistical inference

1 Introduction

The concept of discovery and, more broadly, scientific progress, is intricately linked with scientific collaboration and the successful application of models developed in

https://doi.org/10.32388/UM6NLZ

other disciplines. Numerous studies have explored this connection [1, 2], often emphasizing that while we cannot predict future discoveries, an examination of the historical trajectory within the field can provide us with essential insights [3]. These insights may serve as the foundational ingredients, allowing us to craft a recipe for the art of discovery.

While bridging the logical gaps between scientific concepts and facilitating information diffusion are considered central elements of this recipe [4–8], investigating the emergence of scientific innovation is not straightforward. In formulating a computational theory of discovery [9], the nature of scientific change has been studied from many distinct perspectives, including the philosophy of science, sociology, the history of science, and various quantitative approaches such as scientometrics, citation analysis, information science, and complex network analysis.

A promising direction to advance the computational theory of discovery and principles of knowledge production lies in studying the properties of the natural knowledge network encoded in scientific communication records. The key elements of this network are scientific concepts that collectively define the scientific ontology. The general availability of metadata about electronically published scientific literature makes it one of the most essential and easily accessible sources of information for these studies [10, 11]. Processing the texts of scientific publications and extracting statistics of scientific concepts adds a new dimension to the analysis, enabling a deeper understanding of the structure and dynamics of scientific knowledge [12–14].

Each published scientific document represents well-reasoned scientific research aimed at answering specific questions and introducing new scientific knowledge. Therefore, the occurrence of scientific concepts in the texts of these documents is not completely arbitrary. The frequency (number of times) with which certain concepts appear in a particular document serves the specific goal of elucidating the scientific meaning of the research. We show that the statistical distribution of concept occurrences, commonly referred to as term frequencies (tf), across a collection of documents serves as a valuable source of information for studying the underlying dynamics of the knowledge network via the properties of concepts, akin to thermodynamic systems.

Scientific ontology is dynamic both in terms of the meanings of existing concepts and the emergence of new concepts, reflecting the current state of scientific progress. As new knowledge is incorporated, the meanings of scientific concepts are updated when related novelties are accepted by the scientific community. Each concept has multiple connections to other concepts, with the strongest connections providing the most relevant scientific context. The strength of these connections, which can be measured using mutual information metrics, changes over time in response to the stochastic input from newly published scientific documents.

Studying the properties of such complex information networks usually requires intricate calculations of the topological properties of corresponding graphs. Such calculations can become NP-hard as the size of the network increases. In this research, we propose an approach to analyse network properties whose computational complexity grows linearly with the size of the network. We demonstrate that the current state of each node (scientific concept) in the network, when defined as a state of a thermodynamic system, effectively reflects all past and current network communications. In this approach, nodes process information from the network and adjust their states accordingly. Estimating the state parameters of a node, even from a historical perspective, requires fewer computational resources. The computational costs grow linearly with the number of nodes, rather than exponentially as in topological calculations.

The proposed approach allows us to apply the reasoning and methods developed in statistical physics for over a century to analyze purely information systems during information transfer and knowledge production. The key connection between thermodynamics and information science lies in recognizing the fact that the entropy of any system, physical or not, is not its intrinsic property but a function of the variables we choose to define its states. Different choices of variables will correspond to varying entropies for the same system [15]. This resolves the longstanding discussion about whether Shannon and Boltzmann's entropy are the same quantity. The only distribution of variables in which Gibbs-Shannon entropy equals thermodynamic entropy is the generalized Boltzmann distribution [16]. If the above condition is satisfied, we can use thermodynamics to study information systems.

In Section 2, we show that the coarse-grained frequencies of scientific concepts extracted from multiple documents follow the generalized Boltzmann distribution, thereby allowing for a thermodynamic description. Formally, the necessary probability mass function was obtained in [12] using Jaynes' 'MaxEnt' method. The relevant constraints applied in the process of entropy maximization represent the generalized energy E of a system, which, as in physics, represents the cost needed to assemble a system of particles. The cost for a joining particle (concept term frequency tf from a new document) to enter the existing community can be expressed in terms of *economies of scale*, where the cost depends on the current size of the community [17]. Initially, for a small community, energy grows almost linearly with system size. When the system size becomes large, due to the absence of a correlation length scale, the energy of the system does not increase linearly with system size, giving rise to a non-extensive energy function. This view is consistent with the appearance of power laws in critical phenomena, where interactions are effectively long-ranged.

From a thermodynamic point of view, when E is a combination, as in our case, of logarithmic and linear terms, the system interacts with two types of thermal baths. The type of bath with which the system is currently interacting the most depends on the size of the system. State dynamics then reflect the crossover of the corresponding tf distribution from exponential to power law. The size of the system under study can be understood in terms of the number of stable connections with other concepts in the corresponding network, or in other words, the number of its active information channels. This situation is similar to establishing neural links in a brain, where the complexity of interactions increases with the number of connections, leading to changes in the system's overall behavior and properties [18–21]. In physics and social science, this behavior indicates the presence of a phase transition in a diverse range of examples, such as magnetic systems near the Curie point, polymer chains in solution, and the spread of information in social networks [22, 23].

In the realm of thermodynamics and information theory, the statement that 'information is energy' highlights the profound connection between informational and physical systems. Just as energy drives physical processes, information drives the dynamics of complex information networks. This perspective not only bridges the gap between thermodynamics and information theory but also provides a framework for understanding the dynamics of knowledge production and dissemination in scientific communities.

The aim of this research is as follows: (i) Based on the fact that Gibbs-Shannon entropy, as introduced in [12], is the entropy of a thermodynamic system, calculate the internal energy, Helmholtz free energy, temperature, and heat capacity of a scientific concept as a closed thermodynamic system connected to two types of thermal baths; (ii) using a data-driven approach based on frequencies tf of more than 12,000 physical concepts in over 500,000 scientific papers published on the ArXiv preprint server from 2002 to 2018, map the state space of the concept's knowledge network with analogs of thermodynamic diagrams; (iii) examine the dynamics of the state space, focusing on the regions where states are densely populated, which represent the most probable values for the network's thermodynamic parameters.

This paper is organized as follows. Section 2 elaborates on the various types of states used to describe concepts as information systems, based on available data concerning their term frequencies. It introduces definitions for concept information entropy and thermodynamic entropy per scientific document, alongside temperature and internal energy. Section 3 discusses the connection between Helmholtz free energy, thermodynamic work, residual entropy, and the production of knowledge (entropy). Section 4 presents knowledge network state maps in the form of thermodynamic diagrams. Finally, Section 5 provides conclusions and a future outlook.

2 Information and thermodynamic entropy

The relationship between thermodynamic entropy and its information analog, introduced by Shannon, has sparked numerous debates and misconceptions within the scientific community [16, 24, 25]. Thermodynamic entropy, a concept deeply rooted in classical physics, describes the measure of disorder or randomness within a physical system. In contrast, Shannon entropy, originating from information theory, quantifies the uncertainty or unpredictability inherent in a set of data or information [26]. The potential solution to understanding how these concepts are related lies in recognizing that entropy should not be viewed as an inherent property of the system itself but rather as a characteristic of how we describe it [27, 28]. As Caticha stated [29], "*entropy is fundamentally tied to our method of observation and description rather than being an intrinsic property of the system under study.*" More explicitly, entropy is a function of the macroscopic variables chosen to define the macrostate. For instance, different macrostates, based on varying choices of variables, will correspond to different entropy values for the same system. However, entropy is not solely dependent on the macrostate. As demonstrated below, entropy reflects a relationship between different descriptions of the same system. Besides the macrostate, we can also consider the set of microstates and mesostates. The differences in entropy among these states indicate the amount of additional information required to identify the other states.

Having a concept c, we will define the set of its microstates associated with probability $1/N_c(t)$ for a concept to be found in the document, where $N_c(t)$ is the number of documents that contain concept c and was published until date t. The uniform probability distribution of the concept microstates and the corresponding maximal value of the entropy per document $S_{micro} = \ln N_c(t)$ corresponds to the absolute minimum information we have about the system in this representation [30, 31].

The more detailed text analysis of the documents enables the collection of additional information about concept occurrences, or more commonly, concept frequencies denoted as tf. Grouping documents based on extracted frequency values $tf = k, k \in \mathbb{Z}^+$ associates the probabilities for a concept to be cited exactly k times with a set of *mesostates*. The probability of a particular frequency is given by the relation $P_c(k,t) = N_c(k,t)/N_c(t)$, where $N_c(k,t)$ is the number of documents that mention a concept k times up to time t. The corresponding Shannon entropy per document is then given by [1].

$$S_{meso}(t) = -\sum_{k=1}^{\max(tf_c)} P_c(k,t) \ln P_c(k,t). \tag{1}$$

The value of the mesostate entropy $S_{meso} < S_{micro}$ reflects the collection of additional information about the state of the concept, thereby reducing state uncertainty. The natural question arises: what is the minimum value of k that should be used to define the corresponding entropy? Martini addressed this question in [12] and concluded that entropy defined with $k > 0$ provides greater descriptive power for topic classification in the concepts network. The choice $k_{\min} = 1$ in our case is motivated by the proposed thermodynamic framework, where it sets the minimum energy level $E = \ln\ 1 = 0$ for the logarithmic bath of a concept as a thermodynamic system.

The Jaynes MaxEnt principle [28] allows us to define a normalized macrostate probability mass function $\Pi_c(k,t)$ of a concept as the one that maximizes the S_{meso} entropy. Applied constraints in the form of the first moment $\langle k \rangle$ and log-moment $\langle \ln k \rangle$ of $P_c(k,t)$ give:

$$\Pi_c(k,t;\beta,\lambda) = \frac{1}{Z}\frac{e^{-\lambda k}}{k^\beta}, \quad Z = \sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^\beta} = Li_\beta(e^{-\lambda}) \tag{2}$$

where Z is the polylogarithm Li of order β and argument $e^{-\lambda}$ for the Lagrangian multipliers $\lambda > 0$ and β [12]. The corresponding macrostate entropy is then given by

$$S_{macro} = \ln Z + \beta\ (\langle \ln k \rangle + \frac{\lambda}{\beta}\ \langle k \rangle) \tag{3}$$

with

$$\langle k \rangle = \frac{Li_{\beta-1}(e^{-\lambda})}{Li_\beta(e^{-\lambda})}, \quad \langle \ln k \rangle = -\frac{\partial_\beta Li_\beta(e^{-\lambda})}{Li_\beta(e^{-\lambda})}. \tag{4}$$

Similar results were obtained in [32], where it was shown that power-law and exponential distributions independently maximize the ratio $Q = S_{macro}/E$ with the exponents β or λ approximated to it [33, 34].

---

[1] The units of entropy in this study are denoted as *nats*, corresponding to base 'e' (Euler's number) logarithms, which are used throughout. While it is more common in the literature to use the base-2 logarithm, in that case, the units of entropy are expressed in *bits*, aligning with the binary digit terminology commonly associated with information theory.

Following the formalism developed in [16], the Eq.(2) can be understood as a generalized Boltzmann distribution of a concept as a system of logical particles each representing a scientific document. The entropy of the macrostate S_{macro} is then related to the thermodynamic entropy S_{therm} as:

$$S_{therm} = k_B S_{macro} = k_B(\beta E + \ln Z), \tag{5}$$

where k_B is the Boltzmann constant, β corresponds to the inverse temperature $1/k_B T$ and E is the average internal energy per document:

$$E = \langle \ln k \rangle + \frac{\lambda}{\beta} \langle k \rangle. \tag{6}$$

In physics, the concept of internal energy E defines heat, and there is no distinction between a system's total and internal energy if a reference frame is attached to a concept. Therefore, the second term $\beta^{-1} \ln Z$ in Eq.(5), which is commonly related to the motion of the system boundaries (volume change), is associated with work exclusively. Since concepts as information entities do not have traditional spatial boundaries, it is nontrivial to calculate work from the change in a system's volume. We would have to define a space with a common reference frame where all concepts interact as separate entities. We will consider that concepts define the surroundings for each other, and each interaction of target concepts with another will add a new dimension; we will end up in a multidimensional space, the dimensions of which will vary for different target concepts. A concept's corresponding volume can then be calculated as a simplex if the distance measure between concepts is taken in the form of normalized mutual information [35].

We can avoid this complication by estimating work from the corresponding system's free energy change. In this case, we will effectively consider all informational interactions between concepts. In the next section, we will use Helmholtz's free energy and Jarzynski's equality to quantify thermodynamic work and relate it to the system's entropy/information production.

3 Residual entropy and free energy

As evident from the previous section, the scientific concept can be represented as a *closed* thermodynamic system, with its energy defined by Eq.(6). The scientific concept as a thermodynamic system appears in a state of equilibrium when Shannon-Gibbs entropy S_{meso} equals Boltzmann entropy S_{macro}. Such concepts were named in [12] as basic or generic and considered as not useful for fine-grained topic detection. Presumably, they can be found in almost every document of a studied collection and represent very broad topics that can be identified with whole scientific fields. The latter is correct if equilibrium is permanent and the entropy of a concept is relatively large in comparison to other concepts. But if equilibrium is instantaneous and entropy is relatively small, it indicates special conditions for a concept when it is related to a hot topic [36]. In general, for the rest of the concepts, the following inequality holds $S_{macro} > S_{meso}$, implying the fact that these concepts can be considered as non-equilibrium thermodynamic systems.

The "age" of a concept, determined by the date of the earliest article in which it appeared within the studied document collection, influences the amount of historical data available regarding its term frequencies. This age generally affects the concept's proximity to thermodynamic equilibrium. While relatively new concepts can achieve a macrostate with maximal entropy, they seldom attain a true steady equilibrium. The macrostate parameters λ and β are time-dependent, with their most probable values indicating the equilibrium state of a concept. In Section 4, we will demonstrate how a concept's age correlates with the values of λ and β and its proximity to equilibrium.

The amount of additional information required to specify the mesostate with respect to a macrostate is defined via residual entropy $R = S_{macro} - S_{meso}$ that can be expressed as the Kullback-Leibler divergence (see proof in [12]) between mesostate and macrostate as a state of instantaneous thermodynamic equilibrium:

$$R(t) = D_{KL}(P_c||\Pi_c) = \sum_k P_c(k,t) \ln \frac{P_c(k,t)}{\Pi(k,t)} \geq 0. \tag{7}$$

The net change of the residual entropy $\Delta R = R_f - R_o$ in a process, when the system changes from some initial to end states (o and f), defines the entropy production in the system [37, 38] and in terms of the entropy change, it is the difference $\Delta R = \Delta S_{macro} - \Delta S_{meso}$. According to the second law of thermodynamics, ΔR must decrease, indicating that the total entropy production is always positive or zero, ensuring the irreversibility of natural processes.

In cases where work is performed on the system, the entropy of the system itself may decrease, implying that ΔR gets larger. However, the total entropy of the combined isolated system (system + surroundings) remains constant or increases, as stated by the second law of thermodynamics. Drawing an analogy to a system in thermodynamic equilibrium, a decrease in entropy production becomes particularly intriguing, as it potentially allows us to quantify the work required to move the system out of equilibrium. In the context of an information system, this 'work' is defined as the additional *useful information* R transmitted into the system, which decreases its meso entropy.

In thermodynamics, the amount of energy that can be converted to work and is not tied up in the entropy of the system is given by Helmholtz free energy:

$$A = E - k_B T S_{macro} = -k_B T \ln Z. \tag{8}$$

The change in Helmholtz free energy ΔA tells us about the maximum obtainable work from a process when the system passes from o to f state. Helmholtz's free energy change ΔA reaches its minimum value at equilibrium, and no work is possible as $\Delta A \to 0$ and $\Delta R \to 0$. Non-equilibrium systems can do work, and the change in Helmholtz free energy (ΔA) reflects the maximum theoretical work obtainable under constant temperature and system volume. This ΔA value can be positive or negative depending on whether the system does work on the surroundings ($\Delta A > 0$) or the surroundings do work on the system ($\Delta A < 0$).

The relation of useful information R to Helmholtz free energy is then followed from Eq.(8):

$$R = \frac{E - A}{k_B T} - S_{meso} \quad (T, V = const). \tag{9}$$

The useful expression to quantify work, that is valid for the systems in- and far-from-equilibrium regime, is given by the recently discovered Jarzynski equality [39, 40], representing the relationship between the difference in free energy ΔA of two equilibrium ensembles and the amount of work W needed to switch between them in a finite amount of time:

$$\langle e^{-\beta W} \rangle = e^{-\beta \Delta A}. \tag{10}$$

Here, $\langle ... \rangle$ indicates an average over multiple repetitions of the process, and $\beta = 1/k_B T$.

Noting that Helmholtz free energy is a function of state from Eq.10, we can measure the work needed to drive the system between o and f states separated by the time interval Δt as proportional to ΔA. In the case of an isothermal process, the analytical expression for the free energy change is simple and given by:

$$\Delta A = A_f - A_o = \Delta E - k_B T(\Delta R - \Delta S_{meso}). \tag{11}$$

In this case, we can also write a simple analytical expression for the efficiency of information flow η_{inf} represented by the dimensionless ratio of information transfer ΔR over total irreversible entropy production ΔS_{meso} [41]. Using Eq.(11), we get:

$$\eta_{inf} = \frac{\Delta R}{\Delta S_{meso}} = \frac{1}{k_B T} \frac{\Delta E - \Delta A}{\Delta S_{meso}} - 1. \tag{12}$$

We can see that for the reversible near-equilibrium processes $\Delta S_{macro} \approx \Delta S_{meso}$, $\Delta S_{macro}/\Delta E = 1/T$ and $\Delta A = 0$ the efficiency of information flow is minimal $\eta_{inf} = 0$ unlike thermodynamic efficiency, which is maximized under these conditions [41].

Although the thermodynamic evolution of a concept as a thermodynamic system may seem like a simplification, and the actual dependence of ΔA on other thermodynamic parameters is more complex, it can be a valid model for concepts in a state of thermodynamic equilibrium. As we will show below, concepts that are used in a sufficiently large number of documents all reach equilibrium at a specific temperature and consequently all have low values of η_{inf}. In contrast, new concepts have larger values of information flow efficiency as their residual and S_{meso} entropy vary significantly during their state evolution.

The thermodynamic analogy suggests a profound connection between the effective use of information by the information system and efficient thermodynamic operation. The system dynamics performs a computation by changing its state as a function of the driving signal, in our case, the concepts tf in each new document. The system's state retains information about past environmental fluctuations, and a fraction of this information is predictive of future ones [42]. The closer the state of a concept is to thermodynamic or 'infodynamic' equilibrium, the more predictive information is available. The system's efficiency is then related to the amount of useful information (or persistent knowledge) it can keep about its environment.

4 Thermodynamic diagrams

In this section, we will calculate the macrostate parameters β and λ from Eq.(2) based on historical data about concepts tf and analyze the concept's state space using thermodynamic diagrams. For each of the 13,945 concepts recorded in our database, we try to find these parameters by numerically fitting the empirical mesostate distribution formed for a specific time interval using the *maximum likelihood* estimation method. We cross-check the results by requiring that for the obtained β and λ the Eq.(4) must hold with a tolerance of $< 10^{-5}$ when $\langle \ln k \rangle$ and $\langle k \rangle$ are obtained directly from empirical data.

Figure 1 presents a heat map illustrating the distribution of concepts across different values of the state parameters β and λ. We can see that the largest number of concepts have λ in the interval from 0 to 0.15 and β from 1 to 2. The maximum concept density in macrostate parameters space shifts from $\beta = 2$ when $\lambda = 0$ to $\beta = 1.6$ when $\lambda = 0.15$. Similar results were previously obtained in [12] where the mean value for the inverse temperature distribution was reported to be $\overline{\beta} \sim 3/2$. We find that this value remains almost constant for the studied period, with $\overline{\beta} = 1.51$ until the end of 2010 to $\overline{\beta} = 1.61$ in 2018.

The difference between states with different β and λ is better understood in terms of probabilities of *information channels* $K = \{k\}$ that form the mesostate distribution. The number of microstates (documents) that contribute to a particular channel k quantifies the history of a system along the specific microscopic phase-space path. The probability for the system to follow this path is then given by $P_c(k,t)$, which changes over time as the system receives new information about new relevant microstates. For the concepts with a large number of microstates, we have an almost sequential filling of information channels starting from $k = 1$, which has the largest probability $P_c(k = 1, t)$ and which is decreasing for $k > 1$ mostly in a power law manner with non-zero probabilities for $k >> 1$.

Our analysis shows that the larger the values of β, the narrower the channel band with small k indexes and consequently lower entropy S_{meso}, S_{macro}, energy E, and temperature. For most concepts, λ, in this case, is small, and consequently, the linear contribution to E in Eq.(6) is small compared to the logarithmic term. From Fig.1, we can see that the number of concepts with the pure logarithmic spectrum, i.e., when $\lambda \to 0$, is increasing towards $\beta = 2$ and then decreasing when $\beta > 2$. The role of a linear term in E becomes more important with the temperature of a system. Our results show that the number of concepts with a combined spectrum rises towards $\lambda = 0.1$ until $\beta \to 1$. For higher temperatures, the linear spectrum becomes dominant.

Our results complement and extend previous research in social collaborations [32] where it was shown that for the pure power-law distributions, if $\beta > 2$, the main contribution to E is made by the low energy (low k) information channels. If $\beta < 2$, more contribution is obtained from high energy (high k) channels, and when $\beta = 2$, all channels contribute evenly. Those conclusions were made assuming an infinite number of information channels k but their number in real-world datasets is not infinite. The exponential cut-off in the macrostate probability mass function $\Pi_c(k,t)$ accounts for this fact by setting it to zero for some $k > k_{max}$. The smaller β, the larger λ is required to keep $\langle k \rangle$ from becoming infinite. When λ is large and β is small, the entropy of the

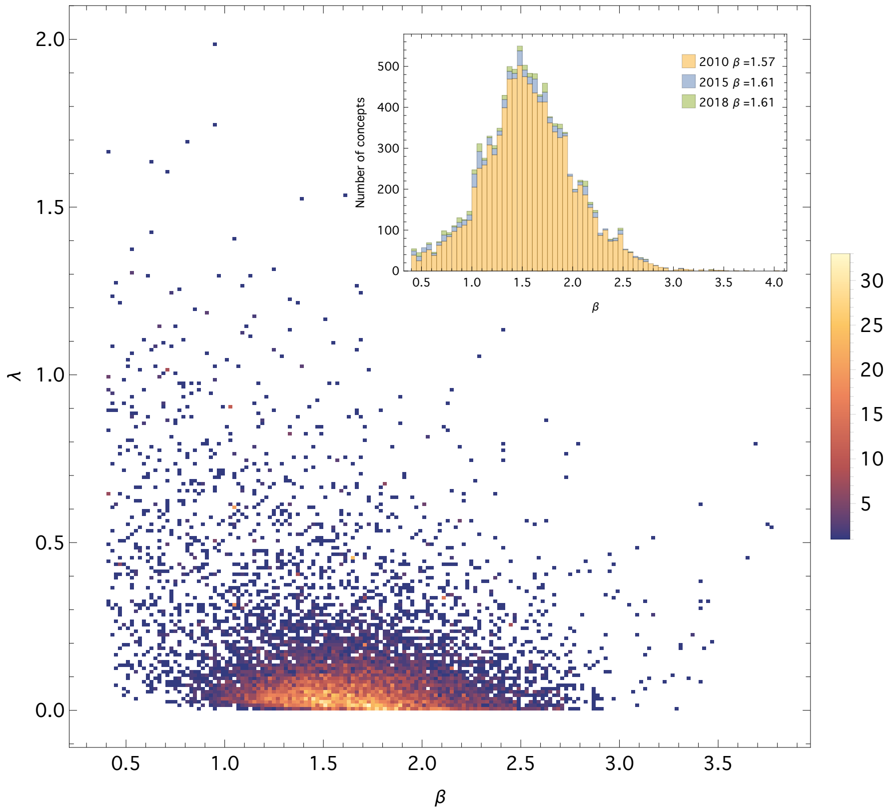

Fig. 1 A heat map illustrating the distribution of macrostate parameters for 11737 scientific concepts. The sub-figure demonstrates the distribution of inverse temperature β (where $\beta > 0.001$) and the evolution of the distribution over three time periods, beginning in 2002 and concluding in 2010, 2015, and 2018. The mean value of β, which is specified, is calculated at the end of each period.

concept is influenced by both the logarithmic and linear contributions, leading to a complex interplay of information channels with varying probabilities.

In each studied period, there is a group of new concepts whose states are so far from equilibrium that no macrostate parameters can be estimated based on the available term frequency (tf) distribution. These concepts represent novelties in the document collection that may later gain a wider range of frequencies and develop into separate topics or become part of existing ones. However, during the studied publication period, these concepts are found in a very small number of documents. As a result, their microstates are mostly characterized by a single $k = 1$ frequency, leading to a mesostate entropy of $S_{meso} = 0$. We exclude such concepts from the macrostate parameter estimation process and will include them back in future analyses when their distributions are sufficient to allow accurate model fitting. We can study different time periods by shifting the initial and final dates, but we prefer to fix the initial date and change the final date, including more and more data. The concepts excluded in shorter time periods due to insufficient data may appear in the extended period analysis since we may now have enough data and consequently the possibility to estimate their thermodynamic parameters. However, we currently have data only up to 2018, so we are unable to shift the final date further. Therefore, in the given example, 2,208 concepts with insufficient data are not analyzed and can only be analyzed if we gain access to data from later years. As illustrated in Fig. 2, the concept 'Anomalous Hall effect' demonstrates this issue: it took four years of data collection, from 2000 (cyan line) to 2004 (orange line), to calculate the system's thermodynamic parameters with the required level of accuracy.

Figure 2 indicates values of the concept's state parameters in 2018 concerning the time of their origin in the texts of the studied collection of documents. We can conclude that states with minimal residual entropy $R \to 0$, associated with the state of thermodynamic equilibrium, are more common for concepts first introduced in the oldest documents, although instantaneous equilibrium can be reached by relatively new concepts also. From Fig. 2(c), we can see that not all 'old' concepts have maximal entropy (minimal R) but many have power-law tf distribution (i.e. $\lambda \to 0$). Concepts that appeared in a state of stationary thermodynamic equilibrium, when R is constant and small for an extended period (marked green in Fig. 2), appeared to have both of these properties – near power-law distribution ($\lambda \to 0.02$) and maximal possible entropy ($R \to 0$). We can assume that this limit is a future of all old concepts with well-established contextual meaning, which is what we can expect in the case of physical systems according to the second law of thermodynamics.

Figures 2(a),(d) show that the free energy A and temperature of concepts in a state of stationary equilibrium tend to certain non-zero values, indicating a stable balance between the entropic forces and the energy landscape governing the frequency distribution of these concepts. Other concepts, appearing in the state of instantaneous equilibrium, have similar values of the (inverse) temperature, close to the most probable $\overline{\beta} = 3/2$ evident from Fig. 1, but for many of them, free energy A has smaller values than for the stationary state concepts.

As we can see from Figs. 2 and 3, the state of a concept is related to the number of documents where it can be found. The 'age' of a concept and the number of such documents are correlated quantities such that older concepts tend to appear in a larger number of documents. We can draw the conclusion that if the number of such documents exceeds 1000, the state of a concept is mostly in equilibrium (having maximum entropy for the particular term frequency distribution, meaning that Shannon-Gibbs and Boltzmann state descriptions yield the same entropy). Its further evolution is more like that of reversible physical systems. Concepts' term frequency distributions tend to become more power-law-like as the number of relevant documents increases. The most frequent concepts are in a stationary state, tend to be the oldest, and have, on average, larger values for entropy and internal energy compared to other equilibrium state concepts. The amount of free energy is generally lower for such concepts than for any others, indicating that they have done more work on their surroundings. Their current free energy change has to be the lowest among other equilibrium concepts.

4.1 Energy-entropy diagram.

To explore the state space of the concepts information network as a collection of interacting thermodynamic systems, we map the states of concepts onto a series of state diagrams. One of the common diagrams is an energy-entropy (E-E) diagram, shown in Fig. 4 and Fig. 5. Dashed blue lines on these diagrams represent the largest S_{macro} entropy for different constant cut-off parameter values $\lambda > 0$ calculated from Eqs.(3),(4), and (6). The solid blue line shows the maximal entropy of a macrostate obtained if the distribution is very close to a power-law (i.e. $\lambda = 0.001$). From the Fig. 4 diagram, we can conclude that the state of instantaneous thermodynamic equilibrium can be approached almost for any value of λ parameter representing the equilibrium metastable states, but most of the concepts reach true stable equilibrium when $\lambda < 0.05$.

In Fig.5, we can see examples of such stable equilibrium concepts as "Diquark" and "Mass," which have almost constant energy and entropy for a period of 15 years. The concept "Diquark" shows a trend common to many equilibrium concepts: a slow temperature drift towards lower temperature and energy at constant λ. Another example of the concept state dynamics is given by the "Anomalous Hall effect" concept. In Fig.5, we observe the combined mesostate (orange line) and macrostate (red line) dynamics for this concept in energy-entropy coordinates. It shows a continuous decrease in residual entropy R and internal energy E during the concept's state evolution. This concept reached equilibrium at a temperature $\beta = 1.5$ in 2018. In 2004, at the beginning of the studied period, the state parameters of the concept exhibited much larger fluctuations, which diminished as the concept approached equilibrium closer to 2018, similar to the trend observed in Fig. 2.

The example of the 'Anomalous Hall effect' concept is representative of most new concepts until they reach equilibrium at maximum possible entropy S and minimal energy E. This behavior aligns with what we would expect from physically closed thermodynamic systems according to the second law of thermodynamics. As suggested by Peng et al. [32], the ratio termed *entropy efficiency*, $Q = S_{macro}/E$, which is maximized for equilibrium states, represents another interpretation of the second law: 'a system would use the minimum energy to produce the same amount of entropy.' In Fig. 6, we show the diagram of states for entropy efficiency Q plotted against inverse temperature.

As we can see from Fig.6, there is an almost linear correlation between the largest possible entropy efficiency Q for a concept and its inverse temperature. Concepts with the smallest R and λ appear to have the highest entropy efficiency, and the concept "Anomalous Hall effect" state dynamics shows an example of how general concepts reach this limit. In Fig.6, if we change S_{meso}/E to S_{macro}/E on the y-axis, the observed linear correlation between Q and β will not change; only the concepts with small λ will become concentrated near the upper dashed line. We can see that entropy efficiency is increasing towards lower temperatures, which may explain the observed state dynamics of non-equilibrium and equilibrium concepts in Fig. 6. The evolution of equilibrium concepts towards higher efficiency is much slower than for non-equilibrium systems, which suggests the existence of different mechanisms of information (energy) exchange for these systems.

4.2 Free energy diagrams.

In this section, we analyze diagrams showing the dependence of Helmholtz free energy on other thermodynamic functions and state parameters. From Fig.7, we can see that non-equilibrium concepts with a close to power-law tf distribution ($\lambda \to 0$) identify macrostates with the largest value of $E - A = TS_{macro} > 0$ with $A < 0$, representing the energy tied up in entropy at a given temperature. As the temperature increases ($\beta \to 0$), this energy grows and becomes divergent below $\beta = 1$. The energy $E - A$ is in general smaller for the equilibrium concepts, especially in cases when A is positive. The stationary equilibrium concepts, which have the smallest average λ observed among other equilibrium concepts, maintain the largest value of TS_{macro}. The difference in TS_{macro} energy for equilibrium and non-equilibrium concepts becomes less pronounced for lower temperatures, after which $\beta = 2.8$ almost all concepts appear in equilibrium.

Both E and A functions have a minimum at specific values of temperature and the λ parameter. The minimum possible value for the internal energy E increases with the system's temperature and λ, and is mostly represented by equilibrium concepts. Across a wide range of temperatures, the energy minimum corresponds to λ values between 0.2 and 0.4 (see Fig.8), indicating that both terms in Eq. 6 contribute to the system's internal energy. In contrast, the minimum of free energy A is mainly represented by non-equilibrium concepts with a close to power-law tf distribution.

The residual energy $TR = TS_{macro} - TS_{meso}$ is what distinguishes different non-equilibrium concepts with a near power-law distribution from equilibrium concepts at the same temperature. This residual energy can be regarded as a form of potential energy that complements the free energy A and is necessary to adjust the probabilities of the information channels k of a given concept's mesostate to achieve maximum entropy. As previously mentioned, the residual entropy R represents the information needed to specify the mesostate based on the system's macrostate description. When a concept reaches an equilibrium state, $R \to 0$, and any net change in residual entropy ΔR signifies information transferred from the system to its surroundings.

We can assume that the system utilizes the potential energy TR for information transfer, which can be viewed as a form of creative potential realized through increasing mutual information with other concepts. This residual energy is used to perform work by creating or strengthening mutual information with other concepts. In this way, the knowledge network increases its dynamic complexity by adopting new knowledge. If mutual information cannot be decreased, then this process is irreversible. However, from the analysis of the dynamics of mutual information between concepts, we know that mutual information can decrease [35]. This process appears to be less probable and reduces the information interaction between concepts. In such cases, a concept gains additional information from the surroundings, which increases its free energy and residual entropy.

As we can see from the example given in Fig. 8 for the observed energy-temperature state evolution of the "Anomalous Hall effect" concept, the observed thermodynamic process initially is not isothermal and non-equilibrium later, as the temperature oscillations subsided, becomes closer to equilibrium.

Figure 7 shows that the limit of the smallest $\lambda < 0.05$ is represented mostly by the non-equilibrium concepts at the early stage of their evolution. In Fig.9, it is shown that these group concepts have another interesting property – the dimensionless ratio A/E, obtained from Eq.9, remains almost constant for the wide range of dimensionless quantity known from [32] as the entropy reduction ratio $1 - TS_{meso}/E$, i.e., the part of the system's energy that is effectively available for performing useful work (or processing information) after accounting for the "cost" of entropy.

A constant A/E suggests that the efficiency of energy conversion (from total energy to useful work) does not drastically change for the near power-law concepts despite variations in entropy and temperature. This might imply a robust mechanism of energy allocation and utilization within the network of concepts. Moreover, this group of concepts has the largest heat capacity (see Fig.10) $C = \partial E/\partial T$ compared to others, indicating a higher capacity to absorb and retain heat energy.

The heat capacity C of the concepts is maximized around the most probable temperature and eventually becomes zero at very high ($\beta \to 0$) or very low ($\beta \to 3.5$) temperatures. For the equilibrium concepts, it shows a variety of different values across the available range of temperatures. Concepts in a state of stable equilibrium (green color on Fig.10) have larger heat capacity and temperature among other equilibrium concepts with small residual entropy R.

The value of a concept's free energy changes along its state path trajectory, reflecting the role of the concept in its interaction with its surroundings. A concept can either give energy to or obtain it from the thermostat it interacts with. In this way, we can say that a concept does work on other concepts, or that a group of concepts does work on this particular concept, changing the direction of information transfer. We find that there is a correspondence between the change in free energy ΔA and the concept's entropy dynamics, which vary the value of its residual entropy ΔR. From Fig. 11, we can see a connection between information transfer and free energy change in our dataset for the period 2017-2018. It shows that if A is increasing, the residual entropy is also increasing, and if free energy is decreasing, the residual entropy is decreasing.

Concepts in a state of stationary equilibrium show the minimal variations of free energy and residual entropy in correspondence to other concepts. Concepts that have a tf distribution close to power-law but have not yet reached an equilibrium state have the largest variations of both A and R. The example of the "Anomalous Hall effect" concept shows the typical stochastic trajectory for the concept state in these variables.

5 Conclusions

In this research, we analyze more than 11,000 scientific concepts, which represent units of scientific ontology in the domains of Astronomy and High Energy Physics, as thermodynamic systems. We demonstrate that the thermodynamic states of these concepts can be estimated from the accumulated historical information about their term frequency distributions. We used as a source a ScienceWISE database of scientific concepts' term frequencies extracted from multiple documents published in the ArXiv:HEP section over the 1992-2018 period. The features of the obtained distributions allow us to associate them with a generalized Boltzmann distribution and represent each concept as a closed statistical ensemble system. We show how to formulate and calculate several types of Shannon entropy S to describe the state of a concept, with the entropy of a macrostate given by the distribution from Eq. (**??**) corresponding to thermodynamic entropy. The proposed approach allows us to formulate known thermodynamic quantities, such as internal energy E, Helmholtz free energy A, and temperature T, to describe the state of a scientific concept.

The use of scientific concepts in document texts is intended to encode and transfer scientific information. Each concept is not an independent unit; it interacts with other concepts, storing complex scientific knowledge. Connections between concepts form a network that stores, processes, and distributes this information. Each node in this network can be represented as a system interacting with multiple other systems, similar to how a thermodynamic system interacts with its surroundings in physics. Consequently, the state of each node reflects the information from all its past interactions.

At each moment, the surroundings for some specific node are a complex thermostat, represented by a subset of concepts it is currently interacting with. The states of concepts in a thermostat can be steady or time-dependent. We assume that steady-state equilibrium concepts (see green dots on Fig.10) are an essential part of the global thermostat that sets the most probable temperature for the whole network irrespective of the stochastic update from external sources.

By analogy with physical systems, such a thermostat increases or decreases the temperature of concepts, eventually bringing their states closer to the global equilibrium temperature. The energy and temperature of each concept are defined by its frequency distribution, which can be inferred from mutual information with its surrounding concepts. The more mutual information we have between a pair of concepts, the closer their frequency distributions and, consequently, other thermodynamic parameters like temperature are.

The structure of the metadata collected about concepts' term frequencies allows us to estimate their current state and trace the dynamics of all thermodynamic variables, revealing the universal properties of their state evolution. We find that the evolution of a concept, as a thermodynamic system, starts from a highly non-equilibrium state with high residual entropy (R). After a period of quasi-equilibrium, where residual entropy is already low, the concept may reach a state of stationary thermodynamic equilibrium where parameters like temperature (T) and internal energy (E) remain constant. The time needed to reach equilibrium varies for different concepts. Our analysis indicates that 1,000 documents is a threshold after which concepts have residual entropy low enough to reach an equilibrium state, though instantaneous equilibrium is possible with fewer relevant documents. We also find that the number of concepts in a state of stationary equilibrium slowly increases after this threshold.

Within the first 1,000 documents of a concept's evolution, as the number of microstates increases, we observe a phase transition from a state described by a linear energy spectrum and exponential term frequency tf distribution to a logarithmic spectrum corresponding to a power-law tf distribution. We show that at this phase transition, the heat capacity of non-equilibrium concepts with close to power-law tf distribution is maximized around the most probable concept temperature $T \sim 3/2$. Another interesting discovered property is that the ratio of the free-to-internal energy for all these concepts is distributed around the most probable value $A/E \sim -0.5$ for a large variation of the residual entropy R.

It is worth noting that for rare concepts, the observed phase transition does not fully align with the process of concepts' thermalization, as they can reach equilibrium before the exponential contribution to the tf distribution can be neglected. However, as the number of relevant documents grows, the energy spectrum for these concepts will also become logarithmic.

We argue that the state space path a concept takes, from complete uncertainty to stationary state thermodynamic equilibrium, reflects the evolution of concept-related scientific knowledge from its absence to a well-defined and stable understanding within the scientific community reflected in the context each concept finally obtains. The exact state space trajectory of a concept is reflected in the amount of useful information ΔR transferred to or from it to other concepts in each period. This representation supports and allows us to quantify Ariel Caticha's definition of information as what changes our beliefs [15]. In our case, useful information changes our current scientific knowledge by updating the context for the concept in a network and identifying the next state of a concept as an information entity.

We can measure useful information in terms of the amount of work needed to drive the concept along its state path. This work is justified by the amount of information or combined energy spent by many researchers on explaining a term's context and providing all necessary connections with other scientific concepts. Following the thermodynamic framework, we can relate work to a change in the system's free energy ΔA. Jarzynski's equality provides the necessary relation between these quantities and allows us to measure the amount of useful information associated with scientific concepts as the total difference between the free energy between some initial and final states. From Fig. 11, we can see that the sign of ΔA changes along the concept's state path trajectory, reflecting similar changes in residual entropy ΔR. When a concept does work ($\Delta A > 0$), its residual entropy change ΔR is positive; when surrounding concepts do work on the concept ($\Delta A < 0$), ΔR is negative.

These results demonstrate the universal nature of context evolution for scientific concepts, applicable across different collections of concepts and scientific domains. This thermodynamic approach provides a powerful tool for understanding and quantifying the dynamics of scientific knowledge production and dissemination.

References

[1] Chen, C.: Searching for intellectual turning points: Progressive knowledge domain visualization. Proceedings of the National Academy of Sciences of the United States of America **101**, 5303–5310 (2004) https://doi.org/10.1073/pnas.0307513100

[2] Crane, D.: Invisible Colleges; Diffusion of Knowledge in Scientific Communities. University of Chicago Press, Chicago, (1972)

[3] Jong, T., Joolingen, W.: Scientific discovery learning with computer simulations of conceptual domains. Review of Educational Research **68** (1998) https://doi.

org/10.3102/00346543068002179

[4] Lazega, E., Burt, R.: Structural holes: The social structure of competition. Revue Française de Sociologie **36**, 779 (1995) https://doi.org/10.2307/3322456

[5] Burt, R.: The Social Capital of Structural Holes, (2001). https://doi.org/10.1093/oso/9780199249145.003.0002

[6] Burt, R.: Structural holes and good ideas. American Journal of Sociology **110**, 349–399 (2004) https://doi.org/10.1086/421787

[7] Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chávez, C., Wojick, D.: Population modeling of the emergence and development of scientific fields. Scientometrics **75**, 495–518 (2008) https://doi.org/10.1007/s11192-007-1888-4

[8] Liben-Nowell, D., Kleinberg, J.: Tracing information flow on a global scale using internet chain-letter data. Proceedings of the National Academy of Sciences **105**(12), 4633–4638 (2008) https://doi.org/10.1073/pnas.0708471105 https://www.pnas.org/doi/pdf/10.1073/pnas.0708471105

[9] Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., Pellegrino, D.: Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics **3**(3), 191–209 (2009) https://doi.org/10.1016/j.joi.2009.03.004 . Accessed 2023-01-17

[10] Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Reviews of Modern Physics **74**(1), 47–97 (2002) https://doi.org/10.1103/RevModPhys.74.47

[11] Newman, M.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America **98**, 404–9 (2001) https://doi.org/10.1073/pnas.021544898

[12] Martini, A., Cardillo, A., Rios, P.D.L.: Entropic selection of concepts unveils hidden topics in documents corpora. ArXiv (2018) arXiv:1705.06510 [physics.soc-ph]

[13] Prokofyev, R., Demartini, G., Boyarsky, A., Ruchayskiy, O., Cudré-Mauroux, P.: Ontology-Based Word Sense Disambiguation for Scientific Literature. In: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M.Y., Weikum, G., Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval vol. 7814, pp. 594–605. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36973-5_50 . Series Title: Lecture Notes in Computer Science. http://link.springer.com/10.1007/978-3-642-36973-5$_5 0 Accessed 2023-07-14$

[14] Palchykov, V., Gemmetto, V., Boyarsky, A., Garlaschelli, D.: Ground truth? Concept-based communities versus the external classification of physics manuscripts. EPJ Data Science **5**(1), 28 (2016) https://doi.org/10.1140/epjds/s13688-016-0090-4 . Accessed 2023-07-14

[15] Caticha, A., Giffin, A.: Updating Probabilities. AIP Conference Proceedings **872**(1), 31–42 (2006) https://doi.org/10.1063/1.2423258 https://pubs.aip.org/aip/acp/article-pdf/872/1/31/11812479/31_1_online.pdf

[16] Gao, X., Gallicchio, E., Roitberg, A.E.: The generalized Boltzmann distribution is the only distribution in which the Gibbs-Shannon entropy equals the thermodynamic entropy. The Journal of Chemical Physics **151**(3), 034113 (2019) https://doi.org/10.1063/1.5111333 . Accessed 2023-07-18

[17] Peterson, J., Dixit, P., Dill, K.: A maximum entropy framework for nonexponential distributions. Proceedings of the National Academy of Sciences of the United States of America **110** (2013) https://doi.org/10.1073/pnas.1320578110

[18] Bear, M., Connors, B., Paradiso, M.A.: Neuroscience: Exploring the Brain, Enhanced Edition: Exploring the Brain, Enhanced Edition. Jones & Bartlett Learning, ??? (2020). https://books.google.pl/books?id=m-PcDwAAQBAJ

[19] Kandel, E.R., Schwartz, J.H., Jessell, T.: Principles of Neural Science, Fourth Edition. McGraw-Hill Companies,Incorporated, ??? (2000). https://books.google.pl/books?id=yzEFK7Xc87YC

[20] Bullmore, E., Sporns, O.: Complex brain networks: Graph theoretical analysis of structural and functional systems. Nature reviews. Neuroscience **10**, 186–98 (2009) https://doi.org/10.1038/nrn2575

[21] Beggs, J.: The criticality hypothesis: How local cortical networks might optimize information processing. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences **366**, 329–43 (2007) https://doi.org/10.1098/rsta.2007.2092

[22] Kardar, M.: Statistical Physics of Fields. Cambridge University Press, ??? (2007)

[23] Clauset, A., Shalizi, C., Newman, M.: Power-law distributions in empirical data. SIAM Review **51** (2007) https://doi.org/10.1137/070710111

[24] Thims, L.: Thermodynamics Information Theory: Science's Greatest Sokal Affair

[25] Bera, M.N., Winter, A., Lewenstein, M.: Thermodynamics from Information. In: Binder, F., Correa, L.A., Gogolin, C., Anders, J., Adesso, G. (eds.) Thermodynamics in the Quantum Regime vol. 195, pp. 799–820. Springer,

Cham (2018). https://doi.org/10.1007/978-3-319-99046-0_33 . Series Title: Fundamental Theories of Physics. http://link.springer.com/10.1007/978-3-319-99046-$0_3 3 Accessed 2024-01-08$

[26] Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**, 379–423 (1948). Accessed 2003-04-22

[27] Wallace, D.: Philosophy of Physics: A Very Short Introduction. Oxford University Press, Oxford (2021)

[28] Jaynes, E.: Information theory and statistical mechanics i. Physical Review **106**, 620–630 (1957) https://doi.org/10.1103/PhysRev.106.620

[29] Caticha, A.: Entropy, Information, and the Updating of Probabilities. Entropy **23**(7), 895 (2021) https://doi.org/10.3390/e23070895 arXiv:2107.04529 [physics.data-an]

[30] Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, ??? (2012). https://books.google.pl/books?id=VWq5GG6ycxMC

[31] Caticha, A., Mohammad-Djafari, A., Bercher, J.-F., Bessiere, P.: Entropic inference. In: AIP Conference Proceedings. AIP, ??? (2011). https://doi.org/10.1063/1.3573619 . http://dx.doi.org/10.1063/1.3573619

[32] Peng, H.-K., Zhang, Y., Pirolli, P., Hogg, T.: Thermodynamic Principles in Social Collaborations (2012) https://doi.org/10.48550/ARXIV.1204.3663 . Publisher: arXiv Version Number: 1. Accessed 2023-06-27

[33] Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions draft manuscript. Internet Mathematics **1** (2003) https://doi.org/10.1080/15427951.2004.10129088

[34] Mandelbrot, B.: An infromational theory of the statistical structure of language. Communication theory **486** (1953)

[35] Chumachenko, A., Kreminskyi, B., Mosenkis, I., Yakimenko, A.: Dynamics of topic formation and quantitative analysis of hot trends in physical science. Scientometrics **125** (2020) https://doi.org/10.1007/s11192-020-03610-6

[36] Chumachenko, A., Kreminskyi, B., Mosenkis, I., Yakimenko, A.: Dynamical entropic analysis of scientific concepts. Journal of Information Science **48**(4), 561–569 (2022) https://doi.org/10.1177/0165551520972034 https://doi.org/10.1177/0165551520972034

[37] Strasberg, P., Esposito, M.: Non-Markovianity and negative entropy production rates (2018) https://doi.org/10.48550/ARXIV.1806.09101 . Publisher: arXiv Version Number: 3. Accessed 2023-12-15

[38] Osara, J.A., Bryant, M.D.: Methods to calculate entropy generation. Entropy **26**(3) (2024) https://doi.org/10.3390/e26030237

[39] Jarzynski, C.: Nonequilibrium equality for free energy differences. Phys. Rev. Lett. **78**, 2690–2693 (1997) https://doi.org/10.1103/PhysRevLett.78.2690

[40] Jarzynski, C.: Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. Phys. Rev. E **56**, 5018–5035 (1997) https://doi.org/10.1103/PhysRevE.56.5018

[41] Allahverdyan, A.E., Janzing, D., Mahler, G.: Thermodynamic efficiency of information and heat flow. Journal of Statistical Mechanics: Theory and Experiment **2009**(09), 09011 (2009) https://doi.org/10.1088/1742-5468/2009/09/P09011 . Accessed 2024-06-13

[42] Still, S., Sivak, D.A., Bell, A.J., Crooks, G.E.: Thermodynamics of Prediction. Physical Review Letters **109**(12), 120604 (2012) https://doi.org/10.1103/PhysRevLett.109.120604 . Accessed 2023-12-13

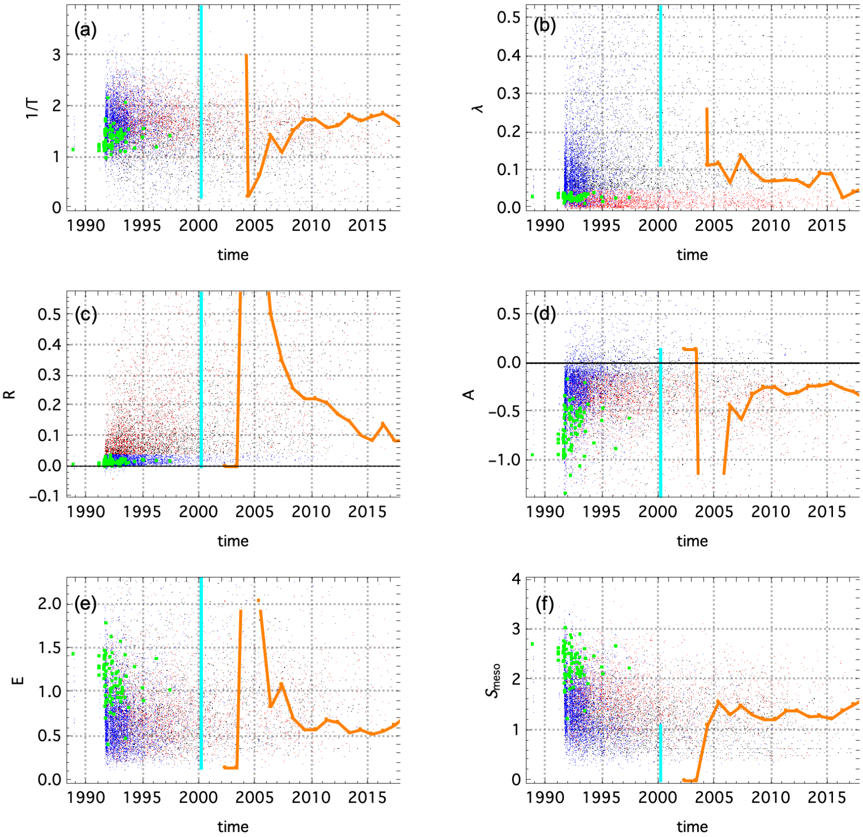

Fig. 2 The thermodynamic state parameters β, λ, along with residual entropy R, free energy A, internal energy E and mesostate entropy S_{meso} for 11,737 physical concepts estimated from tf data for the period from 1991 to 2018 and ordered by the time the concepts were first mentioned in the particular document in the studied collection (ArXiv document repository started accepting documents from August 14, 1991). Blue dots indicate concepts that appeared in a state of instantaneous thermodynamic equilibrium in 2018, defined as having minimal residual entropy $R < 0.04$ (maximal entropy). Green dots denote concepts in a stationary thermodynamic equilibrium state, where residual entropy $R < 0.04$ remains constant for the period from 2000 to 2018. Red dots highlight concepts with near power-law term frequency distribution ($\lambda < 0.04$), while black represents all other concepts that do not fit into the mentioned parameter range. The state evolution calculated with one year time step for the concept 'Anomalous Hall effect' illustrates a four-year gap between the cyan (date of first occurrence) and orange line, during which the macrostate parameters β and λ could not be calculated with the required precision due to insufficient data on the concept's term frequency. The figures demonstrate a general trend in the state dynamics of concepts towards equilibrium (small R) and a finite temperature ($\beta \sim \frac{3}{2}$). Power-law distributions for term frequencies appear more probable for 'old' concepts with sufficient historical data collected.

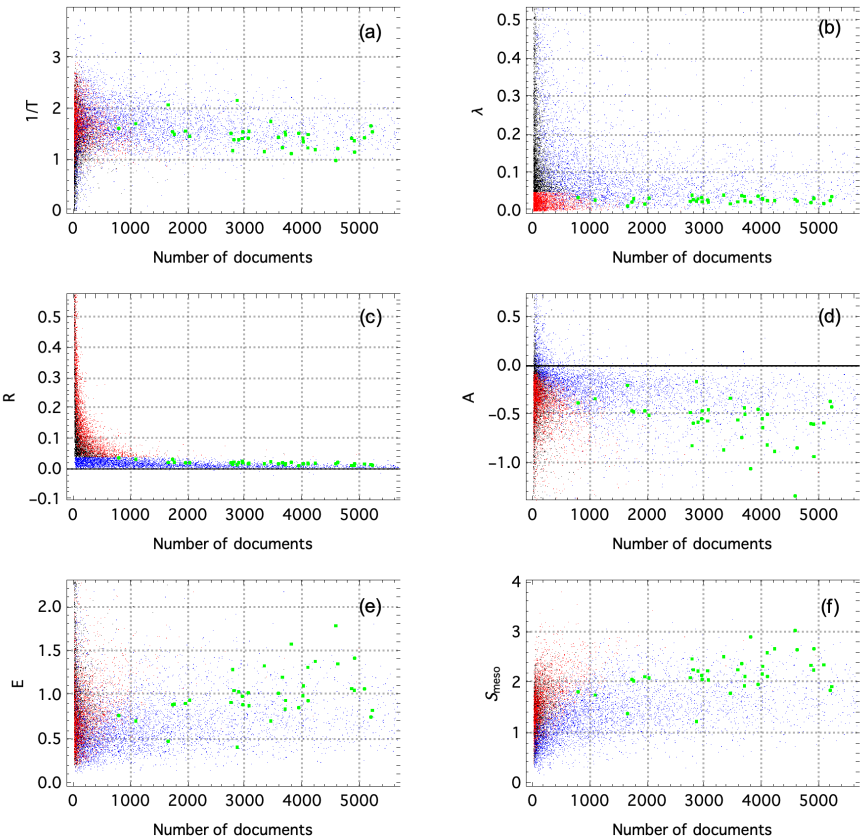

Fig. 3 The values of state parameters and thermodynamic quantities such as residual entropy R, free energy A, internal energy E and mesostate entropy S_{meso} shown as a function of the number of relevant documents from which concepts were extracted. Non-equilibrium concepts are mostly present in less than 1000 documents with a very wide diapason of values for the λ parameter. Concepts found in a larger number of documents reach the state of equilibrium, and the most frequently used concepts reach the stationary state (green dots). The mesoentropy S_{meso} of concepts, as well as internal energy E, is increasing with the number of documents, while temperature is maintained near its mean value $\beta \sim 3/2$.

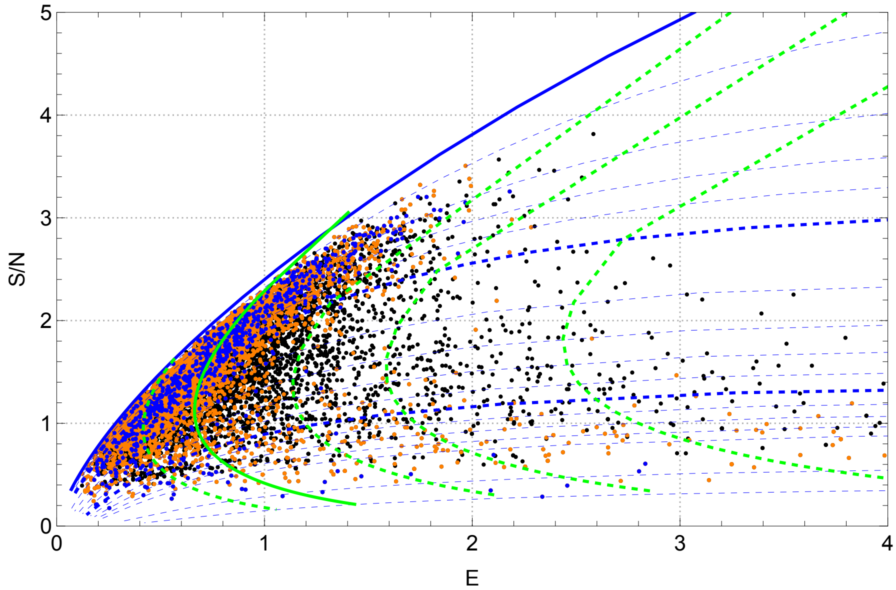

Fig. 4 States of 11737 physical concepts as of the end of 2018 are depicted on the entropy-energy diagram. The blue lines (solid and dashed) correspond to the $S(E(\lambda = c, \beta))$ function for specific values of $c = \{0.001, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.5, 2\}$ from top to bottom. The green lines represent isotherms for $\beta = \{0.5, 1, 1.5, 2\}$ from right to left, with the solid green line for $\beta = 1.5$. Concepts for which $R > 0.04$ are depicted with black dots, the ones with $R < 0.04$ and $R < 0.005$ are depicted as orange and blue dots, respectively.

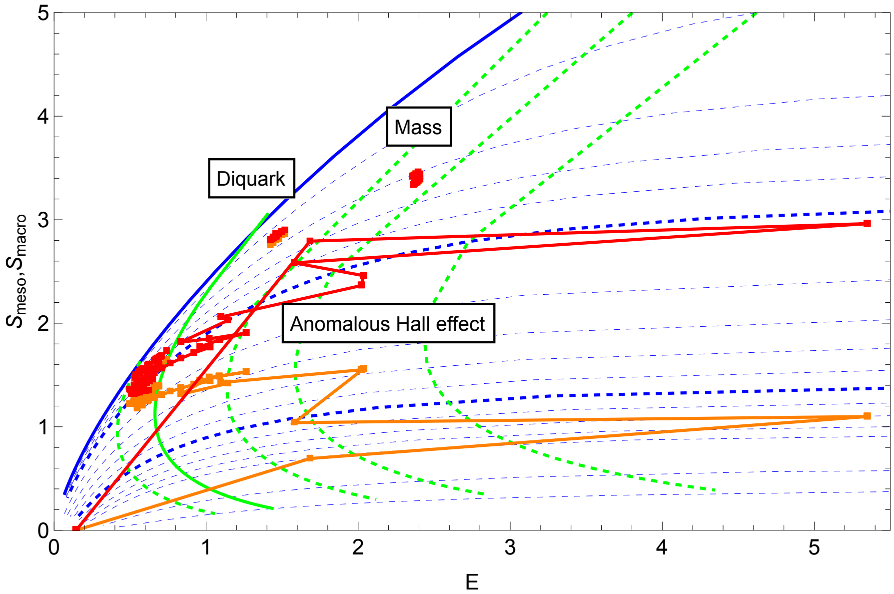

Fig. 5 Examples of equilibrium ("Mass" and "Diquark") and non-equilibrium ("Anomalous Hall effect") concepts state dynamics for the period from 2004 until 2018. Orange and red lines stand for mesostate and macrostate evolution respectively. The blue lines (solid and dashed) correspond to the $S(E(\lambda = c, \beta))$ function for specific values of $c = \{0.001, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.5, 2\}$ from top to bottom. The green lines represent isotherms for $\beta = \{0.5, 1, 1.5, 2\}$ from right to left, with the solid green line for $\beta = 1.5$.

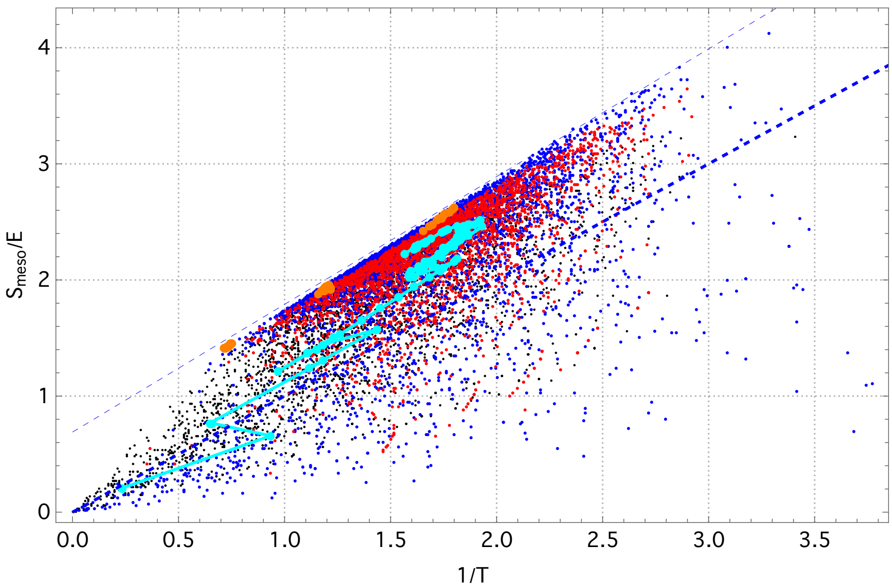

Fig. 6 Temperature diagram for entropy efficiency $Q = S_{meso}/E$, where concepts in instantaneous thermodynamic equilibrium are represented by the blue dots, and concepts with near power law tf distribution are red. Historical values for equilibrium concepts "Mass", "Diquark", "Energy" are orange, and for non-equilibrium "Anomalous Hall effect" are cyan. Dashed lines are linear functions for $Q = \beta$ (bottom dashed line) and $Q = 0.69 + 1.1\beta$ (upper dashed line).

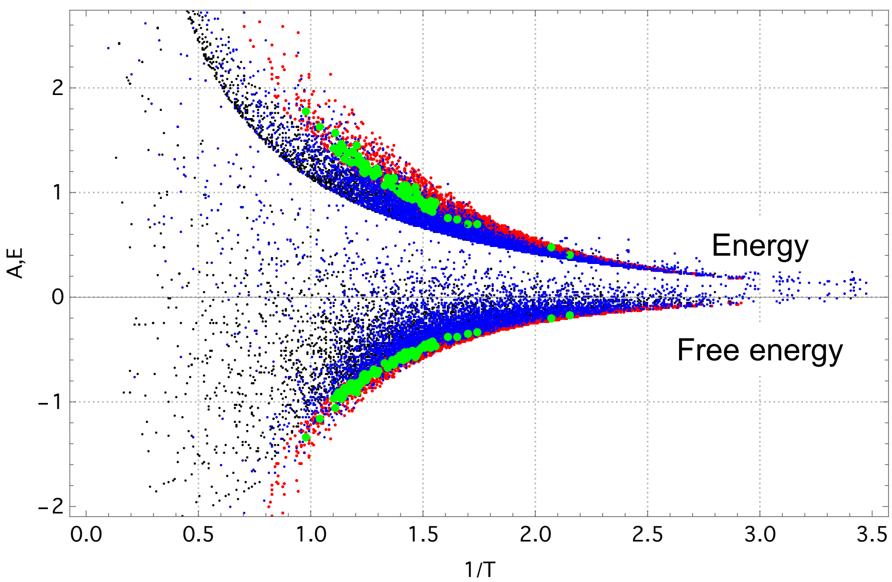

Fig. 7 Free energy A and average internal energy per document E temperature diagram for concepts with $\lambda < 0.05$ (red dots), with $R < 0.04$ (blue dots), and all other concepts (black dots). Green color identifies concepts in a state of stable equilibrium whose residual entropy $R < 0.04$ does not change in the period from 2002–2018.

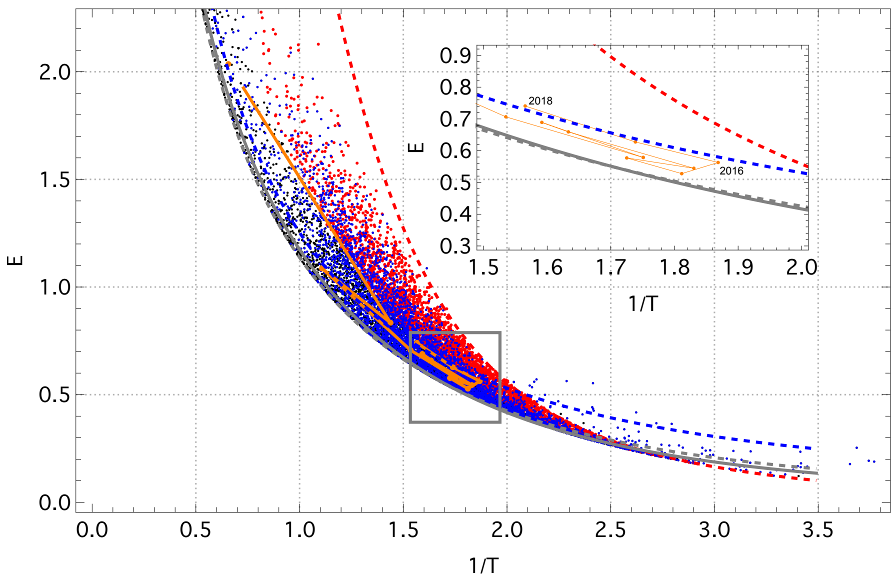

Fig. 8 Energy-temperature (inverse) diagram. The red and blue dashed lines represent the $E(\beta)$ dependence for fixed $\lambda = 0.001$ and $\lambda = 0.8$, respectively. Gray lines correspond to $\lambda = 0.2$ and $\lambda = 0.3$ (dashed) and highlight the minimum $E(\beta)$ function over a wide range of temperatures. The orange line shows the state dynamics of the "Anomalous Hall effect" concept.

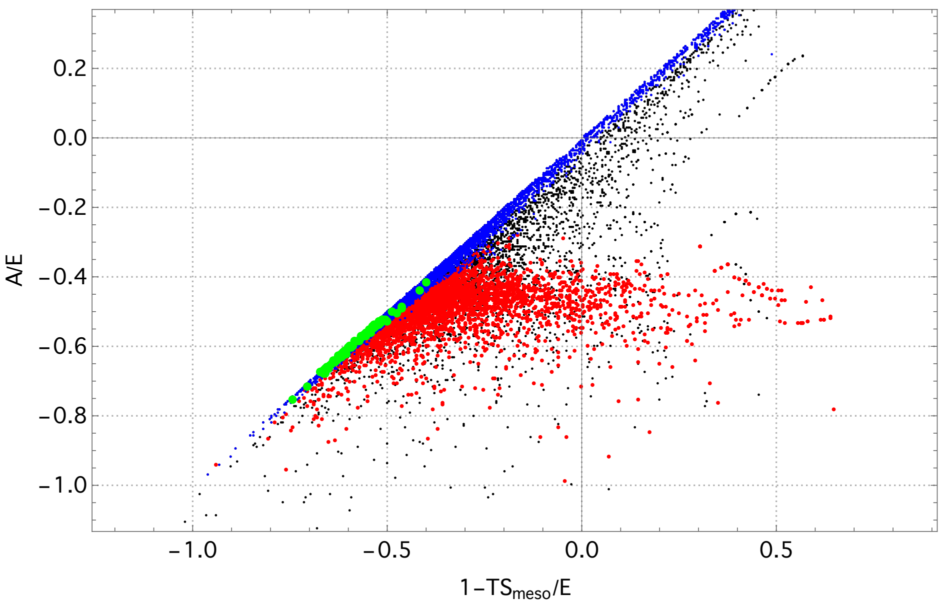

Fig. 9 Correlation between the free energy to internal energy ratio and entropy reduction ratio for four states of concepts: concepts with $\lambda < 0.05$ (red dots), with $R < 0.04$ (blue dots), concepts in a stable equilibrium (green dots), and all other concepts (black dots).

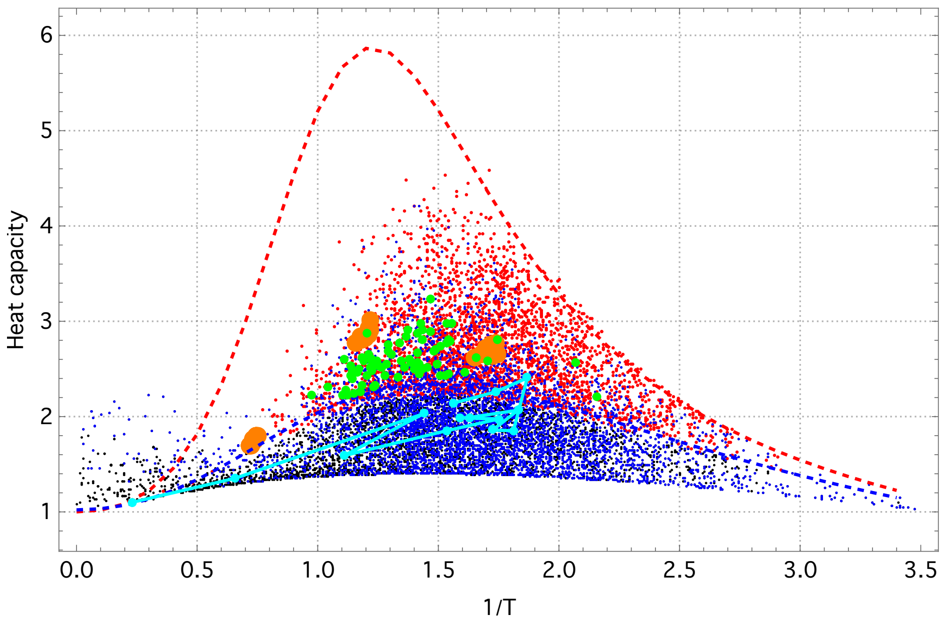

Fig. 10 Heat capacity of concepts from temperature is shown for the several types of concepts: concepts with $\lambda < 0.05$ (red dots), with $R < 0.04$ (blue dots), concepts in a stable equilibrium (green dots), and all other concepts (black dots). The heat capacity dynamics is calculated for the equilibrium concepts (orange color) and non-equilibrium concept "Anomalous hall effect" (cyan color line). The derivative $\partial E/\partial T$ is calculated using Eqs.(6 and (4)) for the constant $\lambda = 0.001$ (red dashed line) and $\lambda = 0.0042$ (blue dashed line)

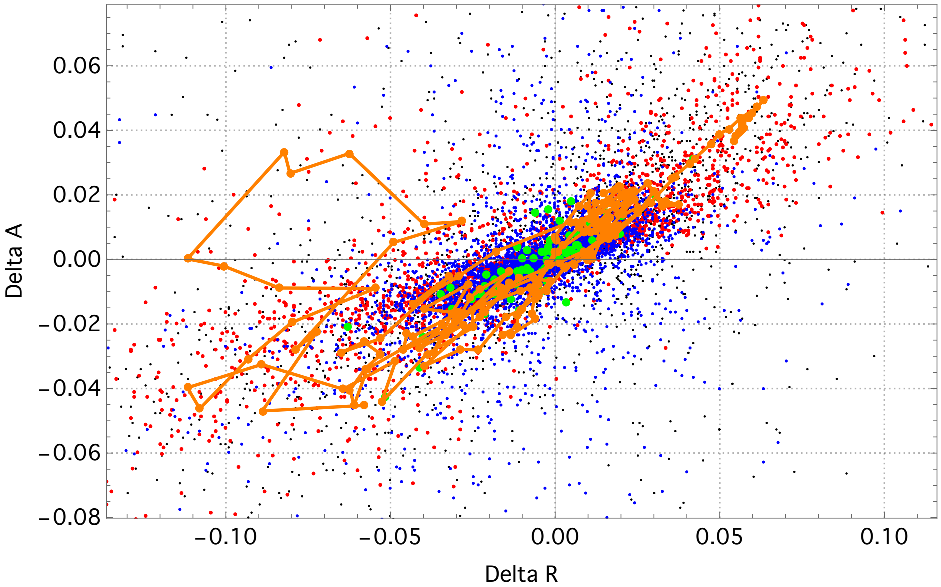

Fig. 11 Free energy change from entropy production for the period 2017-2018 for concepts in a spatial and steady equilibrium (blue and green dots respectively), concepts whose distribution is close to power law with $\lambda < 0.05$ (red dots). Stochastic dynamics of the concept "Anomalous Hall effect" for the period 2002-2018 with a monthly time step is shown by the orange line.