BERGISCHE
UNIVERSITÄT
WUPPERTAL

Social dilemmas in science

A computational approach to collective action problems and trade-offs in collective cognition

Supervisors:
Prof. Thomas Heinze
Jun. Prof. Radin Dardashti

Lucas Gautheron

BERGISCHE
UNIVERSITÄT
WUPPERTAL

Dissertation im Fach

Soziologie

mit dem Titel

Social dilemmas in science. A computational approach to collective action problems and trade-offs in collective cognition.

zur Erlangung des akademischen Grades

Dr. rer.soc.

durch die Fakultät für Human- und
Sozialwissenschaften
der Bergischen Universität Wuppertal

vorgelegt von
Lucas Gautheron

aus
Paris, France

Wuppertal, im März 2025

This thesis explores social dilemmas in science and collective cognition more generally, using high-energy physics as a primary example. Through multiple computational case-studies, it examines how scientists in this field navigate dilemmas related to specialization, adaptation, cooperation, and coordination.

The core contribution of this research is the development of an "adaptive multi-agent system" framework for understanding and modelling collective cognition at large. This framework integrates concepts from social epistemology, game theory, and cognitive and complexity sciences in order to investigate how rationally bounded agents, operating within structured social networks, resolve social dilemmas arising in scientific contexts. Three central tenets guide this approach: (1) collective cognition gives rise to emergent social dilemmas with significant epistemic implications, (2) understanding these dilemmas requires a computational framework drawing from a wide range of disciplines, and (3) to better understand collective cognition, we need computational approaches that are both formal and empirical.

Three case studies in high-energy physics illustrate different aspects of the framework. The first study examines the growing divide between phenomenologists and theorists, particularly in relation to the theory of supersymmetry. The second analyzes how physicists collectively balance specialization and adaptation in a changing scientific landscape, using Inverse Optimal Transport to assess how cognitive costs shape research trajectories. The final case study investigates the trade-offs and dilemmas involved in the diffusion of scientific conventions by applying a novel statistical physics approach to a sign convention in physics.

This thesis offers new perspectives for the philosophy of science, the sociology of science, and computational social science. First, it shows how the notion of collective cognition offers new insights on longstanding issues such as scientific underdetermination and the relation between the cognitive and the social in science. In addition, this thesis contributes to fill a gap between qualitative and quantitative social studies of science, through novel implementations of qualitative insights into quantitative analyses, and by demonstrating the potential of computational models to assist the formation of new philosophical or sociological concepts. Finally, this thesis shows how inverse problems can bridge otherwise disconnected formal and empirical traditions in computational social science. It also suggests that recurrent patterns of self-organization in socio-cognitive systems could be understood functionally as solutions to social dilemmas, opening up new research opportunities in the science of complex systems.

Keywords

Philosophy of science, sociology of science, social epistemology, science of science
Collective cognition, collective intelligence, collective adaptation, collective action
Computational social science, game theory, agent-based modeling
Bayesian inference, inverse problems, statistical physics
Natural language processing, topic modeling
Complex networks, network analysis

Contents

Chapter A

Introduction

High-energy physics is perhaps the most spectacular achievement of modern science, and much of our fascination for it owes to the marvelous complexity of particle physics instruments. The Large Hadron Collider, the largest particle accelerator ever built, is the culmination of this scientific enterprise. It is deployed underground, through a 27-kilometer circular tunnel situated as far as 175 meters below the surface. In both directions of this tunnel, protons are accelerated at nearly the speed of light and made to collide at various focal points, at a rate of about one billion collisions per second. Around each focal point lies a sophisticated machinery of absurd proportion; the ATLAS experiment, for instance, is a 25-meter wide and 46-meter long bundle of sensors and wires weighing no less than 7 000 tons. As these sensors are traversed by the products of particle collisions, they produce electric signals that are transformed into numerical data. In the process, the *instrumental* complexity becomes *computational*: the experiments at the LHC record about 90 millions of gigabytes of data per year [1]. The ATLAS experiment itself requires ∼3 000 computers for reconstructing the physical processes underlying the sensor data, and ∼30 000 more computers scattered around the globe for simulation and data analysis [2]. Numerical simulations of the ATLAS detector alone require modelling about 300 millions objects. Complexity, however, is not confined to experiments: in fact, as one enters the realm of theory, high-energy physics also strikes its *mathematical* complexity. It is indeed difficult to think of areas of mathematics that have no part to play in it, and, in fact, many of them have developed from the needs of fundamental physics!

However, perhaps the most fascinating form of complexity manifested through high-energy physics is its *social* complexity. ATLAS gathers around 3 000 collaborators from all over the world, and the Large Hadron Collider as a whole involves roughly 10 000 scientists. Beyond these large-scale collaborations, our theoretical knowledge of the physical world has been refined and transmitted over generations and generations by thousands of scientists. All other forms of complexity (whether instrumental, computational, mathematical, and so forth) would not have come to fruition without achieving a tremendous magnitude of cooperation, spanning across borders and extending over decades. Cooperation, however, and especially at such scales, brings its own set of challenges – we will call them *dilemmas*. For instance, particle physicists are confronted with major decisions

which eventually shape the field for decades, such as: "should we build one large experiment, or two smaller ones?". Such issues, we will find, are by no means idiosyncratic to high-energy physics, in spite of its apparent singularity: many prominent methodological challenges faced by physicists are in fact experienced across the sciences. Nevertheless, because of its size, high-energy physics exacerbates the sorts of issues that collectives must overcome in order to produce complex knowledge generally. Thus, while the large-scale experiments described above were meant to probe the fundamental laws of the physical world, we might as well use them as probes of the laws of collective minds, too.

The present thesis makes a small and humble step towards this ambitious goal by outlining the contours of a collective cognition approach to science through multiple case-studies of high-energy physics research. This proposal rests on three tenets. The first is that collective cognition (broadly speaking) gives rise to emergent phenomena that manifest as non-trivial social dilemmas with significant epistemic implications in the context of scientific inquiry. The second tenet is that, in order to understand collective cognition in a scientific setting, we should expand our theoretical toolkit beyond philosophy and sociology of science, by appealing to an "adaptive multi-agent system" framework, integrating concepts and insights from cognitive science, game theory, social epistemology, cultural evolution, and complex systems (among others). The third tenet is methodological; it contends that, to understand how agents navigate dilemmas and trade-offs in contexts of collective cognition, we should appeal to computational methods, both formal and empirical. To support and illustrate such an approach, the thesis articulates three case-studies of high-energy physics, whose stringent division of labor and large-scale collaborations exacerbate the sort of dilemmas that may generally arise in contexts of collective cognition.

This introductory chapter is organized as follows. §A.1 argues for a shift in focus from individual to collective cognition in science studies, drawing beyond social epistemology and the sociology of scientific knowledge. §A.2 argues that collective cognition entails dilemmas of various kinds (such as collective action problems) with important implications for scientific inquiry. §A.3 introduces an "adaptive multi-agent system framework" for investigating such dilemmas by computational means, formal and empirical. Then, §A.4 shows how the three papers included in the present thesis illuminate multiple social dilemmas arising in the context of high-energy physics by progressively leveraging multiple dimensions of this framework. §A.5 summarizes the insights of these three papers for computational social science, the sociology and philosophy of science, and the joint research program in which the thesis was conducted. I also reflect on the future perspectives for the approach developed in this thesis in science studies and computational social science.

A.1 From individual cognition to collective cognition

What is collective cognition? According to Cosma Shalizi, it refers to "[those] forms of individual cognition [that] are enhanced by communication and collaboration with

other intelligent agents", of which "modern science" provides "the most spectacular and important instance" [3]. Beyond their individual cognitive abilities, humans crucially rely on collective cognition, and their survival depends extensively on their aptitudes for social and cultural learning [4, 5]. In fact, it has been argued that even our capacity to reason – to which we might want to attribute much of our species' success – is fallible when used internally and may have instead evolved for social purposes such as persuasion and argumentation[1] [6]. Similarly, empirical studies have shown that efficient learning can emerge in collectives when individuals follow simple social heuristics rather than sophisticated reasoning [7]. Generally speaking, humans are truly remarkable for their propensity to produce increasingly complex "cumulative culture" over generations, via the cultural transmission of know-how, technology, and even scientific knowledge [8, 9]. Cultural and technological innovation is fundamentally a collective process – a byproduct of the "collective brain" [10] – resulting from successive iterations and improvements without necessarily requiring causal understanding of the resulting solutions [11], such that individual cognition may play a limited role. As a result, technological and cultural complexity scales with population size [12, 13]: in other words, only "collective brains" a certain size can achieve and maintain a given level of complexity.

Some have gone further and argued that all intelligence is fundamentally collective, as it emerges from the collective behavior of lower-level parts (e.g. neurons) at every level in the hierarchy of biological systems [14][2]. Taken together, these perspectives suggest that collective behavior is fundamentally constitutive of scientific knowledge, rather than an accidental aspect of it, in contrast to philosophical views that neglect or dismiss the causal role of social processes in the emergence of rationality [17]. For instance, Laudan has claimed that "whenever a belief can be explained by adequate reasons, there is no need for, and little promise in, seeking out an alternative explanation in terms of social causes" [18][3]. Such a "methodological" stance, however, would leave us entirely ignorant of how (and to what extent) "rationality" can arise as an emergent feature of collective behavior (as in the examples listed above). As Giere puts it, "from the perspective of distributed cognition, what many regard as purely social determinants of scientific belief can be seen as part of a cognitive system [...] There is no longer a sharp divide [...] The cognitive and the social *overlap* [my emphasis]"[4][5]. Consequently, the success of science

[1] In particular, [6] discusses the implications of this theory for the philosophy of science, and underplays the role of individual genius.

[2] See also Hayek ([15, 16]), who believed that complex distributed systems such as the mind or society at large could not achieve complete oversight over their "own" distributed process.

[3] One way to reject this opposition (or "division of labor", cf. [18]) between rational and sociological explanations in science studies, is to consider that the behavior of scientist collectives can and should *always* receive sociological explanations that are independent from the truth or falsity of what scientists conceive as knowledge. This strategy, which implies that the same sociological processes determine scientists' beliefs irrespective of whether these are "true" or "false", was the path taken by the advocates of the "strong programme" in the sociology of science. This "social constructivist" (or reductionist) program has been criticized for promoting relativism and undermining trust in science. By contrast, the present thesis is interested in the mechanisms through which rationality can emerge from the social.

[4] Note that Giere has also proposed to distinguish distributed cognition from mere collective cognition. The latter, in particular, may include non-human parts [20] (as in Latour's Actor-Network Theory).

[5] Similarly, before Giere, Thagard suggested regarding science as "distributed computing", in an at-

cannot be explained without making reference to its collective dimension. One might say that the collective *is* the cognitive system.

However, this collective dimension simultaneously gives rise to emergent challenges that scientists must properly navigate in order for "the whole" (i.e. the collective) to truly exceed "the sum of its parts" (i.e. the individuals). Understanding these phenomena demands a shift in focus from individual cognition to collective cognition. How can we achieve such a move? A promising possibility is by extending traditional approaches to the social dimension of science by developing accounts of short-sighted and "cognitively bounded" individuals that can nevertheless collectively achieve impressive epistemic outcomes. This includes how collectives can mimic Bayesian learning, even when individuals do not behave like sophisticated Bayesian agents [7]. Interestingly, we might find some inspiration as far reaching as animal studies, given that collective cognition [22], and even cumulative culture [23, 24], are phenomena observed across multiple species, including those whose individual cognitive abilities are not particularly impressive. Perhaps provocatively, we might say that there are things we can learn about science by looking at collectives of scientists in the same way we look at ant colonies and bee hives.

Before we make such a drastic move, we may begin by turning to extant approaches to the social dimension of science. Social epistemology, in particular, has specifically developed from the recognition that there is more to say about truth-seeking activities in groups compared to truth-seeking activities among isolated epistemic agents. Emblematic of this line of thought is the "independence thesis", which states that individual rationality and group rationality are partially independent, such that, for instance, a learning strategy that is optimal for agents learning in isolation, may fail for agents learning in group [25]. To support the independence thesis, formal social epistemologists have produced evidence in the form of computer simulations of agent-based models. This has shown that some level of conservatism in the face of new empirical evidence could improve group learning [26]. Additionally, social epistemology has emphasized the non-triviality of judgment aggregation in groups, often in the form of impossibility theorems [27]. Both the independence thesis and the literature on judgment aggregation have provided justification for the view that "more is different" [28], and that collective cognition is characterized by emergent phenomena that should be studied in their own right. In particular, refusing to acknowledge these emergent features may, according to the independence thesis, lead to erroneous normative claims about scientific methodology. This general view is increasingly acknowledged by philosophers of science; in fact, many of them have adopted this line of research, and several emblematic works around in the independence thesis are explicitly models of *scientific* inquiry. These ideas have even influenced philosophers of science interested in metaphysics and ontology; for example, Stanford argues that a commitment to scientific realism requires some confidence that scientists can freely explore the space of "conceivable alternatives" to our best theories, but that it may not be the case at present, due to the incentive structures of contemporary scientific institutions [29].

---

tempt to clarify the relationship between the social and the cognitive in science [21].

Social epistemology, however, cannot be the sole input in our understanding of collective cognition. Most veristic approaches in social epistemology assume strong individual rationality to begin with, while neglecting cognitive constraints on individual reason ("bounded rationality"), and are not so interested in rationality as an emergent (rather than merely improved) feature of the collective. Conceptually, the present thesis is sympathetic to prior proposals to regard science as "distributed computation" [21] or "distributed cognition" [19]. Yet, these proposals have not been fully articulated, formalized, or concretely applied. This thesis contributes an attempt to resolve this gap. In doing so, I will not draw from a single disciplinary corpus; as I explore "social dilemmas" in science and high-energy physics in particular, I find that many perspectives must be brought together in their analysis. I am most interested in perspectives that are amenable to computational applications or have normative potential. In particular, I focus on dilemmas and trade-offs with significant methodological implications.

A.2 Dilemmas and trade-offs in collective cognition

To engage our discussion of dilemmas and trade-offs in collective cognition, let us start with a curious observation about ATLAS and CMS, the two largest experiments from the Large Hadron Collider at CERN. Interestingly, these two experiments pursue nearly identical research goals [30], and yet they actively maintain their independence, by developing independent research instruments and strategies, and even by restricting communications between their teams at certain times. This might seem like a wasteful duplication of efforts, given that both experiments are considerably large and expensive, gathering about 3 000 collaborators each. This decision, however, addresses a fundamental social dilemma between independent and cooperative learning, as we will see below. In a nutshell, the duplication of efforts at ATLAS and CMS may be costly, but it increases the probability that at least one experimental design is successful (in addition to providing a means of independently reproducing discoveries made in each experiment) [30].

More generally, collective cognition gives rise to emergent social dilemmas and trade-offs between competing objectives. By focusing on these dilemmas, the present thesis proposes to directly tackle features of collective cognition that are universal (observed throughout a wide array of situations much beyond high-energy physics or even science), non-trivial (susceptible of being improperly addressed), and bear significant consequences on epistemic outcomes. We can distinguish two broad kinds of social dilemmas (i-ii).

First (i), social dilemmas are generally conceived as *collective action problems* [31]. These arise when individuals would collectively benefit from cooperating – by coordinating their efforts and dividing labor in certain ways –, but struggle to do so, for instance because they lack information, proper incentives, or central oversight. As discussed below (§A.3.4), collective action problems pervade collective cognition and occasionally conduce to epistemic failures, unless they are corrected with devices such as norms or institutions. Prominent collective action problems include free-riding (when individuals exploit the collective instead of contributing to its welfare) and lack of coordination (when individuals'

actions are mutually inconsistent).

Second (ii), even in presence of central institutions effectively promoting collective action, collective epistemic enterprises imply emergent trade-offs between disadvantages that must be balanced with each other in non-trivial ways. This requires an expanded understanding of social dilemmas. A very well known example of epistemic trade-off is the balance between exploration (the costly search for potentially superior solutions to a problem) and exploitation (the immediate appeal to known solutions, potentially at the expense of unknown superior alternatives) [32]. This trade-off, of course, occurs even among isolated epistemic agents. However, in the context of collective epistemic enterprises, the allocation of cognitive labor among individuals (who does what) introduces new degrees of freedom responsible for additional trade-offs. For instance, collectives must balance independent learning – in which individuals or small groups explore diverse alternatives in parallel by restricting cooperation and exchanges of information –, and cooperative learning – in which individuals divide the cognitive labor or communicate extensively. While independent learning can be slow and unsuitable for achieving elaborate solutions, cooperative learning can be sub-par [33] due to higher coordination costs [34] or a lower diversity of alternatives explored [26, 35]. Therefore, it is often reasonable to strike a balance between independence and cooperation [36]. In science, small teams (which can pursue independent learning in parallel) and large teams (which rely on cooperative learning) play complementary roles [37], reflecting the comparative advantages of smaller versus higher degrees of cooperation. This helps us understand why CERN decided to duplicate major efforts between ATLAS and CMS. While each experiment requires a large amount of internal cooperation to function, it is necessary to engineer some level of independence in order to avoid being stuck with one suboptimal and potentially dysfunctional experimental design. The strategy of pursuing two experiments with identical goals (rather than just one, or more than two) is a solution (or compromise) to this particular dilemma.

Let us stress again, however, that such dilemmas extend much beyond the case of large "Big Science" collaborations. For instance, in mathematics, where scientific collaborations are relatively infrequent and small [38], increasing levels of specialization are undermining the coordination of research efforts [39], which illustrates a more universal trade-off between diversity and coordination [34]. Indeed, collectives must sometimes decide whether to focus on a single cognitive task, which limits the amount of knowledge they may gain, or instead to divide their attention among multiple tasks, which constrains the magnitude of social learning and cooperation that may be achieved. This trade-off is the focus of Chapter 1. Chapter 2 investigates another trade-off, between specialization and adaptation in a collective setting, and Chapter 3 investigates three trade-offs involved in conventions, including the tension between social, temporal, and contextual consistency. While every chapter involves a case-study of high-energy physics, the dilemmas under investigation are each much more broadly relevant for collective cognition. Before discussing the contribution of each chapter in more detail (which will be done in §A.4), let us lay out the framework underlying all three papers.

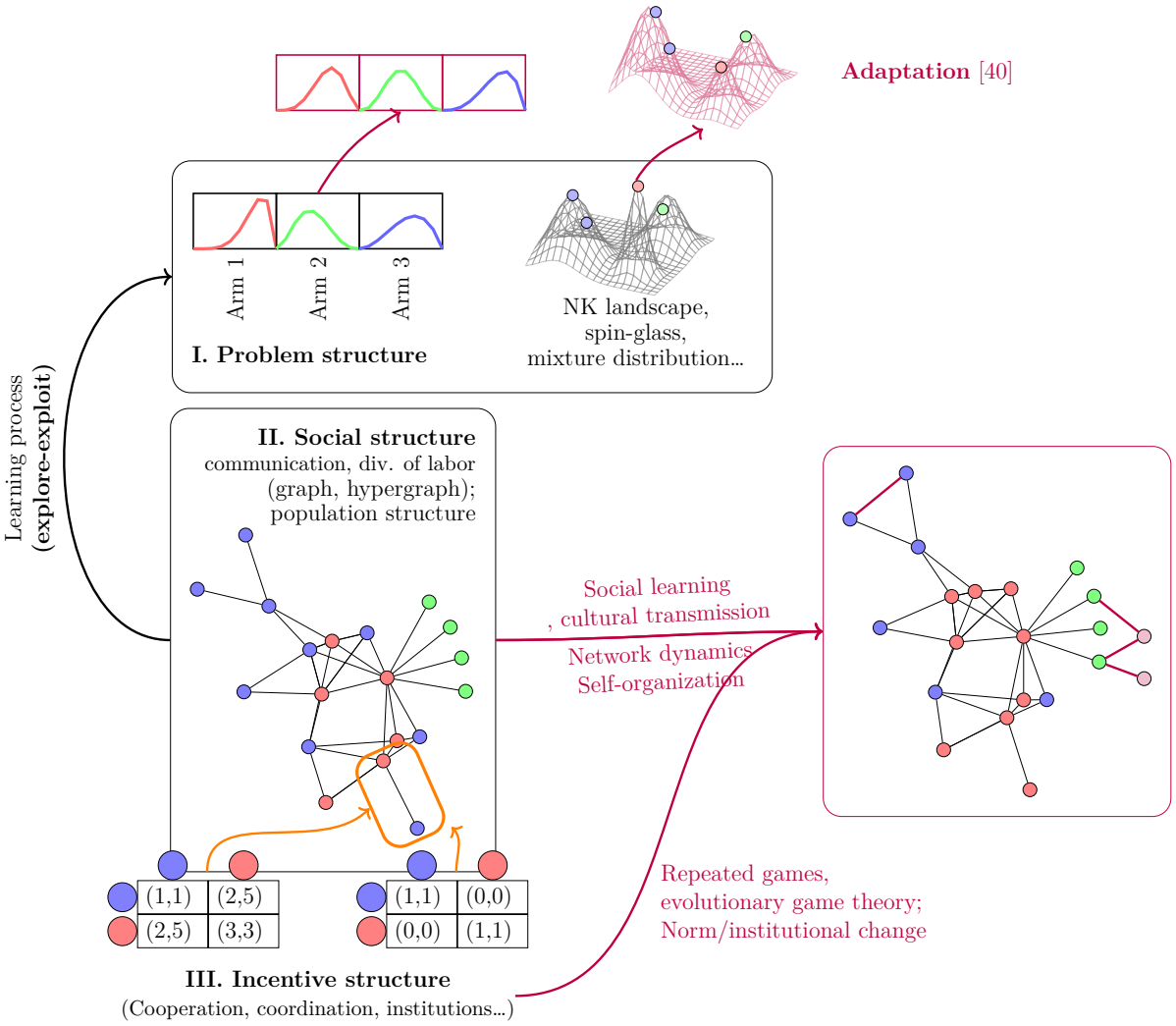


Figure A.1: **Collective cognition in adaptive multi-agent systems**. This formal framework contains three ingredients. (I) **The problem structure** defines the cognitive task navigated by the collective, typically formalized as multi-armed bandits (left) or complex landscapes over large spaces of solutions (right). (II) The **agents** themselves (the nodes in the graph), arranged according to a **social structure**, defined by the flows of information among agents or other relevant relationships, which are typically encoded as graphs. (III) Interactions between agents are influenced by **incentives structures**, shaped by cooperation and coordination problems, social norms, and institutions. As shown in purple, all these components (I-III) co-evolve in ways that confer the system its adaptive nature.

A.3 An adaptive multi-agent system approach to collective cognition

The present thesis proposes a computational approach to multiple social dilemmas arising in collective cognition. In this section, I enumerate several perspectives relevant to collective cognition, including the special case of scientific inquiry. I then combine them into an "adaptive multi-agent system framework" [3], as represented in Fig. A.1. This framework captures essential universal features of intelligent collectives in a way that is directly amenable to computational analyses. It includes three components; (I) the structure of the problems that collectives aim to solve (§A.3.1); (II) the multi-agent system itself, composed of rationally bounded individuals (§A.3.2) organized and connected through a social structure (§A.3.3); and (III) the incentive structure shaping the interactions between individuals and their learning strategies (§A.3.4).

This framework aims to provide a blueprint that can be used as a starting point for any analysis of collective cognition, including scientific inquiry, regardless of the discipline. This is not to ignore the extent of the differences among the sciences [41] [6]; however, the phenomena described below tend to arise universally across all the sciences, since there are none that do not involve collective cognition.

A.3.1 Problems and learning in collective cognition

The functional role of collective cognition is to perform cognitive tasks. Therefore, computational approaches to collective cognition generally begin by assuming a model of the cognitive task at hand (see item (I) in Fig. A.1). We may highlight three recurrent formal strategies for modeling cognitive tasks in social epistemology, management science, learning theory, and adjacent fields. The first approach is to model individuals' knowledge as probabilistic beliefs, such that individuals seek evidence (on their own or with the help of others) to update and improve their own subjective priors about the truth-value of various statements (such as the truth of a scientific hypothesis). However, the aim of collective cognition is not merely to compile observations and produce truth-statements. Fortunately, many alternatives approaches dispense with the notion of "truth". One of them is the appeal to multi-armed bandits (MABs). MABs are "slot machines" with $n > 1$ arms taken to represent alternative theories or solutions. Each time an individual pulls the lever of slot machine i, they receive a random reward R_i drawn from an unknown distribution that varies across arms. MABs are the prototypical model of explore-exploit dilemma, which arise when individual must choose between exploring the space of solutions (in this case, by trying out multiple arms to learn their reward distribution, even those that seem inferior) or exploiting their present knowledge (by pulling the arm that seems best at a certain time, even though it might in fact be suboptimal). Finally,

[6] Cetina [41] argues that there are substantive differences among fields, and suggests that "distributed cognition" is rather specific to high-energy physics. But there is no difference of nature between high-energy physics and other fields in that respect, only a difference of scale.

another popular approach to modeling epistemic problems are "epistemic landscapes". These models (inspired from adaptive fitness landscapes in evolutionary biology) assume that individuals explore the space of potential solutions to an epistemic problem, looking for the solutions of maximal "fitness" (the best solutions). If the space of solutions is two-dimensional, then their fitness can be represented along a third dimension (height), thus forming a surface (or landscape) of solutions [42]. However, fitness landscape models often involve a high-dimensional spaces, with a flexible number of dimensions [43]. The degree of complexity (or ruggedness) of these landscapes can also often be adjusted, such that more complex problems require deeper exploration and are more difficult to solve.

Few attempts have been made to map actual problem spaces to these formal complex landscape models. The closest precedent is [44], in which a fitness landscape of religious systems – a cultural rather than epistemic landscape – is reconstructing empirically using historical cultural data. In Chapter 3, I use this approach to recover the complex cultural fitness landscape of a collection of typesetting conventions in scientific papers. In science studies, however, it is much more common to recover latent "epistemic spaces" from corpora of scientific texts by computational means. For instance, word embeddings are now a widespread approach for locating scientific concepts into abstract high-dimensional spaces, in which the "distance" between words is a measure of semantic dissimilarity [45, 46]. Alternatively, topic modelling is a widespread approach for mapping out documents in a multi-dimensional conceptual space [47, 48]. Both techniques (word embeddings and topic models) are leveraged simultaneously in Chapter 2 in order to measure the research portfolios of physicists throughout time. Chapter 1 also uses topic modeling to explore the dynamics of the multiple contexts in which supersymmetry arises in high-energy physics.

Whatever the structure of the problem, a population of *epistemic agents* (e.g. scientists) engages with it through *learning strategies* which specify how individuals search and process information[7]. In collective cognition, these strategies build upon the information gathered by others.

A.3.2 Cognition among rationally bounded agents

How can collectives be intelligent, when they gather relatively short-sighted individuals? In order to understand how and when collective cognition subdues individual cognition, we must better understand the limitations of individual cognition. Our cognitive capacities are notoriously bounded, for multiple reasons: first, we have access to limited information and/or limited computational capacities for processing this information; we "suffer" from an array of cognitive biases, such as confirmation bias [49]; and we often rely on imperfect heuristics to achieve "satisficing" rather than optimal solutions [50]. Most crucially, our cognitive limitations prompt us to *specialize* – by concentrating our intellectual resources into bounded epistemic domains – and divide cognitive work among individuals, which in turns raises multiple challenges for scientific inquiry. Specialization is the focus of chapters 1 and 2, which respectively discuss the implications of the disunity

[7]These include, for instance, greedy searches for MABs, hill climbing for complex landscapes, etc.

of physics research and the tension between adaptation and specialization.

Chapter 2, in particular, measures the "cognitive costs" of moving across research areas in high-energy physics, and proposes a measure of "intellectual capital", i.e., a proxy for individuals' cognitive resources. It shows that physicists' collective adaptive patterns (in response to new evidence or theoretical innovations) are fundamentally constrained by their prior knowledge. This explains the occurrence of path-dependency in scientific exploration, and to some extent justifies some degree of conservatism in science.

Cognitive factors also surface in Chapter 3, which discusses the trade-offs involved in scientific conventions. In particular, the paper shows that, presumably in order to avoid cognitive costs, individuals tend to stick to their favorite scientific convention, even if it is inferior in a given context. Incidentally, this observation is achieved by appealing to an item-response model, a popular modeling strategy in cognitive psychology. Generally speaking, cognitive science provides insights about the individual agents' themselves, as well as some basic rules about their psychology and how they interact with other agents. In addition, bounded rationality plays an important role in collective behavior given that agents' partial and imperfect knowledge of the intentions of others has important effects on the way they may interact and/or their ability to achieve cooperation. Finally, as discussed in §A.5.3, the inverse problem methods used in chapters 2 and 3 both provide measures of the efficiency (or degree of rationality) of utility-maximizing agents.

A.3.3 Social structure in collective cognition

The performance of collectives generally depends on their social structure. This includes communication structures, which determine the flow of information among individuals [51], in addition to collaboration and decision structures. In computational approaches, social structure is most prominently formally encoded by graphs (or networks), which define "dyadic" relationships between pairs of agents[8]. In science, such relationships can encode, for instance, who collaborates with whom [52], who sends e-mails to whom [53], who cites whom, or more generally, who communicates evidence to whom [51, 54]. Formal approaches to collective cognition (in social epistemology, for example) typically compare the outcomes of various network *topologies* [54, 55], capturing features of complex networks such as clustering (i.e. local community structure), small-world properties (the existence of relatively short paths between any two individuals), or hubs (highly connected individuals). Empirical studies of social networks in collective cognition have similarly explored the relationship between the performance of teams and their social structure [33, 56]. Others have investigated the correlation between individual performance and position in a network [57–59]. Chapter 2 contributes to this literature by measuring the effect of scientists' social capital on their ability to adapt to changing circumstances. Beyond network structure, social structure includes social identities, group membership, and demographic variables that can affect collective dynamics as well. Indeed, Chapter 3 finds

[8]However, higher-order structures (such as collaborations) may be more adequately represented by hypergraphs, which define simultaneous relationships between two or more individuals.

that the propagation of scientific conventions can be influenced both by social network infrastructure (e.g. via the imitation of peers, or formally speaking one's neighbors in the social network) and membership to a particular research area (or "disciplinary matrix" [60]). Additionally, group membership can sustain homophilic behavior and in turn drive discrimination, marginalization, and belief polarization [61–63].

A.3.4 Incentives and collective action problems in collective cognition

In collective cognition, epistemic agents are generally self-interested. But, how can agents pursuing their own interests nevertheless succeed in producing reliable knowledge to the benefit of the public? One answer is that institutional incentives in science are articulated around a "credit" economy incentivizing the production of truth [64, 65]. This includes implicit social norms, such as the priority rule, according to which credit should be given to the first discoverers of a truth [66]. The interests of scientists, however, are determined not only by their institutional environment, but also by how the behavior of others. Such dependencies are well described by the framework of game theory, which has been notoriously applied to wide-ranging situations such as conflicts [67], the evolution of cooperation [68], or conventions [69]. Game theory formalizes the collective action problems that arise when individuals would collectively benefit from cooperating in some way, but struggle to do so due to a lack of information or conflicting individual incentives. Incidentally, collective cognition often involves collective action problems, since the degree to which a collective achieves epistemic progress depends on each individual's willingness to contribute to the task [70].

A first kind of collective action problem arises when individuals "free-ride" by exploiting the information and knowledge accumulated by groups instead of contributing their own [71, 72]. For instance, using a multi-armed bandit approach, [72] shows that scientists may be incentivized to remain conservative towards the dominant paradigm, which is detrimental to the exploration of potentially superior alternatives. Generally speaking, collective intelligence is a public good that must occasionally be protected by enforcing cooperation [70]. In addition to "over-exploitation", anti-cooperative behavior notoriously raises issues of deception and trust, which also require strong social norms or institutional safeguards.

Another example of collective action problems are coordination problems, which arise when individuals are better off agreeing to a particular course of action among multiple reasonable possibilities. Collectives can perform worse than isolated individuals for certain problem-solving tasks when they cannot properly coordinate their efforts [73]. In high-energy physics, this has major implications, since large-scale experiments such as particle colliders require high levels of coordination; therefore, physicists must agree about which experiment to perform collectively, even though multiple experiments could reasonably be deemed equally interesting. For instance, whether CERN should "replace" the Large Hadron Collider with larger a circular electron-positron collider, or a linear

accelerator instead, is in a way "conventional" since both possibilities have their own epistemic merits. Indeed, "conventions", per Lewis [69], are solutions to coordination problems. Chapter 3 explores trade-offs and dilemmas affecting the scientific conventions using the example of a perfectly innocuous sign convention in high-energy physics.

A.3.5 Adaptive and evolutionary processes in collective cognition

So far, the picture I have painted – a (I) fixed problem, explored by a (II) population of agents interacting along a rigid social structure and influenced by (III) constant background incentives – is overly static. In reality, these three components are co-evolving, whether change is driven by exogenous factors (e.g. changes in the institutional environment) or endogenous factors (for instance, agents adjusting their behavior in response to innovations). Below, we consider several ways in which multi-agent systems can exhibit evolutionary or "adaptive" behavior.

First, the problem structure itself (Fig. A.1, §I) may evolve over time, reflecting transformations [40] which, in the context of science, may be driven by theoretical and technological disruptions or shifting institutional incentives. In a computational framework, this sort of configuration can be modeled with time-varying landscapes or multi-armed bandits, which prompt agents to adapt. The magnitude of environmental change influences the optimal solutions to the exploration/exploitation dilemma: for instance, unstable environments prompt more exploration than stable ones. In addition, not all agents respond identically to environmental changes, due to path dependency [74]. In particular, changes in scientists' research interests are a byproduct of both the environment *and* their prior research interests, as shown in Chapter 2.

In addition to responding to changes in their environment, epistemic agents can adjust their strategy (e.g. their research focus) based on socially learned and culturally transmitted information. While communication enhances the diffusion of knowledge, it can also trigger erratic information cascades, herd behavior and ambulance chasing [75, 76], or the nucleation of scientific "bubbles" [77]. Furthermore, the social structure itself (the relationships between agents) may change over time, both due to environmental pressure and endogenous dynamics (self-organization). The latter are typically modeled by random walks (when agents locally explore their social network) or preferential attachment (when agents prefer to form relationships with more central individuals) leading to "the rich get richer" types of dynamics [58, 78]. Chapter 1 reveals conceptual transformations in the citation network across theory and phenomenology in response to epistemic change in high-energy physics, which clearly illustrates how such change (Fig. A.1, §I) impacts the communication structure through which scientists exchange knowledge and information (Fig. A.1, §II).

Finally, the incentive structure can also shift due to institutional change, which itself can alter the problem structure by inverting the priority order of different problems [40]. Social and scientific norms themselves may evolve overtime (Fig. A.1, §III), as potentially

described by the frameworks of evolutionary game theory or repeated games [79, 80]. For instance, evolutionary models can account for the emergence [81] and discontinuation [82] of poor statistical practices in science as well as the prevalence of risk-taking and conservatism in research [83].

In addition, an often overlooked aspect of scientific change (but generally important for collective cognition at large) is the generational and demographic dynamics that can affect science over longer time scales. For instance, demographic processes can result in maladaptive losses of knowledge and skills [84, 85] from which it can be difficult to recover. This issue is particularly salient in large-scale physics experiments: their planning and construction can span over decades (the Future Circular Collider at CERN is not expected to start before the mid 2040s), and their operation will require a new generation of physicists to be trained before crucial knowledge and know-how is lost[9].

In general, all of these evolutionary and adaptive changes are unfolding simultaneously and interacting with each other via complex feedback loops. While a detailed account of the mechanisms driving their co-evolution is beyond the scope of the present thesis, following [40], we may summarize the general idea by stating that collective cognition involves *adaptive* multi-agent systems, in reference to the theory of complex adaptive systems [86]. Their adaptive dimension stems from the co-evolution of the three main ingredients of the framework (Figure A.1): the problem structure, the social structure, and the incentive structure.

A.3.6 Bridging two traditions in computational social science

This thesis proposes a computational approach to collective cognition. However, we may distinguish two relevant traditions in computational social science; although there is some overlap and interactions between the two, these traditions have developed their own journals, communities, and methods/topics of inquiry. This thesis makes an effort to bring these two traditions together, in chapters 2 and 3 specifically (in contrast, Chapter 1 falls more into the traditional digital humanities literature).

The first tradition (in chronological order) finds its origins in the sciences of computation and complexity, and relies essentially on formal models and computer simulations. These include agent-based models, which are typically used the explore surprising patterns of collective behavior that can emerge from simple sets of rules dictating the ways agents interact with each other. Illustratory (and often considered seminal) of this line of approach is Thomas Schelling's model of racial segregation in cities, which demonstrates that spatial segregation can arise from small homophilic preferences [87]. This computational tradition has since been embraced by social epistemologists, and it constitutes the principal mode of inquiry for computational philosophy of science [88]. Most cultural evolutionary perspectives on science rely on this approach [65]. Computer simulations

[9]Again, this sort of challenge goes beyond large-scale physics; for what to train the next generation of scientists, when the future is so uncertain, is a universal problem across the sciences. In fact, scientific institutions generally address this problem by maintaining a broad range of knowledge and skills, whose significance might increase in the future.

can dispense with unrealistic assumptions introduced for practical purposes[10], which enables the exploration of more complex and realistic models. Nevertheless, ABMs and formal models of multi-agent systems necessarily remain limited idealizations of real-life scenarios, and what we may learn from this approach is a matter of debate [89]. This is especially true when no attempt is made to connect such models to empirical data, as is generally the case with models of the social organization of science [90].

The second tradition in computational social science is much more recent, and, unlike the previous one, much more empirical and data-driven. In a nutshell, this tradition appeals to computational means, such as natural language processing or network analysis, and generally rudimentary statistical approaches, such as statistical testing, either to describe naturalistic data[11] using exploratory analyses, or to test specific hypotheses (e.g.: is X correlated with Y? Does A cause B?). This new paradigm was made possible by the increasing availability of numerical datasets (whether it includes natively numerical Big Data or digitalized corpora), methodological innovations in computer science and machine learning, and increasingly widespread access to computing resources. This approach to computational social science has somewhat eclipsed the former tradition (not without causing frustration among certain communities [91]).

The present thesis attempts to combine these two approaches. Formal computational models are valuable because they clarify the theoretical assumptions underlying a particular reasoning, and they sometimes enable predictions. Additionally, they allow the transfer of insights across systems: as Smaldino puts it, "when you know that a system in question involves features and constraints similar to those in models you have seen before, insight into how they operate can follow" [92]. This appears very clearly in chapters 2 and 3. The first paper relies on Optimal Transport in order to describe collective adaptation in science. Optimal Transport is a mathematical optimization framework[12] that aims to find the most economic way of displacing matter or goods, by minimizing transportation costs. This "model transfer" [94] invites us to view the tension between specialization and adaptation as one between the imperative to adapt to new research opportunities and the imperative to minimize cognitive costs. The second paper applies the Ising model to the diffusion of scientific conventions. The Ising model is a physical model introduced in 1920 to account for the spontaneous magnetization of ferromagnetic materials as arising from purely local microscopic interactions between neighboring "spins" [95]. In social systems, it can explain the emergence of collective behavior due to micro-level interactions, and it has become widely popular in that context as a result [96]. In Chapter 3, I show that the Ising model can also capture the structure of coordination games on complex networks, as well as the competing effect of local and global mechanisms of coordination that may

[10] As Miller and Page put it, "we want to study models with a few agents, rather than those with only one or two or infinitely many. We want to understand agents that are neither extremely brilliant nor extremely stupid, but rather live somewhere in the middle" [86, p. 7]. Yet such silly assumptions have long been necessary for the sake of mathematical intelligibility.

[11] As opposed to experimental data.

[12] First introduced in 1781 by French mathematician Gaspard Monge in a military setting [93], and later refined by Leonid Kantorovich in the context of economic planning

simultaneously shape individuals' attitudes towards multiple possible conventions.

In this thesis, these computational models are not merely used to formulate theoretical assumptions or generate insights transferable across systems. They are also used to extract *empirical* information from real systems. To this end, I rely on so-called inverse problems, which consist in inferring the rules underlying a set of behavioral observations. For instance, in Chapter 2, I appeal to Inverse Optimal Transport, an emerging topic in computer science, to infer the underlying migration cost matrix (the cost of moving from one research area to another) minimized by high-energy physicists' adaptive patterns in response to changes in their field. In Chapter 3, I solve the inverse Ising problem to measure the magnitude of the contribution of local and global coordination processes shaping physicists' preferences towards a sign convention. I further show that once these contributions have been measured, they may be used as summary statistics to assess the relative plausibility of various models of preference-formation, including models of cultural transmission. To this end, I appeal to "simulation-based inference" [97] with deep-learning [98], a recent approach for performing Bayesian inference about agent-based models from empirical data in spite of their computational complexity. Generally speaking, inverse problems are a promising strategy for bridging formal and empirical computational social science.

The framework outlined above can serve as a blue print for formal and empirical investigations of social dilemmas in collective cognition. Not every aspect of it has to be involved in every approach, but starting from any problem-situation in collective cognition, one can start by looking at Figure A.1 and ask which ingredients of this framework are relevant.

A.4 Application to social dilemmas in high-energy physics

The present thesis investigates social dilemmas in high-energy physics by leveraging multiple aspects of the framework just outlined. It proceeds in three chapters. Chapter 1 is much more similar to traditional digital humanities paper, and therefore stands aside from the other two. It was undertaken at a time when the project was more focused on a specific theory in fundamental physics. Chapters 2 and 3, by contrast, go further and further in implementing the perspective of the adaptive multi-agent system approach to collective cognition.

Chapter 1 explores the tension between unity and pluralism in high-energy physics through a rudimentary quantitative case-study of supersymmetry. While this theory has been extremely influential in the field over the past 40 years, it has disappointingly failed to materialize at the Large Hadron Collider, which has led to a situation of crisis. However, Chapter 1 shows that supersymmetry is unequally appraised throughout the field. As I demonstrate using quantitative analyses of scientific literature, high-energy physics

is divided in two theoretical traditions, phenomenology and pure-theory, carried out by distinct communities, with their own theoretical "language". While interest in supersymmetry has been declining rapidly among phenomenologists since the start of the LHC, as many of them lost faith that it would ever materialize, reference to supersymmetry has remained more steady among theorists, who value its mathematical properties regardless of the empirical support it has received. Finally, I investigate the "trading zone" [99] between phenomenology and theory by looking into the concepts that have sustained exchanges between them throughout time. To this end, I explore the keywords that travel in the citation network, which convey the communication structure (Fig. A.1,§II) uniting the field. I find that while supersymmetry historically played a significant role in tying these two "subcultures" [99], it is now being superseded in the "trading zone" by other concepts such as black holes, dark matter, and gravitational waves. More generally, I find a growing disconnect between particle collider phenomenology and theory. The field is now facing a dilemma: it must decide whether to promote unity, by seeking developments that can sustain fruitful trades between particle phenomenology and theory, or to embrace a pluralist approach, even if that implies a growing disconnect between two traditions with a history of close cooperation.

Chapter 2 focuses on the tension between specialization and adaptation. While scientists must specialize by concentrating their intellectual resources into rather narrow cognitive domains, they must also remain able to adapt to changing circumstances, which may prompt the acquisition of new knowledge. To understand how scientists navigate this trade-off, I study the trajectory of a cohort of $\sim$2 000 high-energy physicists between 2000 and 2020. This time period is particularly interesting, since the Large Hadron Collider and new experimental opportunities (such as direct dark matter searches and gravitational wave astronomy) have profoundly reshaped the landscape of experimental opportunities (Fig. A.1,§I). Using an embedding topic model trained on 180 000 scientific abstracts, I measure the research portfolios of the physicists (i.e. how they have divided their attention across 15 research areas) before and after the start of the LHC. While the cohort's research interests have been rather stable, dark matter research has doubled, to the detriment of neutrino physics and the physics of the electroweak sector, which is the physical domain explored at the LHC. This shows that the cohort has adapted to shifting incentives increasingly favoring dark matter over particle collider phenomenology. In addition, using Inverse Optimal Transport, I show that the observed collective patterns of change are structured by learning costs: the cohort has adapted in a way that tends to minimize cognitive learning costs, which demonstrates that specialization constrains collective adaptation. Finally, I investigate the effect of multiple variables on scientists' individual ability to adapt. In particular, I show that the diversity of their intellectual and social capital is associated with larger magnitude of change. By contrast, "power" (the magnitude of social capital (Fig. A.1,§II) is associated with more stable research agendas.

Chapter 3 investigates multiple dilemmas and trade-offs involved in the propagation and adoption of scientific conventions. The first trade-off concerns the imperatives of

social consistency (driven by coordination costs, Fig. A.1, §III), sequential consistency (driven by the cost of switching between different conventions), and contextual consistency (driven by maladaptation costs, Fig. A.1,§I) that individuals must balance when choosing between competing conventions. The second trade-off is the competition between local processes (propagating locally throughout a social network, Fig. A.1,§II) and global processes (exogenous to social networks) in the propagation of conventions. Finally, the third trade-off is the balance between decision optimality and decision costs in cases where individuals must resolve conflicting preferences about which convention to use. I develop a statistical physics approach to these dilemmas, which I then apply to a sign convention in high-energy physics (the metric signature). I find evidence that social, sequential, and contextual consistency all influence scientists' attitude towards this convention. Using an Ising model approach, I also find evidence for both local and global processes in their diffusion, although global effects (driven by scientists' primary research area) seem to dominate. I then show that the magnitude of local and global processes measured with the Ising model can be used as summary statistics for comparing the relative plausibility of more realistic models of the formation of scientists' preferences with simulation-based inference. This in particular allows us to rule out purely global processes of cultural transmission. Finally, I find that scientists appeal to leadership and seniority to resolve conflicting preferences about which convention to use in collaborations, which suggests that decision costs prime over optimality (e.g. collective satisfaction) for this specific convention.

A.5 Implications and opportunities across disciplines

Each of these chapters make contributions to the specific topics they address – the disunity of physics, adaptation in science, and conventions –, which are developed at length within each paper. In addition, as a whole, these three chapters also suggest broader implications and perspectives for multiple disciplines, which I propose to discuss in this section. First, this thesis leads to a generalized notion of collective constraint that can unify diverse views about conventionality and underdetermination in the philosophy of science (§A.5.1). Second, this thesis illustrates how the gap between qualitative and quantitative approaches in social studies of science can be addressed in a bidirectional way, not just by implementing concepts into quantitative work, but also by using computational model as a source of inspiration in the formation of concepts (§A.5.2). Third, this thesis shows how inverse problems can bridge two otherwise disconnected traditions in computational social science, while suggesting a new range of explanations for recurrent patterns of self-organization in collective systems (§A.5.3).

A.5.1 Implications for the philosophy of science

Undetermination and collective constraints in science

Scientific underdetermination is a major theme in epistemology and philosophy of science [100]. In a nutshell, underdetermination arguments (which come in different flavors) generally purport that *evidence* only loosely constrains our beliefs. A popular variant of underdetermination famously defended by Quine [101], and known as *epistemological holism*, contends that the truth-value of statements cannot be determined in absolute terms, independently of our beliefs about other statements. Consequently, scientific evidence can only refute *collections* of beliefs at best [100]. Underdetermination is often perceived as a challenge to rationality, which implies *constraints* on what constitutes rational beliefs. This is why, in order to refute claims about scientific practice that he perceived as social reductionists, Galison elaborated a defense of the robustness of science, explicitly articulated around the notion of constraints [102]. These constraints, as per Galison, can come in many kinds – experimental, phenomenological, or theoretical, and short-lived or long-term, depending on how far their impact on science goes [102, 103]. Generally speaking, they put some limits on underdetermination, such that not anything goes. However, as Chapter 3 shows by investigating conventions in science, many constraints on rational behavior are not *individual* but *collective* (or perhaps *holistic*). Collective constraints arise when the attitude towards a particular choice among several alternatives is constrained *only* in relation to other choices. In the case of conventions, collectives constraints can be i) social, ii) sequential, and iii) contextual. Let us take for instance left-hand versus right-hand driving: whether I should drive to one side or the road or the other is not constrained in itself, unless it is specified how others will behave. Such situations – which fall into social constraints (Table A.1a) – are known as a coordination problems, and David Lewis based his study of conventions around a game theoretical account of such problems [69] . Certain conventions, however, are only constrained *sequentially* (Table A.1b). For instance, assume you must choose between different keyboard layouts (QWERTY, QWERTZ, AZERTY, etc.). It does not matter so much what choice someone makes as long as his or her choice remains consistent throughout time. Finally, certain choices are only constrained contextually (Table A.1c), that is, in relation to a set of other interconnected choices. Unit systems are a good example; depending on the task at hand, it might make more sense to measure lengths in millimeters as opposed to light-years, although there is no universally correct or superior choice. Chapter 3 performs an empirical analysis of a sign convention in high-energy physics and finds that all of social, sequential, and contextual constraints jointly influence the attitude of individuals towards conventions. The empirical analysis is built upon a mixed game-theoretic and statistical physics framework that formalizes the notion of collective constraint in both an utilitarian (Table A.1) and probabilistic languages. This paper is a contribution to prior philosophical literature in two respects. First, it shows that Lewis' [69] account of conventions, which is focused on their social dimension, must be expanded to include temporal and contextual consistency. It is quite remarkable that the

	$x_j =$ ●	$x_j =$ ●
$x_i =$ ●	$(1,1)$	$(0,0)$
$x_i =$ ●	$(0,0)$	$(1,1)$

(a) **Social consistency**. Alice and Bob are better off if they agree on either ● or ● .

	$x_{t+1} =$ ●	$x_{t+1} =$ ●
$x_t =$ ●	1	0
$x_t =$ ●	0	1

(b) **Sequential consistency**. Alice is better off if she consistently chooses ● or ● .

	$y =$ ●	$y =$ ●
$x =$ ●	1	0
$x =$ ●	0	1

(c) **Contextual consistency**. Alice is better of if she chooses either ● ● or ● ● .

Table A.1: Collective consistency as coordination games involving Alice ( ) and Bob ( ), or Alice alone. Each table represents a payoff matrix associated with a "collective" choice.

same game theoretic account can describe the three dimensions of collective consistency (Table A.1). Conversely, this work supports a form of holism such as that of Quine (by establishing that conventions are collective, i.e. holistic constraints), and simultaneously suggests that epistemic holism must also incorporate social and temporal dimensions. More specifically, if multiple collections of beliefs are logically consistent, it might make sense to choose the one more aligned to the present social setting, or to choose one that minimizes cognitive change – even if this reinforces path dependence in our exploration of the space of ideas. This means individuals may choose one system (or web) of compatible beliefs over another *conventionally*, but in a nevertheless rational way – by addressing the reasonable needs to coordinate their beliefs at a social level and to minimize cognitive efforts.

Underdetermination and pursuit-worthiness

Underdetermination is traditionally discussed in the context of *acceptance*, which affects the assessment of the truth-value of theoretical statements. However, another pervasive debate in philosophy of science concerns whether methodology underdetermines scientific *developments*; and as Thomas Nickles has argued, underdetermination also affects *pursuit*, that is, our assessment of which research directions are worth exploring next [104, p. 167, §17]. The underdetermination of pursuit is particularly severe a priori, since science is by essence open-ended: it is precisely the goal of scientific inquiry to explore a range of possibilities. Nevertheless, not every decision of pursuit in science is equally reasonable – again, there has to be "constraints" on what should be the next step of inquiry. This thesis provides several examples of collective constraints on the underdetermination of pursuit in high-energy physics.

First, collective constraints can be *social* constraints, which arise when a certain choice is constrained only in relation to choices made by others, as exemplified by coordination problems (Table A.1a). In assessments of scientific pursuit worthiness, such problems may arise when an important level of cooperation is required to explore a particular research direction, even if this direction is somewhat arbitrary (or "conventional" [69]). This is obvious in large-scale particle-physics experiments, which entail the commitment of a

large collective to one particular research effort among other reasonable alternatives. For instance, Table A.1a may describe the choice between two equally promising experiments, that may only succeed if both Alice and Bob partake in it.

Additionally, *sequential* constraints arise when certain choices are constrained only in relation to past decisions. For instance, in Table A.1b, regardless of which choice Alice makes, she benefits from remaining committed to it. In science, a major driver of such types of constraints is the need for scientists to capitalize on prior knowledge and achievements, for cognitive or material reasons. Chapter 2 demonstrates that patterns of collective adaptation in high-energy physics over the past decades have indeed been structured by the minimization of individual "efforts" given the aggregate patterns of change observed among a cohort of physicists (Fig. A.2). It suggests that the recent shift towards dark matter in particle physics can be understood as a strategy for making the best of particle physicists' prior knowledge, in a context where further exploration of the high-energy frontier in particle accelerators becomes increasingly less attractive. By contrast, if one seeks purely empirical justifications for the pursuit of dark matter, it might be more difficult to fully appreciate the rationality of its pursuit over that of competing approaches such as modified gravity. In addition, sequential constraints strongly influence large-scale particle-physics experiments. For instance, the underground Large Hadron Collider reuses a significant amount of infrastructure from prior experiments, including the tunnel of its predecessor (LEP), which is itself built upon its ancestors. Sequential constraints capture the fact that progress is impossible if one constantly overthrows prior scientific capital (whether this entails knowledge or material resources). Sequential constraints give rise to path dependency, but are nevertheless necessary for conducting science.

Finally, *contextual* constraints arise when a certain choice is only constrained in relation to other choices for epistemic or instrumental reasons. While the context includes the surrounding system of beliefs [101], it may also include non-epistemic values and axiological commitments. Chapter 1 shows that physicists diverge in their assessment of the pursuit-worthiness of supersymmetry, since this theory is increasingly unlikely to support phenomenological progress while remaining crucial in highly theoretical developments in quantum gravity. Therefore, the pursuit-worthiness of supersymmetry is not constrained (or determined) in absolute terms: it can only be assessed in relation to other choices (e.g. whether to require immediate or foreseeable phenomenological implications) that may themselves be subject to revision.

From the point of view of the philosophy of science, the "collective cognition" perspective hints at additional criteria for heuristic appraisal that may provide further constraints and justifications for our scientific conduct (as in the case of dark matter). These are especially valuable in high-energy physics, which suffers from a scarcity of empirical evidence, and subsequently from higher levels of epistemic underdetermination [105]. In addition, the approach developed in this thesis suggests normative implications for the allocation and divison of labor in science. For instance, the Optimal Transport [106] approach developed in Chapter 2 assumes the existence of a collective constraint on the (supposedly

optimal) distribution of research efforts across research topics (which, roughly speaking, specifies how much effort should be devoted to each of multiple topics). By incorporating the need to minimize cognitive learning costs, Optimal Transport can translate this collective constraint into individual constraints by deriving an optimal allocation of individual labor (the gray arrows in Fig. A.2): for instance, individuals who have skills and knowledge relevant to particle phenomenology are better equipped to search dark matter compared to others[13].

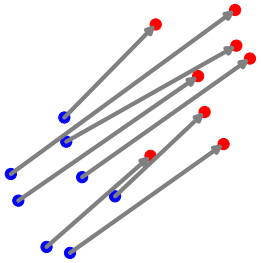


Figure A.2: **Scientists' adaptive patterns in the epistemic space**, as they relocate from the blue nodes to the red nodes. Given the target distribution of research efforts (in red), Optimal Transport provides the assignments (gray arrows) that minimize the distance traveled in the epistemic space, which measures the cognitive effort of each scientist.

A.5.2 Implications for the sociology of science

There is a significant gap in the "world-views" of qualitative and quantitative science studies, which focus on different phenomena and revolve around distinct ontologies [108]. This thesis proposes to address this gap in two different ways. One of them (the top-down approach) is to provide novel ways of operationalizing central qualitative notions from the sociology of science into quantitative analyses of scientific literature. I illustrate this strategy with two such concepts: Galison's trading zones, and Bourdieu's capital. However, another way to establish connections between qualitative and quantitative research is to draw qualitative insights from computational models (the bottom-up approach). Computational models, indeed, can serve as a source of inspiration in the process of generating concepts for the social sciences.

The top-down approach: from concepts to quantitative measures

The concept of trading zone was introduced by Galison [99] in order to account for how physicists with highly different expertise (and jargon) achieve mutual understanding. Since then, this concept has spurred many works in social studies of science [109]. Yet, perhaps surprisingly, Chapter 1 is the first quantitative operationalization of the concept. In this paper, I show that citations in science are a proxy for the "trades" occurring between different scientific cultures. I use this approach to explore the concepts that have contributed to sustaining trades between different areas of high-energy physics throughout time, by locating the keywords most frequently involved in citations across these research areas. This reveals, among other things, a major shift from collider physics to astrophysics in the trading zone between pure theory and phenomenology.

[13] See [107] for an exploration of the relation between game theory and the original Monge problem in Optimal Transport.

Pierre Bourdieu has tremendously influenced the sociology of science, and many quantitative sociologists of science and bibliometricians have naturally sought to translate major concepts of his work into quantitative measures [110], including the notion of capital. Per Bourdieu, capital (the assets that individuals may leverage in a competitive setting) can come in many forms: intellectual, social, symbolic, etc. Despite the influence of this notion in quantitative studies of science, this thesis reveals gaps in previous bibliometric approaches to Bourdieu's notion of capital. For instance, Chapter 2 investigates the effect of the diversity of scientists' intellectual capital on their ability to adapt. Prior works have not provided satisfactory measures for the diversity of scientists' intellectual capital, which prompted me to develop a novel information-theoretical approach based on embedding topic modeling. In addition, while [111] acknowledged that scientists' social capital is multi-dimensional, featuring at least two partially independent components – diversity and power –, their own operationalization of each of these dimensions of social capital turned out to be inappropriate. In particular, their measure for diversity is based on scientists' degree-centrality in their social network (e.g. their amount of collaborators), irrespective of the similarity or dissimilarity of their collaborators' profiles. In response, Chapter 2 proposes a measure for the diversity of social capital that captures the intellectual diversity that one has access to through their social network. Chapter 2 then shows that these two dimensions of social capital (diversity and power) have opposite effects on the magnitude of change in scientists' research interests over time: diversity enhances change, while power enhances stability. This observation would not have been possible had we limited ourselves to the measure of diversity from [111].

The bottom-up approach: from computational models to qualitative concepts

Quantitative studies of science need not be subordinate to high-level qualitative theory; instead, we may use computational approaches as a source of inspiration in the formation of qualitative concepts. Chapter 3, for instance, uses the Ising model to explain the formation of physicists' preferences towards a sign convention. This model comes from statistical physics, where it was introduced to describe "spin systems" and spontaneous magnetization in certain materials. In such materials, atomic spins are influenced by two forces, which are the two components of the Ising model: the local effect of their neighbors, and the effect of the external magnetic field in which they are plunged. Chapter 3 builds upon these two components of the model and proposes a similar distinction between local processes in the diffusion of conventions (which spread throughout a social network) and global processes (external to the social network). While local processes of coordination can include imitation or strategic adaptation, global processes include formal institutions and central authorities (or disciplinary matrices, in the context of science). Therefore, the distinction not only has a clear meaning in the Ising model, but also a significance for social systems. In fact, the strong entrenchment of the divide in a robust model, which can extract quantitative information about real systems, gives us some confidence that the concept designates something real – to paraphrase Hacking [112], "if I can [measure

it], then [it is] real".

Another fruitful example is the notion of collective consistency (already discussed in our discussion of underdetermination), which regroups social consistency (the coordination of individuals' behavior), sequential consistency (the coordination of consecutive actions in a mutually coherent way), and contextual consistency (the collective consistency of simultaneous beliefs or actions). While these seem like highly heterogeneous imperatives, Chapter 3 demonstrates that they can be modeled in similar ways. I have shown that one can appeal to a unified game-theoretic description of these three constraints (Table A.1). Alternatively, one may appeal to a unified probabilistic definition: collective consistency at large involves constraints on a *joint* probability distribution (i.e. $p(x_1, \ldots, x_n)$), without entailing constraints on the marginal probability distributions $p(x_1), \ldots, p(x_n)$. In the case of left-hand versus right-hand driving, for instance, it is more probable that two individuals driving past each other will adopt the same behavior – $[p(L, L) = p(R, R)] > [p(L, R) = p(R, L)]$ – but this does not say which outcome they should prefer on average – such that $p(x_1 = L) = p(x_1 = R)$. In terms of the Shannon entropy H, conventionality implies that $H(x_1, \ldots, x_n) \underset{n \to +\infty}{\ll} \sum_i H(x_i)$: on average, the joint-strategy of the agents is comparatively much constrained than their individual strategies.

The Ising model used in Chapter 3 is a simple account of how such purely collective constraints may surface[14], and is directly tied to two-player, two-action games, such as those represented in Table A.1 [114]. For instance, sequential consistency can be modeled by a Markov chain, where transitions between inconsistent actions introduce "switching costs". However, such a model has the same structure as the Ising model on a one-dimensional lattice. Additionally, contextual consistency (i.e. the consistency of a set of beliefs and cultural practices) can also be reasonably captured by an Ising model. In fact, the Ising model can adequately model cultural fitness landscapes, as previously shown by [44], and as we demonstrate again in Chapter 3 through the empirical reconstruction of a fitness landscape of multiple conventional choices. If collective consistency, in its different forms, can be modeled using the same basic formal components, we may be more willing to consider that the notion, in spite of its breadth, exhibits high coherence. In the end, the finding that social and contextual constraints can be made sense of using the exact same framework suggests that it does provide an account of why and how, as Giere puts it [19], "the social and the cognitive overlap". This confirms that computational models can provide conceptual clarification.

[14]The Ising model is simple because it only assumes pairwise interactions, which provides enough complexity to characterize a broad class of systems. Indeed, seemingly higher-order correlations often emerge from simpler pairwise interactions [113].

A.5.3 Implications for complexity and computational social science

Inverse problems for social science

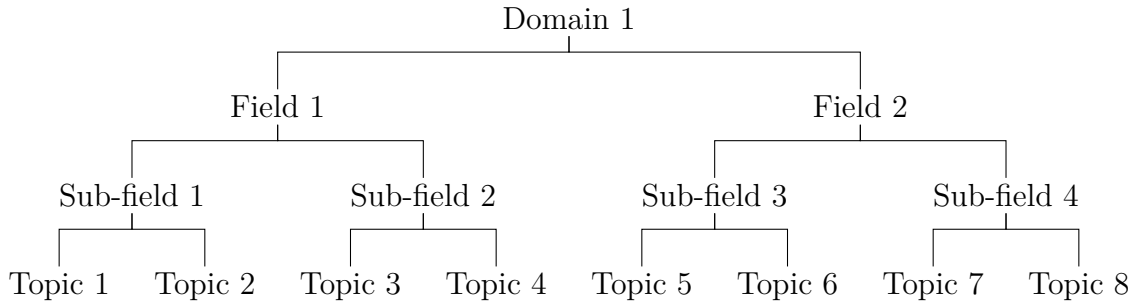
This thesis makes several methodological contributions to the field of computational social science, in particular through novel applications of inverse problems under Bayesian inference. For instance, Chapter 2 is the first empirical application of Probabilistic Inverse Optimal Transport [115], and shows the promise of Optimal Transport as an alternative to traditional models of mobility such as the gravity model [116]. In addition, Chapter 3 is the first paper to apply the inverse Ising problem in order to retrieve the structure of an underlying coordination game from behavioral data. This effort follows very recent theoretical work drawing the connection between the Ising model and coordination games [114] that had not been applied empirically before. I also show that the Ising model can disentangle endogenous collective behavior (emerging out of local interactions throughout a network) from exogenous collective behavior (arising from factors transcending the network) using a set of behavioral observations. While the Ising model is rather formal and unrealistic, as I demonstrate, solving the inverse Ising problem can help compare the plausibility of more realistic agent-based models of collective behavior by providing summary statistics measuring the contribution of these two general processes of coordination. This suggests exciting research directions at the intersection between "social physics" [117] and recent developments in simulation-based inference with deep-learning [97]. Generally speaking, as discussed above, inverse problems establish a bridge between often disconnected traditions in computational social science A.3.6.

To further stress the value of the inverse problems in computational social science, let us emphasize a finding of Chapters 2 and 3 that has only been briefly discussed in both of these papers. Both Inverse Optimal Transport and the inverse Ising problem effectively retrieve utility functions out of behavioral data (which are respectively shaped by migration costs and coordination costs). That is, these methods enable us to learn what agents are "optimizing" for. Fortunately, these do not imply or require that agents are perfectly efficient and rational. In both cases, these methods can in fact simultaneously measure the degree of efficiency and rationality of the agents. In Inverse Optimal Transport, inefficiency is captured by an entropic regularization term (which was historically introduced for completely different reasons). In the inverse Ising model, the inverse temperature β – which is also related to entropy – is a measure of efficiency. This parameter is related to the degree of rationality of individual agents in a coordination game context [114]. It follows from the assumption that agents follow a noisy best-response strategy (the so-called logit rule, which is a simple model of bounded rationality).

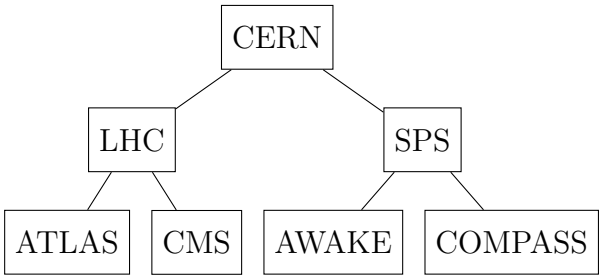
Social dilemmas and self-organization

In collective cognition, socio-epistemic systems "adapt" to their environment by "*self-organizing*" into rich complex structures featuring recurrent patterns [118]. One such

pattern is *modularity*, which arises when different parts of a system "specialize" into different functions [119]. Another example is *hierarchies*. Take this very thesis, for example. It is organized hierarchically, into chapters, which are divided into sections, which are themselves divided into subsections, that are themselves broken up into paragraphs. This structure has a functional role: it serves to indicate the nested imbrication of modules (ideas and arguments) into submodules, and considerably speeds up the search of relevant information. What is the function of hierarchies, in socio-epistemic systems in particular? I suggest hierachies in socio-epistemic systems may serve to resolve multiple of the social dilemmas identified throughout this thesis. For instance, Chapter 2 focuses on the trade-off between specialization and adaptation, and found that the "cost" of shifting attention from one topic to another is shaped by the probability that an author holds knowledge in these two topics (i.e., by their cognitive proximity). Incidentally, Chapter 2 finds hints that knowledge is distributed hierarchically among physicists, who are more likely to hold expertise in topics that belong to a same area (and so forth, across all levels of the hierarchy; see Figure A.3a). This implies that it is easier to migrate across subtopics within the same topic than to migrate across broader topics. A hierarchical distribution of knowledge can help relieve the tension between specialization and adaptation by taking advantage of shared background knowledge between "siblings" at each level of the hierarchy. For instance, physicists may switch from collider physics to dark matter physics, which is relatively easy; if necessary, they may leave high-energy physics for another kind of physics research, which might be a bit harder, but who still salvaged some amount of background knowledge.



(a) Example of hierarchical knowledge structure.



(b) Hierarchical structures at CERN.

Figure A.3

Another social dilemma that is partially relieved by appealing to hierarchical structures is the trade-off between independent and cooperative learning. Scientists must rely on cooperation, in order to elaborate complex solutions and take advantage of each others' knowledge, but they also need to preserve some independence, in order to explore alternative strategies in parallel. If we take the example of CERN again, we can see that its infrastructure follows some sort of hierarchical structure, by running multiple accelerators, which themselves run multiple experiments (Figure A.3b). Such a structure directly relieves the tension between independent and cooperative learning: both the Large Hadron Collider (LHC) and the Super Proton Synchrotron (SPS) benefit from a certain amount of shared infrastructure, but they nevertheless play complementary roles. At the same time, each accelerator diversifies its own scientific output by running multiple experiments (ATLAS and CMS, among others, at the LHC; AWAKE and COMPASS, among others, at the SPS), which themselves benefit from some form of cooperation by relying on the same accelerators. The hierarchical structure in Figure A.3b therefore directly addresses the need to diversify research (by pursuing multiple independent efforts) and yet coordinate these efforts (to reap the benefits of cooperation). Similarly, preliminary findings suggest that the community structure of theoretical physics is arranged into hierarchical levels that correlate with levels in the linguistic hierarchy of knowledge in the field (see Figure A.4). Such correspondence may indicate the adaptive co-evolution of epistemic and social structures, in such a way that alleviates the tension between coordination and specialization.

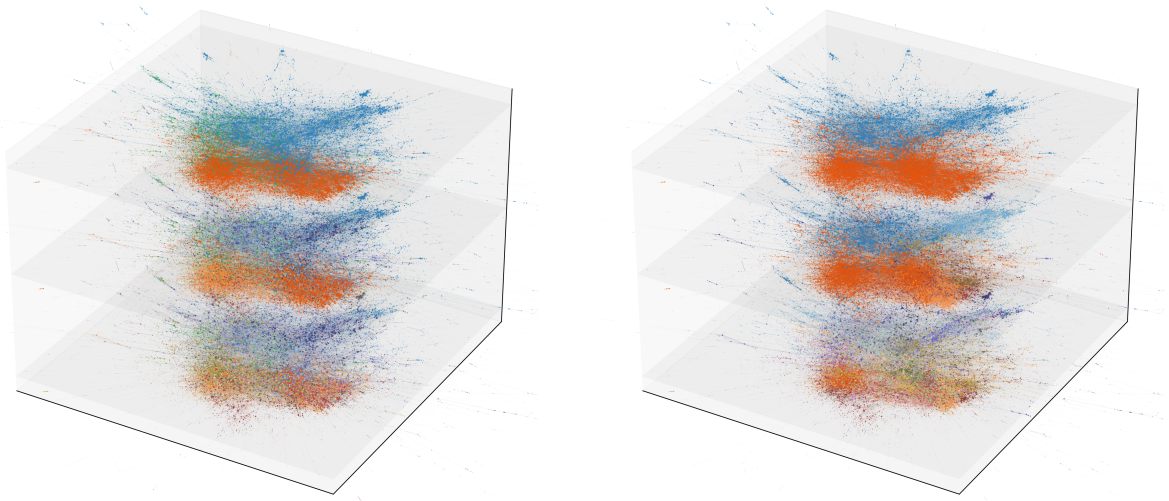
To conclude, one might say that further investigation of social dilemmas arising in contexts of collective cognition may help us understand structural patterns emerging across a range of complex systems, including (but not limited) to modularity and hierarchy.

A.5.4 Implications for scientific and institutional change

This thesis was conducted as part of the research training group on *Transformations of Science and Technology since 1800*. In the last section of this chapter, I propose to reflect upon the contribution of the present thesis to the research program of our group. Before I elaborate, let me just stress that our collective is committed to an interdisciplinary perspective on science; the present thesis clearly exhibits this orientation, since it gathers very diverse perspectives and finds implications across multiple fields (Figure A.5). In what follows, I discuss the other dimensions of our research program.

Topics, processes, and institutions

Our research group proposes to acknowledge that transformations in science involve three inter-related layers: *topics* (the intellectual product of scientific research), *processes* (the making of science), and *institutions*. To understand how science has evolved over the past centuries, we must acknowledge that these dimensions have co-evolved in interaction with each other. In this respect, the adaptive multi-agent system framework outlined in this introduction provides a flexible account of the co-evolution of these three dimensions: in



(a) Hierarchical linguistic structure. Colors indicate the most frequent topic among each author's publications. Topics are recovered by performing unsupervised clustering on the sentence embeddings derived from the abstracts of $D = 228\,748$ publications [120].

(b) Hierarchical community structure. Colors indicate each author's community, across three levels of the hierarchy in the degree-corrected hierarchical stochastic block model [121].

Figure A.4: Partition of authors in the co-authorship network, across different levels of the hierarchy of topics (left) and communities (right). Excerpt from a work in progress with Mike D. Schneider.

Fig. A.1, the problem structure and the solutions explored by the epistemic agents represent the *topics*; the agents' learning strategies, in contrast, represent the *processes*; and the incentive structures represent the *institutions* (whether these are formal institutions, or more or less implicit social norms). As I have shown, these three dimensions interact with each other via complex mechanisms such as feedback loops, which may be formalized in terms of complex adaptive systems [86, 40]. Therefore, historians, sociologists, and philosophers of science might find useful resources in this approach to scientific change, acknowledging the complex interplay between topics, processes, and institutions. More specifically, Chapter 1 shows how division of labor in high-energy physics (*processes*) can lead to self-reinforcing epistemic divergences between different communities (*topics*). Chapter 2 shows how the experimental landscape, which is itself shaped by institutions redefines the scientists' research agendas (processes). Finally, each of the three dimensions of conventions introduced in Chapter 3 (contextual, sequential, and social consistency) can be associated with one layer: contextual consistency demands that *topics* be mutually coherent systems of beliefs; sequential consistency demands that *processes* involve coherent sequences of actions; and social consistency expresses a drive for *institutions* coordinating individuals' behavior. Therefore, the formal model of conventions introduced in Chapter 3 directly provides an account of the interactions between these three layers.

Historical institutionalism and modes of change in science

Another aim of our research group has been to explore potential implications of "historical institutionalism" (HI) for the understanding of scientific change. Historical institutionalism is a framework from political science that acknowledges that institutional arrangements cannot be understood without reference to their history. At the core of historical institutionalism lies the concept of path-dependence, according to which prior arrangements can significantly constrain future developments. Institutional change nevertheless takes place through two broad kinds of patterns, namely *punctuated equilibria* and *gradual change*. In the former, institutions enjoy long periods of stability (equilibria) separated by short periods of significant transformations with long-lasting effects (critical junctures). However, institutional transformations may also result from more continuous and incremental change with significant cumulative effects overtime. Our group proposes to investigate whether a typology of gradual institutional change developed by historical institutionalists [122] can account for transformations in science more generally. This includes displacement (when former rules are abandoned and replaced by new ones), layering (when new rules are added), and drift (when old rules acquire a new meaning under new circumstances). Our group proposes to apply this framework not just to formal *institutions* but also to the other dimensions of science discussed above, that is, *topics* and *processes*. Previous works have shown that this typology of change could account for the shift from high-energy physics to photon science among organizations such as DESY and SLAC [123–125] in response to adaptive pressures. Chapter 2 goes further and applies this typology of incremental change to characterize the adaptation strategies of *individual* scientists. Indeed, as noticed by [126], scientists tend to revise their research agendas gradually, in order to "retain" the benefits of their expertise while progressively engaging in new opportunities. This, however, entails multiple forms of intellectual change. First, scientists can repurpose their prior knowledge (e.g. prior concepts) to new objectives, which I call *conversion*, following the terminology of HI. However, adaptation can sometimes prompt scientists to expand their knowledge by acquiring new concepts or techniques for their research (*layering*). Finally, it may be that scientists have to abandon certain kinds of knowledge altogether in the face of new circumstances (*displacement*). Fundamentally, gradual change appears to be a rather universal adaptation strategy arising from the need to adjust large accumulations of capital (intellectual, social, material, institutional, …) to new realities. It appears in different forms in high-energy physics – for instance, the CERN reused the tunnel from a prior experiment (LEP) for the Large Hadron Collider (*conversion*), which also relies on prior accelerators such as the Super Proton Synchrotron as injectors *layering*.

Institutions, adaptation, and evolution

While evolutionary models have proven successful in describing certain aspects of technological change, it remains an open question whether evolutionary theory provides a satisfactory account of institutional change. For it to be the case, institutions should

undergo processes of variation and selection that result in the "gradual accumulation of beneficial changes" [127]. I have just made the argument that there is a parallel to be drawn between the forms that "variation" take in institutional change and in technological or scientific change. By contrast, [127] believe that the analogy is limited for several reasons: institutional variation is too constrained (since it requires social coordination, only a few "equilibria" can be explored), and the fitness of institutional arrangement depends on how many people adhere to them, which diminishes the efficiency of the selection process. Nevertheless, processes such as group-level selection could help explain how institutions and norms become gradually more adaptive over time [128].

In any case, studies of institutional change may benefit from a complex adaptive system perspective coupled with evolutionary insights. Take for instance the idea of "niche construction", according to which species do not just passively adapt to their environment, but also reshape it in a co-evolutionary process [129]. [130] have proposed to view institution-building as a process of social niche construction, whereby humans establish a background of "stable incentives" that render long-term strategic planning possible. In a nutshell, institutions reshape our cultural environment in a way that promotes stability and enables adaptive learning. In the long run, however, such stability can lead to runaway growth. The larger the niche becomes, the more individuals may rely on it, until it reaches its carrying capacity and becomes unstable. This is well exemplified by the exponential growth of science in the course of the 20th century [131]. Such growth is not sustainable: scientific institutions are increasingly overpopulated, which encourages the emergence of harmful behavior (e.g. "publish or perish", and other practices that can undermine the credibility of science) as competition intensifies. While it might seem counter-intuitive that the same processes that promote stability can eventually lead to instability, such kinds of dynamics are ubiquitous and well understood in the framework of complex adaptive systems.

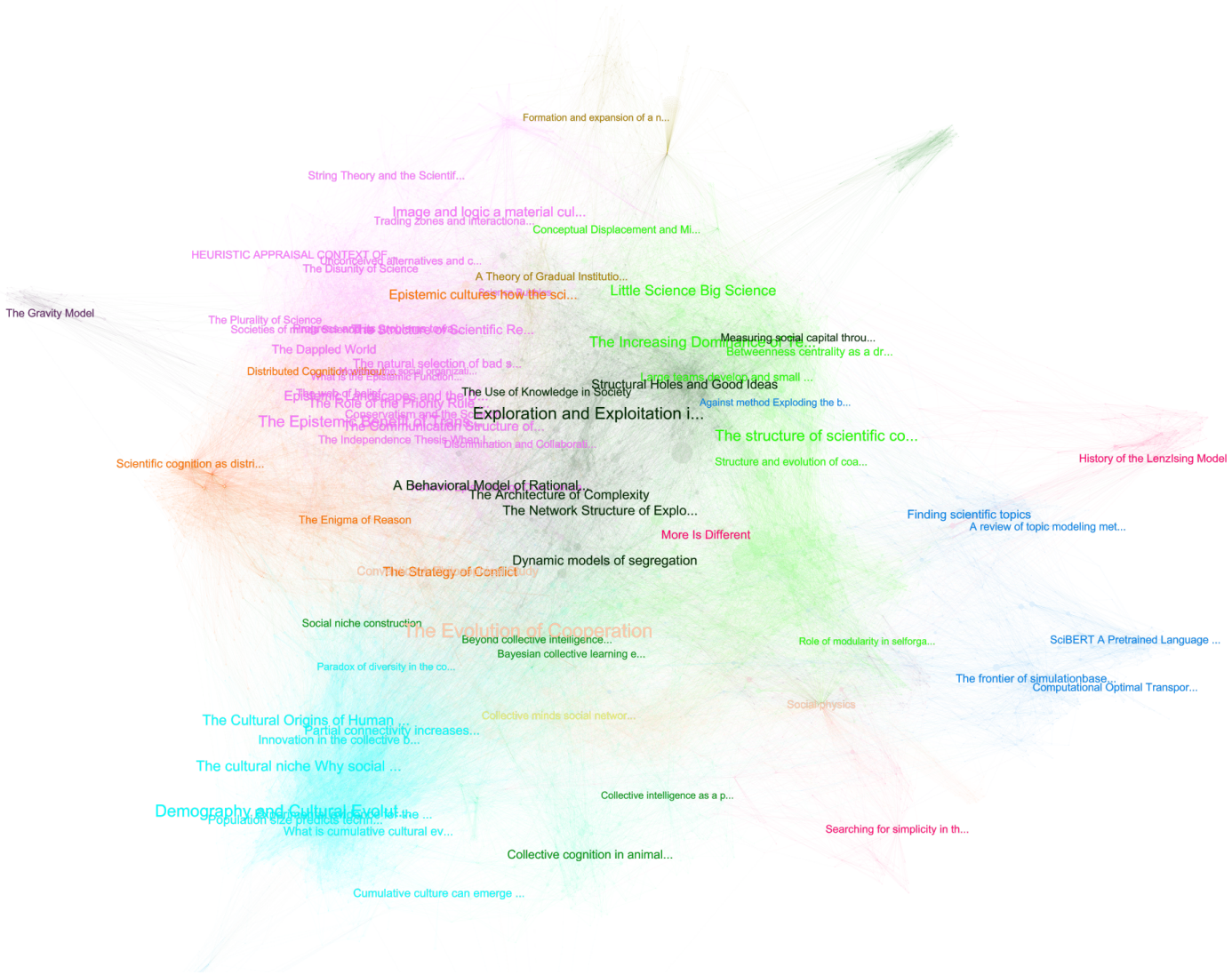


Figure A.5: Citation network of the references listed in this introduction and their forward/backward citations. Clusters include philosophy and social epistemology of science (pink), cognitive and decision science (orange), cultural evolution (cyan), collective intelligence (dark green), statistical mechanics (purple), computer science (dark blue), quantitative studies of science (light green), organizational and institutional science (kaki). The black cluster at the center mostly entails innovation and network science literature.

Bibliography

[1] A. Lopes and M. L. Perrey. *FAQ-LHC: The Guide*. Brochure. CERN, 2022. URL: https://cds.cern.ch/record/2809109/files/CERN-Brochure-2021-004-Eng.pdf (visited on 01/11/2025).

[2] A. Collaboration. *ATLAS Fact Sheet*. Brochure. CERN, 2012. URL: https://cds.cern.ch/record/1457044/files/ATLAS%20fact%20sheet.pdf (visited on 01/11/2025).

[3] C. Shalizi. *Workshop "Collective Cognition: Mathematical Foundations of Distributed Intelligence"*. Santa Fe Institute, 2002.

[4] R. Boyd, P. J. Richerson, and J. Henrich. "The cultural niche: Why social learning is essential for human adaptation". In: *Proceedings of the National Academy of Sciences* 108 (2011), pp. 10918–10925. URL: http://dx.doi.org/10.1073/pnas.1100290108.

[5] J. Henrich. *The Secret of our Success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University press, 2016.

[6] H. Mercier and D. Sperber. *The Enigma of Reason*. Harvard University Press, 2017. URL: http://dx.doi.org/10.4159/9780674977860.

[7] P. Krafft, E. Shmueli, T. L. Griffiths, J. B. Tenenbaum, and A. ". Pentland. "Bayesian collective learning emerges from heuristic social learning". In: *Cognition* 212 (2021), p. 104469. URL: http://dx.doi.org/10.1016/j.cognition.2020.104469.

[8] M. Tomasello. *The cultural origins of human cognition*. Harvard university press, 2009.

[9] A. Mesoudi and A. Thornton. "What is cumulative cultural evolution?" In: *Proceedings of the Royal Society B: Biological Sciences* 285.1880 (2018), p. 20180712. URL: http://dx.doi.org/10.1098/rspb.2018.0712.

[10] M. Muthukrishna and J. Henrich. "Innovation in the collective brain". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1690 (2016), p. 20150192. URL: http://dx.doi.org/10.1098/rstb.2015.0192.

[11] M. Derex, J.-F. Bonnefon, R. Boyd, and A. Mesoudi. "Causal understanding is not necessary for the improvement of culturally evolving technology". In: *Nature Human Behaviour* 3.5 (2019), pp. 446–452. URL: http://dx.doi.org/10.1038/s41562-019-0567-9.

[12] M. A. Kline and R. Boyd. "Population size predicts technological complexity in Oceania". In: *Proceedings of the Royal Society B: Biological Sciences* 277.1693 (2010), pp. 2559–2564. URL: http://dx.doi.org/10.1098/rspb.2010.0452.

[13] M. Derex, M.-P. Beugin, B. Godelle, and M. Raymond. "Experimental evidence for the influence of group size on cultural complexity". In: *Nature* 503.7476 (2013), pp. 389–391. URL: http://dx.doi.org/10.1038/nature12774.

[14] F. J. Benjamin, R. O. Kaaronen, C. Moser, W. Rorot, et al. "All intelligence is collective intelligence". In: *Journal of Multiscale Neuroscience* 2.1 (2023), pp. 169–191.

[15] F. Hayek. "The Use of Knowledge in Society". In: *The American Economic Review* 35.4 (1945), pp. 519–530. URL: https://www.jstor.org/stable/1809376.

[16] F. A. Hayek. *The Counter-Revolution of Science: Studies on the Abuse of Reason*. Glencoe, Illinois: The Free Press, a corporation, 1952.

[17] R. C. Jennings. "Truth, rationality and the sociology of science". In: *The British Journal for the Philosophy of Science* 35.3 (1984), pp. 201–211.

[18] L. Laudan. *Progress and its problems: Towards a theory of scientific growth*. Vol. 282. Univ of California Press, 1978.

[19] R. Giere. "Scientific cognition as distributed cognition". In: *The Cognitive Basis of Science*. Cambridge University Press, 2002, pp. 285–299. URL: http://dx.doi.org/10.1017/cbo9780511613517.016.

[20] R. N. Giere. "Distributed Cognition without Distributed Knowing". In: *Social Epistemology* 21.3 (2007), pp. 313–320. URL: http://dx.doi.org/10.1080/02691720701674197.

[21] P. Thagard. "Societies of minds: Science as distributed computing". In: *Studies in History and Philosophy of Science* 24.49-67 (1993), pp. 251–257.

[22] I. D. Couzin. "Collective cognition in animal groups". In: *Trends in cognitive sciences* 13.1 (2009), pp. 36–43.

[23] T. Sasaki and D. Biro. "Cumulative culture can emerge from collective intelligence in animal groups". In: *Nature Communications* 8.1 (2017). URL: http://dx.doi.org/10.1038/ncomms15049.

[24] C. Gunasekaram, F. Battiston, O. Sadekar, C. Padilla-Iglesias, et al. "Population connectivity shapes the distribution and complexity of chimpanzee cumulative culture". In: *Science* 386.6724 (2024), pp. 920–925. URL: http://dx.doi.org/10.1126/science.adk3381.

[25] C. Mayo-Wilson, K. J. S. Zollman, and D. Danks. "The Independence Thesis: When Individual and Social Epistemology Diverge". In: *Philosophy of Science* 78.4 (2011), pp. 653–677. URL: http://dx.doi.org/10.1086/661777.

[26] K. J. S. Zollman. "The Epistemic Benefit of Transient Diversity". In: *Erkenntnis* 72.1 (2009), pp. 17–35. URL: http://dx.doi.org/10.1007/s10670-009-9194-6.

[27] F. Dietrich and C. List. "Arrow's theorem in judgment aggregation". In: *Social Choice and Welfare* 29.1 (2007), pp. 19–33.

[28] P. W. Anderson. "More Is Different: Broken symmetry and the nature of the hierarchical structure of science." In: *Science* 177.4047 (1972), pp. 393–396. URL: http://dx.doi.org/10.1126/science.177.4047.393.

[29] P. K. Stanford. "Unconceived alternatives and conservatism in science: the impact of professionalization, peer-review, and Big Science". In: *Synthese* 196.10 (2015), pp. 3915–3932. URL: http://dx.doi.org/10.1007/s11229-015-0856-4.

[30] P. Jenni, T. S. Virdee, L. Pontecorvo, and S. Liyanage. "Chasing Success: The ATLAS and CMS Collaborations". In: *Big Science, Innovation, and Societal Contributions*. Oxford University PressOxford, 2024, pp. 22–55. URL: http://dx.doi.org/10.1093/oso/9780198881193.003.0003.

[31] B. Jann and W. Przepiorka. "Introduction". In: *Social dilemmas, institutions, and the evolution of cooperation*. De Gruyter, 2017, pp. 3–10. URL: http://dx.doi.org/10.1515/9783110472974-001.

[32] J. G. March. "Exploration and Exploitation in Organizational Learning". In: *Organization Science* 2.1 (1991), pp. 71–87. URL: https://doi.org/10.1287/orsc.2.1.71.

[33] M. Derex and R. Boyd. "Partial connectivity increases cultural accumulation within groups". In: *Proceedings of the National Academy of Sciences* 113.11 (2016), pp. 2982–2987. URL: http://dx.doi.org/10.1073/pnas.1518798113.

[34] R. Schimmelpfennig, L. Razek, E. Schnell, and M. Muthukrishna. "Paradox of diversity in the collective brain". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1843 (2021). URL: https://doi.org/10.1098/rstb.2020.0316.

[35] P. E. Smaldino, C. Moser, A. Pérez Velilla, and M. Werling. "Maintaining Transient Diversity Is a General Principle for Improving Collective Problem Solving". In: *Perspectives on Psychological Science* 19.2 (2023), pp. 454–464. URL: http://dx.doi.org/10.1177/17456916231180100.

[36] D. Lazer and A. Friedman. "The Network Structure of Exploration and Exploitation". In: *Administrative Science Quarterly* 52.4 (2007), pp. 667–694. URL: http://dx.doi.org/10.2189/asqu.52.4.667.

[37] L. Wu, D. Wang, and J. A. Evans. "Large teams develop and small teams disrupt science and technology". In: *Nature* 566.7744 (2019), pp. 378–382. URL: http://dx.doi.org/10.1038/s41586-019-0941-9.

[38] S. Wuchty, B. F. Jones, and B. Uzzi. "The Increasing Dominance of Teams in Production of Knowledge". In: *Science* 316.5827 (2007), pp. 1036–1039. URL: http://dx.doi.org/10.1126/science.1136099.

[39] R. L. Morris. "Increasing specialization: Why we need to make mathematics more accessible". In: *Social Epistemology* 35.1 (2021), pp. 37–47.

[40] M. Galesic, D. Barkoczi, A. M. Berdahl, D. Biro, et al. "Beyond collective intelligence: Collective adaptation". In: *Journal of The Royal Society Interface* 20.200 (2023). URL: http://dx.doi.org/10.1098/rsif.2022.0736.

[41] K. K. Cetina. *Epistemic cultures: How the sciences make knowledge*. Harvard University Press, 1999.

[42] M. Weisberg and R. Muldoon. "Epistemic Landscapes and the Division of Cognitive Labor". In: *Philosophy of Science* 76.2 (2009), pp. 225–252. URL: http://dx.doi.org/10.1086/644786.

[43] F. A. Csaszar. "A note on how NK landscapes work". In: *Journal of Organization Design* 7.1 (2018). URL: http://dx.doi.org/10.1186/s41469-018-0039-0.

[44] V. M. Poulsen and S. DeDeo. "Inferring Cultural Landscapes with the Inverse Ising Model". In: *Entropy* 25.2 (2023), p. 264. URL: http://dx.doi.org/10.3390/e25020264.

[45] T. Mikolov. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* 3781 (2013).

[46] I. Beltagy, K. Lo, and A. Cohan. "SciBERT: A pretrained language model for scientific text". In: *arXiv preprint arXiv:1903.10676* (2019).

[47] T. L. Griffiths and M. Steyvers. "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101 (2004), pp. 5228–5235. URL: http://dx.doi.org/10.1073/pnas.0307752101.

[48] I. Vayansky and S. A. Kumar. "A review of topic modeling methods". In: *Information Systems* 94 (2020), p. 101582. URL: http://dx.doi.org/10.1016/j.is.2020.101582.

[49] M. Fernández Pinto. "Methodological and Cognitive Biases in Science: Issues for Current Research and Ways to Counteract Them". In: *Perspectives on Science* 31.5 (2023), pp. 535–554. URL: http://dx.doi.org/10.1162/posc_a_00589.

[50] H. A. Simon. "A behavioral model of rational choice". In: *The quarterly journal of economics* (1955), pp. 99–118.

[51] K. J. S. Zollman. "The Communication Structure of Epistemic Communities". In: *Philosophy of Science* 74.5 (2007), pp. 574–587. URL: http://dx.doi.org/10.1086/525605.

[52] M. E. J. Newman. "The structure of scientific collaboration networks". In: *Proceedings of the National Academy of Sciences* 98.2 (2001), pp. 404–409. URL: http://dx.doi.org/10.1073/pnas.98.2.404.

[53] A. Wüthrich. "Characterizing a collaboration by its communication structure". In: *Synthese* 202.5 (2023). URL: http://dx.doi.org/10.1007/s11229-023-04376-z.

[54] K. J. Zollman. "Network epistemology: Communication in epistemic communities". In: *Philosophy Compass* 8.1 (2013), pp. 15–27.

[55] I. Momennejad. "Collective minds: social network topology shapes collective cognition". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1843 (2021). URL: http://dx.doi.org/10.1098/rstb.2020.0315.

[56] Y. Cao, Y. Dong, M. Kim, N. G. MacLaren, et al. "Effects of network connectivity and functional diversity distribution on human collective ideation". In: *npj Complexity* 2.1 (2025). URL: http://dx.doi.org/10.1038/s44260-024-00025-9.

[57] R. S. Burt. "Structural holes and good ideas". In: *American journal of sociology* 110.2 (2004), pp. 349–399.

[58] A. Abbasi, K. S. K. Chung, and L. Hossain. "Egocentric analysis of co-authorship network structure, position and performance". In: *Information Processing & Management* 48.4 (2012), pp. 671–679. URL: http://dx.doi.org/10.1016/j.ipm.2011.09.001.

[59] Y. Liu, M. Zhang, G. Zhang, and X. You. "Scientific elites versus other scientists: who are better at taking advantage of the research collaboration network?" In: *Scientometrics* 127.6 (2022), pp. 3145–3166. URL: http://dx.doi.org/10.1007/s11192-022-04362-1.

[60] T. S. Kuhn. *The Structure of Scientific Revolutions*. 2nd edition, with postscript. Chicago: University of Chicago Press, 1970.

[61] H. Rubin and C. O'Connor. "Discrimination and collaboration in science". In: *Philosophy of Science* 85.3 (2018), pp. 380–402.

[62] S. Fazelpour and H. Rubin. "Diversity and homophily in social networks". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44. 2022.

[63] H. Rubin. "Structural causes of citation gaps". In: *Philosophical Studies* 179.7 (2022), pp. 2323–2345.

[64] P. Kitcher. *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press, USA, 1993.

[65] J. Wu, C. O'Connor, and P. E. Smaldino. "The Cultural Evolution of Science". In: *The Oxford Handbook of Cultural Evolution*. Oxford University Press, 2023. URL: http://dx.doi.org/10.1093/oxfordhb/9780198869252.013.78.

[66] M. Strevens. "The Role of the Priority Rule in Science". In: *Journal of Philosophy* 100.2 (2003), pp. 55–79. URL: http://dx.doi.org/10.5840/jphil2003100224.

[67] T. C. Schelling. *The strategy of conflict*. Rev. ed. Harvard University Press, 1960, 1980.

[68] R. Axelrod and W. D. Hamilton. "The evolution of cooperation". In: *science* 211.4489 (1981), pp. 1390–1396.

[69] D. Lewis. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press., 1969.

[70] N. E. Leonard and S. A. Levin. "Collective intelligence as a public good". In: *Collective Intelligence* 1.1 (2022). URL: http://dx.doi.org/10.1177/26339137221083293.

[71] O. Tchernichovski, S. Frey, N. Jacoby, and D. Conley. "Incentivizing free riders improves collective intelligence in social dilemmas". In: *Proceedings of the National Academy of Sciences* 120.46 (2023). URL: http://dx.doi.org/10.1073/pnas.2311497120.

[72] E. Kummerfeld and K. J. S. Zollman. "Conservatism and the Scientific State of Nature". In: *The British Journal for the Philosophy of Science* 67.4 (2016), pp. 1057–1076. URL: http://dx.doi.org/10.1093/bjps/axv013.

[73] T. Dreyer, A. Haluts, A. Korman, N. Gov, et al. "Comparing cooperative geometric puzzle solving in ants versus humans". In: *Proceedings of the National Academy of Sciences* 122.1 (2024). URL: http://dx.doi.org/10.1073/pnas.2414274121.

[74] P. Pierson et al. *Path dependence, increasing returns, and the study of politics*. Minda de Gunzburg Center for European Studies, Harvard University, 1997.

[75] M. Backović. *A Theory of Ambulance Chasing*. 2016. URL: https://arxiv.org/abs/1603.01204.

[76] P. Grim, J. Kavner, L. Shatkin, and M. Trivedi. "Philosophy of Science, Network Theory and Conceptual Change: Paradigm Shifts as Information Cascades". In: *Complex Systems in the Social and Behavioral Sciences: Theory, Method, and Application*. Ed. by E. Elliot and L. D. Kiel. Ann Arbor: University of Michigan Press, Forthcoming.

[77] D. B. Pedersen and V. F. Hendricks. "Science Bubbles". In: *Philosophy & Technology* 27.4 (2013), pp. 503–518. URL: http://dx.doi.org/10.1007/s13347-013-0142-7.

[78] P. Liu and H. Xia. "Structure and evolution of co-authorship network in an interdisciplinary research field". In: *Scientometrics* 103.1 (2015), pp. 101–134. URL: http://dx.doi.org/10.1007/s11192-014-1525-y.

[79] J.-F. Mertens. "Repeated Games". In: *Game Theory and Applications*. Elsevier, 1990, pp. 77–130. URL: http://dx.doi.org/10.1016/B978-0-12-370182-4.50009-X.

[80] B. Skyrms. *Evolution of the social contract*. Cambridge University Press, 2014.

[81] P. E. Smaldino and R. McElreath. "The natural selection of bad science". In: *Royal Society Open Science* 3.9 (2016), p. 160384. URL: http://dx.doi.org/10.1098/rsos.160384.

[82] P. E. Smaldino and C. O'Connor. "Interdisciplinarity can aid the spread of better methods between scientific communities". In: *Collective Intelligence* 1.2 (2022), p. 263391372211318. URL: http://dx.doi.org/10.1177/26339137221131816.

[83] C. O'Connor. "The natural selection of conservative science". In: *Studies in History and Philosophy of Science Part A* 76 (2019), pp. 24–29. URL: http://dx.doi.org/10.1016/j.shpsa.2018.09.007.

[84] J. Henrich. "Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the Tasmanian case". In: *American antiquity* 69.2 (2004), pp. 197–214.

[85] M. Muthukrishna, B. W. Shulman, V. Vasilescu, and J. Henrich. "Sociality influences cultural complexity". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1774 (2014), p. 20132511. URL: http://dx.doi.org/10.1098/rspb.2013.2511.

[86] J. H. Miller and S. E. Page. *Complex adaptive systems: an introduction to computational models of social life: an introduction to computational models of social life*. Princeton university press, 2009.

[87] T. C. Schelling. "Dynamic models of segregation". In: *Journal of mathematical sociology* 1.2 (1971), pp. 143–186.

[88] P. Grim and D. Singer. "Computational Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta and U. Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University, 2024.

[89] D. Frey and D. Šešelja. "What Is the Epistemic Function of Highly Idealized Agent-Based Models of Scientific Inquiry?" In: *Philosophy of the Social Sciences* 48.4 (2018), pp. 407–433. URL: http://dx.doi.org/10.1177/0048393118767085.

[90] C. Martini and M. Fernández Pinto. "Modeling the social organization of science: Chasing complexity through simulations". In: *European Journal for Philosophy of Science* 7.2 (2016), pp. 221–238. URL: http://dx.doi.org/10.1007/s13194-016-0153-1.

[91] C. Cioffi-Revilla. "The scope of computational social science". In: *Handbook of Computational Social Science, Volume 1*. Routledge, 2021, pp. 17–32.

[92] P. E. Smaldino. *Modeling social behavior: mathematical and agent-based models of social dynamics and cultural evolution*. Princeton University Press, 2023.

[93] G. Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences* (1781), pp. 666–704.

[94] C. Herfeld. "Model transfer in science". In: *The Routledge Handbook of Philosophy of Scientific Modeling* (2024), pp. 270–284.

[95] S. G. Brush. "History of the Lenz-Ising Model". In: *Reviews of Modern Physics* 39.4 (1967), pp. 883–893. URL: http://dx.doi.org/10.1103/RevModPhys.39.883.

[96] M. W. Macy, B. K. Szymanski, and J. A. Hołyst. "The Ising model celebrates a century of interdisciplinary contributions". In: *npj Complexity* 1.1 (2024). URL: http://dx.doi.org/10.1038/s44260-024-00012-0.

[97] K. Cranmer, J. Brehmer, and G. Louppe. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062. URL: http://dx.doi.org/10.1073/pnas.1912789117.

[98] S. T. Radev, M. Schmitt, L. Schumacher, L. Elsemüller, et al. *BayesFlow: Amortized Bayesian Workflows With Neural Networks*. 2023. URL: https://arxiv.org/abs/2306.16015.

[99] P. Galison. *Image and Logic: a material culture of microphysics*. Chicago: University of Chicago Press, 1997.

[100] K. Stanford. "Underdetermination of Scientific Theory". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta and U. Nodelman. Summer 2023. Metaphysics Research Lab, Stanford University, 2023.

[101] W. V. O. Quine and J. S. Ullian. *The web of belief*. Vol. 2. Random House New York, 1978.

[102] P. Galison. "Context and Constraints". In: *Scientific practice : theories and stories of doing physics*. Ed. by J. Buchwald. Chicago: University of Chicago Press, 1995. Chap. 2.

[103] P. Galison. "Theory Bound and Unbound: Superstrings and Experiments". In: *Laws of Nature: Essays on the Philosophical, Scientific and Historical Dimensions*. Ed. by F. Weinert. Berlin, New York: De Gruyter, 1995, pp. 369–408. URL: https://doi.org/10.1515/9783110869859.369.

[104] T. Nickles. "Heuristic Appraisal: Context of discovery of justification?" In: *Revisiting Discovery and Justification*. Ed. by J. Schikore and F. Steinle. Kluwer Academic Publishers, 2005, pp. 159–182. URL: https://doi.org/10.1007/1-4020-4251-5_10.

[105] R. Dawid. *String theory and the scientific method*. Cambridge University Press, 2013.

[106] G. Peyré and M. Cuturi. "Computational Optimal Transport: With Applications to Data Science". In: *Foundations and Trends in Machine Learning* 11.5–6 (2019), pp. 355–607. URL: http://dx.doi.org/10.1561/2200000073.

[107] J. v. Neumann. "1. A Certain Zero-sum Two-person Game Equivalent to the Optimal Assignment Problem". In: *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, 1953, pp. 5–12. URL: http://dx.doi.org/10.1515/9781400881970-002.

[108] D. Kang and J. Evans. "Against method: Exploding the boundary between qualitative and quantitative studies of science". In: *Quantitative Science Studies* 1.3 (2020), pp. 930–944. URL: https://doi.org/10.1162/qss_a_00056.

[109] H. Collins, R. Evans, and M. Gorman. "Trading zones and interactional expertise". In: *Studies in History and Philosophy of Science Part A* 38.4 (2007), pp. 657–666.

[110] M. Schirone. "Field, capital, and habitus: The impact of Pierre Bourdieu on bibliometrics". In: *Quantitative Science Studies* 4.1 (2023), pp. 186–208. URL: https://doi.org/10.1162/qss_a_00232.

[111] A. Abbasi, R. T. Wigand, and L. Hossain. "Measuring social capital through network analysis and its influence on individual performance". In: *Library & Information Science Research* 36.1 (2014), pp. 66–73. URL: https://doi.org/10.1016/j.lisr.2013.08.001.

[112] I. Hacking. *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press, 1983.

[113] G. J. Stephens, L. C. Osborne, and W. Bialek. "Searching for simplicity in the analysis of neurons and behavior". In: *Proceedings of the National Academy of Sciences* 108 (2011), pp. 15565–15571. URL: http://dx.doi.org/10.1073/pnas.1010868108.

[114] F. Zimmaro, S. Galam, and M. A. Javarone. "Asymmetric games on networks: Mapping to Ising models and bounded rationality". In: *Chaos, Solitons & Fractals* 181 (2024), p. 114666. URL: http://dx.doi.org/10.1016/j.chaos.2024.114666.

[115] W.-T. Chiu, P. Wang, and P. Shafto. "Discrete probabilistic inverse optimal transport". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3925–3946.

[116] J. E. Anderson. "The Gravity Model". In: *Annual Review of Economics* 3.1 (2011), pp. 133–160. URL: http://dx.doi.org/10.1146/annurev-economics-111809-125114.

[117] M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, et al. "Social physics". In: *Physics Reports* 948 (2022), pp. 1–148. URL: http://dx.doi.org/10.1016/j.physrep.2021.10.005.

[118] H. A. Simon. "The Architecture of Complexity". In: *Proceedings of the American Philosophical Society* 106 (1962), pp. 467–482.

[119] B. A. Siebert, C. L. Hall, J. P. Gleeson, and M. Asllani. "Role of modularity in self-organization dynamics in biological networks". In: *Physical Review E* 102.5 (2020), p. 052306.

[120] M. Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[121] T. P. Peixoto. "Hierarchical Block Structures and High-Resolution Model Selection in Large Networks". In: *Physical Review X* 4.1 (2014). URL: http://dx.doi.org/10.1103/PhysRevX.4.011047.

[122] J. Mahoney and K. Thelen. "A Theory of Gradual Institutional Change". In: *Explaining Institutional Change: Ambiguity, Agency, and Power*. Cambridge University Press, 2009, pp. 1–37.

[123] O. Hallonsten and T. Heinze. "From particle physics to photon science: Multi-dimensional and multi-level renewal at DESY and SLAC". In: *Science and Public Policy* 40.5 (2013), pp. 591–603. URL: https://doi.org/10.1093/scipol/sct009.

[124] O. Hallonsten and T. Heinze. "Formation and expansion of a new organizational field in experimental science". In: *Science and Public Policy* 42.6 (2015), pp. 841–854.

[125] T. Heinze, O. Hallonsten, and S. Heinecke. "Turning the Ship: The Transformation of DESY, 1993–2009". In: *Physics in Perspective* 19.4 (2017), pp. 424–451. URL: https://doi.org/10.1007/s00016-017-0209-4.

[126] M. Mulkay. "Conceptual Displacement and Migration in Science: A Prefatory Paper". In: *Science Studies* 4.3 (1974), pp. 205–234. URL: https://doi.org/10.1177/030631277400400301.

[127] C. Molho, J. Peña, M. Singh, and M. Derex. "Do institutions evolve like material technologies?" In: *Current Opinion in Psychology* 60 (2024), p. 101913. URL: http://dx.doi.org/10.1016/j.copsyc.2024.101913.

[128] P. Richerson, R. Baldini, A. V. Bell, K. Demps, et al. "Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence". In: *Behavioral and Brain Sciences* 39 (2014). URL: http://dx.doi.org/10.1017/S0140525X1400106X.

[129] L. Fogarty and N. Creanza. "The niche construction of cultural complexity: interactions between innovations, population size and the environment". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1735 (2017), p. 20160428. URL: http://dx.doi.org/10.1098/rstb.2016.0428.

[130] T. Yamagishi and H. Hashimoto. "Social niche construction". In: *Current Opinion in Psychology* 8 (2016), pp. 119–124. URL: http://dx.doi.org/10.1016/j.copsyc.2015.10.003.

[131] D. J. D. S. Price. *Little Science, Big Science*. Columbia University Press, 1963. URL: http://dx.doi.org/10.7312/pric91844.

Chapter 1

How research programs come apart: the example of supersymmetry and the disunity of physics

QSS

an open access journal

Check for updates

Citation: Gautheron, L., & Omodei, E. (2023). How research programs come apart: The example of supersymmetry and the disunity of physics. *Quantitative Science Studies*, *4*(3), 671–699. https://doi.org/10.1162/qss_a_00262

DOI: https://doi.org/10.1162/qss_a_00262

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00262

Supporting Information: https://doi.org/10.1162/qss_a_00262

Received: 10 February 2023
Accepted: 19 May 2023

Corresponding Author:
Lucas Gautheron
lucas.gautheron@gmail.com

Handling Editor:
Vincent Larivière

The MIT Press

RESEARCH ARTICLE

How research programs come apart: The example of supersymmetry and the disunity of physics

Lucas Gautheron[1,2] and **Elisa Omodei**[3]

[1]Interdisciplinary Centre for Science and Technology Studies (IZWT), University of Wuppertal, Wuppertal, Germany
[2]Sciences Po, médialab, Paris, France
[3]Department of Network and Data Science, Central European University, Vienna, Austria

Keywords: citation networks, high-energy physics, scientific pluralism, topic models, trading zones

ABSTRACT

According to Peter Galison, the coordination of different "subcultures" within a scientific field happens through local exchanges within "trading zones." In his view, the workability of such trading zones is not guaranteed, and science is not necessarily driven towards further integration. In this paper, we develop and apply quantitative methods (using semantic, authorship, and citation data from scientific literature), inspired by Galison's framework, to the case of the disunity of high-energy physics. We give prominence to supersymmetry, a concept that has given rise to several major but distinct research programs in the field, such as the formulation of a consistent theory of quantum gravity or the search for new particles. We show that "theory" and "phenomenology" in high-energy physics should be regarded as distinct theoretical subcultures, between which supersymmetry has helped sustain scientific "trades." However, as we demonstrate using a topic model, the phenomenological component of supersymmetry research has lost traction and the ability of supersymmetry to tie these subcultures together is now compromised. Our work supports that even fields with an initially strong sentiment of unity may eventually generate diverging research programs and demonstrates the fruitfulness of the notion of trading zones for informing quantitative approaches to scientific pluralism.

1. INTRODUCTION

This paper focuses on High-Energy Physics (HEP), the field of physics concerned with the fundamental entities of nature, and "supersymmetry," a symmetry between the two basic types of particles in nature. The idea of supersymmetry has brought together many of the most significant developments in the field throughout the past 50 years, all the way from the highly abstract world of string theorists, deep down to the machinery of underground particle colliders. However, none of the discoveries that supersymmetry promised have materialized as expected; as much as supersymmetry may be necessary to theorists seeking to unify the forces of nature into a coherent picture, it is increasingly plausible that it will not be of much use to experimentalists looking to find new particles. Throughout this case study, therefore, our work exhibits the disunity of science, by demonstrating that even scientific fields with a strong "sentiment" of unity, such as HEP (Wilson, 1986), can eventually fail to coordinate various research efforts. Our paper is guided by the idea that empirical case studies, although seemingly narrow in scope, do enrich our understanding of the nature of scientific enterprise (in this case, the nature of the coordination of diverse scientific cultures), and that quantitative studies of

science should provide conceptually informed tools for carrying out such case studies, preferably in ways that can be generalized for a variety of contexts.

We start by presenting Galison's notions of subcultures and trading zones, which is the framework for studying the plurality of science and the dynamics of interactions between scientific fields that underlies our investigation (Section 2.1). We will then provide the necessary background knowledge for understanding the context of our case study before laying out our hypotheses: that theory and phenomenology, over the historical period considered (1980–2020), are to be regarded as two distinct theoretical subcultures within HEP; that supersymmetry generated diverse research programs, some being phenomenological and some being more theoretical; and that supersymmetry significantly contributed to sustaining successful trades between theory and phenomenology until it was put in doubt by experimental data (Section 2.2). We then elaborate our motivation for addressing these hypotheses through quantitative methods (Section 2.3). Then, Section 3 details the quantitative methods that were deployed in order to address each of the three claims put forward in the introduction. It starts with a description of the data on which our analysis rests and how it was collected (Section 3.1). Section 3.2 elaborates quantitative methods for assessing the level of semantic and social autonomy of certain categories (subcultures), and applies these methods to the two theoretical subcultures in HEP. Section 3.3 elaborates a methodology based on topic models in order to address the "plasticity" and "plurality" of supersymmetry, which can in principle be applied to all "boundary objects," (i.e., those objects that can be traded between distinct subcultures while preserving and sustaining their distinctness). Finally, Section 3.4 provides a quantitative model for locating "trading zones" or more broadly concepts that enhance trades between subcultures (or scientific disciplines in general), and applies the model to the exchanges between the theoretical subcultures of HEP. Section 4 reveals and interprets the results of these analyses. Finally, Section 5 explores the consequences of this work, both for our case study (supersymmetry within HEP) and for the more general question of the plurality of science from a quantitative perspective.

2. BACKGROUND

2.1. Subcultures and Trading Zones: Galison's Approach to the Plurality of Science

If science is a unified enterprise, what is the nature of the relationship between fields as diverse as physics, biology, psychology, and economics? Can we translate all the concepts of these disciplines into a basic (say, physical) scientific language, as Carnap proposed? Or, are all these fields so incommensurable and autonomous that it is impossible to translate their respective entities, laws, and explanations from one's language to another's, as proponents of a pluralistic view defend (e.g., Cartwright, 1999; Dupré, 1983; Suppes, 1978)? Disciplines themselves can be so diverse, too, that the nature of what makes their own unity is not necessarily obvious. For instance, the nature of the unity of physics has been the matter of much debate, with sometimes serious political implications: Reductionist views (which imply that HEP is the most fundamental, because it supposedly entails any higher-level theory) were mobilized to justify the funding of large particle physics facilities (Cat, 1998), potentially to the detriment of more "useful" projects, as certain condensed matter physicists argued (Martin, 2018, Ch. 9). Instead, the latter argued that macroscopic systems have emergent properties that cannot be derived from "fundamental" laws. They were most often proponents of a "methodological" form of unity (Martin, 2018, p. 233), according to which the field is bound together by shared norms and conceptual tools (Cat, 1998, p. 267), rather than by relations of logical deduction from the most fundamental to the least fundamental theories. This view provided an

intellectual and philosophical basis for elevating the prestige of condensed matter physics (Martin, 2018, pp. 148–149), thus putting condensed matter and high-energy physicists on a more equal footing.

Even within the subfield of particle physics, there is a strong contrast between theorists and experimentalists. In fact, the nature of the relationship between the objects manipulated by, say, experimentalists (for instance, tracks within a cloud chamber, or electrical signals from a sensor) and the more abstract entities manipulated by theorists (e.g., "quarks," "gluons," or "strings") has been the subject of much philosophical debate. Inheriting a positivist view, some would grant experiment a more fundamental status, by defending its ability to provide robust empirical statements that could dictate theoretical change. Others, such as Kuhn, argued that empirical statements cannot be isolated from a theoretical paradigm and emphasized the "primacy" of theory (Galison, 1988)[1]. It is in order to overcome this debate about the relationship between experiment and theory within the context of physics that Galison originally developed his concepts of subcultures and trading zones (Galison, 1987, 1997). However, these notions may apply more generally whenever distinct scientific communities attempt to overcome difficulties to communicate and achieve coordination (Collins, Evans, & Gorman, 2010, p. 8). Consequently it is useful in a much broader range of contexts than the narrow case of physics; for instance, it is generally useful for studying the dynamics of interactions between disciplines in science[2]. Below, we propose a brief summary of the concepts of subcultures and trading zones and the rationale for their introduction.

The notion of subcultures was introduced by Galison (1987, 1988) to account for two characteristics of HEP: First, that it is subject to a strong division of labor, such that "theory," "experiment," and "instrumentation" are carried out by different groups of people (Galison, 1987, p. 138), with their own skill sets and bodies of knowledge; and second, that each of these "subcultures" is partially autonomous (i.e., none of them is completely subordinate to the others). We can highlight two tangible components of such subcultures: a social component—the community of practitioners—and a linguistic component—the language specific to each community.

For Galison, then, the question is what makes these subcultures part of a "larger culture" (physics), while retaining that their successful coordination is a "contingent matter" (Galison, 1997, p. 18); and his answer is "trading zones." Trading zones allow knowledge to be exchanged across different subcultures, inasmuch as the practitioners of distinct communities can locally agree on the usefulness of certain constructs despite the distinctiveness of their respective languages, commitments, aims, and methodologies. That trading occurs within "zones" captures the fact that the exchange procedure is "local" rather than "global," such that subcultures working out trades with each other can retain much of their autonomy in the process.

What kinds of goods may be subject to these "trades"? Examples of tradable goods are "boundary objects," that is, "objects that are both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a

[1] For instance, in his historical account of the discovery of quarks, Pickering (1984, p. 411) endorses the Kuhnian view: "To attempt to choose between old- and new-physics [gauge] theories on the basis of a common set of phenomena [experimental facts] was impossible: the theories were integral parts of different worlds, and they were incommensurable." Instead, Galison emphasizes the relative continuity and robustness of experimental "facts," across theoretical changes.

[2] For example, Kemman (2021) describes Digital History as a trading zone.

common identity across sites" (Star & Griesemer, 1989, p. 393)[3]. Trading zones may give rise to a purposefully crafted interlanguage that allows for further communication and coordination (a "pidgin"). If the interlanguage grows, it may turn into a full-blown language (a "creole"); this signals the emergence and stabilization of a new scientific discipline of its own.

Arguably, this is the process through which "phenomenology"—a subfield of HEP at the boundary between theory and experiment—has developed (Galison, 1997, p. 837). However, we may wonder whether phenomenology is still merely dedicated to bridging the gap between the theoretical and experimental cultures, or whether it acquired enough autonomy to depart from the supremacy of abstract theory (e.g., by relying on independent sources of inspiration for its own enterprise rather than by seeking to establish connections between high theory and experiment). In the following section we will suggest treating "theory" and "phenomenology" in HEP as two distinct subcultures, such that they may both enjoy considerable autonomy and eventually fail to coordinate their developments—thus extending the distinction made by Galison between theory, experiment, and instrumentation.

2.2. Supersymmetry Across Theory and Phenomenology

2.2.1. Theory and phenomenology as distinct subcultures within HEP

HEP involves a complex web of mathematical and technical knowledge concerning the details of the often abstract underlying theories, the behavior of the instruments that are assembled within sophisticated experiments, statistical notions for the analysis of the data derived from these experiments, etc. As a result of this complexity, there is a strong division of labor within HEP, and we can even distinguish two different groups within the theorists themselves. Although "pure" theorists (we will call them "theorists," in accordance with the terminology within the field) are driven by "the abstract elaboration of respectable theories," phenomenologists (the second kind of theorists) are often more concerned with "the application of less dignified models to the analysis of data and as a guide to further experiment" (Pickering, 1984), or at least more concerned with experimental consequences rather than with high theory. This division is itself strong enough that these two kinds of physicists can generally receive different training and diverge early in their careers, although some physicists—usually prominent ones—have expertise in both these domains and are able to sustain exchanges between the two. Therefore, in the present paper, we will make the following claim:

> **Claim 1**: Over the historical range considered (1980–2020), categories "theory" and "phenomenology" in HEP should be regarded as distinct subcultures with their own bodies of knowledge, ontologies, and methodologies, and are carried out by different people.

It is not controversial in itself that "theory" and "phenomenology" are different matters in HEP; these are now distinct categories within the HEP literature and it is not uncommon for physicists to label themselves as "theorists" or "phenomenologists" depending on their specialization. However, our claim goes further by stating that the nature of their work is *so* distinct that it should not be assumed *a priori* that they can sustain fruitful connections; per Galison, we should not expect *a priori* that subcultures are bound to cooperate flawlessly under any circumstance; we should instead remain open to the possibility that they may fail to produce

[3] In the context of physics, Darrigol's theoretical modules (Darrigol, 2007, p. 214), or multipurpose scientific instruments (Shinn & Ragouet, 2005, pp. 179–182), may be other examples of such tradable goods.

constructs of shared value within the contexts of their respective enterprises. There may not even be one single overarching goal that is equally shared and sought after by HEP theorists and phenomenologists, and it is even less certain that their respective methods should equally contribute to achieving their goals at any time[4]. In the following subsection, we will propose that supersymmetry exemplifies the contingent ability of high-energy physicists to coordinate their respective methods and goals in a successful way. It does so because the story of supersymmetry is that of a partial failure, rather than that of a total success. Although successful cases of cross-fertilization across fields are valuable to illustrate the notion of trading zones, that science (and even physics itself, as Galison claims, against a symbiotic view of theory and experiment) is disunified is better exemplified by those cases where scientific cultures attempt and fail to establish coordination. The dramatic story of supersymmetry provides such an example.

2.2.2. Supersymmetry as a tradable good between theory and phenomenology

Supersymmetry is a symmetry that relates the two fundamental kinds of particles that arise in nature: fermions and bosons. It was postulated simultaneously and independently by several physicists in the early 1970s, who were each motivated by very different goals[5]. Supersymmetry rapidly gathered substantial attention from the theoretical community. The reasons were manifold, but they were clearly theoretical rather than empirical, as early reviews of the topic show[6]. First, symmetry principles play a fundamental role in HEP, and supersymmetry was an especially attractive symmetry because of its peculiar properties. Second, supersymmetry can naturally give rise to gravity, as was observed by Volkov and Akulov (1973), suggesting that it could lead to a consistent theory of quantum gravity. This feature of supersymmetry gave birth to an entire research program, "supergravity," which then spanned several decades[7]. Third, although quantum field theory is prone to mathematical difficulties due to divergences appearing in the perturbative calculations of certain quantities, in many instances, such infinities were suppressed in supersymmetric theories.

However, as appealing as it was to theorists, supersymmetry posed a number of empirical difficulties. First, supersymmetry establishes a symmetry between bosons and fermions; and yet, at first it was not at all clear which of the bosons and fermions should have been related to each other by this symmetry. Moreover, if supersymmetry were perfectly realized in nature, the particles it relates should have identical masses, which was also in contradiction with the data. This contradictory situation was well summarized by Witten (1982) in his *Introduction to supersymmetry*:

> [Supersymmetry] is a fascinating mathematical structure, and a reasonable extension of current ideas, but plagued with phenomenological difficulties. [...] Supersymmetry is a very beautiful idea, but I think it is fair to say that no one knows what mysteries of nature (if any) it should explain.

[4] Galison (1995) provides a distinction between two kinds of theorists similar to the one we propose to make here, resting on the recognition that these two groups rely on very different sets of constraints as guides towards theoretical progress.

[5] For a history of early supersymmetry, see Kane and Shifman (2000).

[6] Fayet and Ferrara (1977), Freedman (1979), and Taylor (1984) provide a good overview of the main arguments for supersymmetry in its early days, all of which are highly theoretical.

[7] Later on, supersymmetry proved even more interesting to theorists, by improving the consistency of string theory, and by supporting the conjectured AdS/CFT correspondence, yet another major development in quantum gravity research.

Still, efforts to incorporate supersymmetry into a theory consistent with the data were undertaken over several years, and they culminated in what is now called the Minimal Supersymmetric Standard Model (MSSM) (Dimopoulos & Georgi, 1981; Fayet & Ferrara, 1977). The MSSM is the result of reconciling the achievements of the Standard Model of Particle Physics (SM) (the best theoretical account available at the time and still today) with the requirement of supersymmetry. This, however, has very undesirable consequences. Compared to the SM, the MSSM introduces 105 additional unspecified parameters, so that supersymmetry can accommodate a large range of observations, and has little predictive power in general (Parker, 1999, p. 1). In particular, although supersymmetry predicts the existence of many new particles (the "superpartners"), there is *a priori* little chance that these particles will have just the right properties to be discoverable in experiments. If not, supersymmetry may be of high value to theorists (because of its mathematical properties, and its promise to achieve a coherent account of quantum gravity), while being of low value to phenomenologists who are interested in building predictive models that can lead to the discovery of new particles or phenomena[8].

Yet, in 2011, supersymmetry was perceived across the field as the theory beyond the SM that was most likely to manifest itself in experiments (Mättig & Stöltzner, 2019, 2020). Arguably, the reason why it became highly credible and valuable to phenomenologists as well was that it could solve the so-called "naturalness" problem of the standard model on the condition that it was discoverable. In parallel to these developments around supersymmetry, there was indeed increasing recognition that an explanation was required as to why the mass of the Higgs boson (an important piece of the SM) could be many orders of magnitude below the mass scale at which the unification of forces is assumed to take place. It was also realized that supersymmetry could provide an answer to this "naturalness" problem (Veltman, 1981; Weinberg, 1979; Witten, 1982), *but* only as long as the masses of the superpartners (the particles predicted by supersymmetry) are not too high, so that they should be discoverable in future experiments[9]. In light of this, supersymmetry became of very high value to phenomenologists and experimentalists as well, rather than just a mathematical toy for the theorists to play with[10].

This situation is summarized in Figure 1. As theorists work out a path towards their goals (e.g., the unification of forces, or the formulation of a consistent theory of quantum gravity), they rely on theoretical heuristics such as renormalizability, symmetry principles, and consistency requirements. (Galison, 1995). In that context, supersymmetry emerges as a very valuable concept. Phenomenologists, on the other hand, try to work out a path towards the discovery of "new physics" (evidence for new phenomena unaccounted for by the SM) by

[8] Supersymmetry suffers from other disadvantages. For instance, many parameters of the theory imply certain phenomena to extents that have not been observed, such as baryon and lepton number violation, or flavor-changing neutral currents (Weinberg, 1995, pp. 201–209, 235–240), which requires *ad hoc* explanations as to why, although allowed by the model, these mechanisms do not occur in nature.

[9] One can put other constraints on the MSSM, by requiring that supersymmetry explains dark matter, or that it ensures the convergence of the "couplings" that measure the strength of the fundamental forces at different length scales, which suggests it should play a role in the unification of these forces. However, as Giudice and Romanino (2004) put it, "the unification and dark-matter arguments [for supersymmetry] are not in general sufficient to insure that new physics be within the LHC discovery reach, contrary to the naturalness criterion."

[10] The naturalness argument also provides a "narrative" that connects what theorists are concerned with (the details of the theories at energy scales unattainable in the experiment) to what experimentalists can probe. As Borrelli (2015, p. 76) puts it, "the strength of the naturalness narrative is largely due to its flexibility, which allows it to become a unifying factor in the high-energy community and to bridge the gap between theorists and experimenters."

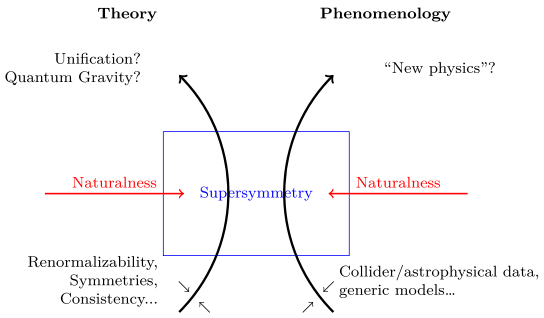


Figure 1. Supersymmetry in the trading zone between theory and phenomenology. Theorists and phenomenologists have different aims and methodologies, and whether they can both positively appraise a particular construct is not guaranteed. In the case of supersymmetry, it is the naturalness requirement that ensures that the MSSM is so valuable to both subcultures. As a result, supersymmetry enhances a trading zone between these two cultures.

relying instead on more generic models and constraints derived from experimental data (e.g., from particle colliders or astrophysical observations). It is the naturalness requirement that makes supersymmetry valuable to phenomenologists as well, by strengthening the belief that supersymmetric particles should have masses that are low enough to be discoverable. In this way, supersymmetry effectively enhances the "trading zone" between theorists and phenomenologists: Both communities can acknowledge its value in spite of the vast differences in their aims, methods, and objects of inquiry.

It is now time to introduce the last (but not the least) player in our drama: the Large Hadron Collider (LHC). Operating since 2010, the LHC is the largest physics experiment ever built. By performing particle collisions at the highest energies ever achieved, it promised to discover supersymmetric particles, provided that they had the properties prescribed by the naturalness problem that supersymmetry should solve. However, no such discovery has been made, which suggests that the "naturalness problem" was unwarranted (Giudice, 2018). If there is no naturalness problem, then, supersymmetry is left unconstrained again; there is no guarantee that supersymmetric particles will ever be discovered; and its phenomenological value plunges back to the depths from which it surfaced. Therefore we will put forward the following claim, which will also be evaluated in the present paper:

> **Claim 2**: Supersymmetry occurs in a variety of partially independent contexts within HEP, some of which belong to "theory" and some of which belong to "phenomenology," and these applications of supersymmetry have responded differently to the LHC's failure to find supersymmetric particles.

Furthermore, we hypothesized that supersymmetry should be losing its ability to sustain trades between theory and phenomenology. Therefore, we will evaluate the following claim:

> **Claim 3**: Supersymmetry sustained trades between theory and phenomenology in HEP until it was challenged by the LHC's failure to observe the particles predicted by supersymmetry.

If theorists and phenomenologists fail to share a similar appraisal of supersymmetry, then this may pose a serious problem for the field: This would imply that theorists' research programs can persist despite their low value to phenomenologists, and conversely that

experimental input has little to offer to theorists; if that is the case, then the unity of HEP would indeed be fragilized. Therefore, addressing claims 1–3 (1—that theory and phenomenology are partially autonomous subcultures of HEP; 2—that supersymmetry arises in distinct, autonomous contexts, which responded differently to the absence of supersymmetric particles at the LHC; and 3—that the value of supersymmetry for bridging together subcultures of physics has decreased as a result of the failure of phenomenological supersymmetry) should contribute to answering the questions of what makes and unmakes unity in HEP.

2.3. Towards a Quantitative Assessment of Subcultures and Trades

In the following, we propose an array of quantitative methods implementing several dimensions of Galison's framework for addressing the plurality of science, which evaluate the claims put forward above. To this end, we will rely on authorship data (for investigating the social entrenchment of theory and phenomenology as distinct subcultures), semantic analyses (for investigating the linguistic divide between these subcultures as well as the plurality of supersymmetry research), and citation data (in order to locate "trading zones" within the field). To our knowledge, this is the first attempt to implement Galison's framework into a quantitative analysis of scientific literature. Of course, the plurality of science and the coordination between scientific fields have already been addressed quantitatively in numerous publications. In the context of physics research, for instance, Battiston, Musciotto et al. (2019) have evaluated the ability of physicists to publish in various subfields. In particular, they demonstrate that high-energy physicists are among the most specialized physicists (i.e., they have a high probability of publishing only in their primary subfield), although their work does not distinguish between the various kinds of high-energy physicists, which will be done in the present paper. There remains to address the linguistic component of the divide between these subcultures, in particular theory and phenomenology, and to this end we will propose a novel strategy based on semantic data (titles and abstracts of the literature).

As for the analysis of the plurality of supersymmetry-related research in HEP, we will develop a topic model approach in order to identify clusters of concepts that are most likely to be associated with supersymmetry in the literature, and we will explore the dynamics of supersymmetry research throughout time.

Finally, we will assess the intensity of trades between theoretical subcultures and locate the concepts that facilitate these trades. Yan, Ding et al. (2013) proposed a quantitative assessment of dependency relations between scientific disciplines based around a metaphor with international trade, by measuring quantities such as "exports," "imports," or "self-dependence" of various fields throughout time based on citation data. However, this work does not investigate what exactly allows these trades to happen (e.g., which concepts sustain them). This requires combining citation data with semantic information about papers' concepts, as achieved by Raimbault (2019), who proposed measures of interdisciplinarity built upon such data. Similarly to Yan et al. (2013), we will assess the self-dependence of experiment, phenomenology, and theory in HEP based on the citation network. However, we will also evaluate the ability of different concepts (such as supersymmetry) to sustain trades across subcultures throughout time by combining semantic and citation data.

More broadly, this work will add to quantitative studies of science literature, by helping to fill a gap that has come to the attention of the community. As stressed by Leydesdorff, Ràfols, and Milojević (2020), Kang and Evans (2020), and Bowker (2020), quantitative and qualitative studies of science have mostly diverged in their goals and "world views," urging the need to

"bridge the gap" between them. We propose, therefore, a bridge connecting these two forms of scientific study. First, we demonstrate that quantitative methods can address questions raised by the philosophy, history, and sociology of physics. Moreover, we show that concepts from qualitative science studies can give structure to quantitative methods, in line with the call by Heinze and Jappe (2020) to inform quantitative analyses with "middle-range theories" (of which Galison's trading zones are an example). As a result, our methods are in principle meaningful in any context where such a theory is valid—whenever scientific cultures attempt to achieve coordination—well beyond the case study proposed in this paper.

3. METHODS

3.1. Data

Our data consist of the scientific literature on HEP and the semantic, authorship, and citation information that it entails, which is of interest for our questions.

The data were retrieved using the Inspire HEP database (Moskovic, 2021). Inspire HEP is a platform dedicated to the HEP community and is maintained by organizations that include CERN, DESY, Fermilab, and SLAC. It aggregates publications from the HEP literature, and maintains a list of institutions and collaborations involved in the community, while also publishing job offers. It replaced Spires in 2012[11].

The database is fed by an automatic aggregator that retrieves articles from multiple sources[12] including a number of databases (Astrophysics Data System, arXiv, etc), research institutions (CERN, DESY, Fermilab, IHEP, IN2P3, SLAC), and scientific editors, such as the American Physics Society or Springer.

Inspire then aggregates data from these sources with automated crawlers, and it performs manual curation for completion or error-correction[13], including author name disambiguation. This database has a strong yet untapped potential for quantitative analyses. However, only contents related to HEP are subject to a systematic effort of collection and curation, and the data should be used preferably in analyses whose scope is limited to HEP, thereby making it unsuitable for studying interactions between HEP and other fields of physics (e.g., condensed matter physics).

The database includes data about the contents of the literature (title, summary, sometimes keywords), the authors (name, unique identifier, institutional affiliations), dates corresponding to different events related to each paper, associated experiments, and references of the articles. The only data pertaining to the contents of the articles that are consistently available and that we have used in the present paper are titles and abstracts. Articles are categorized according to a classification scheme compatible with that of the arXiv preprint platform. This scheme includes categories such as `Theory-HEP`, `Experiment-HEP`, `Phenomenology-HEP`, and `Astrophysics`. Categories of papers published on arXiv.org are extracted directly from the platform (where they are defined by the authors, while being subject to moderation and controls). Categories of papers not published on arXiv.org are now assigned manually by

[11] "Physicists, start your searches: INSPIRE database now online," *Symmetry*, May 24, 2012, https://www.symmetrymagazine.org/breaking/2012/05/24/physicists-start-your-searches-inspire-database-now-online.

[12] Melissa Clegg, "INSPIRE Content Sources," May 30, 2020, https://help.inspirehep.net/knowledge-base/inspire-content-sources/.

[13] Stella Christodoulaki, "Content Policy," March 4, 2020, https://help.inspirehep.net/knowledge-base/content-policy/.

curators. Categories of papers inherited from the ancestor of Inspire HEP (Spires) and absent from arXiv.org were derived according to a mapping between Spires' classification and the current arXiv-based classification. In this paper, we rely mostly on three categories that entail most of the HEP literature: `Theory-HEP`, `Phenomenology-HEP`, and also `Experiment-HEP`, which typically entails papers that report empirical results such as statistical analyses of experimental data. A portion of the articles between the years 1990 to 1995 was not categorized, which led to some issues with the data collection process, as described in the Supplementary materials. For this reason, our longitudinal analyses will focus on later years, which does not prevent us from addressing our research questions. The analysis of subcultures spans the years 1980 to 2020. The years prior to 1980 could also have been interesting for this analysis, but the corresponding data was of lower quality.

3.2. Social and Semantic Analysis of Subcultures of HEP

The first claim that we seek to establish is that "theory" and "phenomenology" should both be regarded as distinct subcultures within physics. There are two components to subcultures: a linguistic one (they should have vocabularies that are distinct enough to signal complementary bodies of knowledge) and a social one (they should correspond to distinct groups of people). Therefore we will proceed twofold. First, we will demonstrate that theory and phenomenology manipulate vocabularies that are so distinct that we can predict with reasonable accuracy whether a paper belongs to one of these categories based on the words present in its abstract; our predictive model will then be used to unveil the ontological differences between these subcultures. Second, we will show that these categories from the literature are associated with different communities.

3.2.1. The semantic divide between theory and phenomenology

If it is possible to tell whether a paper is theoretical or phenomenological based on the words it contains, then this implies that these categories use partially distinct vocabularies (i.e., that each of these two categories has its own "language") in a way that allows papers from one category to be distinguished from those from another. If that is the case, we can then examine the nature of the linguistic divide between theory and phenomenology in order to better understand their differences. In what follows, we apply this strategy using statistical methods, based on the classification of HEP literature provided by Inspire HEP[14]. Although we are more interested in the divide between "theory" and "phenomenology," we also include "experiment" (which Galison himself labeled as a subculture of its own) in our analysis in order to emphasize its differences with phenomenology.

To establish whether we can predict which articles d belong to any of the categories $c \in$ {Experiment, Phenomenology, Theory}, we will build a simple linear logistic regression using a bag-of-words as the predictive features. In this approach, the corpus is represented by a matrix $B = (b_{d,i}) \in \mathbb{R}^{D \times V}$, where D is the number of documents, V is the size of the vocabulary, and $b_{d,i}$ is the number of occurrences of the word (or expression[15]) i in the document d. This representation excludes a lot of semantic information that results from the knowledge of the ordering of the words and the structure of sentences within the documents; it is in line with our goal to find out whether the vocabularies of each category are so distinct that the mere presence or

[14] As this classification relies on a manual assignment of the different categories, any potential linguistic divide between them cannot be the byproduct of some algorithmic bias.

[15] We also include some n-grams in the model (i.e., expressions of several words), provided they follow certain predefined syntactic patterns (e.g., "adjective + noun").

absence of certain words can be used to infer the category of a document. We perform a normalization of the bag-of-words prior to the regression by applying the tf-idf transformation[16] to ($b_{d,i}$), resulting in a normalized bag-of-words which we will name ($b'_{d,i}$). More specifically, our predictive model is defined as:

$$P(d \in c) = \text{logit}^{-1}\left(\beta_c + \sum_{i=1}^{V} \beta_{ci} b'_{di}\right) \quad (1)$$

This model is then trained on $N = 100{,}000$ articles of our database from 1980 to 2020 that belong to any of the following categories: `Experiment-HEP`, `Phenomenology-HEP`, and `Theory-HEP`[17]. The vocabulary used in the regression is the V expressions (n-grams, up to four words long) among those that belong to predefined syntactic patterns[18], that have the highest "unithood" as measured in Omodei (2014)[19]. The size of the vocabulary V is chosen to be a round number that is just high enough to reach about the maximum accuracy of the model, as evaluated on the test set (which consists of 10,000 articles not present in the training set). The accuracy of the predictions of the model is evaluated using the same test set. The coefficients β_{ci} are then analyzed to extract the words that are the most discriminatory between "theory" and "phenomenology," thus revealing the most salient differences. For that, we retrieve those expressions i that maximize $\beta_{\text{th},i} - \beta_{\text{ph},i}$ and $\beta_{\text{ph},i} - \beta_{\text{th},i}$. Because of the inverse document frequency transformation applied prior to the regression, expressions that are more common are favored by this selection process.

3.2.2. The social divide between theory and phenomenology

What does it mean to say that theory and phenomenology have a "demographic component," as Galison (1987, p. 138) puts it, regarding theory and experiment in HEP? It means that these categories of the literature are supplied by distinct groups of people: "theorists" and "phenomenologists." Therefore, we will investigate whether it is the case that experimental, phenomenological, and theoretical papers are published by three distinct groups of physicists, such that these physicists usually contribute mostly to just one of these categories. Again, "experiment," which is a paradigmatic example of subculture in Galison's view, is also included in our analysis. It will be useful to assess whether the distinction between phenomenology and theory is comparable to the distinction between theory and experiment (the one initially stressed by Galison).

Let N_{ij} be the number of articles coauthored by a physicist i that belong to the category $j \in$ {theory, phenomenology, experiment}, and N_i the total number of articles coauthored by i. Let us assume $N_{ij} \sim \text{Binomial}(N_i, p_{ij})$, where p_{ij} is the latent probability that a paper from physicist i belongs to the category j[20]. Because the researchers coauthored widely varying amounts of

[16] For a definition of the tf-idf transformation, and information theoretic justifications of its relevance, see Beel, Gipp et al. (2015) and Robertson (2004). We use scikit-learn's implementation of the inverse-document frequency transformation, which is idf = 1 + log(1/f), where f is the fraction of documents in which a word occurs. It differs from the "textbook" definition log(1/f) because of the regularization term (+1).

[17] The fit is performed with the scikit-learn python library (Pedregosa, Varoquaux et al., 2011) using L2 regularization.

[18] We choose a subset of the syntactic patterns used to analyze the Association for Computational Linguistics Anthology Corpus in Omodei (2014).

[19] The "unithood" measures "the degree of strength or stability of syntagmatic combinations or collocations" (Kageura & Umino, 1996).

[20] As these categories are not mutually exclusive in our database (an article may belong to more than one of them), a multinomial process would not be a good fit.

publications (ranging from a few papers to hundreds), we assumed that the latent probabilities p_{ij} were described by the following model:

$$N_{ij} \sim \text{Binomial}(N_i, p_{ij})$$
$$p_{ij} \sim \text{Beta}(\alpha_j, \beta_j)$$
$$\alpha_j, \beta_j \sim \text{Exponential}(1)$$

The binomial process assumes that each physicist can be imputed a constant latent fraction of papers in each category. The beta prior is a flexible distribution over probabilities, which can be either unimodal or bimodal. The exponential prior over α and β is agnostic regarding these two possibilities, and its exact shape does not significantly matter, considering the amount of available data. Most crucially for us, this model allows us to combine information from researchers with many papers and researchers with very few papers; for those with few papers, the estimation of the latent probabilities is more influenced by the shape of the beta distribution. The model was fitted to 2,500 researchers randomly sampled among those with more than three publications in HEP for 1980–2020. In order to evaluate the social entrenchment of these categories, we verify that most physicists contribute mostly to just one of these categories.

3.3. Assessing the Plurality of Supersymmetry Research with Topic Models

Our second claim pertains to the plurality of supersymmetry research. In this section, we present our methodology for assessing the plurality of supersymmetry-related research, by recovering the contexts (i.e., the topics in which supersymmetry occurs) and by evaluating the extent of their independence and how they responded to the results of the LHC. More broadly we provide a methodology for investigating scientific "objects" akin to "boundary objects" in that they are "plastic enough to adapt to local needs and constraints of the several parties employing them" (Star & Griesemer, 1989, p. 393), by unveiling the plurality and autonomy of the contexts in which such objects may arise.

3.3.1. Model

To evaluate in which contexts supersymmetry arises within the HEP literature, we have chosen to subdivide the literature into subtopics using an unsupervised probabilistic topic model, namely the Correlated Topic Model (CTM; Blei & Lafferty, 2007). We do not use conventional classifications such as the Physics and Astronomy Classification Scheme (PACS) codes from the American Institute of Physics (AIP), because they were not available for the whole data set – PACS codes were only available starting from 1995, and only for a subset of the papers, which may not be representative of the whole. Besides, PACS codes are too numerous (more than 5,000 categories)[21] for our purposes. Therefore, we opted to extract the topics in the literature using unsupervised topic models instead.

Probabilistic topic models generally assume that each document of a corpus is a mixture of variable proportions of a certain amount of topics, each of these topics having their own vocabulary distribution. When trained on a corpus, such models simultaneously learn the "topics" in the corpus (and their vocabulary), as well as the relative contribution of each topic to each document of the corpus. These models have demonstrated their ability to capture the semantic information contained within the scientific and academic literature, as shown in

[21] "Full list of PACS numbers," *Physics-Uspekhi*, https://ufn.ru/en/pacs/all/.

previous work[22], even from abstracts alone (Syed & Spruit, 2017); as a result, this technique has seemingly taken precedence over network-based semantic maps (Leydesdorff & Nerghes, 2016, Figure 1). Although co-occurrence networks may have more conceptual bearing in the STS tradition, we have preferred topic models for their intrinsic ability to capture the polysemy of certain words (e.g., "supersymmetry"), in terms of the probabilities that such words can arise in different contexts (i.e., topics).

In particular, we have chosen the CTM for its ability to capture correlations between topics. In this model, the contribution of a topic z to a document d, $P(z|d)$, is assumed to be drawn from a hierarchical model involving a correlated multivariate distribution (Blei & Lafferty, 2007):

$$\vec{\beta}_d \sim \mathcal{N}\left(\vec{\mu}, \sum\right) \tag{2}$$

$$P(z|d) = \frac{\exp \beta_{d,z}}{\sum_{i=1}^{k} \exp \beta_{d,i}} \tag{3}$$

Through the covariance matrix $\sum$, the CTM is able to learn correlations between topics, and therefore to account for the fact that some topics are more likely to occur together within one document. Moreover, our intuition is that using the CTM allows the derivation of a more realistic topic distribution for short texts such as abstracts, for which the small numbers of words only moderately inform the prior topic distribution. Most importantly, this model allows us to directly assess the level of independence between the topics derived by the model, which is important for assessing the autonomy of the contexts in which supersymmetry arises.

The model is trained on N = 120,000 articles randomly sampled from those between 1980–2020 that belong to any of the categories `Theory-HEP`, `Phenomenology-HEP`, `Experiment-HEP`, and also `Lattice` (a theoretical approach to HEP, with ties to both theory and phenomenology, and in which we expected supersymmetry to potentially arise as well). The procedures for extracting the input vocabulary and for choosing the hyper-parameters are described in detail in Sections S3.1 and S3.2, respectively, of the Supplementary material. Two methodological contributions can be highlighted. First, we included informative n-grams matching predefined syntactic patterns in the vocabulary in order to preserve more semantic information. Second, we made a prudent and balanced use of perplexity and topic coherence measures in order to recognize the advantages and limitations of both these kinds of measures for assessing the quality of topic models and choosing the best hyperparameters. The procedure resulted in the extraction of 75 topics (see Section 3.4 of the Supplementary material for the full list).

3.3.2. Interpretation and validation

Once the model was trained, we manually assigned a label to each topic, by inspecting and interpreting their top-words and the categories from the PACS classification of the physics literature that were most correlated to each topic[23]. Informing our interpretation of each topic

[22] Notable examples are Griffiths and Steyvers (2004), Hall, Jurafsky, and Manning (2008), and Nichols (2014); see Malaterre, Chartier, and Pulizzotto (2022) for a more recent application in the context of History and Philosophy of Science, and Allen and Murdock (2022) for an assessment of the potential and limitations of these methods in the field.

[23] We used pointwise mutual information (see Eq. 1 in Section S3.3) measure of correlation.

with these correlations rather than the sole top-words help overcome issues associated with the interpretation of fat-tailed topic-word distributions based on a handful of top-words (Allen & Murdock, 2022; Chang, Boyd-Graber et al., 2009). We failed to provide a meaningful label for some topics, but this had little impact on the rest of the analysis. Finally, in order to assess the meaningfulness of the metrics produced by the model (the document-topic distributions and the topic-word distributions), we performed an additional validation procedure using the PACS classification of the literature and the input of independent experts (see Supplementary material, Section S3.3).

In Section 4.2, the model is applied to a number of tasks: the evaluation of the contexts (i.e., topics) in which supersymmetry occurs in the literature, the extent of the correlation between these contexts, and finally the trends in research involving supersymmetry since the start of the LHC.

3.4. Locating Trades Across Scientific Cultures

In this section, we elaborate a longitudinal methodology for locating trades between scientific cultures, which we use to assess the ability of supersymmetry to enhance trades between the theoretical and phenomenological cultures of HEP throughout time. Trading zones can manifest themselves in a myriad of ways, some of which are readily prone to a quantitative analysis. For instance, citing the example of quantum chromodynamics, a theory of the strong interaction, Galison notes that "the contact between the experimenters and the phenomenological theorists had grown to the point where Andersson [a theorist] and Hofmann [an experimentalist] could coauthor a *Physics Letter*" (Galison, 1997, p. 655). In that sense, a paper coauthored by scientists from different cultures is indicative of a trading zone, such that coauthorship data can in principle be used to probe trades across scientific cultures. Another manifestation of trading zones can be found in the citation network, which encodes exchanges of knowledge across publications, and sometimes across subcultures. Indeed, that a phenomenological publication, for instance, cites a theoretical paper indicates that phenomenologists can acknowledge the value and significance of certain theoretical constructs (that are present in this specific paper) in their enterprise. Although in principle both the citation networks and the collaboration networks could be used for our purpose, the present analysis will rely on the former. Indeed, the citation graph preserves more information about the directionality of the exchanges involved, thus supporting the trade metaphor in Yan et al. (2013). Intuitively, it is also less vulnerable to nonepistemic factors, as is the case with authorship (e.g., physicists authoring papers they did not contribute to, as is frequent in large collaborations in the field). In addition, for validation purposes, we show in the Supplementary material (Section S4 and Figure S3) that the citation network can indeed reveal the relative autonomy (self-reliance) of HEP subcultures but also the special role of the phenomenological subculture in sustaining the unity of HEP by channeling trades across theory and experiment (which hardly communicate directly otherwise). This further supports the use of the citation graph use as a means of locating trades.

To assess the ability of supersymmetry to facilitate trades between theorists and phenomenologists, we develop a method that combines two important aspects of Galison's trading zones: their locality and their linguistic component (the "interlanguage"). In particular, we look for scientific concepts that are most likely to be involved in trades between these subcultures throughout time. To this effect, we perform the analysis on a subset of the citation graph, such that the nodes are limited to theoretical and phenomenological papers, excluding cross-listed papers (those that belong to both these categories). For each of these two theoretical

cultures, we derive a list of informative keywords from the abstracts of the papers by extracting n-grams ($n \geq 2$) matching certain syntactic patterns. We retain the top N keywords (sorted by decreasing unithood) such that at least 95% of the abstracts contain at least one of the N keywords; this yields $N = 1{,}370$ keywords specific to the phenomenological culture and $N = 1{,}770$ keywords specific to the theoretical culture. From this we derive a bag of words b_{ik} for each publication such that $b_{ik} = 1$ if keyword k is present in abstract i, and $b_{ik} = 0$ otherwise. We then evaluate the probability that the keyword occurs in an abstract given that the paper is involved in a trade between a theoretical and a phenomenological paper at a time t, which we write $P(b_k = 1|\text{trade}_{i\to j}, t)$. We consider trades in both directions: phenomenological papers citing theoretical papers (th→ph), then in a second process theoretical papers citing phenomenological papers (ph→th). To what extent supersymmetry helps sustain the trading zone between these theoretical cultures is roughly measured by $P(b_k = 1|\text{trade}, t)$ for those keywords k related to supersymmetry. In this analysis, we explore 3.7 million citations appearing in papers published between $t = 2001$ and $t = 2019$ (covering similar ranges prior and after the start of the LHC). We included all cited papers from 1980 onwards (180,000 total)[24]. However, because cross-listed papers, which we excluded, have become much more common in the database starting from 2010 for spurious reasons (a change in the classification procedure), we ran a separate analysis to assess the robustness of our results. In this second analysis, we included cross-listed papers and assigned them only one category based on their authors' primary subfield (the subfield to which they contribute the most). We found both analyses to produce similar results. In the following we report the results obtained by excluding cross-listed papers.

4. RESULTS

4.1. Theory and Phenomenology as Distinct Subcultures

Let us now examine our first claim that "theory" and "phenomenology" should be regarded as distinct subcultures within HEP. The claim requires that these categories mobilize distinct bodies of knowledge that manifest themselves through distinct vocabularies. As shown in Table 1, it is indeed possible to predict with reasonable accuracy whether a paper belongs to either one of these categories based on the vocabulary in its abstract. The accuracy is higher than 90% for "theory" and reaches 86% for "phenomenology," far above what one would obtain from assigning the most probable class irrespective of the contents, purely based on their frequency. This conclusion holds throughout the whole historical period considered (see Sections S2 in the Supplementary material). This supports the existence of a linguistic divide between these two theoretical cultures over the years 1980 to 2020.

Our model also unveils the expressions that are most capable of discriminating between theory and phenomenology, as shown in Table 2. One striking difference between theory and phenomenology appears to be the importance of space-time related concepts in theory ("space-time," "geometry," "manifold," "dimension," "coordinate," etc.). The objects (entities) of interest also differ, which signals an ontological divergence: On the pure theory side, "black hole[s]" and "strings" are prominent entities, whereas particles ("quark," "neutrino," "gluon," "hadron," "nucleon," etc.) belong to the realm of phenomenology. Among those terms most specific to phenomenology but absent in pure theory, we also find the notions of model ("mssm," "standard model"), and effective field theories ("effective theory," "chiral

[24] It is unlikely that recent papers would cite publications from before 1980.

Table 1. Accuracy of the model for predicting which categories HEP papers belong to. The precision of the model for each category is estimated based on the test corpus. For reference, the accuracy of a naive model that assigns the most likely class irrespective of any information about the papers is given (baseline). The size of the vocabulary used for the predictions is set to 500 words and expressions

	Theory	Phenomenology	Experiment
Model accuracy	91%	86%	92%
Baseline	55%	51%	84%

perturbation theory") which are approximate theories emerging from more fundamental theories. Moreover, the mention of "experimental data" is a distinctive feature of phenomenology: Theory is not directly committed to establishing a connection with empirical results. Interestingly, one aspect of supersymmetry (the MSSM) appears as markedly phenomenological, whereas "supergravity" is specifically theoretical.

Similarly, keywords that discriminate the most between experiment and phenomenology are shown in Table 3. They confirm the theoretical ("model," "scenario," "effective theory," "implication") and computational ("estimate," "approximation," "contribution," "numerical result," "correction") nature of phenomenology, as opposed to the empirical, "fact-based" dimension of experiment ("measurement," "search," "experiment," "event," "result," "evidence," "data").

What about the "demographic component" of the divide between theory and phenomenology? Do these categories have social counterparts? The results of our social analysis are shown in Figure 2. Figure 2 is a ternary diagram in which each red dot represents a physicist and is positioned according to the relative prevalence of each category (among experiment, phenomenology and theory) among the papers they authored or coauthored. The majority of the dots are clustered near vertices, which means that most physicists dedicate themselves to mostly one of these categories. In particular, the inner part of the ternary diagram, which corresponds to physicists with balanced contributions to each category, is almost empty. We do find that some authors are scattered along the experiment-phenomenology edge and the phenomenology-theory edge; still, our results suggest that the category of phenomenology does feature a "demographic" counterpart as well, although it is more porous than experiment or pure theory. Therefore, phenomenologists do, to some extent, constitute a social group distinct from that of theorists (and experimentalists); however, phenomenology seems to play a special role in sustaining some form of cooperation between experimentalists and theorists.

Table 2. Vocabulary specific to phenomenology (left column) versus theory (right column)

Vocabulary specific to phenomenology	Vocabulary specific to theory
quark, lhc, qcd, neutrino, experimental data, mssm, dark matter, extra dimension, parton, phenomenology, gluon, color, mixing, standard model, electroweak, collider, nucleon, effective theory, sensitivity, new physic, high energy, hadron, chiral perturbation theory, next-to-leading order, impact, neutrino mass, resonance, signal, process, accuracy, collaboration, distribution, flavor, decay, effective field theory, determination, violation, evolution, account, meson, element, baryon, higgs, contribution, gamma	algebra, manifold, geometry, space time, partition, modulus space, gravity, theory, branes, correspondence, central charge, deformation, action, chern-simons, duality, string, horizon, supergravity, ad, quantum, space-time, yang-mills, coordinate, entropy, conformal field theory, sitter, field, construction, surface, dimension, boundary, transformation, black hole, solution, mechanic, space, conjecture, type, class, quantization, dirac, formulation, background, connection, massless

Table 3. Vocabulary specific to phenomenology (left column) versus experiment (right column)

Vocabulary specific to phenomenology	Vocabulary specific to experiment
experimental data, qcd, mssm, quark, dark matter, lhc, color, phenomenology, gluon, plasma, new physic, heavy ion collision, account, inflation, parton, evolution, high energy, factorization, effect, implication, scenario, potential, approach, contribution, electroweak, process, mixing, model, estimate, numerical result, accuracy, integral, approximation, neutrino, unification, higgs, possibility, bound, neutrino mass, calculation, case, early universe, sensitivity, generator, extra dimension	detector, sample, measurement, search, upper limit, confidence, experiment, atlas, target, cm, luminosity, event, proton-proton collision, evidence, resolution, fraction, result, gev, first time, beam, expectation, yield, tev, world, top quark, branching, range, technique, muon, limit, study, construction, data, reaction, recent result, mev, system, investigation, section, paper, observation, respect, differential cross section, electron, experimental result

Overall, we find that 81% of high-energy physicists publish more than 80% of their papers in just one of these categories, which is clear evidence of specialization.

Our quantitative analysis supports our claim that, at least over the years 1980 to 2020, theory and phenomenology should be regarded as distinct subcultures with partially distinct languages. Consequently, strategies ought to be devised for them to properly communicate and coordinate their efforts, as long as physicists believe it to be necessary or worthwhile. It follows that their unity should not be assumed; instead, why a trading zone may be successfully worked out remains to be explained. Before we turn to the ability of supersymmetry to sustain the coordination between these subcultures, we will address the plurality of supersymmetry research itself.

4.2. The Plurality of Supersymmetry

In this section, we apply our methods to address our second claim regarding the plurality of supersymmetry research and the recent decline in phenomenological supersymmetry research as a response to LHC results.

Topic models are able to link one word to several topics, thus allowing us to unveil different aspects of supersymmetry (i.e., different contexts[25] in which this concept may occur). For three words w that explicitly refer to supersymmetry ("supersymmetry," "supersymmetric," "susy"[26]), we evaluated the probability $P(z|w)$ that these words occur in the context of a topic z according to

$$P(z|w) = \frac{P(w|z)P(z)}{P(w)} \tag{4}$$

where $P(w|z)$ is frequency of the term w within the topic z, $P(z)$ is the marginal probability of topic z, and $P(w)$ is the overall term-frequency of w. The five most probable topics for each ofthe words "supersymmetry," "supersymmetric," and "susy" are shown in Figure 3 (for the other topics, the probability $P(z|w)$ for $w \in$ {"supersymmetry," "supersymmetric," "susy"} is residual). We can see that each of these terms may indeed occur in relation to a variety of topics: "supersymmetric theories" (which entail supersymmetry in string theory, or

[25] Like Allen and Murdock (2022), we caution that these "topics" may not be as coherent as the common understanding of the word may suggest and that they should really be understood as different "contexts," although we use both terms interchangeably below.

[26] Short for "supersymmetry."

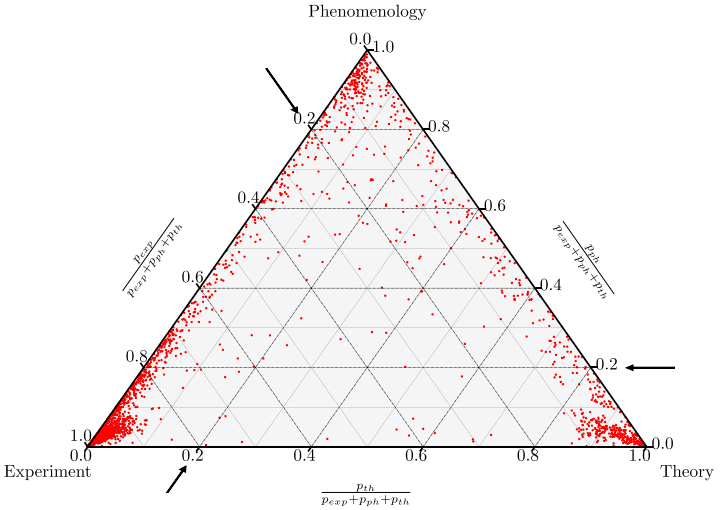


Figure 2. Relative fraction of articles from any of the categories "Experiment," "Phenomenology," and "Theory" for 2,500 high-energy physicists. Each physicist among those sampled is represented by a red dot on the diagram, positioned according to the estimate of ($p_{i,exp}$, $p_{i,ph}$, $p_{i,th}$), the probability that any of his articles belong to those three categories. The dashed lines, along the direction of the arrows, form a grid along which one can read the relative importance of each category for every physicist $\left(\frac{p_{ij}}{\sum_k p_{ik}}\right)$. Physicists near the vertices of the triangle contribute almost exclusively to one category; those near an edge contribute quasi-exclusively to two categories. Most physicists are located near a vertex, thus contributing to mostly one category.

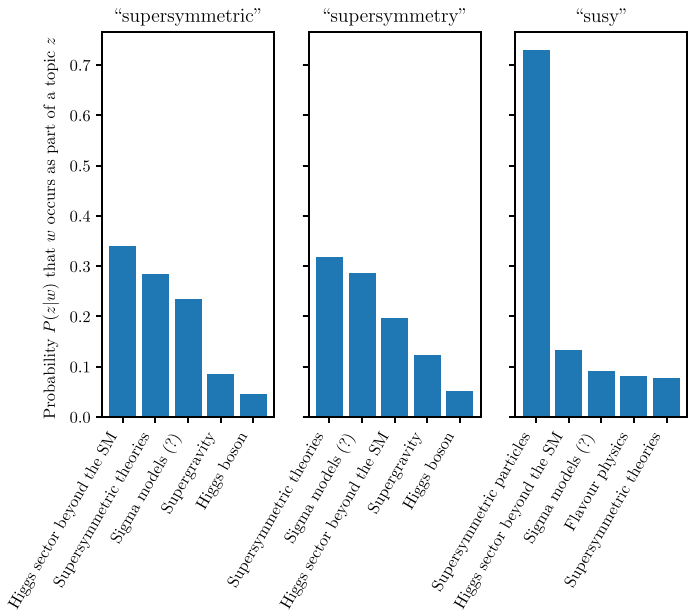


Figure 3. The many uses of supersymmetry. For three terms *w* refering to supersymmetry ("supersymmetric," "supersymmetry," and "susy"), the five topics *z* that are most likely to have led to their occurrence and theirrespective conditional probability $P(z|w)$ are shown. "Supersymmetry" and "supersymmetric" have similar distributions, and mostly occur within theoretical topics. "Susy"'s topic distribution is much more peaked, and most often occurs within phenomenological topics.

supersymmetric gauge theories in general), "sigma models (?)," "Higgs sector beyond the SM," "supergravity," "Higgs boson," "supersymmetric particles," and "flavor physics." The meaning of the "sigma models" context is unclear, although it comprises most occurrences of terms relating to superspaces and superfields. These concepts are directly tied to supersymmetry. They arise from the abstract extension of space by introducing extra anticommuting coordinates. That supersymmetry spans across distinct topics constitutes evidence for the diversity of its uses. It is also notable that several of these topics are in fact dominated by supersymmetry ("supersymmetric theories," "supergravity," and "supersymmetric particles"). This stresses the importance of supersymmetry in the HEP literature.

Moreover, although all these words ("supersymmetry," "supersymmetric," and "susy") should refer to the same concept, we find that they are in fact related to different topics: "supersymmetry" seems to encompass more theoretical aspects of supersymmmetry (e.g., supergravity) but "susy" is more likely to occur in relation to supersymmetric particles (phenomenological supersymmetry). In fact, we find that 60% of papers mentioning "supersymmetry" belong to theory (versus ~40% to phenomenology) and only 30% of papers mentioning "susy" in their abstract belong to "theory" (versus 70% to phenomenology).

That these topics are at least partially independent can be assessed by inspecting the covariance matrix $\sum$ of the CTM from which they were derived. We therefore compute the correlation matrix[27] between the seven topics most commonly associated with supersymmetry; the results can be found in Figure 4. Overall the correlations are close to 0, which suggests that these topics are rather independent, with a few exceptions. In particular, pairs of topics that belong to the same kind (theoretical or phenomenological) are moderately correlated; pairs of topics that are directly tied to supersymmetry (e.g., supergravity and phenomenological supersymmetry) but of different nature (in this case, theoretical and phenomenological, respectively) are less correlated. Further visual evidence is provided in Figure S2 in Section 3.5 of the Supplementary material.

From these results, one can see that supersymmetry is itself a diverse concept. It arises in a variety of partially independent contexts. In particular, theoretical and phenomenological aspects of supersymmetry are quite independent. How have these different aspects of supersymmetry evolved after the negative results of the searches for supersymmetric particles at the LHC?

To address this question, we evaluate the evolution of supersymmetry research in HEP from the first results of the LHC (2011) until today. For that, similarly to Hall et al. (2008), we assess the relative importance $P(z|y)$ of each topic z for every year y from 2011 to 2019:

$$P(z|y) = \frac{1}{D_y} \sum_{d \in y} P(z|d) \tag{5}$$

where D_y is the number of articles first submitted in year y. We then selected the three topics with the highest increase (rising topics) and decrease (declining topics) in magnitude over this period. For that, $P(z|y)$ was fitted to a linear time trend ($P(z|y) = a_z y + b_z$), discarding topics for which the correlation was not significant (i.e., $R = 0$ is excluded from the 99% CI). Then the topics were sorted according to the best fit value of a_z, the rate of increase of its magnitude per

[27] The Pearson correlation coefficients R_{ij} can be deduced directly from the covariance matrix Σ of the CTM model—cf. Eq. 2—according to $R_{ij} = \frac{\sum_{ij}}{\sqrt{\sum_{ii}\sum_{ij}}}$.

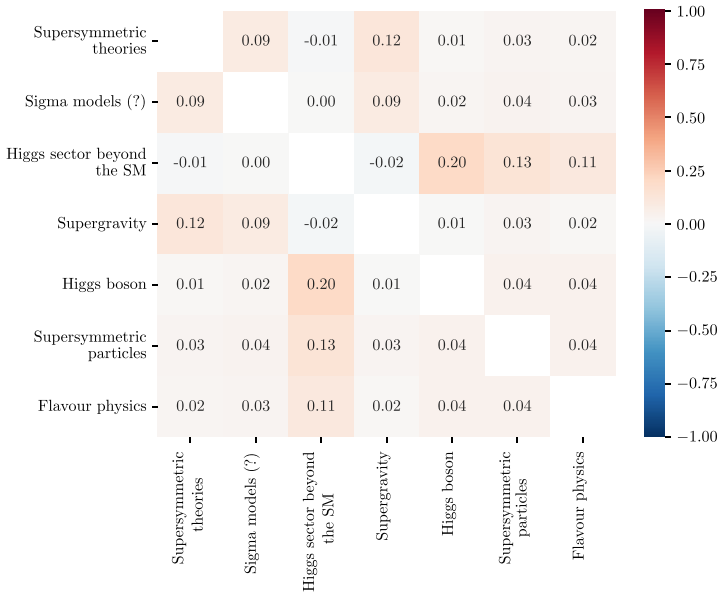


Figure 4. Correlation between the topics most associated to supersymmetry. The Pearson correlation ranges between −1 (perfect anticorrelation) and 1 (perfect correlation). A correlation close to 0 means that a pair of topics is partially independent (i.e., that they can arise or not in variable proportions in a paper).

year (similarly to what was done in Griffiths and Steyvers (2004)). We apply the procedure to all papers mentioning at least one of the words "supersymmetric," "supersymmetry," or "susy" in their title or abstract in the years following the start of the LHC. The results are shown in Figure 5.

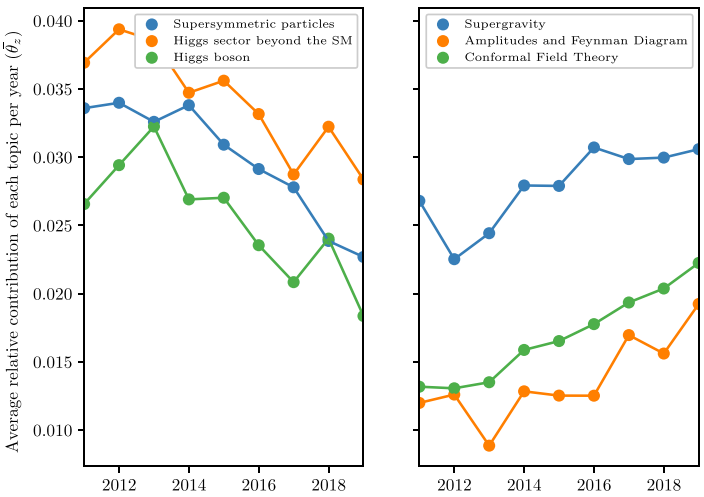


Figure 5. Declining and rising topics among those that mention supersymmetry since the first results of the LHC (2011–2019). On the left, the three topics that are declining the fastest: "Supersymmetric particles," "Higgs sector beyond the SM," and "Higgs boson." On the right, the three fastest rising topics are "Supergravity," "Amplitudes and Feynman Diagrams," and "Conformal Field Theory."

According to these results, the most rapidly declining topics among articles that mention supersymmetry are Higgs-sector related topics and phenomenological supersymmetry (i.e., phenomenological aspects of supersymmetry). By contrast, two of the (relatively) increasingly active topics are very theoretical (in particular, Supergravity and Conformal Field Theory). To understand these dynamics, it is therefore necessary to distinguish theoretical supersymmetry from phenomenological supersymmetry. As physicist Mikhail Shifman argued in an early assessment of the first results of the LHC in 2012,

> [Theoretical supersymmetry] is an example of a complete success story. I use the word 'theoretical' to differentiate from 'phenomenological' supersymmetry ... which ... at the moment has a rather murky status. Theoretical supersymmetry proved to be a powerful tool with which to deal with quantum field theory, especially at strong coupling, a regime which was considered intractable for decades .... Progress in this line of research ... is absolutely steady. (Shifman, 2012, p. 6)

Shifman's assessment strikingly converges with the patterns that emerge from our analysis. Topic models reveal the plurality of supersymmetry in HEP. They support that supersymmetry arises in different contexts, some theoretical and others phenomenological. They allowed us to demonstrate that these "faces" of supersymmetry have responded differently to the absence of evidence for supersymmetric particles at the LHC. Indeed, although phenomenologists find supersymmetry to be much less valuable in the light of the most recent experimental findings, theorists may still rely on it for their own endeavor. This supports that cultures can "trade" certain concepts (according to Galison's terminology) while retaining much of their autonomy, including in their own appraisal of the usefulness of these concepts[28].

In the next section, we investigate the contribution of supersymmetry to sustaining the trading zone between these theoretical traditions throughout time.

4.3. Supersymmetry in the Trading Zone Between Theory and Phenomenology

Which concepts sustain trades within HEP? As proposed in Section 3.4, we measure the ability of certain concepts (keywords) to sustain trades through time in terms of the probability that each of these concepts occurs in citations across theory and phenomenology. The results are shown in Figures 6 and 7.

Both these figures show the probability of occurrence of the five most common keywords (left side) and the five most common supersymmetry-related keywords (right side) involved in trades across these subcultures (excluding redundant keywords). Figure 6 shows those probabilities for trades where phenomenological papers draw from theoretical papers. Three main trends are revealed: the fall of trades involving extra-dimensions (hypothesized spatial dimensions beyond the four space-time dimensions for which we have direct evidence); the increase in trades involving black holes; and, directly relevant to our third claim, the decline of trades involving supersymmetry, despite a short increase after the start of the LHC in 2010. Interestingly, in the early 2000s, "supersymmetric model[s]" had a tradability on par with that of the keywords most involved in these trades. Moreover, "low energy" appears to be one of the most frequent keywords in phenomenological imports of theoretical papers, which makes sense because the low-energy limit of theories of, say, strings and quantum gravity is what matters most from a phenomenological standpoint (it is what can be observed). Turning to Figure 7—

[28] "... trading partners can hammer out a local coordination, despite vast global differences." (Galison, 1997, p. 783)

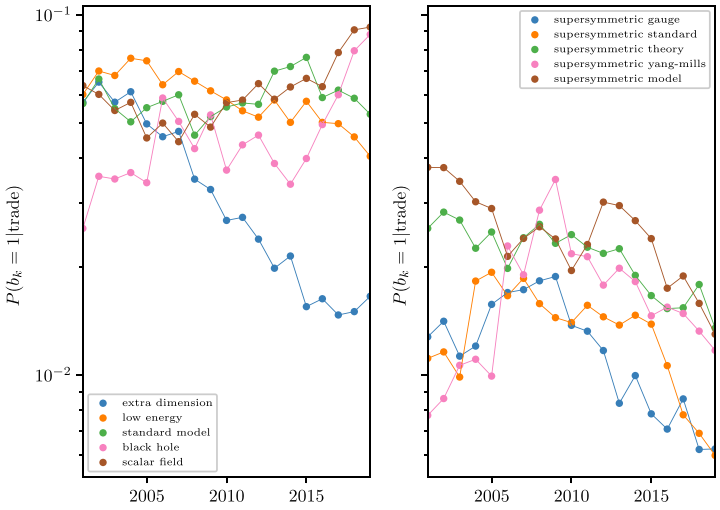


Figure 6. Inside the trading zone: Probability that certain keywords appear in the abstract of a theoretical paper involved in a trade (a phenomenological paper citing a theoretical paper). To the left, the five keywords are those with the highest peak probability of occurrence; to the right are the five keywords with the highest probability of occurrence among supersymmetry-related keywords. Redundant keywords (whose normalized pointwise mutual information with a more frequent keyword exceeds 0.9) are excluded.

trades involving theoretical references—we get an even more striking picture of the demise of "extra dimensions," which were involved in about 30% of the trades in 2001 and went down to 5% only. Similarly, "weak-scale," which refers to the domain of phenomena targeted by the LHC, has become much less frequent in the "trading zone" (from ∼10% of trades to ∼2%).

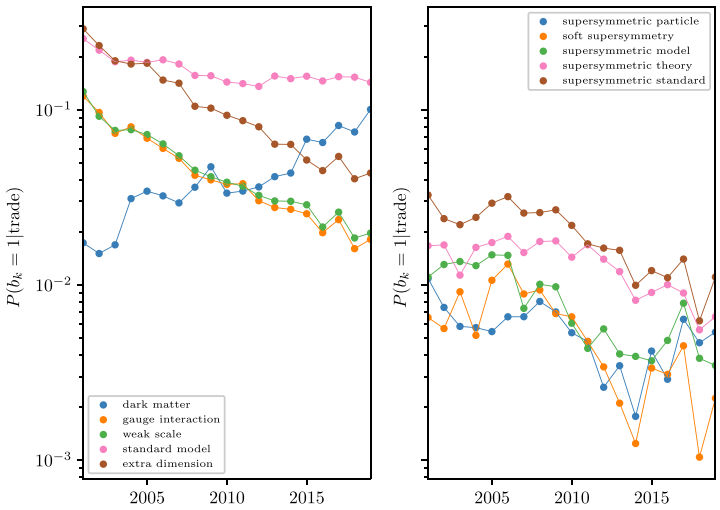


Figure 7. Inside the trading zone: Probability that certain keywords appear in the abstract of a phenomenological paper involved in a trade (a theoretical paper citing a phenomenological paper). To the left, the five keywords are those with the highest peak probability of occurrence; to the right are the five keywords with the highest probability of occurrence among supersymmetry-related keywords. Redundant keywords (whose normalized pointwise mutual information with a more frequent keyword exceeds 0.9) are excluded.

This suggests that phenomenological models dedicated to this domain of phenomena have become much less useful to the "theoretical" subculture over time. On the other hand, "dark matter"[29] is increasingly common in the phenomenological papers theorists draw from. This suggests dark matter is deemed valuable for the theoretical enterprise as well. This figure also confirms the overall decline of supersymmetry in the trading zone, thus providing further support to our third claim: Supersymmetry does not connect developments from the theoretical programs to progress in the phenomenological program as much as it did prior to the LHC. Astrophysical topics, on the other hand, seem to be taking up an increasing proportion of the trading zone.

5. DISCUSSION

5.1. Conceptually Informed Methods in Quantitative Science Studies

Before exploring the implications of this case study, we want to emphasize that Galison's conceptual framework has been a fruitful guide for our quantitative approach. The linguistic component of his notion of subculture led us to build a bag-of-words model for measuring the extent of the divide between two theoretical cultures and for unveiling the concepts that are specific to these cultures as well as their methodological and ontological differences. The social autonomy of these subcultures, too, can be readily quantified from authorship data. Furthermore, the notion of a trading zone invited us to explore citations quantitatively (as a proxy of scientific "trades") while devising ways to determine their "location" in the semantic space. We also found that topic models can reveal the plurality of contexts in which a concept may arise, and how the dynamics of these contexts compare throughout time. Although we have applied our topic model approach to supersymmetry, in principle, it can be applied to any kind of "boundary object," understood in the broad sense of a shared notion that allows some coordination to be achieved while preserving the distinctness of the scientific cultures at play. In the end, these methods illuminated our study of supersymmetry in HEP, and provided further grounds for Galison's claim that unity is a contingent matter.

5.2. Unity Challenged?

The two theoretical subcultures we have distinguished—"pure" theory and phenomenology—no longer seem to value supersymmetry equally. Supersymmetry indeed fails to provide equally satisfying solutions to the heterogeneous commitments of high-energy physicists, which poses a challenge to the unity of the field. Indeed, the example of supersymmetry shows that what drives theoretical progress may not drive phenomenological progress—in contrast with the expectations of the community regarding supersymmetry prior to the LHC as surveyed by Mättig and Stöltzner (2019, 2020)—and developments in these subcultures may become quite orthogonal.

Of course, supersymmetry is not the only channel of coordination between the theoretical and phenomenological cultures in their search for "new physics." Another channel, for instance, has been the notion of extra dimensions (see Figures 6 and 7), which dominated trades in the early 2000s to an extent we did not expect before conducting this analysis. Extra dimensions are required by string theory, but they are also subject to trades with

[29] Dark matter refers to the observation that a significant fraction of the mass of the universe is currently unexplained.

phenomenologists interested in their observable consequences. However, no evidence for extra dimensions was found at the LHC. This further supports that the goals that drive theoretical research programs, such as string theory (like the search for a quantum description of gravity), may not serve the phenomenologists' agenda so well.

Eventually, the LHC provided "a test of the unity of physics"[30], and its verdict was ruthless. In the future, will the field strive to regain unity (possibly to the detriment of certain research programs), or will the socially entrenched divergences between these "cultures" of HEP prevail? We may assume that the challenge is merely transitory, and that theorists will eventually move to other theories, which will be more successful from an empirical or phenomenological standpoint. However, the divergence between these theoretical cultures has become axiological (Camilleri & Ritson, 2015; Laudan, 1984), in the sense that they prioritize different epistemic goals[31]; and this divergence may persist as long as their differences in aims persist; as Galison puts it, "there is no teleological drive towards ever-greater cohesion," and "fields previously bound [may] fall apart" (Galison, 1997, p. 805). As illustrated in Figure 1, the aim of the theorists is to achieve the unification of the fundamental forces and a coherent theory of quantum gravity. By contrast, the aim of phenomenologists is to guide the experiment towards promising directions where evidence of "new physics" may be found. Both these aims may seem well-founded; however, there is no reason to expect that a simultaneous solution can be worked out. The apparent failure of supersymmetry to provide such a simultaneous solution does not undermine by itself the relevance of the "theorists"' aims, nor does it undermine the methodology they deploy for addressing their goals (e.g., their trust in certain theoretical constraints, cf. Galison, 1995). It does, however, challenge the belief that such methods can provide grounds for progress to the field *as a whole*; indeed, unification and quantum gravity might eventually not provide much reliable guidance to the experimental side. Conversely, it can very well be that the details of the theory "at high energy," where quantum effects matter to gravity, cannot be extrapolated from our knowledge of the low-energy theory (i.e., the one that we can probe in our experiments). As a result, Dawid (2013) argues for recourse to meta-empirical assessment of theories in theoretical physics, given that empirical input underdetermines the directions of potential progress in quantum gravity. Disagreements in the aims of a scientific enterprise may not always be resolved on purely epistemic grounds, and a resolution, provided it occurs, may involve some sort of negotiation instead. As long as theorists believe in the feasibility of their aims, they may pursue these aims even if it further isolates them from other cultures[32]. Alternatively, they could decide that the schism should be resolved; as Galison puts it, distinct scientific cultures "can … understand that the continuation of exchange is a prerequisite to the survival of the larger culture of which they are part" (Galison, 1997, p. 803). Meanwhile, the trading zone between theory and phenomenology is shifting from collider physics to astrophysics (whether it concerns dark matter, black holes, or the early universe).

[30] Wilson (1986, p. 29) (cited in Cat, 1998, p. 292) used this expression in reference to the now aborted Super-Conducting Supercollider, also in reference to supersymmetry.

[31] Laudan (1984) refers to disagreements in the goals of scientific inquiry as *axiological* disagreements. Camilleri and Ritson (2015), for instance, have argued that certain controversies around string theory could be understood in terms of an instance of axiological disagreement.

[32] More drastically, Cao and Schweber (1993) expressed the view that the theories at different energy scales (i.e., corresponding to different ranges of phenomena) are irreducible, and they argued for a "pluralist view of possible theoretical ontologies" while challenging the possibility of achieving a "ultimate stable theory of everything" (pp. 69–71). According to this view, the plurality of ontologies in physics is not an accident but the result of partially disconnected "phenomenological domains" through which knowledge cannot be deduced from one another. For criticisms of this view, see Castellani (2002) and Rivat and Grinbaum (2020).

5.3. Trading Zones as a Means to Sustain Diversity

More generally, the example of HEP and supersymmetry demonstrates how disunity can be endogenously produced in the fabric of science. Even initially tightly bound scientific cultures can diverge into quite distinct and autonomous programs, with different ontologies, methodologies, and aims, as new domains of inquiry open up (e.g., quantum gravity) and warrant new modes of knowing. The extent of the coordination between disciplines will in general depend on epistemic factors (depending on how fruitful certain "trades" turn out to be), but also on nonepistemic factors: For instance, it may depend on the institutional setting, or whether such exchanges are incentivized or "coerced" (Collins et al., 2010).

Paradoxically, it can be noted that trading zones can stabilize the heterogeneity of cultures within a field, by sustaining the practitioners' beliefs that, in spite of the large differences in what they are doing, their respective efforts somehow support each other. If that is the case, there is no perceived need for a profound realignment of their respective practice. Trading zones can contribute, therefore, to a mutual process of legitimization of heterogeneous scientific practices, which is not necessarily tantamount to further ontological unity. To further emphasize that, it is useful to come back to the example of HEP, and most particularly that of string theory, a highly theoretical research program driven by the pursuit of a consistent theory of quantum gravity. String theorists such as Matt Strassler have argued that even if string theory did not directly provide testable predictions to phenomenologists and experimentalists, it generated mathematical tools that could be useful to their practice, such as for predicting the behavior of a quark-gluon plasma (Ritson, 2021). Consequently, phenomenologists may have a low appraisal of string theory in terms of its ability to generate models for testing its assumptions about nature, while still recognizing the usefulness of what string theorists do for them, as some of their work is effectively "applicable." As Ritson and Camilleri (2015) put it, "if string theory has proved so useful for branches of physics whose scientific status is not in question, it can be argued it forms a legitimate part of physics." Supersymmetry itself may be experiencing the same fate, considering that "supersymmetry as a tool for exploring gauge dynamics at strong coupling ... is taking precedence over phenomenology" (Shifman, 2020, pp. 7–8). Such trades do support the usefulness of the theoretical program to other endeavors, without necessarily implying further integration of the subcultures of HEP (ontological unity), just like successful interdisciplinary work does not necessarily amount to further integration of disciplines (Grüne-Yanoff, 2016).

5.4. Limitations and Future Work

Before concluding, we would like to hint at several directions for future work that could overcome certain limitations of the present methodology and further inform the question of the disunity of science.

First, none of our semantic methods distinguished between different kinds of words, that is, which words refer to, say, methods (such as computation techniques) rather than entities (e.g., strings, particles). It would be interesting to evaluate to what extent the coordination between theoretical cultures involves ontological or mere methodological trades, depending on whether the constructs of high theory are referred to as the proper description of nature or as mere mathematical tools, and how this may have changed throughout time. This might uncover evidence for a shift from an ontological to a more methodological coordination between the subcultures of HEP, as the arguments for supersymmetry and string theory as "tools" rather than accurate accounts of the natural world suggest.

Another direction of future work involves the topic model approach. Although the topic model used in this work yielded seemingly acceptable results overall, some topics were difficult to interpret. In that respect, we made several improvements compared to previous works, by training the model on not just single words but also *n*-grams matching specific and presumably semantically informative syntactic patterns and by informing our interpretation of topics using correlations with a standard classification (rather than the top-words only). Yet, further improvements could be made. First, vocabulary selection could be enhanced by a better handling of mathematical expressions, such as by parsing LaTeX formulas. The NLTK library picked up some of these expressions, and as they captured some information about the documents, we did not exclude them from the vocabulary; however, this way of proceeding does not preserve the underlying mathematical structure, although it may be valuable to distinguish references to, say, specific particles, or certain symmetry groups, based on their mathematical notations. We may also want the model to learn to discard uninformative words such as *result*, *parameter*, or *model*. In our case, we found such vague words to be clustered in three topics that we labeled as *jargon* which correlated very poorly with the standard classification (see Tables S1 and S2 in the Supplementary material), but they should ideally not emerge as distinct topics on par with more meaningful topics. To this end, we may want to build on Griffiths, Steyvers et al. (2004), which provides a model that is able to distinguish between "semantic" and purely "syntactic" clusters of words without prior knowledge of the language. A more critical limitation of topic models pertains to the challenge of hyperparameter tuning, considering that it is unclear which performance metric should be maximized in the process. Although we proposed a procedure for choosing these parameters that accounts for known limitations to the reliability of perplexity or topic coherence metrics, nonparametric methods may provide a better answer to this fundamental issue (Gerlach, Peixoto, & Altmann, 2018).

Finally, the historical scope of our analysis was limited by our database. In particular, we were only able to analyze the theory/phenomenology divide over a restricted time range (1980–2020), and we could not reveal how such a divide has historically emerged. By contrast, Galison has proposed a number of explanations for the earlier decoupling of theory and experiment, such as increased specialization and the increased time scales of experiments (Galison, 1987, p. 138).

ACKNOWLEDGMENTS

We would like to thank Olivier Darrigol for his extremely insightful comments regarding the concept of trading zones; Arianna Borrelli for commenting on the work that led to this publication; Thomas Heinze and Radin Dardashti for their comments and continuous support; Alexander Blum for his consideration and his suggestions; and Elizabeth Zanghi for her corrections. Finally, we thank all reviewers for their in-depth comments.

FUNDING INFORMATION

The authors acknowledge support from the Open Access Publication Fund of the University of Wuppertal.

COMPETING INTERESTS

The authors have no conflicts of interest.

DATA AVAILABILITY

All the data and code used to derive the results of this paper can be accessed from the following repository using the DataLad software (Halchenko, Meyer et al., 2021): https://github.com/lucasgautheron/trading_zones_material. A permanent archive of the code is available at the following URL: https://doi.org/10.5281/zenodo.7970559.

REFERENCES

Allen, C., & Murdock, J. (2022). LDA topic modeling: Contexts for the history and philosophy of science. In G. Ramsey & A. De Block (Eds.), *The dynamics of science: Computational frontiers in history and philosophy of science*. Pittsburgh, PA: University of Pittsburgh Press. https://doi.org/10.2307/j.ctv31djr2f.9

Battiston, F., Musciotto, F., Wang, D., Barabási, A.-L., Szell, M., & Sinatra, R. (2019). Taking census of physics. *Nature Reviews Physics*, *1*(1), 89–97. https://doi.org/10.1038/s42254-018-0005-3

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2015). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, *17*(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0

Bennett, A., Misra, D., & Than, N. (2021). Have you tried neural topic models? Comparative analysis of neural and non-neural topic models with application to COVID-19 Twitter data. *arXiv:2105.10165*. https://doi.org/10.48550/arXiv.2105.10165

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics*, *1*(1), 17–35. https://doi.org/10.1214/07-AOAS114

Borrelli, A. (2015). Between logos and mythos: Narratives of "naturalness" in today's particle physics community. In H. Blume, C. Leitgeb, & M. Rössner (Eds.), *Narrated communities—Narrated realities* (pp. 69–83). Brill. https://doi.org/10.1163/9789004184121_006

Bowker, G. C. (2020). Numbers or no numbers in science studies. *Quantitative Science Studies*, *1*(3), 927–929. https://doi.org/10.1162/qss_a_00054

Camilleri, K., & Ritson, S. (2015). The role of heuristic appraisal in conflicting assessments of string theory. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *51*, 44–56. https://doi.org/10.1016/j.shpsb.2015.07.003

Cao, T. Y., & Schweber, S. S. (1993). The conceptual foundations and the philosophical aspects of renormalization theory. *Synthese*, *97*(1), 33–108. https://doi.org/10.1007/BF01255832

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press. https://doi.org/10.1017/CBO9781139167093

Castellani, E. (2002). Reductionism, emergence, and effective field theories. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *33*(2), 251–267. https://doi.org/10.1016/S1355-2198(02)00003-5

Cat, J. (1998). The physicists' debates on unification in physics at the end of the 20th century. *Historical Studies in the Physical and Biological Sciences*, *28*(2), 253–299. https://doi.org/10.2307/27757796

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09)* (pp. 288–296).

Collins, H., Evans, R., & Gorman, M. (2010). Trading with the enemy. In M. Gorman (Ed.), *Trading zones and interactional expertise: Creating new kinds of collaboration*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262014724.003.0003

Darrigol, O. (2007). The modular structure of physical theories. *Synthese*, *162*(2), 195–223. https://doi.org/10.1007/s11229-007-9181-x

Dawid, R. (2013). *String theory and the scientific method*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139342513

Dimopoulos, S., & Georgi, H. (1981). Softly broken supersymmetry and SU(5). *Nuclear Physics B*, *193*(1), 150–162. https://doi.org/10.1016/0550-3213(81)90522-8

Dupré, J. (1983). The disunity of science. *Mind*, *92*(367), 321–346. https://doi.org/10.1093/mind/XCII.367.321

Fayet, P., & Ferrara, S. (1977). Supersymmetry. *Physics Reports*, *32*, 249–334. https://doi.org/10.1016/0370-1573(77)90066-7

Freedman, D. Z. (1979). Review of supersymmetry and supergravity. In *19th International Conference on High-Energy Physics* (pp. 535–547). https://cds.cern.ch/record/870690

Galison, P. (1987). *How experiments end*. Chicago, IL: University of Chicago Press.

Galison, P. (1988). History, philosophy, and the central metaphor. *Science in Context*, *2*(1), 197–212. https://doi.org/10.1017/S0269889700000557

Galison, P. (1995). Theory bound and unbound: Superstrings and experiments. In F. Weinert (Ed.), *Laws of nature: Essays on the philosophical, scientific and historical dimensions* (pp. 369–408). De Gruyter. https://doi.org/10.1515/9783110869859.369

Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago, IL: University of Chicago Press.

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, *4*(7), eaaq1360. https://doi.org/10.1126/sciadv.aaq1360, PubMed: 30035215

Giudice, G. F. (2018). The dawn of the post-naturalness era. In *From my vast repertoire: Guido Altarelli's legacy* (pp. 267–292). World Scientific. https://doi.org/10.1142/9789813238053_0013

Giudice, G. F., & Romanino, A. (2004). Split supersymmetry. *Nuclear Physics B*, *699*(1–2), 65–89. https://doi.org/10.1016/j.nuclphysb.2004.08.001

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101, PubMed: 14872004

Griffiths, T. L., Steyvers, M., Blei, D., & Tenenbaum, J. (2004). Integrating topics and syntax. In L. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems*. Cambridge, MA: MIT Press.

Grüne-Yanoff, T. (2016). Interdisciplinary success without integration. *European Journal for Philosophy of Science*, *6*(3), 343–360. https://doi.org/10.1007/s13194-016-0139-z

Halchenko, Y., Meyer, K., Poldrack, B., Solanky, D., Wagner, A., ... Hanke, M. (2021). DataLad: Distributed system for joint

management of code, data, and their relationship. *Journal of Open Source Software*, *6*(63), 3262. https://doi.org/10.21105/joss.03262

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 363–371). Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613763

Heinze, T., & Jappe, A. (2020). Quantitative science studies should be framed with middle-range theories and concepts from the social sciences. *Quantitative Science Studies*, *1*(3), 983–992. https://doi.org/10.1162/qss_a_00059

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J. L., & Resnik, P. (2021). Is automated topic model evaluation broken? The incoherence of coherence. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 2018–2033). https://proceedings.neurips.cc/paper/2021/hash/0f83556a305d789b1d71815e8ea4f4b0-Abstract.html

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, *3*(2), 259–289. https://doi.org/10.1075/term.3.2.03kag

Kane, G. L., & Shifman, M. (Eds.). (2000). *The supersymmetric world: The beginnings of the theory*. World Scientific. https://doi.org/10.1142/4611

Kang, D., & Evans, J. (2020). Against method: Exploding the boundary between qualitative and quantitative studies of science. *Quantitative Science Studies*, *1*(3), 930–944. https://doi.org/10.1162/qss_a_00056

Kemman, M. (2021). The trading zones model. In *Trading zones of digital history* (pp. 39–61). De Gruyter. https://doi.org/10.1515/9783110682106-002

Laudan, L. (1984). *Science and values: The aims of science and their role in the scientific debate*. University of California Press.

Lee, M. Fenstermacher, D., Schneider, J., & Cendrero Almodóvar, J. (2022). bab2min/tomotopy: Python package of Tomoto, the Topic Modeling Tool. *Zenodo*. https://zenodo.org/record/6869125. https://doi.org/10.5281/zenodo.6869125

Leydesdorff, L., & Nerghes, A. (2016). Co-word maps and topic modeling: A comparison using small and medium-sized corpora. *Journal of the Association for Information Science and Technology*, *68*(4), 1024–1035. https://doi.org/10.1002/asi.23740

Leydesdorff, L., Ràfols, I., & Milojević, S. (2020). Bridging the divide between qualitative and quantitative science studies. *Quantitative Science Studies*, *1*(3), 918–926. https://doi.org/10.1162/qss_e_00061

Malaterre, C., Chartier, J.-F., & Pulizzotto, D. (2022). Topic modeling in HPS. In G. Ramsey & A. De Block (Eds.), *The dynamics of science*. Pittsburgh, PA: University of Pittsburgh Press. https://doi.org/10.2307/j.ctv31djr2f.13

Martin, J. D. (2018). *Solid state insurrection: How the science of substance made American physics matter*. Pittsburgh, PA: University of Pittsburgh Press. https://doi.org/10.2307/j.ctv5j02c7

Mättig, P., & Stöltzner, M. (2019). Model choice and crucial tests. On the empirical epistemology of the Higgs discovery. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *65*, 73–96. https://doi.org/10.1016/j.shpsb.2018.09.001

Mättig, P., & Stöltzner, M. (2020). Model landscapes and event signatures in elementary particle physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *69*, 12–25. https://doi.org/10.1016/j.shpsb.2019.07.003

Moskovic, M. (2021). The INSPIRE REST API. *Zenodo*. https://doi.org/10.5281/zenodo.5793383

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741–754. https://doi.org/10.1007/s11192-014-1319-2

Omodei, E. (2014). *Modeling the socio-semantic dynamics of scientific communities* (Thesis). Ecole Normale Supérieure – ENS PARIS.

Parker, M. A. (1999). Searches for supersymmetry at the Large Hadron Collider. In *1999 International Europhysics Conference on High-Energy Physics* (pp. 814–817).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pickering, A. (1984). *Constructing quarks: A sociological history of particle physics*. Chicago, IL: University of Chicago Press.

Raimbault, J. (2019). Exploration of an interdisciplinary scientific landscape. *Scientometrics*, *119*(2), 617–641. https://doi.org/10.1007/s11192-019-03090-3

Ritson, S. (2021). Constraints and divergent assessments of fertility in non-empirical physics in the history of the string theory controversy. *Studies in History and Philosophy of Science Part A*, *90*, 39–49. https://doi.org/10.1016/j.shpsa.2021.08.016, PubMed: 34555653

Ritson, S., & Camilleri, K. (2015). Contested boundaries: The string theory debates and ideologies of science. *Perspectives on Science*, *23*(2), 192–227. https://doi.org/10.1162/POSC_a_00168

Rivat, S., & Grinbaum, A. (2020). Philosophical foundations of effective field theories. *European Physical Journal A*, *56*(3), 90. https://doi.org/10.1140/epja/s10050-020-00089-w

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520. https://doi.org/10.1108/00220410410560582

Shifman, M. (2012). Frontiers beyond the standard model: Reflections and impressionistic portrait of the conference. *Modern Physics Letters A*, *27*(40), 1230043. https://doi.org/10.1142/S0217732312300431

Shifman, M. (2020). Musings on the current status of HEP. *Modern Physics Letters A*, *35*(7), 2030003. https://doi.org/10.1142/S0217732320300037

Shinn, T., & Ragouet, P. (2005). *Controverses sur la science: Pour une sociologie transversaliste de l'activité scientifique*. Éditions Raisons d'Agir.

Star, S. L. & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, *19*(3), 387–420. https://doi.org/10.1177/030631289019003001

Suppes, P. (1978). The plurality of science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1978*, 3–16. https://doi.org/10.1086/psaprocbienmeetp.1978.2.192459

Syed, S., & Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 165–174). IEEE. https://doi.org/10.1109/DSAA.2017.61

Taylor, J. G. (1984). A review of supersymmetry and supergravity. *Progress in Particle and Nuclear Physics*, *12*, 1–101. https://doi.org/10.1016/0146-6410(84)90002-4

van der Maaten, L. J., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Veltman, M. J. G. (1981). The infrared–ultraviolet connection. *Acta Physica Polonica B*, *12*(5), 437–457.

Volkov, D., & Akulov, V. (1973). Is the neutrino a Goldstone particle? *Physics Letters B*, *46*(1), 109–110. https://doi.org/10.1016/0370-2693(73)90490-5

Weinberg, S. (1979). Gauge hierarchies. *Physics Letters B*, *82*, 387–391. https://doi.org/10.1016/0370-2693(79)90248-X
Weinberg, S. (1995). *The quantum theory of fields: Supersymmetry* (Vol. 3). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139644167
Wilson, R. R. (1986). The sentiment of the unity of physics. *Physics Today*, *39*(7), 26–30. https://doi.org/10.1063/1.881034
Witten, E. (1982). Introduction to supersymmetry. In *19th International School of Subnuclear Physics: The Unity of the Fundamental Interactions* (pp. 305–371). https://doi.org/10.1007/978-1-4613-3655-6_7
Yan, E., Ding, Y., Cronin, B., & Leydesdorff, L. (2013). A bird's-eye view of scientific trading: Dependency relations among fields of science. *Journal of Informetrics*, *7*(2), 249–264. https://doi.org/10.1016/j.joi.2012.11.008

Supplementary materials for "How research programs come apart"

Lucas Gautheron · Elisa Omodei

S1 Data collection

Our goal was to collect the whole HEP literature from 1980 to 2020 from the public Inspire HEP API (Moskovic, 2021). For that, we collected metadata for all articles through automated search requests, category per category, and year per year. This strategy was intended to abide with the limitations of the API, in terms of matching entries per search request. However, it appeared that many articles in years 1990 to 1995 were not categorized, and therefore our collection strategy missed many HEP articles from this period. In order to recover these articles, we gathered all articles that were referenced in publications collected through the first batch but which were missing. This methods fails to recover articles that were not cited in any article from the first batch. More importantly, the lack of categories means that selecting all HEP papers during the problematic time period will require unlabeled articles to be manually or automatically classified. Although there are ways to circumvent these issues and to assess their potential implications, we have decided to narrow down several analyses to years 2001 onwards in the present work.

S2 Text-classifier performance stability

The categories (`Theory-HEP`, `Phenomenology-HEP` and `Experiment-HEP`) that we trained our classifier (3.2) to predict have been assigned in different ways in the Inspire HEP database. Although a majority were categorized based on arXiv's classification system, some papers were not, especially those published before arXix was introduced (in the early 1990s). It might seem unclear whether these classification procedures are consistent and revealing of distinct underlying cultures. In order to demonstrate that it is the case, in Figure S1, we show that the performance of the text-classifier is nonetheless roughly stable throughout the period considered (1980–2020). To this end, we subdivide this time-range in bins of five years and perform k-fold cross-validation using each five year bin for the validation set (and the papers from the other bins for the training set). Accuracy remains high and approximately stable over the years 1980 to 2020; therefore, these various classification procedures, and the underlying identity of each of these subcultures, must be rather consistent over this period.

S3 Topic model

S3.1 Data and vocabulary selection

The model is trained on $N = 120,000$ articles randomly sampled from those in the 1980-2020 period that belong to any of the categories `Theory-HEP`, `Phenomenology-HEP`, `Experiment-HEP`, and `Lattice`. Titles and abstracts of each papers are concatenated in order to maximize the textual content used for training. Very short texts (less than 100 characters) are removed.

Before applying the model, we performed a number of pre-processing steps on the abstracts with the goal of maximizing the amount of useful information in the training data. This procedure, largely inspired from Omodei 2014 and implemented with the use of the NLTK library (Bird et al., 2009), is as follows:

Address(es) of author(s) should be given

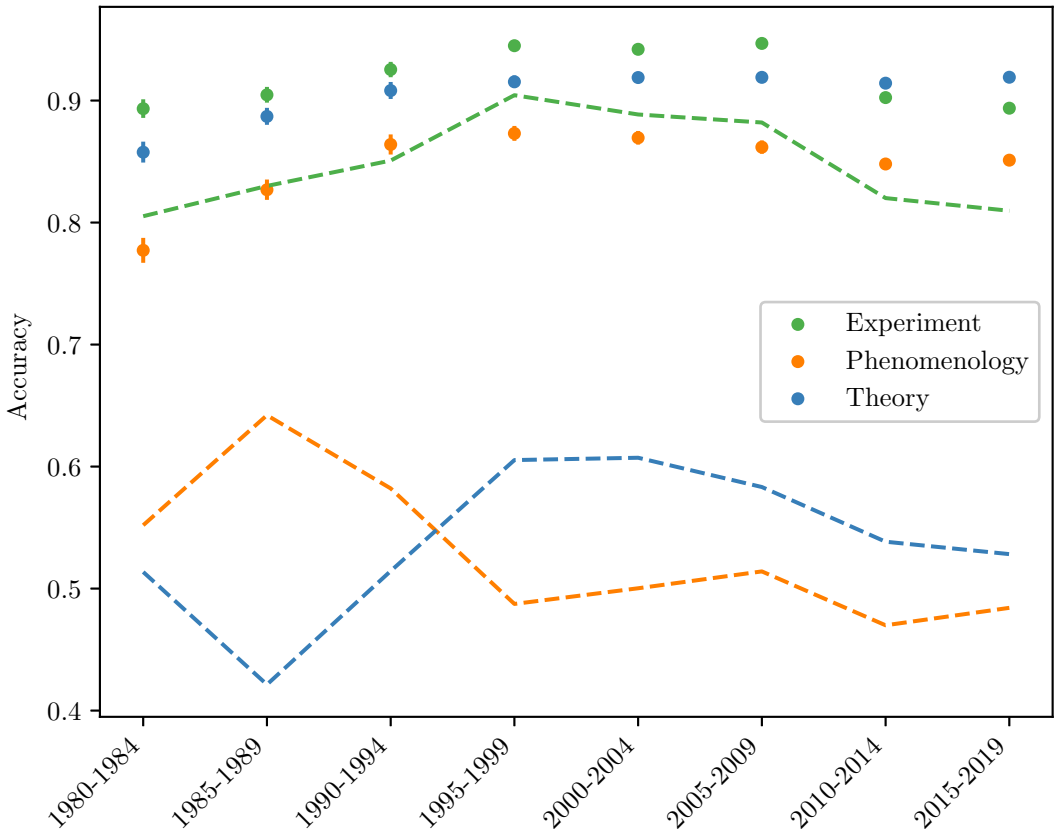


Fig. S1 Accuracy of the text-classifier from Section 3.2 as a function of the papers' years of publication. Error-bars represent the 95% confidence interval. Dashed lines show the accuracy of the baseline model (which may vary only due to variations in the frequency of each category, since the baseline model always predicts the most common class). The accuracy is roughly constant across time for each of the three categories, despite significant variations in the frequency of each class.

- Tokens (words separated by punctuation or spaces) are extracted from the text and transformed to lower-case.
- All single nouns and adjectives are retrieved from these tokens.
- We also retrieve all n-grams that match specific syntactic patterns (e.g. "adjective+noun+noun", such as "supersymmetric standard model", "effective field theory").
- Single words are lemmatized, i.e. they are normalized to their root (e.g. "symmetries" becomes "symmetry").
- Words and expressions that occur less than 20 times are removed.

First, these steps allow us to reduce noise by removing words that convey little to no information about the topics of the articles (such as stop words). Second, extracting n-grams that matching certain syntactic patterns allows us to preserve some information about the relative position of words within the abstracts – which CTM do not do otherwise – while taking advantage of our prior knowledge of the documents' language. For instance, the word "dark" may convey different meanings depending on whether it occurs immediately before the word "matter", or, alternatively, "energy"; similarly, the occurrence of the expression "dark matter" in a text conveys more information than the simultaneous occurrence of "dark" and "matter" without more knowledge about their relative position.

As a result of this procedure, the vocabulary contains $V = 18{,}658$ "words", with 58 words per article on average.

S3.2 Hyper-parameters

The implementation of the CTM by Tomotopy (Bab2min et al., 2021) has three hyper-parameters: the amount of topics k, an $\vec{\alpha}$ parameter that controls the sparsity of the document-topic distribution ($\theta_{d,i}$), and a $\vec{\eta}$ parameter that controls the sparsity of the topic-word distribution (the vocabulary associated to

each topic). For choosing the amount of topics k, we considered three values that seemed acceptable in terms of interpretability and compliance with the values from the literature: 50, 75 and 100. We assumed $\vec{\alpha}$ and $\vec{\eta}$ to be symmetric, i.e. $\alpha_1 = \alpha_k = \alpha$ and $\eta_1 = ... = \eta_V = \eta$[1]. We considered $\alpha \in \{10^{-2}, 10^{-1}, 1\}$ and $\eta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, according to values encountered in the literature. We then trained the model for each triplet of k, α and η among the candidate values. We rejected all triplets that led to significant overfitting, by comparing the perplexity[2] obtained for the training corpus and that obtained by applying the trained model to a validation set of abstracts unseen during training. Although Chang et al. (2009) have shown that perplexity could be negatively correlated to human judgments about the interpretability of the topics recovered by topic models, we believe it is a suitable metric to discard models that fail to capture meaningful regularities in the data, which is the case of models that show overfitting. Among the remaining models, we then selected the two models with the highest normalized pointwise mutual information coherence, a coherence metric frequently used to assess the consistency of topic models (Hoyle et al., 2021). Topic coherence metrics in general, as stressed by Hoyle et al., are not very strongly correlated with human judgments about the quality of a model; however, we believe they may be useful to discard certain models in order to limit the amount of those that should be inspected manually (since manual inspection is time-consuming and quite subjective). We finally inspect manually the two models with the highest coherence measure, and choose the one with $k = 75$, $\alpha = 0.1$ and $\eta = 0.001$. Our preference for this model stemmed from the fact that it contained more topics than the other remaining model, and that these more numerous topics seemed reasonably consistent.

S3.3 Validation

Since the model infers document-topic distributions and topic-word distributions, we would like to assess the validity of these metrics, i.e. "their ability to measure what they purportedly measure" (Bannigan & Watson, 2009, p. 3240). In order to simultaneously assess both measures, we designed the following protocol. First, we derived the Physics and Astronomy Classification Scheme® (PACS) categories c that were the most correlated to each topic z (this approach is in a sense comparable to that employed in Griffiths and Steyvers 2004, who extracted the topics that were more strongly associated with PNAS categories). For that, we listed the categories c that maximize the pointwise mutual information with each topic z according to:

$$\mathrm{pmi}(z, c) = \log \frac{p(z|c)}{p(z)} \tag{1}$$

Where $p(z)$ is the marginal probability of the topic z, and $p(z|c)$ the probability that a word in a document belongs to a topic z given that the document was assigned the PACS category c. Thefore, $\mathrm{pmi}(z, c)$ measures the increase in probability of a given topic provided that a PACS category is present. The 5 categories most correlated to each topic are given in table S2, which helped inform our choice for each tpic label, in complement to their top-words.

Then, we submitted the lists of PACS categories thus constitued to a human task derived from the methodology of Bennett et al. (2021), as follows:

1. We draw at random a topic z_1 with a probability equal to its marginal probability
2. We draw at random 5 PACS categories $c_1, ..., c_5$ among the 10 most correlated to z_1, as described above.
3. Then, we do any of the following, with equal probability 1/2:
   (a) We draw at random another topic $z_2 \neq z_1$ with probability $\frac{p(z_2)}{1-p(z_1)}$, and we pick at random 5 PACS categories $c_6, ..., c_{10}$ among those most correlated with it.
   (b) Alternatively, we draw $c_6, ..., c_{10}$ from the 5 remaining PACS categories most associated to z_1
4. We submit $c_1, ..., c_5$ and $c_6, ..., c_{10}$ to an expert unaware of the model. The expert is asked to guess whether the two lists of 5 categories were drawn from one and same general topic, or whether they were drawn from two separate topics.
5. The procedure is repeated a certain amount of times. The final score is the fraction of correct responses.

[1] This is common in the literature, but this choice is disputable, cf. Wallach et al. 2009. One implication of symmetric priors is that topics must have comparable probabilities. This also has an impact on the meaning of topics.

[2] Perplexity is the exponential of the average log-likelihood per word, cf. Blei et al. 2003. It measures the improbability of a corpus according to a given model.

The rationale for this method is that good scores should only be achievable provided the topics are rather coherent, and that the document-topic distributions $\theta_{d,i}$ are reasonably accurate. The final average score is 0.74 for 100 guesses from two HEP PhD students, which is significantly better than a random baseline (0.5). This shows that, to some extent, the topic distributions derived for each article correlate with PACS categories that are rather coherent with each other.

S3.4 Topics

Table S1: Most frequent terms for each topic.

	Most frequent expressions
Topic (context)	
Algorithms and calculation techniques	simulation, carlo, monte, lattice, method, correlation, distribution, cluster, generator, statistical, study, function, scaling, size, event
Amplitude of scattering processes	amplitude, contribution, state, interaction, resonance, final, final state, process, exchange, reaction, tree, scattering, double, polarization, level
Amplitudes and Feynman Diagram	amplitude, function, loop, limit, pole, conformal, relation, integral, diagram, correlation, scattering, analytic, block, correlators, feynman
Analyses and measurements from colliders	data, measurement, event, result, detector, experiment, gev, algorithm, analysis, muon, experimental, energy, precision, fit, beam
Annihilation and scattering cross-sections	section, cross, annihilation, photon, energy, scattering, gev, production, total, elastic, process, pair, total cross section, total cross, elastic scattering
Astrophysics	star, wave, nuclear, matter, neutron, collision, gravitational waves, energy, nuclear matter, flow, density, gravitational, relativistic, heavy-ion, equation
Black holes	black, hole, black hole, black holes, horizon, entropy, extremal, radiation, schwarzschild, thermodynamics, black hole solutions, black hole entropy, hawking, charge, kerr
Boundary conditions/non-locality	boundary, condition, boundary conditions, state, tensor, entropy, entanglement, distance, case, surface, general, correlation, boundary condition, term, phys
CP violating processes	cp, asymmetry, violation, parameter, b^0, bound, direct cp, direct, mixing, penguin, decay, constraint, experimental, direct cp violation, effect
Chiral symmetry	chiral, quark, qcd, lattice, chiral symmetry, mass, chiral perturbation theory, chiral perturbation, pion, condensate, baryon, transition, perturbation, flavor, symmetry
Conformal Field Theory	conformal, string, algebra, theory, conformal field, conformal field theory, central, central charge, conformal field theories, charge, operator, open, superconformal, virasoro, representation
Cosmological sources	cosmic, spectrum, scale, energy, ray, universe, radiation, gravitational, cosmological, power, observation, cmb, background, cosmic ray, cosmic rays
Cosmology and gravity	cosmological, gravity, constant, axion, scale, lorentz, universe, cosmological constant, violation, problem, quantum, vacuum, cosmology, time, planck
Cross-sections in colliders	production, section, cross, collision, energy, lhc, rapidity, process, pair, pp, inclusive, differential, fusion, nuclear, gev
Dark matter (particles and direct searches)	dark matter, matter, dark, dm, particle, detection, direct detection, direct, wimp, relic, relic density, density, annihilation, search, candidate
Dark matter in the universe	dark, matter, dark matter, dark energy, model, abundance, energy, sector, constraint, density, candidate, galaxy, universe, cold, scenario
Decay measurements	decay, state, d, meson, stat, syst, $+/-$, $+-$, fraction, final, final state, width, ratio, $pi+$, final states
Detectors	detector, experiment, physic, beam, high, crystal, nuclear, liquid, performance, precision, resolution, high energy, search, target, chamber
Double-beta decay	mass, baryon, decay, scalar, beta, double beta decay, double, double beta, scale, light, neutrinoless, effective, glueball, gev, hierarchy
Early-universe and other cosmological data	constraint, big bang, big, galactic, signal, cosmic microwave, background, axion, bound, galaxy, bang, microwave, halo, detection, dm
Effective Field Theory	field, effective, theory, effective field theory, effective field, noncommutative, action, effective action, scalar, scalar field, potential, effective theory, effective potential, eft, non-commutative
Electromagnetism	magnetic, field, particle, magnetic field, electric, relativistic, electromagnetic, effect, plasma, moment, energy, medium, magnetic fields, external, electromagnetic field
Events in colliders (kinematics?)	production, collision, jet, tev, lhc, collider, event, transverse, large hadron collider, energy, large hadron, hadron, pair, pp, luminosity
Events in colliders (signatures?)	jet, event, lhc, tev, production, cm, pair, atlas, final state, final, collision, data, luminosity, channel, large hadron collider
Experimental investigation of the leptonic sector	decay, search, data, limit, gamma, collider, muon, gev, measurement, signal, experiment, detector, magnetic moment, event, upper
Experimental jargon	result, mass, effect, large, parameter, energy, value, analysis, small, order, region, current, due, contribution, present
Experiments on light	photon, electron, particle, experiment, mi, laser, compton, optical, mo, beam, light, atom, year, math, pulse
Field theory and gravity	scalar, field, scalar field, mode, gravity, massive, scalar fields, gravitational, potential, massless, perturbation, geodesic, background, metric, spacetime

Continued on next page

Table S1: Most frequent terms for each topic.

Topic (context)	Most frequent expressions
Flavor mixing	cp, violation, asymmetry, mixing, matrix, lepton, cp violation, flavor, standard model, model, quark, phase, standard, angle, mass
Flavour physics	mass, lepton, bound, flavour, flavor, decay, neutrino, heavy, scale, generation, violation, light, quark, coupling, number
Form factors	factor, form, nucleon, electromagnetic, pion, electromagnetic form, electromagnetic form factors, momentum, form factors, result, ratio, ^{2}, transfer, nn, form-factors
Gauge Theory	gauge, theory, action, invariance, field, lorentz, transformation, invariant, brst, yang-mills, symmetry, effective action, lattice gauge, massive, covariant
Gauge symmetry breaking/GUTs	symmetry, gauge, su, model, group, theory, breaking, anomaly, fermion, spontaneous, unification, representation, discrete, symmetric, grand
Gravitons and extra-dimensions	gravity, dimension, scalar, extra, field, constant, brane, cosmological, massive, cosmological constant, extra dimensions, scalar field, bulk, graviton, derivative
Hadronic zoo	state, resonance, d, gev, mev, b, channel, e^+e^-, charmonium, narrow, b, molecule, s1, reaction, e^+
Heavy quarks and ions	quark, heavy, hadron, distribution, collision, production, gluon, hadronic, qcd, heavy quark, heavy ion, charm, correlation, ion, heavy ion collisions
Higgs boson	higgs, boson, model, standard model, mass, standard, coupling, gauge, sector, sm, higgs mass, doublet, higgs boson, neutral, scalar
Higgs sector beyond the SM	higgs, model, standard model, standard, boson, electroweak, supersymmetric, lhc, minimal, supersymmetric standard model, collider, tev, mass, scalar, supersymmetric standard
High-energy source fluxes	energy, flux, source, high energy, spectrum, high, event, signal, emission, time, radiation, solar, information, gravitational wave, such
Holographic Principle and dualities	conformal, dual, holographic, boundary, entropy, cft, entanglement, ad, bulk, defect, theory, conformal field, correspondence, conformal field theory, entanglement entropy
Inflation	inflation, perturbation, universe, inflationary, field, scalar, cosmological, inflaton, cosmology, potential, scalar field, initial, evolution, fluctuation, curvature
Lattice calculation techniques	operator, lattice, matrix, fermion, loop, wilson, theory, element, gauge, function, action, calculation, continuum, expansion, method
Lepton/Meson decay	decay, branching, ratio, semileptonic, meson, fraction, asymmetry, mode, measurement, rate, br, nu, semileptonic decays, inclusive, lifetime
Lie algebra	algebra, space, integral, representation, function, group, operator, invariant, form, path, transformation, lie, differential, product, partition
Loops and higher order expansions in Feynman Diagrams	correction, order, one-loop, term, contribution, radiative corrections, approximation, qed, calculation, loop, radiative, logarithmic, effective, expansion, expression
M-theory and theories of everything	theory, gauge, duality, supergravity, string, dual, action, dimensional, type, background, m-theory, reduction, dimension, abelian, field
Matter in Yang-Mills theories	su, symmetry, fermion, gauge, chiral, mass, model, breaking, coupling, boson, flavor, color, composite, quark, dirac
Measurements and analysis of colliders data	data, measurement, uncertainty, experiment, analysis, experimental, fit, determination, systematic, first, theoretical, error, parameter, detector, current
Meson phenomenology	meson, state, resonance, vector, decay, mass, width, mev, pseudoscalar, pion, amplitude, experimental, channel, quark, wave
Neutrino physics	neutrino, oscillation, mass, experiment, majorana, neutrino mass, right-handed, neutrino oscillations, neutrino oscillation, flavor, interaction, supernova, antineutrino, seesaw, sterile
Non-abelian theories	gauge, field, spin, topological, theory, chern-simons, higher spin, abelian, vortex, non-abelian, gauge field, dirac, term, hall, fermion
Partons distributions	qcd, distribution, parton, next-to-leading order, order, function, nlo, gluon, jet, next-to-leading, correction, transverse, momentum, calculation, perturbative
Perturbative QCD	qcd, perturbative, factorization, anomalous, order, contribution, result, function, approach, perturbative qcd, calculation, anomalous dimension, coefficient, kernel, expansion
Phenomenological jargon	state, new, interaction, coupling, physic, strong, problem, particle, theory, recent, such, bound, model, approach, role
QCD calculation techniques	propagator, expansion, lattice, gluon, effective, finite, loop, theory, potential, qcd, numerical, gauge, perturbative, method, regularization
Quantum Chromodynamics (QCD)	rule, sum, qcd, wall, domain, qcd sum rules, viscosity, qcd sum, quark, heavy, shear viscosity, shear, vacuum, condensate, bubble
Quantum Field Theory	theory, field, quantum, equation, solution, classical, dimension, quantum field, class, quantum field theory, problem, space-time, dimensional, two-dimensional, arbitrary
Quantum Systems and Equations of motion	equation, hamiltonian, constraint, system, term, formalism, charge, monopole, dirac, solution, first, second, kinetic, nonlinear, part
Quantum systems and thermodynamics	system, energy, time, quantum, state, fluctuation, density, gas, dynamic, thermal, temperature, phase, casimir, force, surface
Renormalization	renormalization, group, flow, point, coupling, scale, fixed, uv, rg, ir, cutoff, infrared, fixed point, effective, ultraviolet
Scattering of composite particles	scattering, function, data, proton, structure, nucleon, inelastic, distribution, moment, deep, dipole, q^2, inelastic scattering, hera, target

Continued on next page

Table S1: Most frequent terms for each topic.

Topic (context)	Most frequent expressions
Search for BSM physics	physic, new, new physics, standard model, experiment, standard, neutral, search, tau, measurement, current, decay, future, lepton, rare
Sigma models (?)	model, symmetry, supersymmetric, supersymmetry, sigma, term, integrable, lagrangian, algebra, su, group, chiral, deformation, fermionic, sl
Solar neutrinos	neutrino, oscillation, solar, mixing, solar neutrino, angle, atmospheric, neutrino mass, sterile, atmospheric neutrino, experiment, hierarchy, sterile neutrinos, matrix, sterile neutrino
Space-time geometry and gravity	solution, gravity, spacetime, metric, gravitational, ad, geometry, space, flat, curvature, sitter, singularity, general, dilaton, einstein
Spin/angular momentum/polarization	momentum, polarization, asymmetry, angular, spin, distribution, angular momentum, polarized, reaction, transverse, cross, section, beam, production, photon
States of matter	phase, transition, critical, temperature, point, holographic, spectral, order, exponent, behavior, imaginary, critical point, finite temperature, finite, first order
String theory	string, solution, charge, soliton, branes, configuration, topological, type, monopoles, open, flux, bps, tachyon, background, vortex
Supergravity	supergravity, modulus, manifold, type, space, calabi-yau, supersymmetric, geometry, supersymmetry, moduli space, topological, bps, class, curve, iib
Supersymmetric particles	mass, susy, parameter, soft, neutralino, space, scale, mssm, squark, region, scenario, constraint, gluino, gaugino, large
Supersymmetric theories	theory, gauge, supersymmetric, yang-mills, supersymmetry, anomaly, supergravity, duality, chiral, $n = 4$, super, $n = 2$, super yang-mills, branch, su
Theoretical jargon	model, case, structure, limit, new, term, function, such, number, different, method, particular, property, spectrum, approach
Thermodynamics	phase, temperature, transition, potential, density, chemical, finite, finite temperature, matter, chemical potential, critical, high, thermal, order, first order
Top quark	quark, top, top quark, mass, decay, bound, standard model, top quark mass, coupling, new physics, lepton, top quarks, standard, chiral quark, physic
Topology	space, dimension, modulus, string, bundle, manifold, vacuum, extra, moduli space, heterotic, torus, instanton, singularity, compact, theory

Table S2: PACS categories most correlated to the topics derived with the unsupervised model. Correlation is measured as the mutual pointwise information (pmi).

topic	PACS category	pmi
Algorithms and calculation techniques	Lattice theory and statistics	1.39
	Lattice gauge theory	1.17
	Lattice QCD calculations	1.12
	Particle correlations and fluctuations	0.99
	Inelastic scattering: many-particle final states	0.80
Amplitude of scattering processes	Baryon resonances (S=C=B=0)	1.13
	Pion-baryon interactions	1.10
	Meson-meson interactions	1.03
	Nucleon-nucleon interactions	0.93
	Dispersion relations	0.92
Amplitudes and Feynman Diagram	Analytic properties of S matrix	1.66
	Properties of perturbation theory	1.57
	General properties of perturbation theory	1.39
	Dispersion relations	1.04
	Lattice theory and statistics	0.86
Analyses and measurements from colliders	Neutrino-induced reactions	0.96
	Muons	0.89
	Neutrino, muon, pion, and other elementary particle detectors; cosmic ray detectors	0.81
	Pion-baryon interactions	0.79
	Meson production	0.77
Annihilation and scattering cross-sections	Total cross sections	1.60
	Hadron production in e−e+ interactions	1.23
	Meson production	1.11
	Elastic and Compton scattering	1.07
	Electromagnetic processes and properties	1.03
Astrophysics	Collective flow	1.91
	Hydrodynamic models	1.74
	Particle correlations and fluctuations	1.52
	Relativistic heavy-ion collisions	1.38
	Particle and resonance production	1.35
Black holes	Black holes	2.64
	Quantum aspects of black holes, evaporation, thermodynamics	2.59
	Physics of black holes	2.57
	Classical black holes	2.55
	Higher-dimensional black holes, black strings, and related objects	2.38
Boundary conditions/non-locality	Entanglement and quantum nonlocality	1.18
	Theory of quantized fields	0.90
	Foundations of quantum mechanics; measurement theory	0.80
	Conformal field theory, algebraic structures	0.71
	Integrable systems	0.70
CP violating processes	Decays of bottom mesons	1.53
	Determination of Cabibbo-Kobayashi & Maskawa (CKM) matrix elements	1.48
	Bottom mesons (|B|>0)	1.34

Continued on next page

Table S2: PACS categories most correlated to the topics derived with the unsupervised model. Correlation is measured as the mutual pointwise information (pmi).

topic	PACS category	pmi
	Charge conjugation, parity, time reversal, and other discrete symmetries	1.30
	Decays of bottom mesons	1.19
Chiral symmetry	Chiral Lagrangians	1.55
	Chiral symmetries	1.54
	Lattice QCD calculations	1.48
	Light quarks	1.30
	Lattice gauge theory	1.21
Conformal Field Theory	Conformal field theory, algebraic structures	1.72
	Algebraic methods	1.34
	Nonperturbative techniques; string field theory	1.19
	Lattice theory and statistics	1.15
	M theory	0.99
Cosmological sources	Background radiations	1.86
	Observational cosmology (including Hubble constant, distance scale, cosmological constant, early Universe, etc)	1.55
	Neutrino, muon, pion, and other elementary particles; cosmic rays	1.49
	Dark energy	1.29
	Cosmology	1.21
Cosmology and gravity	Lorentz and Poincaré invariance	1.34
	Loop quantum gravity, quantum geometry, spin foams	1.32
	Axions and other Nambu-Goldstone bosons (Majorons, familons, etc.)	1.30
	Dark energy	1.28
	Quantum cosmology	1.26
Cross-sections in colliders	Total cross sections	1.57
	Inclusive production with identified hadrons	1.43
	Particle and resonance production	1.42
	Production	1.40
	Inclusive production with identified leptons, photons, or other nonhadronic particles	1.36
Dark matter (particles and direct searches)	Dark matter	2.36
	Elementary particle processes	1.94
	Neutrino, muon, pion, and other elementary particle detectors; cosmic ray detectors	1.40
	Neutrino, muon, pion, and other elementary particles; cosmic rays	1.18
	Supersymmetric partners of known particles	1.15
Dark matter in the universe	Dark matter	1.86
	Dark energy	1.69
	Elementary particle processes	1.44
	Observational cosmology (including Hubble constant, distance scale, cosmological constant, early Universe, etc)	1.36
	Cosmology	1.27
Decay measurements	Decays of charmed mesons	1.93
	Decays of bottom mesons	1.91
	Determination of Cabibbo-Kobayashi & Maskawa (CKM) matrix elements	1.83
	Decays of J/ , ϒ, and other quarkonia	1.82
	Bottom mesons (|B|>0)	1.82
Detectors	Neutrino, muon, pion, and other elementary particle detectors; cosmic ray detectors	1.48
	Muons	0.99
	Ordinary neutrinos	0.98
	Neutrino interactions	0.91
	Solar neutrinos	0.87
Double-beta decay	Baryons	1.20
	Charmed baryons (|C|>0, B=0)	1.08
	Glueball and nonstandard multi-quark/gluon states	1.03
	Bottom baryons (|B|>0)	0.99
	Hadron mass models and calculations	0.97
Early-universe and other cosmological data	Background radiations	1.57
	Dark matter	1.38
	Axions and other Nambu-Goldstone bosons (Majorons, familons, etc.)	1.28
	Neutrino, muon, pion, and other elementary particles; cosmic rays	1.27
	Elementary particle processes	1.11
Effective Field Theory	Noncommutative field theory	1.89
	Noncommutative geometry	1.77
	Quantum mechanics	0.85
	Nonlinear or nonlocal theories and models	0.82
	Canonical quantization	0.81
Electromagnetism	Hydrodynamic models	1.45
	Collective flow	1.31
	Electric and magnetic moments	1.16
	Relativistic heavy-ion collisions	1.11
	Relativistic wave equations	1.11
Events in colliders (kinematics?)	Limits on production of particles	1.71
	Production	1.60
	Inclusive production with identified leptons, photons, or other nonhadronic particles	1.57
	W bosons	1.53
	Jets in large-Q2 scattering	1.53
Events in colliders (signatures?)	Limits on production of particles	1.69
	Jets in large-Q2 scattering	1.56
	Production	1.45
	Inclusive production with identified leptons, photons, or other nonhadronic particles	1.37
	W bosons	1.35
Experimental investigation of the leptonic sector	Limits on production of particles	1.38
	Electromagnetic decays	1.30
	Decays of J/ , ϒ, and other quarkonia	1.26
	Decays of J/ , ϒ, and other quarkonia	1.19
	Muons	1.18
Experimental jargon	Electromagnetic corrections to strong- and weak-interaction processes	0.35
	Solar neutrinos	0.30
	Electroweak radiative corrections	0.30
	Nucleon-nucleon interactions	0.29
	Neutrino-induced reactions	0.25
Experiments on light	Specific calculations	1.31
	Elastic and Compton scattering	1.26

Continued on next page

Table S2: PACS categories most correlated to the topics derived with the unsupervised model. Correlation is measured as the mutual pointwise information (pmi).

topic	PACS category	pmi
	Electromagnetic processes and properties	1.09
	Axions and other Nambu-Goldstone bosons (Majorons, familons, etc.)	1.09
	Quantum electrodynamics	1.08
Field theory and gravity	Classical general relativity	1.10
	Modified theories of gravity	1.08
	Lower dimensional models; minisuperspace models	1.06
	Fundamental problems and general formalism	1.05
	Classical black holes	1.02
Flavor mixing	Quark and lepton masses and mixing	1.36
	Flavor symmetries	1.30
	Charge conjugation, parity, time reversal, and other discrete symmetries	1.28
	Determination of Cabibbo-Kobayashi & Maskawa (CKM) matrix elements	1.10
	Neutrino mass and mixing	1.06
Flavour physics	Global symmetries (e.g., baryon number, lepton number)	1.04
	Flavor symmetries	1.03
	Non-standard-model neutrinos, right-handed neutrinos, etc.	1.02
	Unification of couplings; mass relations	1.00
	Quark and lepton masses and mixing	0.99
Form factors	Electromagnetic form factors	1.97
	Relativistic quark model	1.34
	Protons and neutrons	1.33
	Hyperons	1.18
	Sum rules	1.18
Gauge Theory	Gauge field theories	1.20
	Lorentz and Poincaré invariance	1.16
	Canonical formalism, Lagrangians, and variational principles	1.10
	Lagrangian and Hamiltonian approach	1.09
	Noncommutative field theory	1.08
Gauge symmetry breaking/GUTs	Unified theories and models of strong and electroweak interactions	1.34
	Unification of couplings; mass relations	1.26
	Spontaneous breaking of gauge symmetries	1.15
	Unified field theories and models	1.14
	Spontaneous and radiative symmetry breaking	0.96
Gravitons and extra-dimensions	Higher-dimensional gravity and other theories of gravity	1.41
	Gravity in more than four dimensions, Kaluza-Klein theory, unified field theories; alternative theories of gravity	1.39
	Modified theories of gravity	1.34
	Lower dimensional models; minisuperspace models	1.08
	String and brane phenomenology	1.04
Hadronic zoo	Decays of J/ , Υ, and other quarkonia	1.92
	Heavy quarkonia	1.73
	Exotic mesons	1.71
	Decays of J/ , Υ, and other quarkonia	1.65
	Mesons with S=C=B=0, mass > 2.5 GeV (including quarkonia)	1.58
Heavy quarks and ions	Particle and resonance production	1.40
	Particle correlations and fluctuations	1.39
	Collective flow	1.38
	Relativistic heavy-ion collisions	1.37
	Fragmentation into hadrons	1.29
Higgs boson	Other neutral Higgs bosons	1.91
	Supersymmetric Higgs bosons	1.87
	Non-standard-model Higgs bosons	1.77
	Extensions of electroweak Higgs sector	1.73
	Standard-model Higgs bosons	1.69
Higgs sector beyond the SM	Other neutral Higgs bosons	1.65
	Supersymmetric Higgs bosons	1.64
	Non-standard-model Higgs bosons	1.60
	Extensions of electroweak Higgs sector	1.55
	Standard-model Higgs bosons	1.37
High-energy source fluxes	Neutrino, muon, pion, and other elementary particles; cosmic rays	1.39
	Neutrino, muon, pion, and other elementary particle detectors; cosmic ray detectors	1.33
	Solar neutrinos	1.28
	Background radiations	0.89
	Ordinary neutrinos	0.74
Holographic Principle and dualities	Entanglement and quantum nonlocality	1.89
	Gauge/string duality	1.53
	Conformal field theory, algebraic structures	1.43
	Higher-dimensional black holes, black strings, and related objects	1.06
	Quantum aspects of black holes, evaporation, thermodynamics	1.02
Inflation	Particle-theory and field-theory models of the early Universe (including cosmic pancakes, cosmic strings, chaotic phenomena, inflationary universe, etc.)	1.80
	Origin and formation of the Universe	1.78
	Observational cosmology (including Hubble constant, distance scale, cosmological constant, early Universe, etc)	1.76
	Background radiations	1.70
	Quantum cosmology	1.67
Lattice calculation techniques	Lattice QCD calculations	1.38
	Lattice gauge theory	1.36
	Lattice theory and statistics	0.80
	General properties of perturbation theory	0.76
	Renormalization	0.74
Lepton/Meson decay	Determination of Cabibbo-Kobayashi & Maskawa (CKM) matrix elements	1.97
	Decays of charmed mesons	1.94
	Decays of bottom mesons	1.89
	Decays of charmed mesons	1.86
	Bottom mesons (|B|>0)	1.81
Lie algebra	Algebraic methods	1.39
	Integrable systems	1.28
	Geometry, differential geometry, and topology	1.19
	Noncommutative geometry	1.03
	Quantum mechanics	0.94

Continued on next page

Table S2: PACS categories most correlated to the topics derived with the unsupervised model. Correlation is measured as the mutual pointwise information (pmi).

topic	PACS category	pmi
Loops and higher order expansions in Feynman Diagrams	Electromagnetic corrections to strong- and weak-interaction processes	1.32
	Electroweak radiative corrections	1.23
	Specific calculations	1.08
	Summation of perturbation theory	1.00
	General properties of perturbation theory	0.98
M-theory and theories of everything	M theory	1.63
	Supergravity	1.34
	Nonperturbative techniques; string field theory	1.27
	Compactification and four-dimensional models	1.22
	D branes	1.13
Matter in Yang-Mills theories	Technicolor models	1.23
	Unified theories and models of strong and electroweak interactions	1.06
	Unification of couplings; mass relations	0.99
	Composite models	0.94
	Spontaneous breaking of gauge symmetries	0.88
Measurements and analysis of colliders data	Determination of Cabibbo-Kobayashi & Maskawa (CKM) matrix elements	0.90
	Solar neutrinos	0.87
	Muons	0.84
	Neutrino, muon, pion, and other elementary particle detectors; cosmic ray detectors	0.75
	Decays of charmed mesons	0.73
Meson phenomenology	Other mesons with S=C=0, mass < 2.5 GeV	1.53
	Hadron mass models and calculations	1.48
	Meson-meson interactions	1.45
	Mesons	1.41
	Glueball and nonstandard multi-quark/gluon states	1.37
Neutrino physics	Ordinary neutrinos	2.04
	Solar neutrinos	1.98
	Non-standard-model neutrinos, right-handed neutrinos, etc.	1.97
	Neutrino mass and mixing	1.94
	Neutrino, muon, pion, and other elementary particles; cosmic rays	1.92
Non-abelian theories	Gauge field theories	1.04
	Magnetic monopoles	1.03
	Canonical formalism, Lagrangians, and variational principles	0.97
	Lagrangian and Hamiltonian approach	0.88
	Noncommutative field theory	0.87
Partons distributions	Summation of perturbation theory	1.62
	Factorization	1.49
	Production	1.46
	Jets in large-Q2 scattering	1.44
	Perturbative calculations	1.43
Perturbative QCD	Factorization	1.17
	Summation of perturbation theory	1.10
	Perturbative calculations	1.03
	Production	0.66
	Heavy quark effective theory	0.65
Phenomenologica l jargon	Foundations of quantum mechanics; measurement theory	0.34
	Axions and other Nambu-Goldstone bosons (Majorons, familons, etc.)	0.31
	Loop quantum gravity, quantum geometry, spin foams	0.30
	Experimental tests of gravitational theories	0.29
	Potential models	0.27
QCD calculation techniques	Gluons	1.29
	General properties of perturbation theory	1.02
	Renormalization	0.96
	General properties of QCD (dynamics, confinement, etc.)	0.94
	Lattice gauge theory	0.89
Quantum Chromodynamics (QCD)	Sum rules	2.24
	Other nonperturbative calculations	1.42
	Bottom baryons (|B|>0)	1.32
	Charmed baryons (|C|>0, B=0)	1.26
	Heavy quark effective theory	1.16
Quantum Field Theory	Foundations of quantum mechanics; measurement theory	1.32
	Quantum mechanics	1.15
	Algebraic methods	1.06
	Canonical quantization	0.97
	Theory of quantized fields	0.95
Quantum Systems and Equations of motion	Canonical formalism, Lagrangians, and variational principles	1.23
	Magnetic monopoles	1.15
	Lagrangian and Hamiltonian approach	1.11
	Relativistic wave equations	1.04
	Canonical quantization	1.00
Quantum systems and thermodynamics	Hydrodynamic models	1.21
	Theory of quantized fields	0.96
	Foundations of quantum mechanics; measurement theory	0.93
	Entanglement and quantum nonlocality	0.90
	Quark-gluon plasma	0.75
Renormalization	Renormalization group evolution of parameters	1.77
	Renormalization	1.46
	General properties of perturbation theory	0.85
	Technicolor models	0.85
	Other nonperturbative techniques	0.81
Scattering of composite particles	Total and inclusive cross sections (including deep-inelastic processes)	1.78
	Photon and charged-lepton interactions with hadrons	1.65
	Elastic and Compton scattering	1.49
	Regge theory, duality, absorptive/optical models	1.35
	Polarization in interactions and scattering	1.32
Search for BSM physics	Muons	1.12
	Decays of K mesons	1.09
	Decays of taus	1.09
	Neutrino, muon, pion, and other elementary particle detectors; cosmic ray detectors	1.07
	Neutral currents	1.05
Sigma models (?)	Integrable systems	1.74

Continued on next page

Table S2: PACS categories most correlated to the topics derived with the unsupervised model. Correlation is measured as the mutual pointwise information (pmi).

topic	PACS category	pmi
	Algebraic methods	1.23
	Supersymmetry	1.09
	Lattice theory and statistics	1.00
	Conformal field theory, algebraic structures	0.99
Solar neutrinos	Solar neutrinos	2.64
	Ordinary neutrinos	2.30
	Neutrino mass and mixing	2.13
	Non-standard-model neutrinos, right-handed neutrinos, etc.	1.98
	Neutrino, muon, pion, and other elementary particles; cosmic rays	1.89
Space-time geometry and gravity	Exact solutions	1.75
	Classical general relativity	1.57
	Einstein-Maxwell spacetimes, spacetimes with fluids, radiation or classical fields	1.53
	Classical black holes	1.51
	Higher-dimensional black holes, black strings, and related objects	1.51
Spin/angular momentum/polarization	Polarization in interactions and scattering	1.80
	Photon and charged-lepton interactions with hadrons	1.47
	Fragmentation into hadrons	1.41
	Inclusive production with identified hadrons	1.35
	Meson production	1.21
States of matter	Quark deconfinement, quark-gluon plasma production, and phase transitions	1.09
	Finite-temperature field theory	1.08
	Gauge/string duality	1.02
	Lattice theory and statistics	0.90
	Quark matter	0.84
String theory	D branes	1.86
	Magnetic monopoles	1.71
	Nonperturbative techniques; string field theory	1.67
	Extended classical solutions; cosmic strings, domain walls, texture	1.52
	Strings and branes	1.46
Supergravity	M theory	1.62
	Supergravity	1.58
	Compactification and four-dimensional models	1.51
	Nonperturbative techniques; string field theory	1.37
	Geometry, differential geometry, and topology	1.30
Supersymmetric particles	Supersymmetric partners of known particles	1.68
	Supersymmetric models	1.35
	Supersymmetric Higgs bosons	1.27
	Unification of couplings; mass relations	0.85
	Non-standard-model Higgs bosons	0.82
Supersymmetric theories	Supersymmetry	1.37
	M theory	1.35
	Supergravity	1.20
	Nonperturbative techniques; string field theory	1.05
	Gauge field theories	1.05
Theoretical jargon	Integrable systems	0.36
	Quantum mechanics	0.36
	Foundations of quantum mechanics; measurement theory	0.33
	Algebraic methods	0.31
	Fundamental problems and general formalism	0.28
Thermodynamics	Quark deconfinement, quark-gluon plasma production, and phase transitions	1.62
	Quark matter	1.61
	Finite-temperature field theory	1.57
	Quark-gluon plasma	1.35
	Other models for strong interactions	1.11
Top quark	Top quarks	1.96
	Neutral currents	1.20
	Limits on production of particles	1.07
	Other neutral Higgs bosons	0.98
	Other gauge bosons	0.97
Topology	Compactification and four-dimensional models	1.40
	Geometry, differential geometry, and topology	1.31
	Nonperturbative techniques; string field theory	1.20
	M theory	1.11
	Strings and branes	1.04

S3.5 Topics and their correlation with categories

Below, we evaluate how topics compare with the classification of the literature. For that, we generated a 2D representation of the semantic space by applying a t-SNE transformation (van der Maaten & Hinton, 2008) on the distance matrix $1 - R_{ij}$, where R_{ij} is the correlation matrix for the 75 topics from the CTM. The t-SNE transformation aims to reduce dimensionality (from 75 to 2) while preserving distances, such that highly correlated topics should appear close to each other on the resulting 2D map. We then colored each topic according to the category (among theory, phenomenology and experiment) that has the strongest association (normalized pointwise mutual information) with this topic. The graph was then rotated such that the x-axis would explain most of the variance in these three categories. Topics related to supersymmetry were emphasized and labeled. The resulting map is shown in Figure S2.

Although the t-SNE transformation does not yield very stable results, it generally appears (as in this figure) that topics most associated with a given category (e.g. theory) appear closer to each other, such that these three categories explain part of the variance in the semantic space. Second, in this representation, the distinction between phenomenological supersymmetry and theoretical supersymmetry is supported by the emergence of two separate clusters of supersymmetry-related topics.

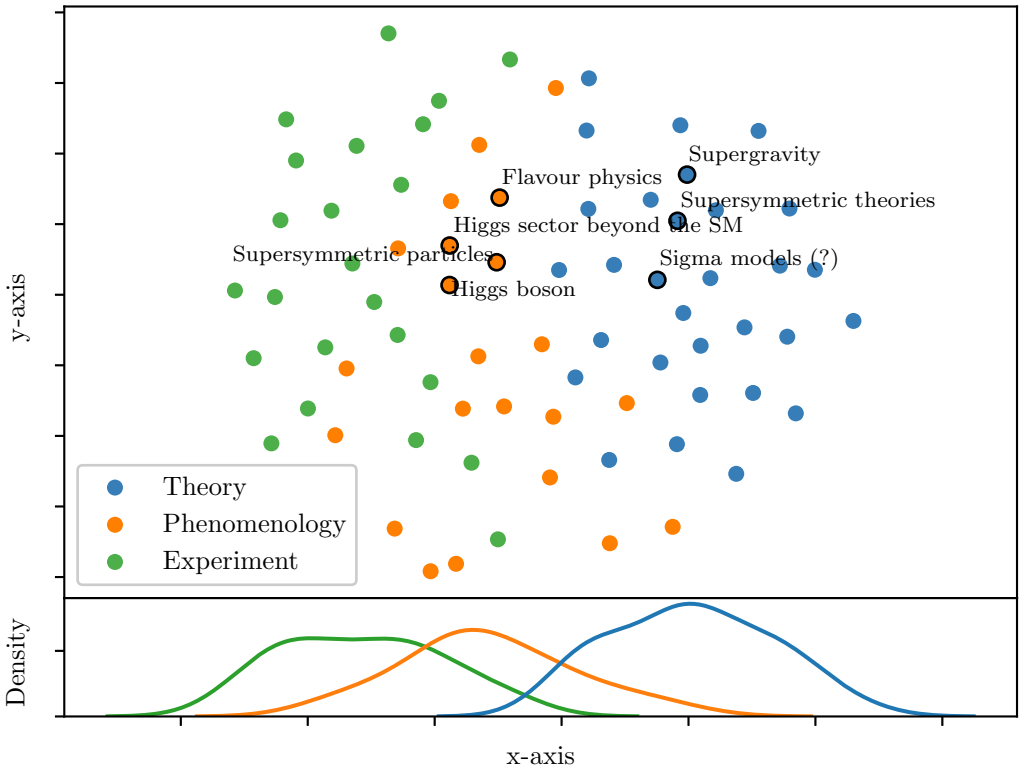


Fig. S2 Semantic map extracted from the topic model, after applying a t-SNE transformation. Each dot represents a topic. Each topic is assigned the category, among theory, phenomenology and experiment, that is most associated with it. Correlated topics appear closer to each other. For each category, the density of topics along the x-axis is shown in the lower plot.

S4 Validity of the citation network for exploring the trading zone

Below, we support the relevance of the citation network as a means of exploring trading zones between scientific cultures by showing it can be used to recover known facts, in particular i) that theory and experiment in HEP do not communicate directly and ii) that phenomenology channels most exchanges across them.

We build a citation network where each node is one paper of the literature and the edge between nodes x and y is assigned a weight $w_{x,y} = 1$ if x cites y and 0 otherwise. From this we can define the amount of citations of papers from the category i to a paper from the category j as:

$$n_{ij} = \sum_{x \in i, y \in j} \frac{w_{xy}}{(\sum_c \mathbb{1}_c(x))(\sum_c \mathbb{1}_c(y))} \tag{2}$$

Where $\mathbb{1}_c(x) = 1$ if x belongs to $c \in \{\text{Experiment, Phenomenology, Theory}\}$, and 0 otherwise. We then normalize n_{ij} by the amount of citations *from* category i, thus yielding the normalized matrix $\tilde{n}_{ij}$. By construction, $0 \leq \tilde{n}_{ij} \leq 1$ is the effective fraction of references from papers of category i to papers of category j. The matrix is built from the citation network between 2001 and 2019. We then verify that $\tilde{n}_{ii}$ is high (papers mostly cite papers from the same category); and that for cross-culture citations ($i \neq j$), $\tilde{n}_{ij} \ll 1$ unless i or j is "phenomenology"; i.e., "trading zones" in the field occur around phenomenology. Evaluating the fraction of citations from papers of a category i that target papers from a category j yields the matrix in Figure S3. In this matrix, borrowing the trade metaphor from Yan et al. (2013), non-diagonal elements represent "imports" (references to publications from other subcultures) and diagonal elements measure the "self-dependence" of each subculture. The results confirm that most citations occur within categories, emphasizing the relative autonomy of each of these subcultures including phenomenology – it is less obvious for experimental papers, which are much more scarce then the others, and cannot cite themselves as much. Moreover the results confirm that most trades involve phenomenology: cross-citations between purely theoretical and experimental papers are very rare (∼1% of their references). Overall, "theory" is highly self-reliant.

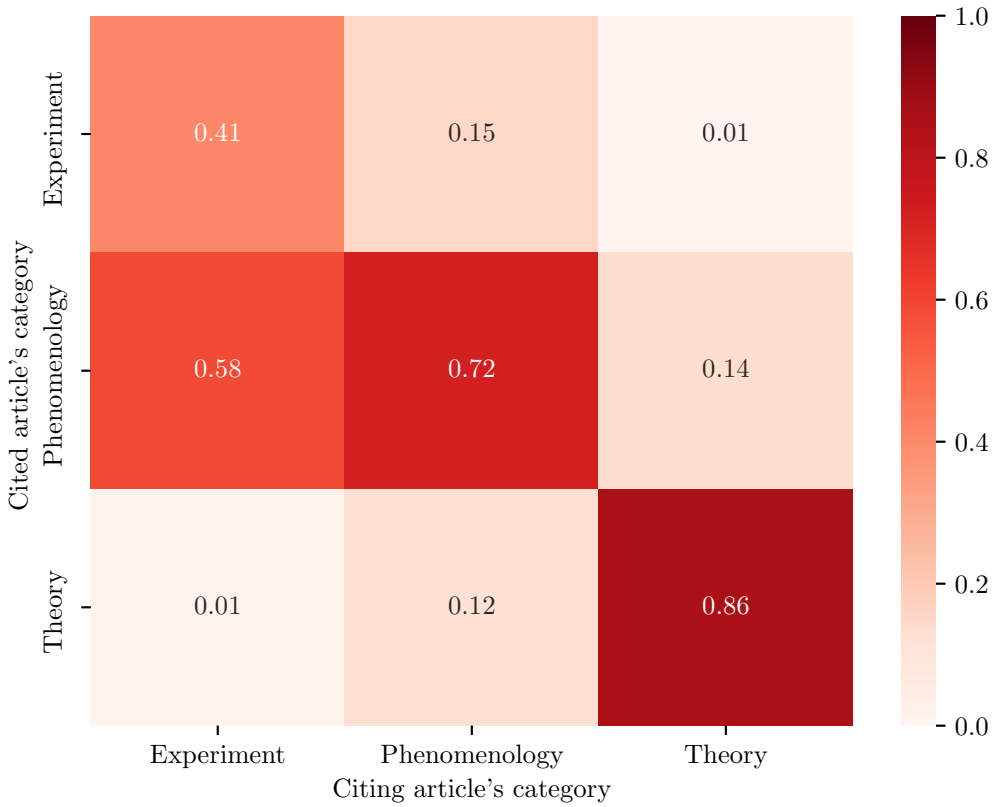


Fig. S3 Origin of the references (citations) in the HEP literature Each matrix element $\tilde{n}_{ij}$ represents the fraction of references from the x-axis category (columns) that target papers from the y-axis category (lines). For instance, 41% of references in experimental papers refer to experimental papers. 15% of citations from phenomenological papers refer to experimental papers. If these categories were completely hermetic, the matrix would equal the identity matrix, which is not the case.

References

Bab2min, Fenstermacher, D., & Schneider, J. (2021). Python package of tomoto, the topic modeling tool. Zenodo.

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237–3243.

Bennett, A., Misra, D., & Than, N. (2021). Have you tried Neural Topic Models? Comparative Analysis of Neural and Non-Neural Topic Models with Application to COVID-19 Twitter Data.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*(Jan), 993–1022.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models, In *Neural information processing systems*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J. L., & Resnik, P. (2021). Is automated topic model evaluation broken? the incoherence of coherence (A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan, Eds.). In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems*.

Moskovic, M. (2021). The INSPIRE REST API. Zenodo.

Omodei, E. (2014). *Modeling the socio-semantic dynamics of scientific communities* (Thesis). Ecole Normale Supérieure.

van der Maaten, L. J., & Hinton, G. E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why Priors Matter (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta, Eds.). In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems*, Curran Associates, Inc.

Yan, E., Ding, Y., Cronin, B., & Leydesdorff, L. (2013). A bird's-eye view of scientific trading: Dependency relations among fields of science. *Journal of Informetrics*, *7*(2), 249–264.

Chapter 2

Balancing specialization and adaptation in a transforming scientific landscape

Gautheron *EPJ Data Science* (2025) 14:2
https://doi.org/10.1140/epjds/s13688-024-00516-8
EPJ Data Science
a SpringerOpen Journal
EPJ.org
RESEARCH Open Access

Balancing specialization and adaptation in a transforming scientific landscape

Lucas Gautheron[1,2*]

*Correspondence:
lucas.gautheron@gmail.com
[1]Interdisciplinary Centre for Science and Technology Studies (IZWT), University of Wuppertal, Wuppertal, Germany
[2]Département d'Études Cognitives, École Normale Supérieure, Paris, France

Abstract

How do scientists navigate between the need to capitalize on their prior knowledge through specialization, and the urge to adapt to evolving research opportunities? Drawing from diverse perspectives on adaptation, this paper proposes an unsupervised Bayesian approach motivated by Optimal Transport of the evolution of scientists' research portfolios in response to transformations in their field. The model relies on 186,162 scientific abstracts and authorship data to evaluate the influence of intellectual, social, and institutional resources on scientists' trajectories within a cohort of 2108 high-energy physicists between 2000 and 2019. Using Inverse Optimal Transport, the reallocation of research efforts is shown to be shaped by learning costs, thus enhancing the utility of the scientific capital disseminated among scientists. Two dimensions of social capital, namely "diversity" and "power", have opposite associations with the magnitude of change in scientists' research interests: while "diversity" is associated with greater change and expansion of research portfolios, "power" is associated with more stable research agendas. Social capital plays a more crucial role in shifts between cognitively distant research areas. More generally, this work suggests new approaches for understanding, measuring and modeling collective adaptation using Optimal Transport.

Keywords: Adaptation; Specialization; Science of science; Cultural evolution; Computational social science; Optimal transport

1 Introduction

Scientists are subject to conflicting incentives. On the one hand, they must work within the realm of their expertise, where they can most effectively exploit their prior knowledge and compete with peers; this conservative preference for familiar research topics is at the root of specialization. On the other hand, scientists are simultaneously compelled to revise their research interests to engage with more promising research areas in order to benefit from more exposure or to secure funding. Thus, in some instances, specialization is at odds with the need to adapt to the decline of certain research opportunities and the growth of new ones. How do scientists navigate the trade-off between specialization (i.e. the concentration of their intellectual resources within a narrow cognitive range) and adaptation (i.e. the need to adjust these resources to new realities)? This conflict differs from the "essential tension" between "tradition" and "innovation" proposed by Kuhn [1], or that between "exploration" and "exploitation" [2], which have both been explored quan-

Springer

titatively in previous works [3–9]. First, "adaptation" is not tantamount to innovation or disruption, for it can be a conformist move (e.g. as a result of a bandwagon effect [10]). Moreover, unlike "exploration", adaptation is not identical to a search strategy in a static landscape [11], but rather the convergence towards a new state more congruent with current realities. Disruptions due to breakthroughs in Machine Learning or challenges due to climate change urge to understand how scientists adapt to changing circumstances. Therefore, the present paper investigates scientists' responses to changes in their field (whether driven by epistemic or institutional factors), and the effect of their capital (intellectual, social, or institutional) on their ability to adapt. Drawing insights on adaptation from cultural evolution and institutional change, we develop an unsupervised Bayesian approach to analyze changes in scientists' research agenda while measuring the effect of "capital" (intellectual or social) on their individual trajectories. The model is applied to a cohort of high-energy physicists between the years 2000 and 2019, a time during which the historical driver of progress in the field – particle accelerators – have been contested by emerging astrophysical experiments, thus transforming the landscape of opportunities.

Our approach reveals trends in the field: the boom of dark matter research – fueled by shifts away from the physics of neutrinos and the electroweak sector – and the partial disintegration of string theory into the study of black holes and holography/dualities. More importantly, this analysis also shows that changes in scientists' research portfolios are shaped by learning costs, as scientific communities adapting to new circumstances address an "Optimal Transport" problem by reallocating research efforts efficiently. Optimal Transport is a mathematical framework initially concerned with the optimal displacement and allocation of resources [12–14], and has since then found wide-ranging applications. We show that it also provides a characterization of scientists' behavior, as driven by the need to maximize the utility of their scientific capital under changing circumstances. Moreover, the comparative analysis shows that two dimensions of social capital, namely "diversity" and "power" [15], have opposite associations with change. While "diversity" of social capital – the extent to which scientists have access to diverse cognitive resources via their collaborators – is correlated with greater change and further diversification of scientists' research interests, "power" – roughly speaking, the size of their network – is associated with more stability in their research interests. Social capital has a stronger association with transfers between research areas that are more cognitively "distant". There is no discernible effect of institutional stability after controlling for academic age (although affiliation data is a bit noisy in the dataset). Overall, we contribute: i) a conceptual account of the features of change in scientists' research interests; ii) a novel methodological approach that introduces a model of scientists' trajectories connected to Optimal Transport and measures of intellectual capital, social capital, diversity, and power; and finally, iii) some empirical evidence from high-energy physics. More generally, this paper addresses the relative lack of empirical works within the body of literature that investigates science as a cultural evolutionary system [16]. It demonstrates that Optimal Transport provides an insightful description of certain aspects of collective adaptation, but also computational tools (such as Probabilistic Inverse Optimal Transport [17] and OT based measures of change) for measuring adaptive behavior, and more generally, mobility in physical and abstract spaces.

In what follows, Sects. 1.1 and 1.2 summarize previous research and lay out the conceptual background on which the analysis rests, and Sect. 1.3 introduces the context of

high-energy physics to which the model is applied. Section 2 elaborates the methodology: the data 2.1, the topic model approach for measuring authors' research portfolios 2.2, the proposed measures of intellectual and social capital 2.3, and the model of scientists' trajectories 2.4. Section 3 presents the results: i) the transfers of attention from one research area to another due to changes in the scientific landscape; ii) the structuring role of learning costs in the observed patterns of adaptation 3.1, and iii) the effect of capital (intellectual and social) and institutional stability on physicists' strategies 3.2.

1.1 Empirical background

Several works have investigated the evolution of scientists' research interests. For instance, by mapping the trajectories of 103,246 physicists over 26 years using the American Physical Society (APS) dataset and its topic classification (the Physics and Astronomy Classification Scheme® (PACS)), Aleta et al. [5] demonstrated that a majority of physicists gradually migrate to entirely different topics by the end of their careers while often staying within the same general area. They reveal differences between subfields of physics, such that "exploitation" (i.e. specialization, as opposed to the "exploration" of new topics) is especially prevalent in particle physics. Using the same data, Jia et al. [4] instead find an exponentially decaying distribution of changes in scientists' interests. Previous works generally agree, however, on the graduality of change in research topics [4–6], as previously observed by Gieryn [18]. Recognizing that scientists typically investigate several research questions in parallel, Gieryn proposed four mechanisms of gradual change, including "accretion" (a problem is added to their "problem set"), "selective substitution" (one problem is replaced by another), and "selective disengagement" (one problem is neglected).

While [4–6] document the structure of changes in scientists' interests, they do not relate these transformations to changes in epistemic and institutional context, or to the scientists' incentives and resources. Tripodi et al. [7] have taken a step in this direction. Using the APS dataset, they show that physicists are more likely to explore areas to which they are connected via their collaborators, and highlight the crucial importance of collaborations in the expansion of research portfolios, especially for the exploration of research areas distant from one's core specialization. However, their work does not primarily address the *transformations* of scientists' research portfolios throughout time – they do not quantify "change" –, and they recognize the need for further longitudinal analyses. Finally, previous works have explored the connection between spatial mobility patterns and scientific mobility using gravity or radiation models [8, 9]. In particular, [8] used such methods to compare the characteristics of two types of scientists, "explorers" as opposed to "exploiters".

The present paper complements previous works on changes in scientists' research in several ways. First, since the focus is on adaptation strategies, the core of our approach is both comparative (as in [7], and unlike other previous works) *and* longitudinal (unlike [7], although their paper includes longitudinal robustness checks). Second, this contribution evaluates previously unexplored aspects, such as the choice between expansion or consolidation of research portfolios and the effect of affiliation stability. Third, this work relates the findings to the epistemic context of the field and its transformations by performing the analysis at a circumscribed scale (high-energy physics). Fourth, this work does *not* rely primarily on the APS dataset and PACS categories, on which most previous works depend [4, 5, 7, 19], or any other pre-existing classification of the literature. Research areas are

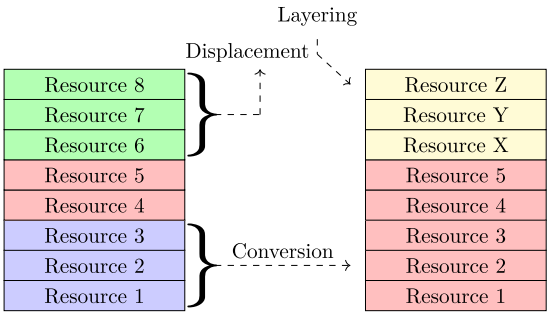


Figure 1 Changes in a scientist's "research portfolio" over time. Colors designate research areas. Resources entail any intellectual or methodological assets that a scientist uses to investigate problems in each area. "Conversion" repurposes knowledge to new goals. "Displacement" is the replacement of certain research interests with little or no transfer of prior knowledge, as illustrated by the green area. Layering is the introduction of new research interests via the addition of new knowledge

clustered using an unsupervised topic model, such that this approach measures linguistic change, which is arguably a more direct proxy of cognitive change. Fifth, this paper is the first application of the Probabilistic Inverse Optimal Transport approach from [17], which provides an alternative to other approaches to mobility (e.g. gravity models). Finally, the proposed approach is grounded in theory, by operationalizing concepts such as capital [20, 21], and by exploiting theoretical insights from diverse approaches to "adaptation".

1.2 Conceptual framework

A central dilemma of adaptation consists in choosing which resources to leverage among those already available (although those may be suboptimal or irrelevant under new circumstances) and which resources to abandon and replace with others (which may be inefficiently costly). By adapting gradually, scientists can strategically retain the benefits of "problem retention" (e.g., the exploitation of "accumulated skills and resources" in one area, or "of an established research network", [18, p. 106]) while progressively investing resources in new research directions. This is illustrated in Fig. 1, which represents the research portfolios of one scientist during two distinct time periods. Cells indicate the resources exploited by the scientist (e.g., concepts, models, methods, etc.) and colors indicate to what problem areas this knowledge is applied. Figure 1 shows how scientists can enter new research areas by repurposing certain resources to new ends [22, 23]. We call this strategy "*conversion*", in reference to the typology of incremental institutional change proposed by Mahoney et al. [24][1].

Not all knowledge can be successfully applied to new research areas: as illustrated in Fig. 1, entering new research areas typically requires "*layering*",[2] that is, the introduction of new concepts, models, or methods, on top of prior knowledge. The acquisition of knowledge entails learning costs, which can be partially avoided by collaborating with experts in the target domain [7]. Another mode of change is *displacement*, when the replacement of one research area for another involves significant neglect of prior knowledge. This may

[1] Indeed, as shown in previous works on the transformations of high-energy physics facilities to photon science instruments [25–27], historical institutionalism can account for gradual adaptations with large cumulative effects taking place in response to scientific and technological change [28]. In this paper, we apply the typology of change to *individuals* rather than organizations.

[2] Again, borrowing the terminology from historical institutionalism.

not be the preferred strategy, since it fails to take advantage of accumulated resources. However, certain knowledge may not apply to a new context, or sometimes there might be reasons to suspend a line of research in order to focus on more promising topics. Overall, we expect that these transformations will manifest themselves as changes in scientists' linguistic behavior, i.e., as changes in the vocabulary of their publications. Generally, we expect an important amount of continuity in linguistic behavior, given the need to minimize cognitive learning costs by capitalizing on prior knowledge.

Scientists manage two kinds of assets when navigating the trade-off between specialization and adaptation: their own prior expertise, and the expertise to which they have access through their social network. Both constitute "capital" [21], i.e. assets that individuals accumulate and leverage in the competitive context of their field. "Capital" (whether "economic", "cultural", "social" or even "symbolic", cf. [21]) defines the scope of scientists' opportunities and therefore their ability to adapt. This paper considers the *intellectual capital* possessed by scientists in the form of scientific knowledge, and *social capital*. Measures that represent these concepts will be proposed, and their effect on the magnitude of transfers of attention across research areas will be evaluated. Emphasis will be put on the divide underlined in [15] between two dimensions of social capital, namely "power" (roughly speaking, network size in the present paper) and "diversity" (of cognitive resources). Group diversity is generally recognized as a factor of adaptation in an evolving environment or in the context of collective problem-solving [29–32]. "Power" is also plausibly associated with higher abilities.

While capital defines scientists' opportunities, it is not sufficient to explain *why* scientists *do* turn to new research areas or not, which also requires understanding actors' incentives and why they must respond to these incentives. Consequently, the present paper also considers the effect of institutional stability and academic age on migrations between research areas, since the need to respond to changes in the epistemic and institutional environment is presumably different for, say, tenured physicists versus postdocs. Moreover, we may assume that younger and older generations play different roles in cultural change and collective adaptation in general [33]. Finally, the effect of productivity will also be considered.

1.3 The case of high-energy physics: navigating a changing epistemic landscape

High-energy physics is a prime example for investigating adaptation in a transforming scientific landscape. As this field relied on the input of increasingly large particle colliders to achieve progress, it has accumulated considerable capital directed towards collider physics in the form of large infrastructure and complex knowledge. These efforts culminated in 2010 with the start of the Large Hadron Collider (LHC), the largest accelerator ever. However, the LHC has found no evidence for anything that was not already predicted by the Standard Model of particle physics, and it is increasingly plausible that no future accelerator could ever find any evidence for new particles, leading to a situation of "crisis" [34]. Although the LHC will continue to take data for years and plans for successors are being discussed [35], some have speculated that particle physics as we know it has come to an end [36, 37], "the proscenium [being] captured by astrophysics and cosmology" instead [37]. However, according to physicist Mikhail Shifman the "pause in accelerator programs we are witnessing now is not necessarily [...] the end of explorations at [high energies]"; instead, such explorations "will continue, perhaps in a new form, with novel devices" [38].

Indeed, new experimental opportunities have emerged in parallel, including gravitational waves astronomy (since 2015 [39]), searches for dark matter of astrophysical origin in underground facilities, and more precise observations of the cosmic microwave background (see Fig. 10, Appendix A.1). Morever, astrophysics seems to be increasingly replacing particle colliders in citations across experimental and theoretical high-energy physics [40]. The present paper investigates how high-energy physicists have adapted in reaction to these transformations.

2 Methods

2.1 Data

Our source is the Inspire HEP database [41]. It aggregates High-Energy Physics (HEP) literature from various sources, including the main scientific publishers and arXiv, and has been used in a few works [40, 42–45]. For the literature on HEP, it is more comprehensive than the often used APS dataset which is limited to a few journals.[3] Moreover, it implements both automatized and manual measures for the disambiguation of author names,[4] thus allowing careers' analyses [44] (nevertheless, occasional misidentifications remain possible). The database also contains data on experiments; consequently, the evolution of the landscape of experimental opportunities can be retrieved (see Fig. 10, Appendix A.1).

The analysis includes all papers from the categories "Theory-HEP" and "Phenomenology-HEP" (inspired from arXiv's categories "hep-th" and "hep-ph"), to which most HEP publications belong, which amounts to $D = 186{,}162$ articles between 2000 and 2019. The minority of purely experimental high-energy physics publications are excluded: such papers are typically authored by thousands of collaborators, and authorship data provide no information about individual experimentalists' specialization. Therefore, this paper documents how theorists and phenomenologists have adapted to the changes outline above.

For the longitudinal analysis, two time periods are considered. An initial phase (2000-2009) is used to infer a reference "research agenda" for each physicist in the cohort, as well as their intellectual and social capital. A late phase (2015-2019) is used to measure how each physicist's research agenda has shifted in comparison to the initial time period, in the context of the changes outlined above. The five-year gap between these two periods allows to measure the cumulative effect of the transformations in the scientific landscape that have unfolded gradually between 2010 and 2015 (had they been sudden, we would not have introduced such a wide gap – see Fig. 10, Appendix A.1), together with the effect of the capital accumulated *prior* to these transformations. Only physicists with ≥ 5 publications during each time period (2000 to 2009, and 2015 to 2019) are included, resulting in a cohort of $N = 2108$ physicists. This study therefore considers physicists that have remained dedicated to high-energy physics, thus revealing adaptation and "survival" strategies *within* HEP, excluding authors that exited the field. This author inclusion rule excludes scientists who publish very irregularly; however, although scientists who continuously publish are a minority, they make up most of the publications in their field.[5] We do

[3] https://journals.aps.org/datasets.

[4] Besides the use of "advance algorithms" of author-disambiguation, Inspire invites scientists to correct their own publication record on the website (https://twiki.cern.ch/twiki/pub/Inspire/WebHome/INSPIRE_background.pdf, June 2014).

[5] Less than 1% of scientists active in the years 1996 to 2011 have published every year during this period, and yet they are responsible for 47% of the publications [46]; the 13% of physicists with ≥ 16 publications between 1985 and 2009 account for 82% of publications [9].

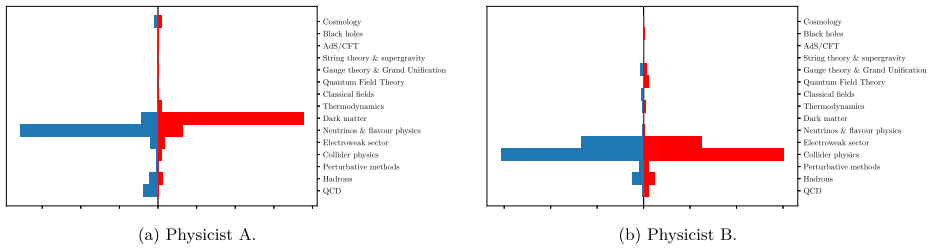
not seek "representativeness", but rather to achieve enough variance to uncover patterns among this particular cohort of productive high-energy physicists. The median academic age of the cohort was 23 years in 2015. 49% of these physicsts have had an affiliation that spanned the entire time period (see Appendix A.2, Fig. 11).

2.2 Measuring research portfolios

Research portfolios are evaluated in terms of the distribution of keywords (n-grams) that belong to each research area within the scientists' publications' abstracts. Instead of reyling on PACS categories and citation data as in most previous works, research areas are extracted with a topic model that recovers latent "topics" within the corpus and their vocabulary distributions, while directly classifying keywords into separate research areas. Arguably, texts provide a more direct access to the kinds of knowledge leveraged by a scientist in their publications: linguistic flexibility implies cognitive flexibility and a low commitment to a specific body of knowledge. Moreover, this approach is applicable to a wider range of situations for which only textual data (even short texts) are available. For instance, PACS codes are only sparsely available in the recent HEP literature. A coarse-grained classification of the literature into $K_0 = 20$ broad "topics" is performed. The number of topics is always somewhat arbitrary, but that topic models give some control over the cognitive scale is a feature, rather than a bug:[6] as Gieryn [18] puts it, "in such analyses [of problem change], empirical findings will in part reflect the defined scope of problem areas", which is itself arbitrary. In our case, we would ideally like our clustering to be just fine-grained enough to measure the impact of the shifts in the landscape of experimental opportunities that we are interested in (the start of the LHC, the rise of new probes of dark matter and black holes, etc.). In this respect, $K_0 = 20$ turns out to be just sufficient to discern the effects of the transformations in HEP discussed in Sect. 1.3 as well as the observed evolution in the popularity of various kinds of experiments shown in Appendix A.1, Fig. 10. Additional models were trained for robustness assessment, setting different values for K_0 (15, 20, 25). More coarse-grained models (using lower values of K_0) are typically less able to observe fine-grained patterns of adaptation to changing experimental opportunities, and the initial $K_0 = 20$ model is better at distinguishing black hole phenomenology from cosmological phenomenology than the most fine-grained model (see Appendix A.3.5, Fig. 14).

We use an embedding model [47], a recent and straightforward approach that relies on pretrained embeddings representations of the n-grams and provides more reliable classifications for heavy-tailed vocabulary distributions than previous models such as Latent Dirichlet Allocation [48]. Given the coarseness of the clustering, Language Models were not deemed necessary. The model is trained on $D = 186{,}162$ abstracts published between 2000 and 2019. Tokens are extracted from the papers' titles and abstracts by filtering n-grams between one and three words matching syntactic expressions susceptible of carrying scientific information (by designating concepts, models, methods, etc.), following the procedure from [40, 49]. Embeddings are learned using a skip-grap model in $L = 50$ dimensions (few are needed given the small size of the vocabulary, $V = 4751$; nevertheless, some analyses are re-iterated with $L = 150$; see Appendix A.3, Fig. 12) [50]). We obtain the topics listed in Appendix A.3, Table 1. Four of the 20 resulting topics regroup keywords

[6] Previous works based on the PACS categories have leveraged the different levels of this hierarchical classification system to investigate different scales.



(a) Physicist A. (b) Physicist B.

Figure 2 Initial (in blue) and late (in red) research portfolios of two physicists. Physicist A shifted their focus from neutrinos physics to dark matter[7]. Physicist B, on the other hand, pursued a remarkably stable research agenda

that do not clearly refer to any specific research area (e.g. "paper", "approach") and correlate poorly with the PACS categories, which suggests their lack of scientific dimension. Consequently, the present analysis only considers the $K = 15$ remaining topics that designate actual research areas (as confirmed by their strong tendency to preferably cite themselves, cf. Appendix A.3.4, and their correlation patterns with the PACS classification, cf. Figure 15, Appendix A.3.6). In order to enhance the robustness of the topic removal process, we made sure all retained topics had a maximal loading on the PACS classification higher than that of all removed topics.

Then, we derive n_{dk}, the amount of keywords in the abstract of d that refer to a research area k (using the method described in Appendix A.3.2), and consequently $X_{a,k}$, the amount of times keywords ("resources", i.e. concepts, models, methods, etc.) in relation to research area k have occurred in papers (co-)authored by a in the initial time-period (2000 to 2009). Mathematically, $X_{a,k} = \sum_{d\in[2000,2009],a\in A_d} n_{d,k}$, where A_d is the set of authors of a publication d. The matrix $Y_{a,k}$ is derived similarly, using publications from the later time period (2015 to 2019). Research portfolios are then normalized into distributions $x_{ak} \equiv X_{ak}/\sum_{k'} X_{ak'}$ and $y_{ak} \equiv Y_{ak}/\sum_{k'} Y_{ak'}$, thus encoding how scientists divided their attention during each period. This approach ensures that research portfolios are evaluated based on the frequency of keywords that belong to each "topic", according to the idea illustrated in Fig. 1. Therefore, this approach captures variations in the prevalence of different kinds of vocabulary (and thus bodies of knowledge) exploited in scientists' publications. For purposes of illustration, Fig. 2 shows the research portfolios of one physicst who migrated from neutrinos to dark matter physics, and of one physicist who maintained their research agenda over the time periods considered.

2.3 Measuring capital

As shown by Schirone [51] in an extensive review of references to Bourdieu in bibliometrics, most mentions of capital focus on symbolic and social capital. Only a dozen works considered cultural/intellectual capital, and none of those proposed a measure that adequately captured the distribution of capital across epistemic domains. Therefore, an alternative unified approach for measuring the distribution of intellectual and social capital is proposed below.

[7] Physicist A's personal website reads: "I am working on particle astrophysics and cosmology. In particular, *I am interested in dark matter problem in the Universe, and how to probe it using annihilation products such as energetic gamma rays and neutrinos.* […] I *started my research career by studying supernova neutrinos from various aspects* [our emphases]." (https://staff.fnwi.uva.nl/s.ando/eng/Research.html). This is therefore an instance of "conversion" of prior knowledge to new purposes, one of the forms of change drawn from historical institutionalism represented in Fig. 1.

2.3.1 Intellectual capital

Intellectual capital is represented by a vector $\boldsymbol{I_a} = (I_{ak})$ that measures the concentration of a's intellectual resources in each domain $k \in \{1, \dots, K\}$. It is constructed in a way similar to $\boldsymbol{x_a}$, summing the contribution of keywords dedicated to each research area in the publications of each author between 2000 and 2009 (thus excluding publications that belong to the outcome research portfolio $\boldsymbol{y_a}$), except that publications are now weighed differently depending on the amount of authors. Indeed, publications with fewer co-authors convey more information about each author's own expertise. The weight is $\frac{1}{|A_d|}$, where $|A_d|$ is the amount of authors of publication d:[8]

$$I_{ak} \propto \sum_{d \in [2000,2009], a \in A_d} \frac{n_{d,k}}{|A_d|} \tag{1}$$

$\boldsymbol{I_a}$ is normalized, such that $\sum_k I_{ak} = 1$; therefore, $\boldsymbol{I_a}$ only captures the ways scientists divide their cognitive resources between research areas, rather than the "absolute magnitude" of their knowledge of each area (by contrast, the measure of semantic capital proposed in [52] measures total knowledge but cannot capture diversity).

2.3.2 Social capital

Many measures of scientists' social capital have been proposed [15, 51], the simplest being the amount of collaborators of a scientist (i.e. degree centrality in the co-authorship network [52]). Other measures revolve around *betweeness* centrality, which captures the extent to which an actor "bridges" a network (e.g. "brokerage", i.e., the ability of an individual to overcome "structural holes" in a social network [53]). Abbasi et al. [15] distinguish two general approaches to social capital, depending on whether the emphasis is placed on "power" versus "diversity". Measures of social capital (as those discussed in [15]) typically represent social capital by single scalars; however, social capital has multiple dimensions. In fact, according to Bourdieu [20], "the volume of social capital possessed by a particular agent [...] depends on the extent of the network of links that he can effectively mobilize, and on the volume of capital (economic, cultural or symbolic) possessed by each of those to whom he is linked". In that respect, social capital can come in different forms depending on the resources being leveraged via one's network. In the following, we focus on the intellectual dimension of social capital, which we represent by a vector $\mathbf{S_a}$ defined as the sum of the intellectual capital of a's collaborators, weighted by the strength of their relationship:

$$\boldsymbol{S}_a \equiv \sum_{c \in \text{co-authors}(a)} w_{ac} \boldsymbol{I}_{c \setminus a} \tag{2}$$

$\boldsymbol{I}_{c \setminus a}$ is the intellectual capital of c, evaluated by excluding papers co-authored with a (in order to disentangle the effect of an author's own knowledge and that available to them via their collaborators). Collaborators outside the cohort are taken into account. The weight w_{ac}, which represents the strength of the relationship between a and c, is defined as:

$$w_{ac} \equiv \max_{d | \{a,c\} \subset A_d} \frac{1}{|A_d| - 1} \tag{3}$$

[8] A justification for this weight is that the probability that a given author has been responsible for introducing any particular concept or method present in the paper is $\mathcal{O}(1/|A_d|)$.

Where A_d is the set of the co-authors of a paper d. This weighing scheme – inspired from [54] – captures the fact that a paper with, say, two co-authors, signals a stronger relationship between the authors than a publication with a dozen authors.[9] However, it does not take into account the recency and frequency of collaborations.

2.3.3 Diversity and power

Measures of "diversity" (and "power") can be readily derived from $\boldsymbol{I_a}$ (and $\boldsymbol{S_a}$). A common measure of diversity is the Shannon entropy H [55]. Let $D(\boldsymbol{I_a}) = \exp H(\boldsymbol{I_a})$ be the diversity of intellectual capital (and $D(\boldsymbol{S_a}) = \exp H(\boldsymbol{S_a})$ that of social capital). Roughly speaking, these are measures of how many research areas scientists have divided their cognitive or social resources among. In the cohort, individuals typically have cognitive resources in several research areas ($\mu_{D(\boldsymbol{I_a})} = 5.6$, $\sigma = 2.0$), and social capital is even more diverse ($\mu_{D(\boldsymbol{S_a})} = 7.8$, $\sigma = 2.1$). In fact, intuitively, scientific collaborations enable individuals to take advantage of their group's diversity. Furthermore, $D(\boldsymbol{I_a})$ and $D(\boldsymbol{S_a})$ are highly correlated ($R = 0.75$); indeed, individuals with more diverse expertise are more able to engage with diverse collaborators. Since the diversity of social capital is mostly expected to enhance individuals' abilities when it exceeds that of their own knowledge, from now on, only *excess* social capital diversity $D^*(\boldsymbol{S_a})$ (defined as the residuals of the linear regression of $D(\boldsymbol{S_a})$ against $D(\boldsymbol{I_a})$, by ordinary least squares) is considered.[10]

The "power" dimension of social capital is evaluated as the *magnitude* of social capital:

$$P(\boldsymbol{S_a}) \equiv \sum_k S_{ak} = \sum_{c \in \text{co-authors}(a)} w_{ac} \tag{4}$$

"Power" is therefore the amount of collaborators weighed by the strength of each relationship. Our measures of diversity and power depart from [15], which conflates diversity with network size and power with performance. By combining semantic and authorship data, our approach assesses diversity more directly.

Alternative measures of diversity and power are considered for robustness assessment. The alternative measure of diversity is based on Stirling's index, and the alternative measure of power uses the notion of brokerage. All these measures are defined and compared in Appendix A.4.

2.4 Modelling trajectories

The model for the late research portfolio $\boldsymbol{Y_a}$ is schematically illustrated in Figs. 3a, 3b, 3c, and a more formal representation is given in Fig. 3d. It captures the idea expressed in Fig. 1 that research portfolios are transformed via strategic transfers of knowledge and attention from one research area to another. Occurrences of keywords that belong to each research area k in papers by a in the late time period, $\boldsymbol{Y_a} \in \mathbb{N}^K$, are assumed to be drawn from a hierarchical multinomial logistic model. $\boldsymbol{Y_a}$ results from a linear combination of the initial research portfolio, $\boldsymbol{x_a}$, and a mixing matrix θ_a that measures the fraction of attention redistributed from each research area to another. θ_a is drawn from a hierarchical process,

[9] Assuming that in a collaboration, each author interacts with a constant amount of co-authors in practice (regardless of the total amount of co-authors), then the probability that they had interactions with one specific co-author in particular is $\propto \frac{1}{|A_d|-1}$.

[10] This approach aims to address the difficulty of determining the direction of the causal relationship between social resources and research interests raised by Tripodi et al. [7].

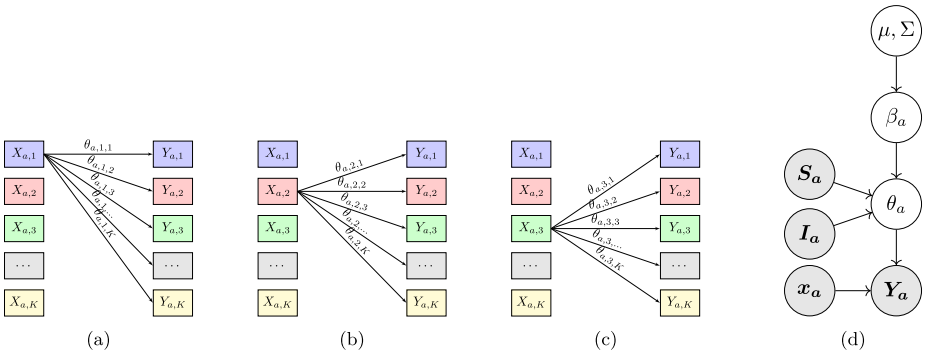


Figure 3 3a,3b,3c: Transfers of attention across research areas. $\boldsymbol{x_a}$ and $\boldsymbol{y_a}$ are the distributions scientist a's attention across research areas in two consecutive time periods. $\boldsymbol{\theta_{ak}} = (\theta_{akk'})$ represents the fraction of the attention devoted by an author a to a topic k redirected to topics $k' \in \{1, \ldots, K\}$, as scientists repurpose, expand or concentrate their knowledge. By definition, $\sum_{k'} \theta_{akk'} = 1$. Fig. 3d: Hierarchical model. θ_a is drawn from a hierarchical process, with intellectual and social capital ($\boldsymbol{I_a}$ and $\boldsymbol{S_a}$) as covariates. Observed variables are represented in gray, latent variables in white

thus capturing the "average" cohort behavior. Formally speaking, $\boldsymbol{Y_a}$ is assumed to derive from a multinomial process involving linear combinations of (x_{ak}):

$$\boldsymbol{Y_a} \sim \text{multinomial}(\sum_{k=1}^{K} x_{ak}\theta_{ak1}, \ldots, \sum_{k=1}^{K} x_{ak}\theta_{akK}) \tag{5}$$

Where $\theta_{akk'}$ is the fraction of attention to a topic k by a that has been redirected to a topic k'. θ is a function of intellectual capital $\boldsymbol{I_a}$ and social capital $\boldsymbol{S_a}$ according to the following generalized linear model:

$$\boldsymbol{\theta_{ak}} = \text{softmax}\left(\beta_{ak1} + \gamma_{k1}I_{a1} + \delta_{k1}S_{a1}, \ldots, \beta_{akK} + \gamma_{kK}I_{aK} + \delta_{kK}S_{aK}\right) \tag{6}$$

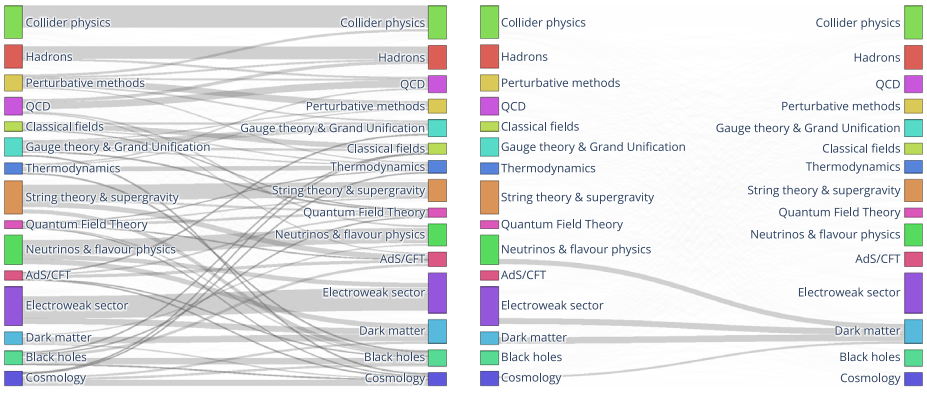
$\delta_{kk'}$ is the effect of the scientists' social capital in a research area k on the magnitude of transfers from k to k'. Similarly, $\gamma_{kk'}$ is the effect of having more expertise in k' (intellectual capital) on shifts from k to k'. High values of the diagonal elements of γ would imply that physicists are more conservative towards research areas in which they concentrate more expertise. The coefficients $\beta_{akk'}$ encode the average behavior of the cohort plus individual deviations to the average behavior that are unexplained by the covariates.[11,12] The

[11] The priors for this hierarchical model are:

$$\beta_{akk'} \sim \mathcal{N}(\mu_{kk'}, \sigma_{kk'}) \text{ for } 1 \leq k' \leq K-1 \text{ and } \beta_{akK} = \mu_{kK} \tag{7}$$
$$\mu_{kk'} \sim \mathcal{N}(\lambda \times \nu_{kk'}, 1) \text{ (average behavior)} \tag{8}$$
$$\delta_{kk'} \sim \mathcal{N}(\delta_0 + \lambda' \times \nu_{kk'}, 1) \text{ (effect of social capital)} \tag{9}$$
$$\gamma_{kk'} \sim \mathcal{N}(0, 1) \text{ (effect of intellectual capital)} \tag{10}$$
$$\sigma \sim \text{Exponential}(1) \tag{11}$$

Where $\nu_{kk'}$ is the fraction of physicists with expertise in k (that is, with more intellectual capital than average in k) who also have expertise in k'. Priors must be thought thoroughly, as certain invariances can lead to identification issues – for instance, shifting μ by a constant does not change the likelihood.

[12] The fit is performed with Stan's Hamiltonian Monte-Carlo sampler (HMC is better behaved than Gibbs for such problems).



(a) Transfers across all research areas. (b) Flows of attention directed towards "dark matter".

Figure 4 (a) Aggregate transfers of attention across research areas, between 2000-2009 (to the left) and 2015-2019 (to the right). Widths of flows are proportional to $\sum_a \chi_{ak}\theta_{akk'}$. Insignificant transfers (that happen less than expected by chance alone assuming uniform mixing) are transparent. (b) For purposes of clarity, the same figure is repeated to the right, highlighting only the flows directed towards "Dark matter"

ability of the model to predict individual trajectories is assessed in Appendix A.5, for various temporal segmentations of the initial and late research portfolios and topic models. The model is better at predicting individual trajectories for larger adaptive responses and longer time-scales.

Our model is strongly connected to Optimal Transport [56, 57], which seeks optimal ways to "transport" an input distribution (say, $\boldsymbol{x} = (x_k)$) to a target distribution (e.g., $\boldsymbol{y} = (y_{k'})$) through "transfers" ($\theta_{kk'}$) across their components while minimizing a cost function $\sum_{k,k'} x_k \theta_{kk'} C_{kk'}$ (where $C_{kk'}$ is a cost matrix) [14]. The difference is that the proposed Bayesian approach estimates transfers $\theta_{kk'}$ by minimizing a likelihood rather than a cost function. However, the connection with Optimal Transport suggests an economic interpretation of the reallocation of research efforts, which will be leveraged in Sect. 3.1 to show that patterns of change in research interests are shaped by learning costs.

3 Results

Figure 4a shows the aggregate transfers of attention at the level of the cohort revealed by the model. The most obvious feature is the remarkable stability of most research areas: indeed, physicists' conservatism toward their research area due to specialization is known to be particularly high in HEP [5]; late research portfolios are largely constrained by prior research interests: they exhibit *path dependence* [11] (using Inverse Optimal Transport, Sect. 3.1 shows that these patterns are structured by learning costs). Conservatism seems especially prevalent in the case of "collider physics", a research area dedicated to knowledge specific to particle accelerators. Nevertheless, "dark matter" has doubled, fueled by a shift away from "neutrinos and flavor physics", and "electroweak sector", a phenomenal domain studied at the LHC (Fig. 4b).[13,14] This confirms that the cohort has responded

[13]The electroweak notably includes Higgs physics, which are very prominent at the LHC, where the Higgs boson was discovered.

[14]The migration of many particle physicists towards dark matter provides an explanation for the persisting schism between two research programs in fundamental physics, namely dark matter particle research and modified gravity. Both research programs seek to explain a shared set of anomalies in astronomical observations, and yet their communities communicate

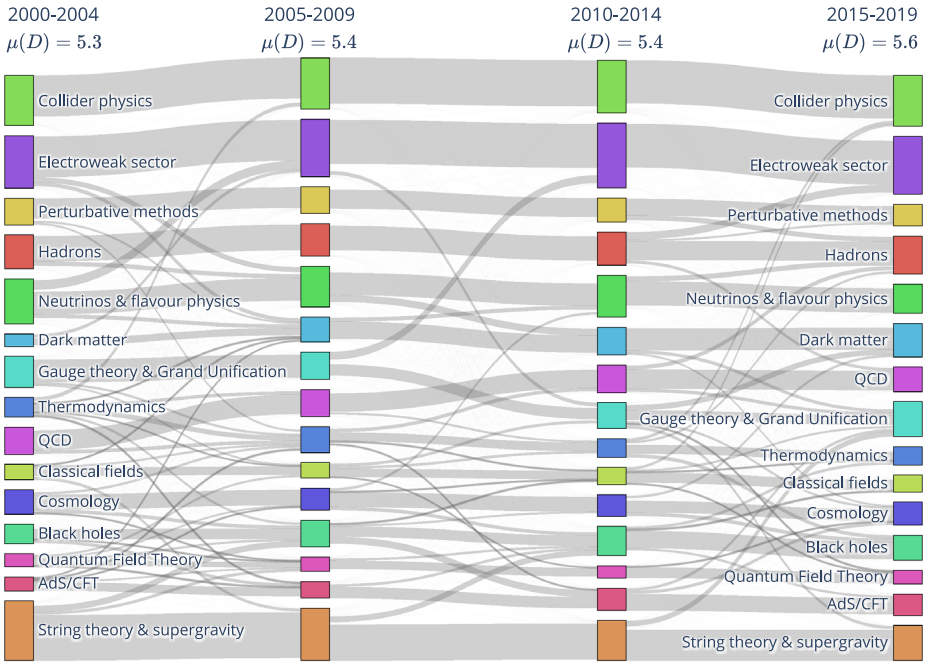


Figure 5 Aggregate transfers of attention throughout the years 2000 to 2019, considering four time-intervals of five years each. $\mu(D)$ is the average portfolio diversity of the cohort during each time-period

to changes in the landscape of experimental opportunities. Moreover, "string theory and supergravity" has declined in favor of "AdS/CFT" (a research program that explores dualities between theories of quantum gravity and certain types of theories of quantum field theory) and "black holes".[15] Of course, this rough description must be considered with caution, as interpreting clusters from topic models is notoriously hard, and these topics in particular can regroup quite heterogeneous research programs.

While the paper focuses on two time periods (2000-2009 and 2015-2019), multiple alternative temporal segmentations can be considered. Figure 5 shows the transfers of research attention of a cohort of physicists across four time-periods of five years each. It reveals that the changes outlined above have unfolded rather gradually. The average diversity of physicists' research portfolio ($\mu(D)$, the average of the exponentiated entropy of $\boldsymbol{x}$) during each five-year time-bin is also shown. It has gradually increased over the years ($P < 10^{-4}$): on average, physicists have *expanded* their portfolio. Interestingly, the average linguistic diversity of each individual paper increased as well (with a confidence level $P < 10^{-4}$) from 2.81 topics per paper (2000-2004) to 2.96 (2015-2019). This means physicists diversified their research portfolios in part by diversifying the knowledge leveraged *within* each of their individual papers (rather than solely by writing multiple papers on separate issues).

very little [58, 59]. Our approach suggests that particle physicists' interest in dark matter is in great part motivated by the fact this is a natural extension of their previous research; particle physicists would therefore not consider the alternative to dark matter (modified gravity), given this topic that would make little use of their expertise.

[15] This converges with physicist Peter Woit's controversial assessment that "string theorists" are no longer doing string theory per se, though they keep identifying themselves as string theorists. As Peter Woit puts it, citing the 2022 "Strings" conference: "one thing that stands out is that the string theory community has almost completely stopped doing string theory."; and, "[presentations' titles] make very clear what the string theory community has found to replace string theory: black holes" (Woit, 2022, https://www.math.columbia.edu/~woit/wordpress/?p=12981).

3.1 The structuring effect of learning costs on scientists' behavior

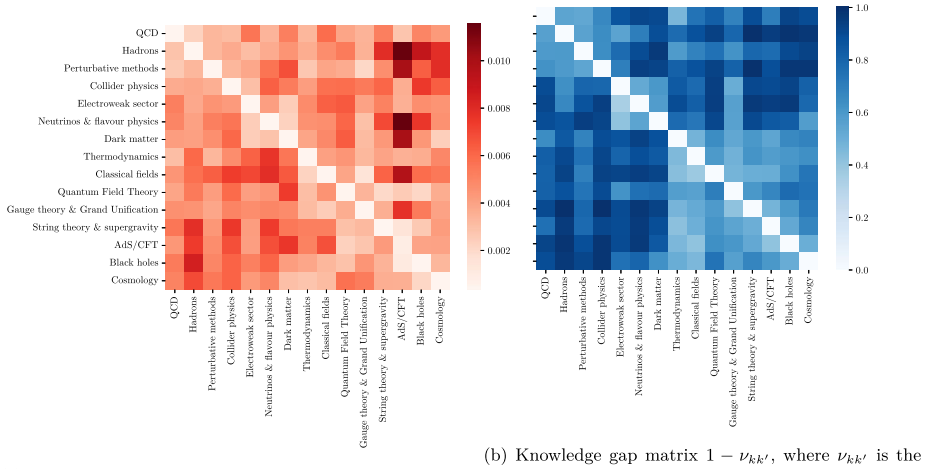
As research priorities change, research efforts must be efficiently reallocated among scientists, given their prior knowledge. Here, we leverage the similarity between our model of scientists' trajectories and Optimal Transport (OT) [56, 57] to formulate the observed behavior in economic terms and demonstrate that the shifts of attention are structured by the minimization of "learning costs", thus providing a first-order explanation of the aggregate patterns of change in high energy physicists' research interests. Optimal Transport is a mathematical framework first introduced by Gaspard Monge for finding optimal ways of displacing piles of sand in a military context [12], and later refined by Leonid Kantorovitch in the context of economic planning [13].

Let $\boldsymbol{x} = \sum_a \boldsymbol{x_a}$ be the cohort's initial distribution of attention across research areas (summing over each author a), and $\boldsymbol{y} = \sum_a \boldsymbol{y_a}$ the late distribution. Let us further assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ can be considered "fixed" by the institutions (scientific leadership, laboratories, funding agencies, etc.) that define scientific priorities throughout time. In order to achieve the distribution of research efforts $\boldsymbol{y}$ given the previous distribution $\boldsymbol{x}$, some scientists must redirect their attention away from certain research areas (for which $y_k < x_k$) and towards more pressing ones (for which $y_k > x_k$). What is the most efficient way to reallocate research efforts and achieve the transition from $\boldsymbol{x}$ to $\boldsymbol{y}$? Intuitively, research areas should be assigned to scientists in a way that requires as few of them as possible to acquire new knowledge – in other words, in a way that minimizes learning costs, given the way knowledge is distributed among individuals. This, we show, can be framed as an Optimal Transport problem. Let $T_{kk'}$ be the coupling matrix that encodes how much attention has been shifted from one research area k to a research area k'. There are two constraints on $T_{kk'}$: $\sum_{k'} T_{kk'} = x_k$, and $\sum_k T_{kk'} = y_{k'}$. These constraints encode the need to adapt (since $\boldsymbol{y} \neq \boldsymbol{x}$). But T must also minimize learning costs (C), which we assume to be linear in T for simplicity:[16] $C = \sum_{k,k'} T_{kk'} C_{kk'}$, where $C_{kk'}$ is a cost matrix. The problem of finding the couplings $T_{kk'}$ that minimize the "transportation" costs (given a cost matrix $C_{kk'}$ and the constraints on $T_{kk'}$) is an instance of Optimal Transport problem [14]. Typical instances of OT problems include how to efficiently transport (say, ore from mines to factories) or the optimal assignment of workers to firms [60]. In our case, the couplings are known (they were recovered by the model from Sect. 2.4), and we want to infer the underlying cost matrix that these couplings minimize. The transfers from a research area k to k' for each individual are simply $x_{ak}\theta_{akk'}$. Summing over individuals yields the coupling matrix $T_{kk'} = \sum_a x_{ak}\theta_{akk'}$, which measures how much attention was shifted away from k and toward k' at the cohort level.

The problem of recovering the cost matrix $C_{kk'}$ given the couplings is an Inverse Optimal Transport problem. Below, this problem is solved using the probabilistic method proposed in [17].[17] This method requires to put a prior on $P(C_{kk'})$ – indeed, infinitely many cost matrices yield the same optimum, and priors are needed to lift this degeneracy. Following [17], we consider a prior such that $\sum_{k,k'} C_{kk'} = \text{cst}$. We assume that

[16] Roughly speaking, this linear assumption entails that scientists from a given research area are equally able to shift attention to another research area. In practice, some scientists have more abilities to switch to a given research area, and the cost will increase non-linearly as more and more scientists are required to make the transition, including those less able to make the switch.

[17] It should be noted that the approach by Chiu et al. [17] does not entail the assumption that scientists' behavior is perfectly minimizing the cost matrix. Indeed, the optimization problem they consider includes an entropic regularization term; while this term is often introduced for numerical reasons, in the case of human behavior, it can be taken to represent inefficiencies and random deviations from the optimum behavior [61].



(a) Cost matrix $C_{kk'}$. The cost matrix represents the cost of shifting a unit of cohort's research attention from k to k'.

(b) Knowledge gap matrix $1 - \nu_{kk'}$, where $\nu_{kk'}$ is the fraction of physicists with expertise in the horizontal research area (k) who also hold expertise in the vertical research area (k').

Figure 6 Learning costs and knowledge gap. The cost of shifting attention from one area to another (left plot) is in great part determined by how frequently individuals hold knowledge in both areas (right plot)

$\mathbb{E}(C_{kk'}) \propto \text{softmax}(\beta \times (1 - \nu_{kk'}))$,[18] where $\nu_{kk'}$ is the fraction of physicists who already held expertise in k' among those who already held expertise in k, and $\beta \sim \mathcal{N}(0,1)$ is the effect of learning costs on C_{kk}. If $\nu_{kk'} \sim 1$ (i.e. $1 - \nu_{kk'} \sim 0$), then shifting attention from k to k' does not entail the acquisition of additional knowledge, and $C_{kk'}$ should be almost zero. If $\nu_{kk'} = 0$ (i.e. $1 - \nu_{kk'} = 1$), any scientist shifting from k toward k' must acquire new knowledge, and the cost should be maximal. If actual behaviors *do* involve the minimization of learning costs, we should observe a strong correlation between $C_{kk'}$ and the "knowledge gap" $(1 - \nu_{kk'})$.

Using the "MetroMC" algorithm proposed in [17], we empirically recover $C_{kk'}$, the underlying cost matrix, as shown in Fig. 6a. We find a strong correlation with the "knowledge gap" (shown in Fig. 6): $R(C_{kk'}, 1 - \nu_{kk'}) = 0.76$, such that $R^2 = 0.58$ (Fig. 17, see also Appendix A.6). Using a finer-grained temporal segmentation (2000-2004 → 2005-2009), the resulting correlation is similar ($R^2 = 0.62$). The empirical cost of shifting research efforts from one research area to another is therefore shaped by learning costs. The derivation of $C_{kk'}$ is potentially useful: one could in principle predict aggregate transfers of attention given a counterfactual target distribution of research efforts $\boldsymbol{y}$ using optimal transport and plugging-in $C_{kk'}$ as the cost matrix [57].

Under changing circumstances, research efforts must be reallocated efficiently. Scientific norms and institutions must address an Optimal Transport problem by providing incentives for scientists to conform to new research imperatives, in a way that factors "learning costs". Consequently, shifts between research areas which entail the acquisition of new knowledge (layering) must be less likely that those which can take advantage of prior knowledge (conversion). In the case of HEP, it does seem that adaptative patterns are structured by this OT problem. Interestingly, the matrices in Fig. 6a and 6b feature blocks (indicative of an underlying hierarchical structure), such that it is easier for scien-

[18] This prior has the merit of simplicity – it is a simple generalized linear model with the desired support.

tists to migrate within than across these blocks. While these observation characterizes the cohort's behavior, drivers of differences among individuals are considered next.

3.2 Individual behavior and the effect of capital

3.2.1 Magnitude of change and capital

We propose a change score c_a that measures how much the research agenda of a scientist has changed between the two time periods under consideration, defined as the total variation distance between their initial and late research portfolios:

$$c_a \equiv \mathrm{d_{TV}}(\boldsymbol{x_a}, \boldsymbol{y_a}) = \frac{1}{2}\sum_k |y_{ak} - x_{ak}| \tag{12}$$

This measure of change is naturally motivated by Optimal Transport: it is the minimal cost of transporting $\boldsymbol{x_a}$ to $\boldsymbol{y_a}$ if the cost matrix has zeros on the diagonal and ones everywhere else ($C_{kk} = 0$ and $C_{kk'} = 1$ for $k \neq k'$.). This measure, however, weighs migrations across different research areas identically, regardless of their cognitive proximity. We will therefore also consider a second measure of change, the "cognitive distance" (d_a), defined as the minimum cost of transporting $\boldsymbol{x_a}$ to $\boldsymbol{y_a}$ [14] induced by the cost matrix empirically recovered (see Fig. 6a) using the Inverse Optimal Transport approach described in the previous section. Another interesting aspect of Optimal Transport is indeed its ability to provide "distances" between distributions that emphasize certain costs in particular.[19]

c_a is comprised between 0 (if research attention has remained identically distributed) and 1 (if the research agenda has been entirely redistributed). Large values of c_a are rare, with 50% of authors lying between 0.21 and 0.40 (Appendix A.7.1, Fig. 18). Although the absolute value of c_a (and d_a) has limited interpretability (it depends on the choice of "cognitive scope" and the duration of the time periods considered, and c_a is never expected to be exactly zero due to random fluctuations and measurement noise), it allows for comparisons between physicists. We evaluate the effect of several factors on c_a: i) the diversity of intellectual capital $D(\boldsymbol{I_a})$; ii) the excess diversity of social capital $D^*(\boldsymbol{S_a})$; and iii) the magnitude of social capital ("power"). We also consider the effect of iv) affiliation stability, represented by a binary variable s_a ($s_a = 1$ if scientist a has at least one affiliation that spans the whole time range considered, and $s_a = 0$ otherwise), the effect of v) academic age, and productivity, estimated from vi) all papers and vii) solo-authored papers. We perform a linear regression of c_a as a function of these variables, adjusting for $Z_a = \arg\max_k x_{ak}$, i.e. physicists' primary research area over the years 2000 to 2009 (see the model specification in Appendix A.7.1).

The results are shown in Fig. 7. The diversity of intellectual capital has a significant positive effect: physicists with resources in many areas tend to revise their research agenda more. There is also evidence of a small but positive effect of diversity of social capital on the magnitude of changes in scientists' research focus (interpreting these results in terms

[19] Given a cost matrix $C_{kk'}$, we can define a measure of the gap from one distribution $\boldsymbol{x}$ to another distribution $\boldsymbol{y}$, as the minimum cost of displacing $\boldsymbol{x}$ to $\boldsymbol{y}$:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \min_{\substack{\theta_{kk'} \\ \theta \mathbf{1} = \mathbf{1}, \theta^T \boldsymbol{x} = \boldsymbol{y}}} \sum_{kk'} x_k \theta_{kk'} C_{kk'} \tag{13}$$

If the cost matrix meets certain properties (such as symmetry), then d is a distance. See [14] for more on the metric properties of OT.

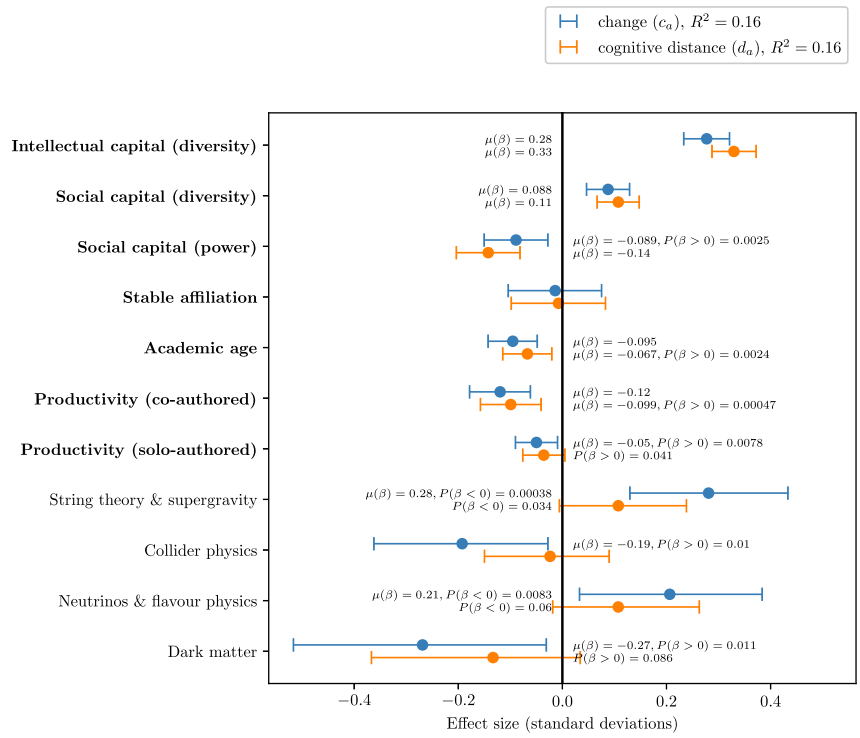


Figure 7 Effect of intellectual capital, social capital, and institutional stability the magnitude of changes in high-energy physicists' interests.. Change (c_d) is the total variation distance and weighs migrations across research areas equally. The cognitive distance (d_a) gives more weight to migrations across cognitively distant research area. Diversity is evaluated as the exponential of the Shannon entropy of the distribution of intellectual and social capital across research areas. The effect of the primary research area during the initial period (2000-2009) was also included as a control. Three research areas have a significant (95% CL) effect. Error bars indicate 95% credible intervals; $\mu(\beta)$ denotes the mean posterior effect size. $R^2 = 0.17$. Continuous variables (diversity, power and c_a) are standardized

of Optimal Transport, we might say that social capital helps overcome cognitive learning costs). However, the *magnitude* of social capital, "power", has a negative direct effect on change. In other words, "power" is associated with stability, and "diversity" with change. It is noteworthy that these dimensions of social capital have opposite effects. More senior physicists are more conservative toward their research agenda, possibly because they experience less incentive to "adapt". This comes in contrast with [6]. The difference could stem in the specificity of high-energy physics, and in the fact that change is measured in linguistics terms (instead of relying on citation patterns) at a rather coarse-grained scale in the present paper. There is no discernible effect of affiliation stability after adjusting for academic age. However, affiliation data is a bit noisy, and this could have the consequence of underestimating the effect of institutional stability relative to that of academic age. Finally, productivity (in co-authored papers) is associated with stability. Overall, entrenchment (age, power, productivity, specialization) all drive stability and conservatism.

Unsurprisingly, both research areas that have shrunk considerably have a significant positive effect on migration scores. "Collider physics" and "dark matter", on the other hand, have a negative effect on the magnitude of change. All effects combined, physicists whose primary category is "Collider physics" are the most conservative, with an average change score 24% lower than the rest of the cohort; the long time-scales of collider experiments

provide stable opportunities to physicists in that area [62, p. 138], thus promoting conservatism. The variance explained remains low ($R^2 = 0.17$): these factors only partially explain differences between individuals.

c_a neglects the cognitive gap between research areas. Using our alternative measure of change that takes into account cognitive distance (d_a), the diversity of intellectual and social capital has slightly larger effects. The robustness of these results is assessed by using different operationalizations of diversity and power in Appendix A.7.3), Table 3. Most findings are stable, except that "brokerage" (unlike the magnitude of social capital, i.e. degree centrality) has no *direct* effect on change (beyond the effect of productivity resulting from co-authored publications).

In addition, in order to exclude the influence of direct collaborations on change, we conducted a second comparative analysis including only scientists' first-authored and last-authored publications in their research portfolios.[20] The effect of the diversity of intellectual and social capital is stable; however, the direct effect of power is reduced (Appendix A.7.3, Tables 5 and 6). Moreover, the analysis was reproduced across different choices of temporal segmentation and using different topic models. The previous findings remain stable.

3.2.2 Diversification versus concentration

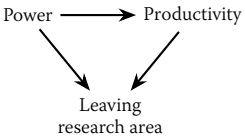
Research portfolios can be altered by two opposite strategies. One is "diversification", i.e. the addition of new research areas. Another is "concentration", i.e. the desertion of research areas to focus increasingly on others (Fig. 1 illustrates how this can happen, whether via "conversion" or "displacement"). Figure 8 shows the effect of the same factors as above on the probability that physicists have i) entered at least one new research area in between the two periods or ii) exited one research area (model description provided in Appendix A.7.2).[21]

The diversity of intellectual capital has a strong positive effect on the probability of exiting a research area; intuitively, scientists with diverse interests can afford to disengage from certain research areas even if this implies to abandon maladaptive prior knowledge ("displacement"). Excess diversity of social capital increases the probability of entering new research areas, but has no discernible effect on the probability of exiting research areas. In contrast, there is some evidence that "power" decreases the probability of leaving a research area. Figure 8 shows the direct effect of power (controlling for productivity due to co-authored papers), which does not pass the 95% significance test; however, the *total* effect of power[22] on the probability of leaving a research area is significantly negative ($\mu(\beta) = -0.24$, $P(\beta > 0) < 10^{-4}$). Presumably, having many collaborators allows scientists to remain committed to many research areas with minimal personal investment,

[20] It is important to emphasize that alphabetical ordering is still very prevalent in this field [63], and therefore this strategy does not address the issue entirely.

[21] A research area k is considered "entered" by a when $x_{ak} < \frac{1}{N}\sum_{a'} x_{a'k}$ and $y_{ak} > \frac{1}{N}\sum_{a'} y_{a'k}$; conversely, a research area is considered exited when $x_{ak} > \frac{1}{N}\sum_{a'} x_{a'k}$ and $y_{ak} < \frac{1}{N}\sum_{a'} y_{a'k}$.

[22] That is, assuming the following causal structure:



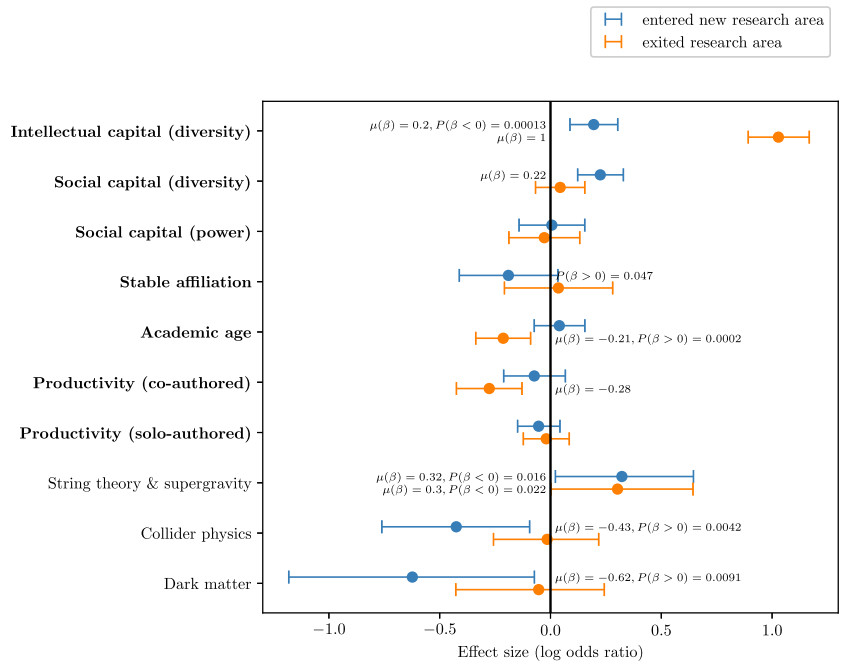


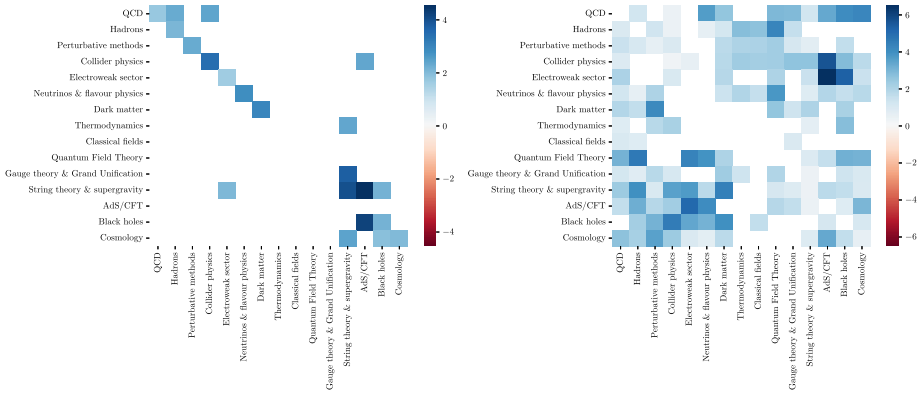
Figure 8 Effect of capital and institutional stability on the probability of entering new research areas or exiting previously explored ones. Bars indicate 95% credible intervals. $\mu(\beta)$ denotes the mean posterior effect size. Continuous variables (diversity and power) are standardized

thus stabilizing their research agendas. Moreover, "power" also decreases the probability of entering new research areas; this suggests that powerful scientists have less incentives to invest resources in new topics. String theory & supergravity has a significant positive effect of both entering and exiting research areas (suggesting physicists frequently "replaced" it with another topic). "Dark matter" has a negative effect on the probability of entering a new research area, possibly because scientists with prior commitment to this research area had less incentive to diversify their research portfolio and more incentive to focus increasingly on this topic given its success over the following years. These conclusions hold as alternative measures of diversity and power are considered (A.7.3, Table 4).

Again, we ran a second analysis considering only scientists' first-authored and last-authored publications. The effects of intellectual and social diversity remain stable. However, the effect of productivity resulting from co-authored papers no longer has an effect on the probability of exiting a research area when only first- and last-authored papers are included in physicists' research portfolios (Table 8, Appendix A.7.3.). The results are generally stable across different topic models and temporal segmentations, except for i) the effect of academic age on the probability of exiting a research area, which is not consistently 95% CL significant; and ii) the effect of the diversity of social capital on the probability of entering a new research area, which is zero in one particular configuration (Tables 7 and 8, Appendix A.7.3.).

3.2.3 Why diversity promotes change

Access to diverse cognitive resources is associated with change. To see why, it is insightful to look into how the concentration of scientists' intellectual capital in each research area k ($\boldsymbol{I_a} = (I_{ak})$) affects their trajectories. In the model introduced in Sect. 2.4, the diagonal co-



(a) Effect of intellectual capital ($\gamma_{kk'}$). (b) Effect of social capital ($\delta_{kk'}$).

Figure 9 Effect of intellectual and social capital on transfers across research areas. Rows represent research areas of origin and columns represent target areas. Effect sizes are expressed in log-odds ($\log[\theta_{akk'}/(1-\theta_{akk'})]$, where $\theta_{akk'}$ is the fraction of attention to k redirected to k') per unit of intellectual or social capital in the target research area k'. Effects that are not significant (at the 95% credible level) are displayed in white for purposes of clarity

efficients of the γ matrix measure the effect of having intellectual resources in a certain area on the commitment to this area. As shown in Fig. 9a, most coefficients on the diagonal are significantly positive: physicists with a strong specialization in a research area tend to preserve their specialization in this area.

The effect of social capital on transfers across research areas ($\delta_{kk'}$, Eq. (6)), is shown in Fig. 9b. Statistically significant effects are always positive: scientists tend to redirect attention to research areas in which they have more collaborators involved, in line with a very recent observation by Venturini et al. [64].

Moreover, we find a strong correlation between the effect of social capital ($\delta_{kk'}$), and $\nu_{kk'} \in [0,1]$, the fraction of physicists with more expertise than average in a research area k' among those who have more expertise than average in k (see Eq. (9), Sect. 2.4). $\delta_{kk'}$ decreases by 1.5 unit of standard deviation on average as $\nu_{kk'}$ goes from 0 (nobody holds expertise in both research areas k and k') to 1 (everyone with expertise in k has expertise in k'). Social capital plays a more important role in shifts between cognitively distant research areas, in line with a previous finding by Tripodi et al. [7]. These general patterns (the association between the concentration of intellectual capital in one area and the commitment to this area, and the increasing effect of social capital with cognitive dissimilarity) are insensitive to the temporal segmentation in place (see Figure 19, Appendix A.9 for similar Figures based on a different segmentation).

4 Discussion

This paper addressed the conflict between specialization and adaptability in science. To this end, an unsupervised Bayesian approach was developed, based on the idea that transformations in the scientific landscape prompt scientists to efficiently repurpose their prior knowledge. The model simultaneously measures transfers of attention across research areas and the effect of various variables on the evolution of scientists' research portfolios, in particular intellectual and social capital.

The model was applied to a cohort of $N = 2108$ high-energy physicists between the years 2000 and 2019. At the macroscopic level, it reveals the decline of neutrinos physics due

to migrations towards the electroweak sector (explored at the LHC) and more importantly towards dark matter. Similarly, many physicists have shifted resources from the electroweak sector towards dark matter. Moreover, string theory & supergravity has started to disintegrate into black holes and AdS/CFT research. The cohort has therefore responded to new experimental opportunities as well as to theoretical developments in quantum gravity.

Then, leveraging the connection between our model and Optimal Transport (OT), we showed that the reallocation of research efforts among scientists is shaped by learning costs. Indeed, under changing circumstances, scientific institutions must address an OT problem by efficiently reallocating research efforts in a way that balances learning costs and the imperative to adapt to new circumstances. This has the effect of enhancing the utility of the scientific capital disseminated amongst scientists as the perceived payoff of certain research areas change. OT structurally explains path dependency, as individuals experiencing pressures to adapt seize the nearest opportunity available to them. OT is also methodologically useful: Inverse OT allows one to derive cost functions from observed behavior, and thus offers a potential way to better connect empirical data with evolutionary agent-based models of scientists' behavior that postulate latent utility functions [16]. Moreover, it has the potential of informing policy by identifying potential bottlenecks if research efforts were to be reallocated in certain ways (as one can use OT to estimate the "cost" of various counterfactual scenarios). Overall, the OT approach illustrates that the adaptability of epistemic communities is constrained by how knowledge is distributed among individuals (specialization).

The longitudinal comparative analysis of physicists' trajectories revealed that the diversity of intellectual and social capital is positively associated with change: diversity promotes adaptability under new circumstances, and therefore diversifying research portfolios is a reasonable strategy when the future is uncertain. However, enforcing diversity can lead to suboptimal allocations of research efforts under stable circumstances [30]. Differences among research areas are found: physicists expert in particle colliders have remained particularly conservative, possibly because they have secured long-term research opportunities (thanks to very large investments in particle accelerators like the LHC). There is also evidence that physicists specialized in dark matter have been consolidating their specialization, presumably because the increasing popularity of the topic encouraged them to double down their investment in this research area. Higher concentrations of intellectual capital in certain research areas generate stronger commitment towards these areas; therefore, specialized scientists are more at risk of being trapped in a sunk cost fallacy as their expertise becomes unsuitable for new circumstances. However, specialized scientists can offset the risks associated with specialization by diversifying their social network. This raises the possibility of free-riding, as scientists are encouraged to focus on what seems most promising at the time and let their peers take the risk of exploring alternatives until their value is established [65]. Additionally, social capital plays an increasingly important role as scientists expand their research agenda further beyond their specialization, as observed by [7], suggesting that collaborations are crucial in overcoming cognitive barriers between research areas. Unlike diversity, "power" is associated with more stable research interests: presumably, cooperation can safeguard individuals from adaptive pressures, and most importantly minimizes the cost of remaining invested in certain areas.

We have described renewal strategies of research portfolios according to a typology of incremental change developed in the context of historical institutionalism, which has been shown to account for how organizations like DESY and SLAC have transitioned from HEP to multi-purpose photon science. Adaptation strategies include the "conversion" of prior knowledge to new purposes, the "layering" of additional research interests via the acquisition of new knowledge, or the "displacement" of former research commitments resulting in a loss of knowledge. The connection with institutional change stems from the fact that institutions can face a challenge similar to that experienced by specialized scientists confronted with transformations in their scientific landscape, in that they too must sometimes adapt and redirect accumulated capital in directions that may not have been foreseen at the time of their "design". The LHC itself has evolved through similar processes of gradual repurposing of prior infrastructure, including the accelerator's tunnels [66]: adaptation prompts individuals (and collectives) to efficiently repurpose capital previously accumulated in different forms (cultural, social, economic, etc.). Following [11], we conclude that a better understanding of collective adaptation benefits from the pooling of diverse insights: while studies of institutional change have documented strategies of gradual adaptation that progressively leverage and repurpose accumulated capital when change is difficult, the literature on cultural evolvability stresses the critical roles of diversity and social learning. In return, this work provides an empirical contribution to the literature that treats science as a cultural evolutionary system [16].

Throughout the paper, various methodological limitations were raised. First, the requirement on the quantity of authors' publications makes the cohort atypical. A second issue, already noted by Gieryn [18], lies in the arbitrariness in the choice of temporal and cognitive scopes for measuring change. In this paper, the choice in temporal segmentation was driven by the time-scale associated with the transformations in the landscape of experimental opportunities. For shorter time-scales, the very notion of research portfolio – as operationalized in this paper – may break down and lose any predictive force. As per the cognitive scope, previous works (e.g. [4, 5, 7]) have typically relied on the hierarchical PACS classification of physics literature, such that cognitive scales were imposed. For this work, the literature was divided into 15 topics that captured the features of change in the epistemic landscape discussed above, i.e. the rise of new probes of the cosmos, but other scales could have been considered. Moreover, although the unsupervised topic model approach is arguably a better proxy of cognitive change, it introduces noise which could in part explain the low predictive power of the models of change in scientists' research agendas. Additionally, the topic model was trained on the entire time range covered by the analysis. Changes in the relationships between topics and in their own vocabulary distributions are not considered, even though they constitute another interesting linguistic dimension of adaptive patterns that would deserve further investigation. Morever, the cost of shifting from one research area to another, is itself time-varying quantity in reality. Finally, one must be cautious before drawing strong causal conclusions from these findings. While the causal pathway "power → collaborations → stabilization of research interests" seems to be a reasonable interpretation of the results, the relationship between diversity and change is less clear; they could both be confounded by a latent trait specific to certain researchers (e.g. the "explorers" in [67]). Moreover, the sample size (N = 2108) is insufficient to explore sophisticated interactions or potential moderators in the comparative analysis.

Appendix: Supplementary material

A.1 Transformations in high-energy physics

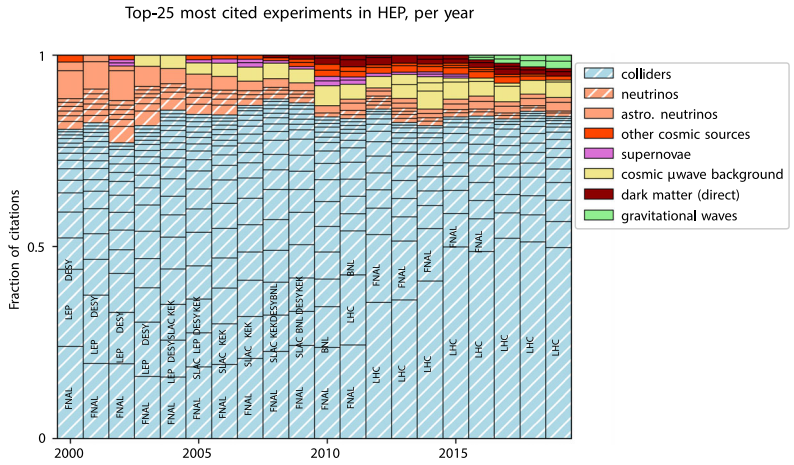


Figure 10 Share of citations of each of the most cited experiments in high-energy physics literature, between 2000 and 2019. Hatched rectangles correspond to experiments observing particles produced in colliders or nuclear reactors; other rectangles correspond to observations of phenomena or particles of astrophysical origin. We use citation and experiment data from the Inspire HEP database [41]. The classification of these experiments is our own

A.2 Cohort characteristics

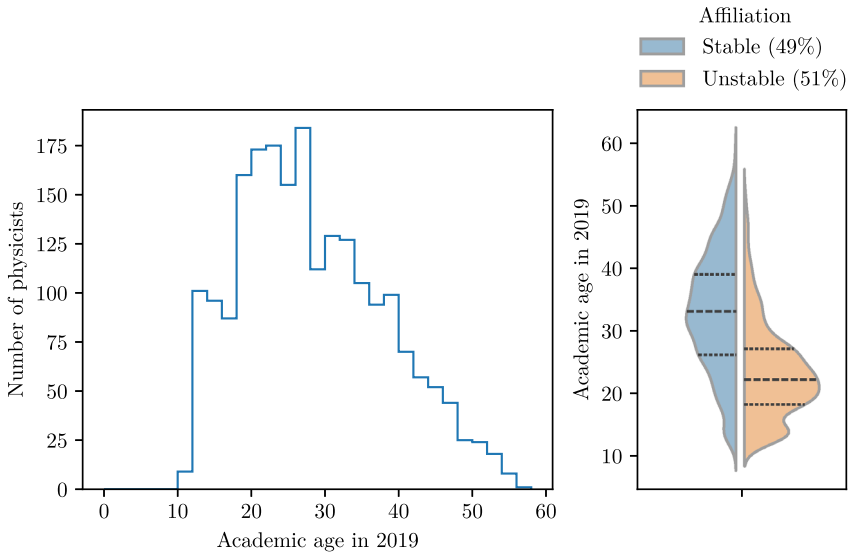


Figure 11 Distribution of academic age (in 2019) and affiliation stability within the considered cohort. Academic age (left plot) is evaluated as the time passed since the first publication. There is a cut-off in the distribution academic age due to the requirement on publication counts between 2000 and 2009. The median academic age is 27 years in 2019. 49% of the cohort has had at least one permanent affiliation spanning the period under consideration (right plot). As expected, permanent affiliations are associated with higher academic ages (bars indicate quartiles)

A.3 Topics

A.3.1 Word embeddings dimension

The word2vec skip-gram models is trained using different values for L the dimension of the embeddings' space. $L = 50$ (the choice made in the present paper) generally lies somewhere between under-fitting and over-fitting. The latter is a concern due to the small sample size (the model is trained on abstracts rather than full-texts).

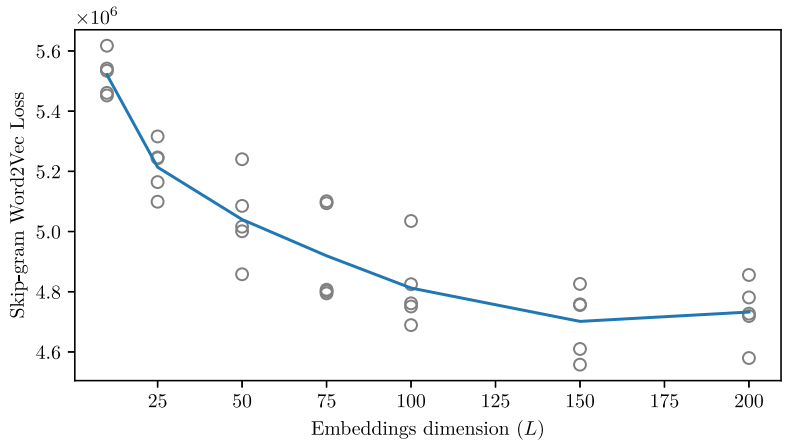


Figure 12 Loss of the word2vec model as a function of the embeddings dimension (L)

A.3.2 Keyword classification

In the Embedding Topic Model used in the present paper – as in topic models in general–, documents are mixtures of several topics; moreover, keywords may belong to several topics. This is a desirable feature: certain concepts serve different purposes depending on the context, and some concepts do not clearly belong to any research area. More importantly, this is crucial to the present approach, given that we must be sensitive to situations in which scientists have repurposed certain concepts to new research goals; i.e., instances where the same resources are applied in a new context (a new research area). For that, each keyword i from a document d is assigned a research area $z_{di} := \arg\max_{k=1,\dots,K} P(z_{di} = k|w_{di}, \boldsymbol{\theta}_d)$, which is the research area most probably associated with the keyword i given the topic distribution of the document $\boldsymbol{\theta}_d$ – i.e. the context. In fact:

$$z_{di} := \underset{k=1,\dots,K}{\arg\max}\, P(z_{di} = k|w_{di}, \boldsymbol{\theta}_d) = \underset{k=1,\dots,K}{\arg\max}\, \frac{P(w_{di}|z_{di} = k)P(z_{di} = k|\boldsymbol{\theta}_d)}{P(w_{di}|\boldsymbol{\theta}_d)} \tag{14}$$

In the process, we discard ambiguous keywords for which $H(z_{di}) \geq \log 2$, where H denotes the entropy of the distribution $P(z_{di} = k|w_{di})$. Either such keywords do not carry any scientific content, or the context is insufficient to disambiguate among the possible research areas to which they might belong.

The average effective amount of topics in documents according to the topic model – measured as $\exp H(\boldsymbol{\theta}_d)$ where H is the Shannon entropy – is 7.5, which is unrealistically high. The filtering and classification procedure reduces the average effective amount of topics to ∼3.0 per document, which is much more informative. This procedure is espe-

cially important for short texts such as abstracts, which poorly constrain the latent topic distribution $\boldsymbol{\theta}_d$.

A.3.3 List of topics

Table 1 Research areas, their top-words, and their correlation with a standard classification (PACS)

Research area	Top words	Most correlated PACS categories
AdS/CFT	boundary, holographic, flow, bulk, critical, conformal, critical_point, boundary_theory, cfts, point, bootstrap, central, conformal_anomaly, strip, fixed, free, entanglement_entropy, conformal_field_theory, criticality, condition	Gauge/string duality (0.27) Conformal field theory, algebraic […] (0.23) Theory of quantized fields (0.15) Theories and models of […] (0.14) Critical point phenomena (0.14)
Black holes	hole, black_hole, gravity, black, horizon, geometry, gravitational, spacetimes, spacetime, curvature, thermodynamics, einstein, schwarzschild, metric, ad, hair, relativity, space_time, observer, graviton	Quantum aspects of black holes, […] (0.51) Classical black holes (0.41) Physics of black holes (0.32) Exact solutions (0.24) Higher-dimensional black holes, […] (0.23)
Classical fields	scalar, general, first, scalar_field, massless, real, explicit, dynamical, second, exact, special, linear, full, symmetric, static, electromagnetic, classical, nonlinear, approximate, non_trivial	Modified theories of gravity (0.19) Lorentz and Poincaré invariance (0.16) Nonlinear or nonlocal theories and […] (0.11) Exact solutions (0.10) Higher-dimensional gravity and […] (0.10)
Collider physics	distribution, collision, production, cross_sections, section, parton, hadron, cross, cross_section, process, hadronic_collision, scattering, correction, fragmentation, partons, kinematics, transverse, impact, event, partonic	Perturbative calculations (0.29) Polarization in interactions and […] (0.28) Inclusive production with […] (0.27) Total and inclusive cross sections […] (0.26) Relativistic heavy-ion collisions (0.25)
Cosmology	constant, cosmological, inflation, cosmic, perturbation, vacuum, universe, inflationary, cosmology, fluctuation, inhomogeneity, tension, lambda, planck, inflaton, cosmological_perturbation, era, epoch, density, background	Particle-theory and field-theory […] (0.59) Cosmology (0.32) Observational cosmology (including […] (0.28) Dark energy (0.25) Background radiations (0.21)
Dark matter	matter, dark, dark_matter, detection, dm, signal, abundance, observation, relic, direct, constraint, candidate, wimp, asymmetric, prospect, dark_matter_particle, center, detectable, cold, contribute	Dark matter (0.74) γ-ray (0.22) Cosmic rays (0.19) γ-ray sources; γ-ray bursts (0.17) Elementary particle processes (0.17)
Electroweak sector	standard, higgs, boson, particle, standard_model, physic, lhc, sm, top, tev, collider, mssm, electroweak, minimal, phenomenology, extension, extra, supersymmetric_model, superpartners, new_particle	Extensions of electroweak Higgs sector (0.34) Supersymmetric models (0.33) Non-standard-model Higgs bosons (0.30) Supersymmetric partners of known […] (0.28) Standard-model Higgs bosons (0.27)
Gauge theory & Grand Unification	dimension, coupling, scale, structure, operator, fermion, value, matrix, number, su, charge, sector, spin, group, topological, anomalous, breaking, anomaly, global, flavor	Unified theories and models of […] (0.22) Unification of couplings; mass relations (0.17) Quark and lepton masses and mixing (0.13) Unified field theories and models (0.12) Field theories in dimensions other […] (0.12)

Table 1 (*Continued*)

Research area	Top words	Most correlated PACS categories
Hadrons	decay, data, channel, bound, resonance, gamma, meson, width, experimental_data, collaboration, kaon, prediction, experimental, measurement, admixture, narrow, process, hadronic_decay, s0, ratio	Decays of bottom mesons (0.30) Decays of J/ψ, Y, and other quarkonia (0.24) Meson-meson interactions (0.21) Decays of bottom mesons (0.20) Bottom mesons (|B|>0) (0.19)
Neutrinos & flavour physics	neutrino, violation, oscillation, flavor, cp, angle, mixing, experiment, lepton, flavour, hierarchy, majorana, cp_violation, beta, leptogenesis, asymmetry, neutrino_mass, neutrino_oscillation, smallness, generation	Neutrino mass and mixing (0.74) Non-standard-model neutrinos, […] (0.41) Ordinary neutrinos (0.30) Neutrino interactions (0.28) Quark and lepton masses and mixing (0.23)
Perturbative methods	amplitude, qcd, loop, diagram, sum, contribution, perturbative, expansion, vertex, rule, light_cone, perturbative_qcd, propagator, approach, correlator, one_loop, evaluation, nonperturbative, kernel, diagrammatic	General properties of perturbation […] (0.25) Other nonperturbative calculations (0.24) Sum rules (0.22) Perturbative calculations (0.21) General properties of QCD […] (0.16)
QCD	quark, chiral, magnetic, baryon, relativistic, moment, qcd, light_quark, strong, heavy, heavy_quark, lattice, magnetic_field, electric, deconfinement, chromodynamics, current, diquarks, plasma, color	Lattice QCD calculations (0.27) Chiral symmetries (0.26) Chiral Lagrangians (0.25) Quark-gluon plasma (0.23) General properties of QCD […] (0.20)
Quantum Field Theory	quantum, group, quantum_field, representation, quantisation, mechanic, quantum_field_theory, transformation, hamiltonians, algebra, finite_dimensional, quantization, commutator, algebraic, arbitrary, operator, qft, invariant, analog, associated	Algebraic methods (0.26) Noncommutative field theory (0.25) Quantum mechanics (0.22) Noncommutative geometry (0.19) Quantum groups (0.18)
String theory & supergravity	string, supersymmetric, superstring, six_dimensional, modulus, super, instantons, supergravity, dyons, n2, mathcaln, superpotentials, heterotic, sigma_models, n1, n4, gauged, space, deformation, compactifications	Supersymmetry (0.31) Strings and branes (0.29) Supergravity (0.29) Compactification and four- […] (0.25) D branes (0.20)
Thermodynamics	potential, effective, interaction, limit, temperature, action, finite, local, freedom, approximation, level, weak, chemical, force, effective_field_theory, lagrangian, finite_temperature, effective_field, degree, effective_theory	Finite-temperature field theory (0.26) Chiral symmetries (0.09) Nuclear forces (0.08) Quark-gluon plasma (0.08) General properties of QCD […] (0.08)
Uninterpretable	approach, method, analysis, recent, calculation, numerical, formalism, study, prediction, sigma, previous, work, theoretical, systematic, comparison, uncertainty, agreement, good, investigation, paper	Lattice QCD calculations (0.07) Baryon resonances (S=C=B=0) (0.05) Other nonperturbative calculations (0.05) Few-body systems (0.05) Lagrangian and Hamiltonian approach (0.05)
Uninterpretable	solution, equation, phase, space, time, system, transition, region, condition, constraint, dynamic, class, background, configuration, wave, range, motion, set, star, instability	Exact solutions (0.14) Nonlinear or nonlocal theories and […] (0.11) Extended classical solutions; […] (0.10) Relativistic wave equations (0.10) Modified theories of gravity (0.09)

Table 1 (*Continued*)

Research area	Top words	Most correlated PACS categories
Uninterpretable	form, correction, momentum, tensor, mode, relation, higher, factor, vector, invariant, formula, angular, part, theorem, spectrum, power, dimensional, invariance, expression, derivative	Electromagnetic form factors (0.17) Protons and neutrons (0.10) Lorentz and Poincaré invariance (0.08) Gauge field theories (0.06) Dispersion relations (0.06)
Uninterpretable	spectrum, low, problem, low_energy, important, high, property, high_energy, small, physical, soft, fundamental, behavior, analytic, behaviour, spectral, dispersion, essential, phenomenon, regime	General properties of QCD […] (0.07) Regge formalism (0.07) Wave propagation, transmission and […] (0.05) Elastic scattering (0.05) Lattice gauge theory (0.05)
Uninterpretable	different, possible, particular, present, various, mechanism, type, example, massive, several, scenario, simple, single, similar, consistent, addition, hand, different_type, interesting, way	Particle-theory and field-theory […] (0.07) Modified theories of gravity (0.07) Field theories in dimensions other […] (0.05) Cosmology (0.05) Dark energy (0.05)

A.3.4 Topic validation using the citation network

In order to validate the consistency and scientific dimension of the topics that were recovered, we verify that papers from a given topic tend to cite more papers from the same topic. Let $N_{k,k'}$ be the amount of citations of articles that belong to topic k' originating from articles that belong to k, and $N = \sum_{k,k'} N_{k,k'}$ the total number of citations. From this matrix, a normalized pointwise mutual correlation $\mathrm{npmi}(k,k')$ is calculated:

$$\mathrm{npmi}(k,k') = \log \frac{N_{k,k'}/N}{(\sum_i N_{k,i}/N)(\sum_i N_{i,k'}/N)} \tag{15}$$

$\mathrm{npmi}(k,k')$ is shown in Fig. 13. It measures how frequent citations from k to k' are, relative to what would be expected if citations were uniformly distributed. The diagonal values are positive, indicating that the topics we retrieved tend to refer to themselves significantly more than expected by chance alone, providing further evidence of their scientific content and coherence.

A.3.5 A comparison of three topic models

Figure 14 compares the ability of three topic models to measure the transformations in high-energy physics research resulting from changes in the landscape of experimental opportunities. In the coarse-grained model ($K_0 = 15$, in the middle), many types of experiments are lumped together into a single topic. It is therefore ill-suited for assessing the impact of the transformations in the landscape of experimental opportunities. The model used throughout the paper is well able to distinguish neutrino research for dark matter research, which have both undergone significant transformations according to Fig. 10. It is also better able to separate black hole phenomenology and cosmology, compared to the fine-grained model ($K_0 = 25$).

A.3.6 Topic validation using the PACS classification

Blue cells show that the research areas recovered from the topic model correlate with the PACS classification (which further confirms their scientific dimension), but also that they

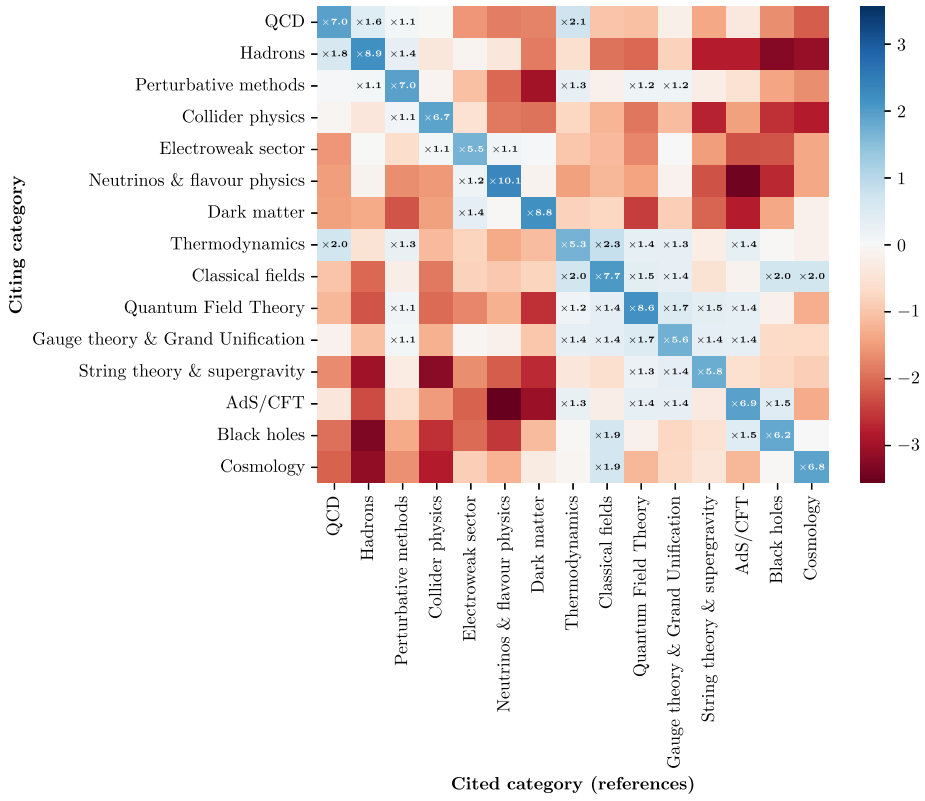


Figure 13 Normalized pointwise mutual information $\mathbf{npmi}(k, k')$ of the citation matrix $N_{k,k'}$. Positive values (in blue) indicate that the y-axis research area cites the x-axis research area more often than expected by chance; negative values indicate that citations occur less than expected by chance. When shown, individual values indicate how many times citations occur compared to chance alone. For instance, papers about Hadrons cite papers about Perturbative methods 1.4 times what would be expected if citations were uniform across research areas

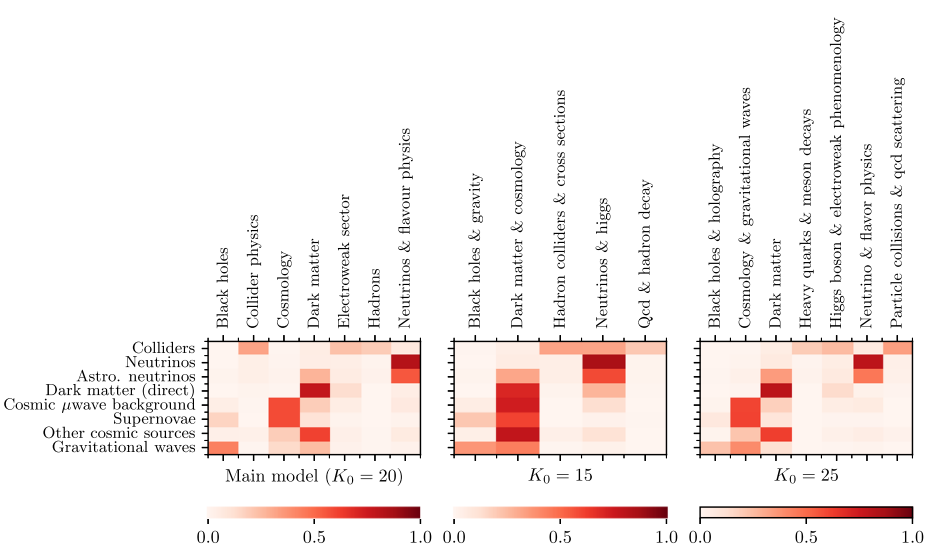


Figure 14 Origin of papers citing each type of experiment, according to three topic models: the model considered throughout the present paper (to the left), a model with fewer topics ($K_0 = 15$), and a model with more topics ($K_0 = 25$). A value of one (dark red) indicates that 100% of the papers citing a certain type of experiment (cf. rows) originate from a given topic (cf. columns)

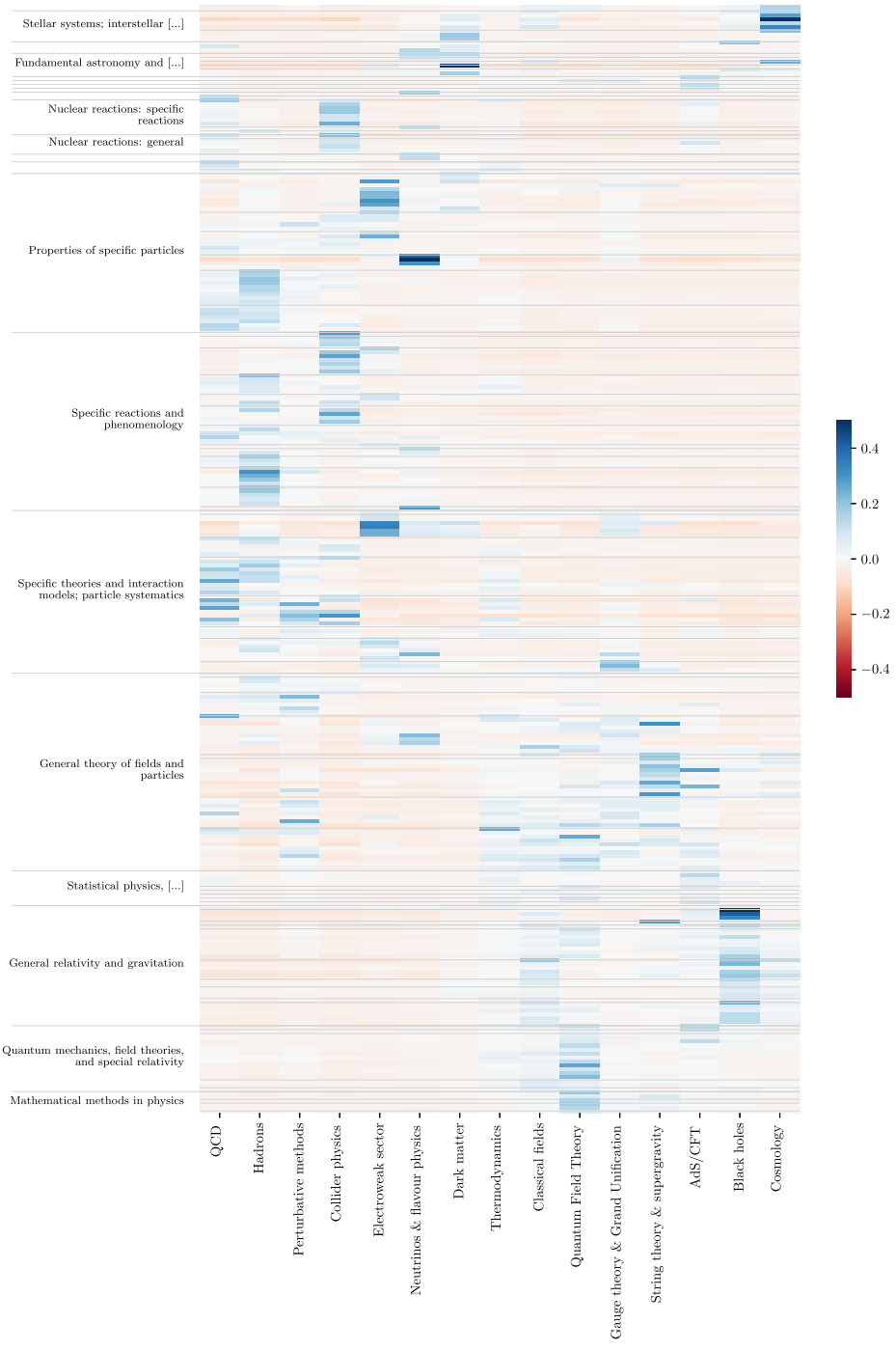


Figure 15 Correlation between PACS categories present in ≥ 100 publications in the corpus (rows), and the research areas recovered by the topic model (columns). Each colored cell indicates the correlation between a leaf category of the hierarchical PACS classification and one of the topics from the topic model

can give a different picture. For instance, each of the topics "dark matter", "black holes" and "cosmology" span over several higher-level categories of the PACS classification (e.g. "fundamental astronomy …" and "specific theories and interaction models …" for dark matter).

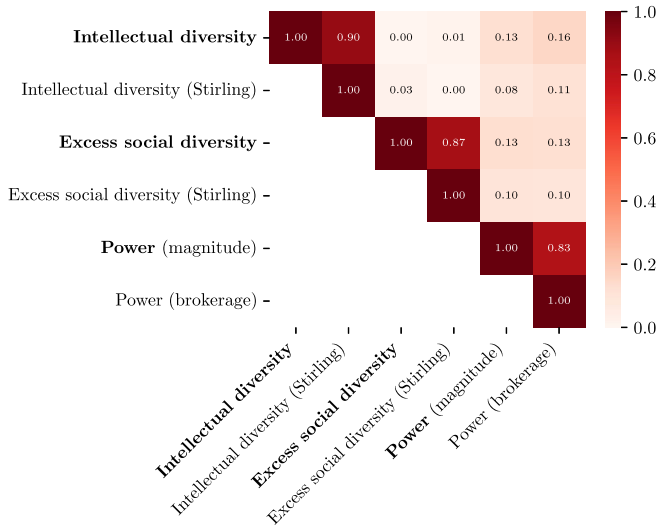


Figure 16 Correlation between different measures of capital. Measures considered in priority in this analysis are shown in bold. Alternative measures are shown for comparison purposes. By construction, excess social diversity is orthogonal to intellectual diversity ($R = 0$)

A.4 Measures of capital

Figure 16 shows the Pearson correlation between different measures of the diversity of intellectual and social capital and of power, as evaluated among the cohort of high-energy physicists. For comparison purposes, an alternative measure of diversity based on Stirling's index [68], with prior applications to studies of interdisciplinarity [69, 70] is evaluated.[23] A measure of brokerage is also considered.[24]

As shown in Fig. 16, the entropic measure of diversity considered in this paper correlates strongly with the Stirling measure. The magnitude of social capital (which is similar to degree centrality) correlates weakly with excess social diversity, thus emphasizing that power and diversity are partially orthogonal aspects of social capital. The magnitude of social capital is strongly correlated with brokerage; indeed, strongly connected scientists, with higher degree centrality, are also those scientists who initiate collaborations between otherwise disconnected scientists, as measured by brokerage.

A.5 Model performance over multiple corpora, temporal segmentations, and topic granularities

The predictive power of the model can be assessed by evaluating the total variation distance between the true distribution $\boldsymbol{y_a}$ and the predicted distribution $\boldsymbol{y_a}^{\text{pred}}$. This perfor-

[23] The Stirling-based diversity measured is evaluated as:

$$D_{\text{Stirling}} = 1 - \sum_{k,k'} d_{kk'} l_{ak} l_{ak'} \tag{16}$$

Where $d_{kk'}$ is the fraction of scientists who have more expertise than average in both k and k' among those that have expertise in one or the other (i.e., a similarity matrix). This follows from previous approaches for measuring research interdisciplinarity [69, 70].

[24] We evaluated brokerage as the amount of pairs of scientists that have collaborated with a given physicist while having no common collaborator except for this physicist. This effectively measures the extent to which this physicist connects otherwise disconnected scientists.

Table 2 Performance of the actual model versus that of the baseline model for various corpora, temporal segmentations, topic model parameters, and authorship criteria

Corpus	Authorship	K	L	Cohort size	Temporal segmentation	Model $\mu(d_{\mathrm{TV}}(\boldsymbol{y_a}, \boldsymbol{y_a}^{\mathrm{pred}}))$	Baseline $\mu(d_{\mathrm{TV}}(\boldsymbol{y_a}, \boldsymbol{x_a}))$
HEP	Any	20	50	2108	2000–2009 2015–2019	**0.306**	0.316
HEP	1st/last	20	50	2108	2000–2009 2015–2019	**0.306**	0.316
HEP	Any	20	50	1836	2000–2004 2005–2009	**0.262**	0.262
HEP	Any	20	50	2530	2005–2009 2010–2014	**0.261**	0.265
HEP	Any	20	50	3816	2010–2014 2015–2019	0.246	**0.244**
HEP	Any	15	50	2375	2000–2009 2015–2019	**0.293**	0.297
HEP	Any	25	50	2109	2000–2009 2010–2019	**0.315**	0.328
HEP	Any	15	50	2069	2000–2009 2015–2019	**0.290**	0.295
HEP	Any	20	150	2169	2000–2009 2015–2019	**0.309**	0.318
ACL Anthology	Any	20	50	578	2002–2011 2012–2022	**0.337**	0.466

mance metric is calculated via 10-fold cross-validation. It is compared to a baseline model that predicts no change in the research agenda ($\boldsymbol{y_a}^{\mathrm{baseline}} = \boldsymbol{x_a}$). The results are shown in Table 2. For the cohort of high-energy physicists, the model performs only marginally better than the baseline, given that individuals have remained quite conservative on average, most of the fluctuations being difficult to predict. Table 2 also considers a cohort of scientists from the ACL anthology corpus of computational linguistics research [71], by running the same pipeline (the measurement of research portfolios during two consecutive time periods using our topic model approach and the training of the trajectory model). Although the data are of significantly lesser quality, research portfolios have undergone much more significant transformations in this dataset (see Appendix A.10, Fig. 20). Consequently, our model performs much better than the baseline for this cohort.

A.6 Learning costs and optimal transport

The MCMC algorithm from [17] is run on 1,000,000 iterations of the "MetroMC" algorithm, using what the authors call a "P1" prior (that is, a prior such that $\sum_{kk'} C_{kk'} = C_0 = \mathrm{cst}$[25]). More precisely, we assume that:

$$P(c_{kk'}|p_{kk'}) = \frac{1}{Z}\frac{1}{\prod_{kk'} c_{kk'}^{1/2}} \exp\left(-\alpha D_{KL}(c_{kk'}||p_{kk'})\right) \text{ with } c_{kk'} = C_{kk'}/C_0 \tag{17}$$

$$\text{and } p_{kk'} = \mathrm{softmax}(\beta(1 - v_{kk'})) \tag{18}$$

[25] We chose the minimum value of C_0 for which the system admitted a solution.

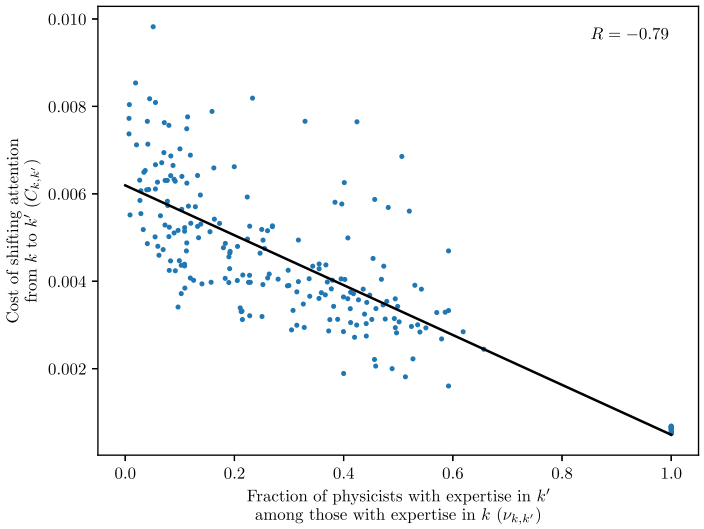


Figure 17 Cost of shifting a unit of cohort's research attention from k to k' as a function of the fraction of physicists with expertise in k' among those with expertise in k ($\nu_{kk'}$)

(17) is sometimes referred to as the entropic prior [72, 73]. The mean posterior values of $C_{k,k'}$ are shown in Fig. 17, as a function of the knowledge gap from k to k'. The knowledge gap $1 - \nu_{kk'}$ is the fraction of experts in k who do not hold significant expertise in k' (ν is shown in Fig. 6b). A significant correlation is found ($R = -0.76$). This is true also for the replication dataset of computational linguistics research (Appendix A.10), for which $R = -0.63$.

A.7 Effect of capital on strategies of change

A.7.1 Model for the magnitude of change

The model for c_a is:

$$
\begin{aligned}
z(c_a) &\sim \mathcal{N}(\mu_a, \sigma) \\
\mu_a &= \beta^{\text{int-div}} z(D(\boldsymbol{I_a})) + \beta^{\text{soc-div}} z(D^*(\boldsymbol{S_a})) + \beta^{\text{power}} z(P(\boldsymbol{S_a})) + \beta^{\text{stability}} p_a \\
&\quad + \beta^{\text{age}} z(a_a) + \beta^{\text{prod}} z(\pi_a) + \mu_{k_a}^{\text{area}} + \mu \\
k_a &= \arg\max_k x_{ak} \\
\beta, \mu &\sim \mathcal{N}(0, 1) \\
|\mu_k^{\text{area}}| &\sim \text{Exponential}(\tau) \\
\tau, \sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

Where $z(\cdot)$ denotes standardized variables.

A.7.2 Model for the probability of having entered/exited a research area

The model for the probability p_a of having entered a new field:

$$
\begin{aligned}
\text{logit}(p_a) &= \beta^{\text{int-div}} z(D(\boldsymbol{I_a})) + \beta^{\text{soc-div}} z(D^*(\boldsymbol{S_a})) + \beta^{\text{power}} z(P(\boldsymbol{S_a})) + \beta^{\text{stability}} p_a \\
&\quad + \beta^{\text{age}} z(a_a) + \beta^{\text{prod}} z(\pi_a) + \mu_{k_a}^{\text{area}} + \mu
\end{aligned}
$$

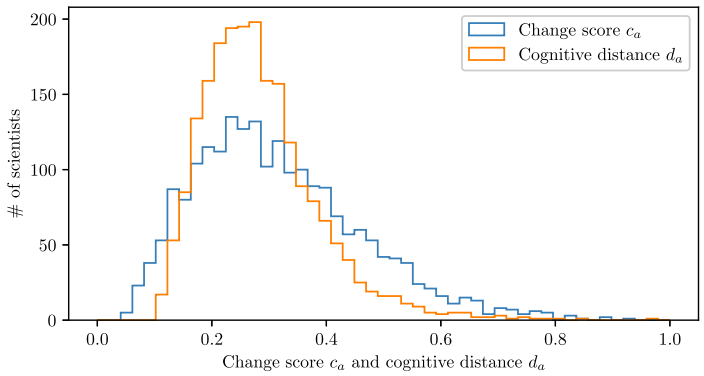


Figure 18 Distribution of change and cognitive distances in the cohort. Higher values correspond to more drastic changes in a scientists' research agenda

$$k_a = \arg\max_k x_{ak}$$

$$\beta, \mu \sim \mathcal{N}(0,1)$$

$$|\mu_k^{\text{area}}| \sim \text{Exponential}(\tau)$$

$$\tau, \sigma \sim \text{Exponential}(1)$$

The same model structure is used for the probability of having exited a research area.

A.7.3 Effect of capital and robustness checks

Table 3 Effect of each variable on (a) the change score and (b) the cognitive distance for each model. The reference model uses entropy as the diversity measure D and the magnitude of intellectual capital as a measure of power P. Values indicate the mean posterior effect size and the 95% credible interval. Significant effects are shown in bold

Predictor	Dep. variable					
	Change score (c_a), model:			Cognitive distance (d_a), model:		
	Reference	D = Stirling	P = Brokerage	Reference	D = Stirling	P = Brokerage
Intellectual capital (diversity)	**+0.28**$^{+0.044}_{-0.044}$	**+0.28**$^{+0.042}_{-0.043}$	**+0.27**$^{+0.044}_{-0.043}$	**+0.33**$^{+0.043}_{-0.042}$	**+0.34**$^{+0.042}_{-0.042}$	**+0.32**$^{+0.043}_{-0.043}$
Social capital (diversity)	**+0.09**$^{+0.04}_{-0.04}$	**+0.07**$^{+0.04}_{-0.04}$	**+0.08**$^{+0.04}_{-0.04}$	**+0.11**$^{+0.04}_{-0.041}$	**+0.09**$^{+0.04}_{-0.04}$	**+0.1**$^{+0.04}_{-0.04}$
Social capital (power)	**−0.09**$^{+0.06}_{-0.06}$	**−0.07**$^{+0.06}_{-0.06}$	−0.02$^{+0.05}_{-0.05}$	**−0.14**$^{+0.061}_{-0.061}$	**−0.12**$^{+0.06}_{-0.06}$	−0.05$^{+0.05}_{-0.05}$
Stable affiliation	−0.01$^{+0.09}_{-0.09}$	−0.009$^{+0.09}_{-0.09}$	−0.0008$^{+0.09}_{-0.09}$	−0.007$^{+0.09}_{-0.09}$	+0.0009$^{+0.09}_{-0.09}$	+0.01$^{+0.09}_{-0.09}$
Academic age	**−0.1**$^{+0.05}_{-0.05}$	**−0.1**$^{+0.05}_{-0.05}$	**−0.1**$^{+0.047}_{-0.047}$	**−0.07**$^{+0.05}_{-0.05}$	**−0.07**$^{+0.047}_{-0.047}$	**−0.08**$^{+0.05}_{-0.05}$
Productivity (co-authored)	**−0.12**$^{+0.058}_{-0.059}$	**−0.12**$^{+0.058}_{-0.058}$	**−0.17**$^{+0.052}_{-0.053}$	**−0.1**$^{+0.06}_{-0.06}$	**−0.1**$^{+0.058}_{-0.056}$	**−0.17**$^{+0.053}_{-0.052}$
Productivity (solo-authored)	**−0.05**$^{+0.041}_{-0.04}$	**−0.05**$^{+0.04}_{-0.04}$	**−0.06**$^{+0.04}_{-0.04}$	−0.04$^{+0.04}_{-0.04}$	−0.03$^{+0.04}_{-0.04}$	**−0.04**$^{+0.04}_{-0.04}$
Hadrons	−0.009$^{+0.2}_{-0.2}$	−0.11$^{+0.18}_{-0.2}$	−0.008$^{+0.2}_{-0.2}$	+0.03$^{+0.1}_{-0.1}$	−0.05$^{+0.1}_{-0.2}$	+0.04$^{+0.1}_{-0.1}$
String theory & supergravity	**+0.28**$^{+0.15}_{-0.15}$	**+0.34**$^{+0.18}_{-0.18}$	**+0.25**$^{+0.15}_{-0.15}$	+0.11$^{+0.13}_{-0.11}$	**+0.21**$^{+0.15}_{-0.15}$	+0.07$^{+0.1}_{-0.1}$
Perturbative methods	+0.12$^{+0.22}_{-0.18}$	+0.06$^{+0.2}_{-0.2}$	+0.13$^{+0.22}_{-0.18}$	+0.15$^{+0.22}_{-0.17}$	+0.1$^{+0.21}_{-0.17}$	+0.16$^{+0.23}_{-0.18}$
Classical fields	−0.25$^{+0.36}_{-0.59}$	−0.21$^{+0.38}_{-0.55}$	−0.23$^{+0.35}_{-0.57}$	−0.19$^{+0.27}_{-0.58}$	−0.17$^{+0.3}_{-0.51}$	−0.16$^{+0.26}_{-0.55}$
Collider physics	**−0.19**$^{+0.16}_{-0.17}$	**−0.34**$^{+0.19}_{-0.19}$	**−0.2**$^{+0.17}_{-0.17}$	−0.02$^{+0.1}_{-0.1}$	**−0.16**$^{+0.15}_{-0.16}$	−0.03$^{+0.1}_{-0.1}$
Neutrinos & flavour physics	**+0.21**$^{+0.18}_{-0.17}$	+0.17$^{+0.2}_{-0.18}$	**+0.18**$^{+0.18}_{-0.17}$	+0.11$^{+0.16}_{-0.13}$	+0.1$^{+0.2}_{-0.1}$	+0.09$^{+0.2}_{-0.1}$

Table 3 (*Continued*)

Predictor	Dep. variable					
	Change score (c_a), model:			Cognitive distance (d_a), model:		
	Reference	D = Stirling	P = Brokerage	Reference	D = Stirling	P = Brokerage
Black holes	$+0.06^{+0.2}_{-0.2}$	$+0.15^{+0.22}_{-0.19}$	$+0.05^{+0.2}_{-0.2}$	$-0.03^{+0.1}_{-0.2}$	$+0.06^{+0.2}_{-0.1}$	$-0.04^{+0.1}_{-0.2}$
Gauge theory & Grand Unification	$-0.07^{+0.3}_{-0.4}$	$-0.05^{+0.3}_{-0.4}$	$-0.06^{+0.3}_{-0.3}$	$-0.02^{+0.2}_{-0.3}$	$+0.003^{+0.3}_{-0.3}$	$-0.008^{+0.2}_{-0.3}$
Dark matter	$\mathbf{-0.27}^{+0.24}_{-0.25}$	$\mathbf{-0.32}^{+0.26}_{-0.27}$	$\mathbf{-0.28}^{+0.24}_{-0.25}$	$-0.13^{+0.17}_{-0.23}$	$-0.18^{+0.2}_{-0.24}$	$-0.13^{+0.17}_{-0.23}$
Thermodynamics	$+0.14^{+0.36}_{-0.26}$	$+0.27^{+0.41}_{-0.32}$	$+0.16^{+0.37}_{-0.27}$	$+0.1^{+0.34}_{-0.2}$	$+0.25^{+0.41}_{-0.29}$	$+0.12^{+0.36}_{-0.21}$
Cosmology	$-0.02^{+0.16}_{-0.17}$	$+0.01^{+0.2}_{-0.2}$	$-0.03^{+0.2}_{-0.2}$	$-0.06^{+0.1}_{-0.2}$	$-0.02^{+0.1}_{-0.2}$	$-0.06^{+0.13}_{-0.17}$
Electroweak sector	$-0.14^{+0.15}_{-0.16}$	$\mathbf{-0.21}^{+0.18}_{-0.18}$	$\mathbf{-0.17}^{+0.16}_{-0.16}$	$-0.04^{+0.1}_{-0.1}$	$-0.1^{+0.13}_{-0.15}$	$-0.07^{+0.1}_{-0.1}$
QCD	$-0.009^{+0.2}_{-0.2}$	$-0.07^{+0.2}_{-0.2}$	$+0.001^{+0.2}_{-0.2}$	$+0.02^{+0.1}_{-0.1}$	$-0.03^{+0.14}_{-0.16}$	$+0.04^{+0.1}_{-0.1}$
Quantum Field Theory	$+0.04^{+0.2}_{-0.2}$	$+0.09^{+0.2}_{-0.2}$	$+0.04^{+0.21}_{-0.19}$	$-0.09^{+0.2}_{-0.2}$	$-0.04^{+0.2}_{-0.2}$	$-0.08^{+0.2}_{-0.2}$
AdS/CFT	$+0.003^{+0.2}_{-0.2}$	$+0.07^{+0.3}_{-0.2}$	$+0.005^{+0.2}_{-0.2}$	$+0.02^{+0.2}_{-0.2}$	$+0.11^{+0.27}_{-0.2}$	$+0.03^{+0.2}_{-0.2}$

Table 4 Effect of each variable on (a) the probability of having entered a new research area and (b) the probability of having exited a research area, for each model. The reference model uses entropy as the diversity measure D and the magnitude of intellectual capital as a measure of power P. Values indicate the mean posterior effect size and the 95% credible interval. Significant effects are shown in bold

Predictor	Dep. variable					
	Entered a new research area, model:			Exited a research area, model:		
	Reference	D = Stirling	P = Brokerage	Reference	D = Stirling	P = Brokerage
Intellectual capital (diversity)	$\mathbf{+0.2}^{+0.11}_{-0.11}$	$\mathbf{+0.17}^{+0.1}_{-0.1}$	$\mathbf{+0.19}^{+0.11}_{-0.11}$	$\mathbf{+1}^{+0.14}_{-0.14}$	$\mathbf{+0.85}^{+0.12}_{-0.12}$	$\mathbf{+1}^{+0.14}_{-0.14}$
Social capital (diversity)	$\mathbf{+0.22}^{+0.1}_{-0.1}$	$\mathbf{+0.18}^{+0.1}_{-0.1}$	$\mathbf{+0.22}^{+0.099}_{-0.1}$	$+0.04^{+0.1}_{-0.1}$	$+0.04^{+0.1}_{-0.1}$	$+0.04^{+0.1}_{-0.1}$
Social capital (power)	$+0.006^{+0.1}_{-0.1}$	$+0.03^{+0.15}_{-0.15}$	$+0.04^{+0.1}_{-0.1}$	$-0.03^{+0.2}_{-0.2}$	$+0.02^{+0.2}_{-0.2}$	$+0.03^{+0.1}_{-0.1}$
Stable affiliation	$-0.19^{+0.22}_{-0.22}$	$-0.18^{+0.22}_{-0.22}$	$-0.19^{+0.22}_{-0.22}$	$+0.04^{+0.2}_{-0.2}$	$+0.06^{+0.2}_{-0.2}$	$+0.04^{+0.24}_{-0.24}$
Academic age	$+0.04^{+0.12}_{-0.11}$	$+0.04^{+0.1}_{-0.1}$	$+0.04^{+0.1}_{-0.1}$	$\mathbf{-0.21}^{+0.12}_{-0.12}$	$\mathbf{-0.21}^{+0.12}_{-0.12}$	$\mathbf{-0.22}^{+0.12}_{-0.12}$
Productivity (co-authored)	$-0.07^{+0.1}_{-0.1}$	$-0.08^{+0.1}_{-0.1}$	$-0.09^{+0.1}_{-0.1}$	$\mathbf{-0.28}^{+0.15}_{-0.14}$	$\mathbf{-0.28}^{+0.15}_{-0.15}$	$\mathbf{-0.31}^{+0.13}_{-0.13}$
Productivity (solo-authored)	$-0.05^{+0.1}_{-0.09}$	$-0.05^{+0.1}_{-0.09}$	$-0.06^{+0.1}_{-0.1}$	$-0.02^{+0.1}_{-0.1}$	$-0.007^{+0.1}_{-0.1}$	$-0.03^{+0.1}_{-0.1}$
Hadrons	$-0.14^{+0.29}_{-0.36}$	$-0.24^{+0.33}_{-0.39}$	$-0.14^{+0.29}_{-0.35}$	$+0.03^{+0.3}_{-0.2}$	$-0.09^{+0.3}_{-0.4}$	$+0.04^{+0.3}_{-0.2}$
String theory & supergravity	$\mathbf{+0.32}^{+0.32}_{-0.3}$	$\mathbf{+0.4}^{+0.35}_{-0.33}$	$\mathbf{+0.32}^{+0.32}_{-0.29}$	$\mathbf{+0.3}^{+0.34}_{-0.3}$	$\mathbf{+0.65}^{+0.39}_{-0.38}$	$+0.28^{+0.34}_{-0.29}$
Perturbative methods	$+0.11^{+0.45}_{-0.35}$	$+0.09^{+0.5}_{-0.4}$	$+0.11^{+0.45}_{-0.35}$	$-0.03^{+0.29}_{-0.34}$	$-0.13^{+0.4}_{-0.48}$	$-0.03^{+0.3}_{-0.3}$
Classical fields	$+0.22^{+1.1}_{-0.6}$	$+0.34^{+1.4}_{-0.7}$	$+0.22^{+1.1}_{-0.6}$	$-0.07^{+0.4}_{-0.6}$	$-0.07^{+0.7}_{-0.8}$	$-0.07^{+0.4}_{-0.6}$
Collider physics	$\mathbf{-0.43}^{+0.33}_{-0.34}$	$\mathbf{-0.61}^{+0.33}_{-0.34}$	$\mathbf{-0.42}^{+0.33}_{-0.34}$	$-0.01^{+0.2}_{-0.2}$	$-0.28^{+0.32}_{-0.37}$	$-0.02^{+0.2}_{-0.2}$
Neutrinos & flavour physics	$+0.08^{+0.3}_{-0.3}$	$+0.04^{+0.3}_{-0.3}$	$+0.07^{+0.3}_{-0.3}$	$-0.21^{+0.25}_{-0.35}$	$-0.31^{+0.35}_{-0.41}$	$-0.21^{+0.26}_{-0.36}$
Black holes	$-0.0006^{+0.3}_{-0.3}$	$+0.06^{+0.4}_{-0.3}$	$-0.003^{+0.3}_{-0.3}$	$+0.08^{+0.4}_{-0.3}$	$+0.43^{+0.53}_{-0.45}$	$+0.08^{+0.4}_{-0.3}$
Gauge theory & Grand Unification	$-0.04^{+0.6}_{-0.6}$	$-0.04^{+0.6}_{-0.7}$	$-0.03^{+0.6}_{-0.6}$	$-0.08^{+0.4}_{-0.6}$	$-0.11^{+0.64}_{-0.79}$	$-0.08^{+0.4}_{-0.6}$
Dark matter	$\mathbf{-0.62}^{+0.55}_{-0.56}$	$\mathbf{-0.68}^{+0.55}_{-0.56}$	$\mathbf{-0.63}^{+0.55}_{-0.56}$	$-0.05^{+0.3}_{-0.4}$	$-0.11^{+0.42}_{-0.5}$	$-0.05^{+0.3}_{-0.4}$
Thermodynamics	$-0.03^{+0.5}_{-0.6}$	$+0.01^{+0.62}_{-0.58}$	$-0.02^{+0.5}_{-0.6}$	$-0.05^{+0.4}_{-0.6}$	$+0.03^{+0.7}_{-0.6}$	$-0.05^{+0.41}_{-0.53}$
Cosmology	$-0.07^{+0.3}_{-0.4}$	$-0.03^{+0.3}_{-0.4}$	$-0.07^{+0.3}_{-0.4}$	$+0.09^{+0.4}_{-0.3}$	$+0.36^{+0.56}_{-0.43}$	$+0.09^{+0.4}_{-0.3}$
Electroweak sector	$+0.05^{+0.3}_{-0.3}$	$+0.009^{+0.3}_{-0.3}$	$+0.05^{+0.3}_{-0.3}$	$-0.003^{+0.2}_{-0.2}$	$-0.04^{+0.3}_{-0.3}$	$-0.009^{+0.2}_{-0.2}$
QCD	$+0.008^{+0.3}_{-0.3}$	$-0.03^{+0.3}_{-0.4}$	$+0.01^{+0.3}_{-0.3}$	$+0.04^{+0.32}_{-0.26}$	$+0.04^{+0.4}_{-0.4}$	$+0.04^{+0.3}_{-0.3}$
Quantum Field Theory	$+0.15^{+0.52}_{-0.38}$	$+0.22^{+0.57}_{-0.41}$	$+0.15^{+0.52}_{-0.38}$	$+0.05^{+0.4}_{-0.3}$	$+0.26^{+0.64}_{-0.44}$	$+0.05^{+0.4}_{-0.3}$
AdS/CFT	$+0.14^{+0.6}_{-0.42}$	$+0.22^{+0.67}_{-0.47}$	$+0.14^{+0.59}_{-0.42}$	$-0.06^{+0.4}_{-0.5}$	$+0.07^{+0.6}_{-0.5}$	$-0.06^{+0.3}_{-0.5}$

Table 5 Summary of the effect of each predictor on the change score (c_d) across topic models and temporal segmentations. Values indicate the mean posterior effect size and the 95% credible interval. Significant effects are shown in bold

Authorship	K_0	L	Temporal segmentation	Intell. capital (diversity)	Soc. capital (diversity)	Soc. capital (power)	Stable affiliation	Academic age	Prod. (co-auth.)	Prod. (solo-auth.)
Any	20	50	2000–2009 2015–2019	**+0.28**$^{+0.044}_{-0.044}$	**+0.09**$^{+0.04}_{-0.04}$	**−0.09**$^{+0.06}_{-0.06}$	−0.01$^{+0.09}_{-0.09}$	**−0.1**$^{+0.05}_{-0.05}$	**−0.12**$^{+0.058}_{-0.059}$	**−0.05**$^{+0.041}_{-0.04}$
1st/last	20	50	2000–2009 2015–2019	**+0.25**$^{+0.055}_{-0.056}$	**+0.09**$^{+0.05}_{-0.05}$	−0.01$^{+0.08}_{-0.08}$	+0.04$^{+0.1}_{-0.1}$	**−0.12**$^{+0.059}_{-0.06}$	**−0.16**$^{+0.075}_{-0.074}$	−0.05$^{+0.05}_{-0.05}$
Any	20	50	2000–2004 2005–2009	**+0.37**$^{+0.045}_{-0.045}$	**+0.12**$^{+0.043}_{-0.043}$	−0.02$^{+0.06}_{-0.06}$	−0.02$^{+0.1}_{-0.1}$	**−0.08**$^{+0.05}_{-0.05}$	**−0.21**$^{+0.059}_{-0.059}$	**−0.06**$^{+0.04}_{-0.04}$
Any	20	50	2010–2014 2015–2019	**+0.36**$^{+0.039}_{-0.039}$	**+0.08**$^{+0.04}_{-0.04}$	**−0.06**$^{+0.051}_{-0.052}$	**−0.11**$^{+0.087}_{-0.086}$	**−0.06**$^{+0.04}_{-0.04}$	**−0.21**$^{+0.051}_{-0.051}$	**−0.04**$^{+0.04}_{-0.04}$
Any	20	50	2000–2009 2010–2019	**+0.37**$^{+0.033}_{-0.033}$	**+0.06**$^{+0.03}_{-0.03}$	**−0.06**$^{+0.05}_{-0.05}$	−0.02$^{+0.07}_{-0.07}$	**−0.04**$^{+0.036}_{-0.035}$	**−0.22**$^{+0.046}_{-0.046}$	**−0.04**$^{+0.03}_{-0.03}$
Any	20	50	2000–2009 2015–2019	**+0.32**$^{+0.044}_{-0.044}$	**+0.09**$^{+0.04}_{-0.04}$	**−0.09**$^{+0.06}_{-0.06}$	+0.02$^{+0.09}_{-0.09}$	**−0.12**$^{+0.047}_{-0.046}$	**−0.15**$^{+0.057}_{-0.057}$	**−0.05**$^{+0.04}_{-0.04}$
Any	15	50	2000–2009 2015–2019	**+0.33**$^{+0.04}_{-0.04}$	**+0.09**$^{+0.04}_{-0.04}$	**−0.12**$^{+0.057}_{-0.057}$	−0.04$^{+0.09}_{-0.09}$	**−0.12**$^{+0.05}_{-0.051}$	**−0.1**$^{+0.054}_{-0.054}$	**−0.06**$^{+0.04}_{-0.04}$
Any	25	50	2000–2009 2015–2019	**+0.35**$^{+0.042}_{-0.042}$	**+0.13**$^{+0.04}_{-0.04}$	**−0.15**$^{+0.058}_{-0.058}$	−0.002$^{+0.09}_{-0.09}$	**−0.15**$^{+0.052}_{-0.052}$	**−0.12**$^{+0.056}_{-0.056}$	**−0.05**$^{+0.04}_{-0.04}$
Any	15	150	2000–2009 2015–2019	**+0.35**$^{+0.044}_{-0.044}$	**+0.07**$^{+0.04}_{-0.04}$	**−0.08**$^{+0.06}_{-0.06}$	−0.02$^{+0.089}_{-0.089}$	**−0.08**$^{+0.05}_{-0.05}$	**−0.1**$^{+0.056}_{-0.057}$	−0.02$^{+0.04}_{-0.04}$
Any	20	150	2000–2009 2015–2019	**+0.34**$^{+0.044}_{-0.043}$	**+0.08**$^{+0.04}_{-0.04}$	**−0.09**$^{+0.06}_{-0.06}$	+0.008$^{+0.09}_{-0.09}$	**−0.1**$^{+0.046}_{-0.046}$	**−0.12**$^{+0.056}_{-0.056}$	**−0.06**$^{+0.04}_{-0.04}$
Any	25	150	2000–2009 2015–2019	**+0.34**$^{+0.044}_{-0.044}$	**+0.06**$^{+0.04}_{-0.04}$	**−0.09**$^{+0.06}_{-0.06}$	−0.009$^{+0.09}_{-0.09}$	**−0.11**$^{+0.047}_{-0.047}$	**−0.13**$^{+0.057}_{-0.057}$	−0.02$^{+0.04}_{-0.04}$

Table 6 Summary of the effect of each predictor on the cognitive distance (d_a) across topic models and temporal segmentations. Values indicate the mean posterior effect size and the 95% credible interval. Significant effects are shown in bold

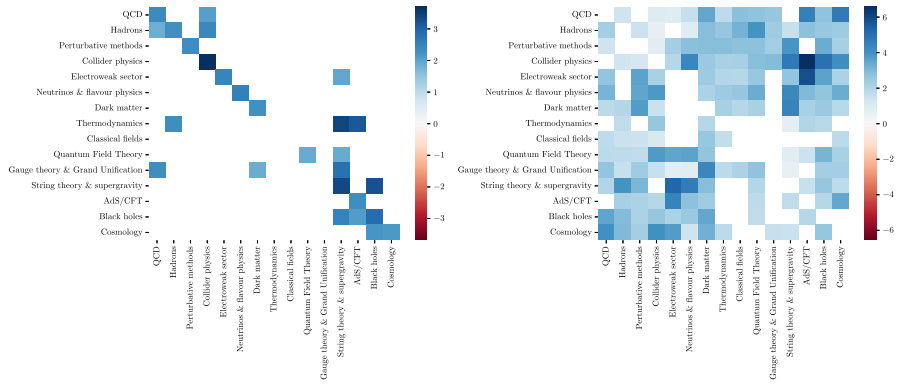
Authorship	K_0	L	Temporal segmen1tation	Intell. capital (diversity)	Soc. capital (diversity)	Soc. capital (power)	Stable affiliation	Academic age	Prod. (co-auth.)	Prod. (solo-auth.)
Any	20	50	2000–2009 2015–2019	**+0.33** $^{+0.043}_{-0.042}$	**+0.11** $^{+0.04}_{-0.041}$	**−0.14** $^{+0.061}_{-0.061}$	−0.007 $^{+0.09}_{-0.09}$	**−0.07** $^{+0.05}_{-0.05}$	**−0.1** $^{+0.06}_{-0.06}$	−0.04 $^{+0.04}_{-0.04}$
1st/last	20	50	2000–2009 2015–2019	**+0.3** $^{+0.056}_{-0.055}$	**+0.1** $^{+0.053}_{-0.052}$	−0.06 $^{+0.08}_{-0.08}$	−0.01 $^{+0.1}_{-0.1}$	**−0.07** $^{+0.06}_{-0.06}$	**−0.12** $^{+0.076}_{-0.075}$	−0.04 $^{+0.05}_{-0.05}$
Any	20	50	2000–2004 2005–2009	**+0.37** $^{+0.044}_{-0.044}$	**+0.13** $^{+0.043}_{-0.044}$	−0.05 $^{+0.06}_{-0.06}$	−0.04 $^{+0.1}_{-0.1}$	**−0.07** $^{+0.05}_{-0.05}$	**−0.19** $^{+0.06}_{-0.06}$	**−0.07** $^{+0.04}_{-0.04}$
Any	20	50	2010–2014 2015–2019	**+0.37** $^{+0.037}_{-0.037}$	**+0.09** $^{+0.04}_{-0.04}$	**−0.07** $^{+0.05}_{-0.05}$	**−0.12** $^{+0.086}_{-0.086}$	−0.03 $^{+0.04}_{-0.04}$	**−0.21** $^{+0.05}_{-0.051}$	−0.04 $^{+0.04}_{-0.04}$
Any	20	50	2000–2009 2010–2019	**+0.4** $^{+0.031}_{-0.031}$	**+0.06** $^{+0.031}_{-0.031}$	**−0.09** $^{+0.05}_{-0.05}$	+0.01 $^{+0.07}_{-0.07}$	**−0.05** $^{+0.035}_{-0.035}$	**−0.19** $^{+0.046}_{-0.046}$	**−0.03** $^{+0.03}_{-0.03}$
Any	20	50	2000–2009 2015–2019	**+0.36** $^{+0.043}_{-0.043}$	**+0.11** $^{+0.04}_{-0.04}$	**−0.14** $^{+0.059}_{-0.06}$	+0.003 $^{+0.09}_{-0.09}$	**−0.09** $^{+0.05}_{-0.05}$	**−0.14** $^{+0.056}_{-0.056}$	**−0.05** $^{+0.04}_{-0.04}$
Any	15	50	2000–2009 2015–2019	**+0.28** $^{+0.041}_{-0.041}$	**+0.1** $^{+0.04}_{-0.04}$	**−0.18** $^{+0.057}_{-0.058}$	−0.04 $^{+0.09}_{-0.09}$	**−0.09** $^{+0.05}_{-0.05}$	−0.05 $^{+0.05}_{-0.06}$	−0.03 $^{+0.04}_{-0.04}$
Any	25	50	2000–2009 2015–2019	**+0.25** $^{+0.043}_{-0.043}$	**+0.12** $^{+0.041}_{-0.04}$	**−0.18** $^{+0.059}_{-0.058}$	+0.06 $^{+0.09}_{-0.09}$	**−0.12** $^{+0.053}_{-0.053}$	**−0.06** $^{+0.06}_{-0.06}$	−0.03 $^{+0.04}_{-0.04}$
Any	15	150	2000–2009 2015–2019	**+0.28** $^{+0.048}_{-0.047}$	**+0.07** $^{+0.04}_{-0.04}$	**−0.09** $^{+0.06}_{-0.06}$	+0.03 $^{+0.09}_{-0.1}$	**−0.07** $^{+0.05}_{-0.05}$	**−0.08** $^{+0.06}_{-0.06}$	−0.02 $^{+0.04}_{-0.04}$
Any	20	150	2000–2009 2015–2019	**+0.27** $^{+0.046}_{-0.045}$	**+0.04** $^{+0.04}_{-0.04}$	**−0.1** $^{+0.062}_{-0.061}$	+0.06 $^{+0.09}_{-0.09}$	**−0.1** $^{+0.05}_{-0.05}$	**−0.09** $^{+0.06}_{-0.06}$	**−0.04** $^{+0.04}_{-0.04}$
Any	25	150	2000–2009 2015–2019	**+0.26** $^{+0.047}_{-0.047}$	**+0.07** $^{+0.043}_{-0.043}$	**−0.15** $^{+0.063}_{-0.064}$	+0.02 $^{+0.09}_{-0.09}$	**−0.08** $^{+0.05}_{-0.05}$	**−0.08** $^{+0.06}_{-0.06}$	−0.003 $^{+0.04}_{-0.04}$

Table 7 Summary of the effect of each predictor on the probability of having entered a research area across topic models and temporal segmentations. Values indicate the mean posterior effect size and the 95% credible interval. Significant effects are shown in bold

Authorship	K_0	L	Temporal segmen1tation	Intell. capital (diversity)	Soc. capital (diversity)	Soc. capital (power)	Stable affiliation	Academic age	Prod. (co-auth.)	Prod. (solo-auth.)
Any	20	50	2000–2009 2015–2019	**+0.2**$^{+0.11}_{-0.11}$	**+0.22**$^{+0.1}_{-0.1}$	+0.006$^{+0.1}_{-0.1}$	−0.19$^{+0.22}_{-0.22}$	+0.04$^{+0.12}_{-0.11}$	−0.07$^{+0.1}_{-0.1}$	−0.05$^{+0.1}_{-0.09}$
1st/last	20	50	2000–2009 2015–2019	**+0.25**$^{+0.14}_{-0.13}$	**+0.19**$^{+0.13}_{-0.13}$	**+0.21**$^{+0.19}_{-0.19}$	+0.12$^{+0.29}_{-0.29}$	−0.07$^{+0.1}_{-0.1}$	**−0.33**$^{+0.18}_{-0.18}$	−0.0006$^{+0.1}_{-0.1}$
Any	20	50	2000–2004 2005–2009	**+0.16**$^{+0.11}_{-0.11}$	**+0.2**$^{+0.11}_{-0.11}$	+0.01$^{+0.15}_{-0.15}$	−0.007$^{+0.2}_{-0.2}$	−0.06$^{+0.1}_{-0.1}$	**−0.14**$^{+0.14}_{-0.14}$	−0.05$^{+0.1}_{-0.1}$
Any	20	50	2010–2014 2015–2019	**+0.24**$^{+0.095}_{-0.095}$	**+0.18**$^{+0.091}_{-0.091}$	−0.06$^{+0.1}_{-0.1}$	+0.05$^{+0.2}_{-0.21}$	+0.07$^{+0.1}_{-0.1}$	**−0.18**$^{+0.12}_{-0.12}$	−0.02$^{+0.09}_{-0.09}$
Any	20	50	2000–2009 2010–2019	**+0.27**$^{+0.083}_{-0.083}$	**+0.27**$^{+0.076}_{-0.076}$	−0.01$^{+0.1}_{-0.1}$	−0.08$^{+0.2}_{-0.2}$	+0.04$^{+0.09}_{-0.09}$	**−0.16**$^{+0.11}_{-0.11}$	−0.004$^{+0.07}_{-0.07}$
Any	20	50	2000–2009 2015–2019	**+0.24**$^{+0.11}_{-0.11}$	**+0.2**$^{+0.1}_{-0.1}$	+0.03$^{+0.15}_{-0.14}$	−0.07$^{+0.2}_{-0.2}$	−0.01$^{+0.1}_{-0.1}$	**−0.19**$^{+0.13}_{-0.14}$	−0.06$^{+0.1}_{-0.09}$
Any	15	50	2000–2009 2015–2019	**+0.13**$^{+0.11}_{-0.1}$	**+0.13**$^{+0.096}_{-0.095}$	−0.005$^{+0.1}_{-0.1}$	−0.03$^{+0.2}_{-0.2}$	−0.04$^{+0.1}_{-0.1}$	**−0.16**$^{+0.13}_{-0.13}$	+0.002$^{+0.09}_{-0.09}$
Any	25	50	2000–2009 2015–2019	**+0.18**$^{+0.12}_{-0.12}$	**+0.17**$^{+0.11}_{-0.11}$	−0.009$^{+0.2}_{-0.2}$	−0.07$^{+0.2}_{-0.2}$	−0.04$^{+0.1}_{-0.1}$	**−0.15**$^{+0.15}_{-0.15}$	−0.02$^{+0.1}_{-0.1}$
Any	15	150	2000–2009 2015–2019	**+0.12**$^{+0.1}_{-0.1}$	**+0.19**$^{+0.095}_{-0.095}$	−0.08$^{+0.1}_{-0.1}$	−0.02$^{+0.2}_{-0.2}$	−0.02$^{+0.1}_{-0.1}$	−0.04$^{+0.1}_{-0.1}$	−0.06$^{+0.09}_{-0.09}$
Any	20	150	2000–2009 2015–2019	**+0.23**$^{+0.11}_{-0.11}$	+0.07$^{+0.1}_{-0.1}$	+0.12$^{+0.15}_{-0.15}$	−0.06$^{+0.2}_{-0.2}$	−0.08$^{+0.1}_{-0.1}$	**−0.21**$^{+0.14}_{-0.14}$	−0.05$^{+0.095}_{-0.093}$
Any	25	150	2000–2009 2015–2019	**+0.31**$^{+0.13}_{-0.13}$	+0.01$^{+0.1}_{-0.1}$	−0.09$^{+0.2}_{-0.2}$	−0.14$^{+0.25}_{-0.25}$	+0.06$^{+0.1}_{-0.1}$	+0.001$^{+0.2}_{-0.2}$	−0.02$^{+0.1}_{-0.1}$

Table 8 Summary of the effect of each predictor on the probability of having exited a research area across topic models and temporal segmentations. Values indicate the mean posterior effect size and the 95% credible interval. Significant effects are shown in bold

Authorship	K_0	L	Temporal segmen1tation	Intell. capital (diversity)	Soc. capital (diversity)	Soc. capital (power)	Stable affiliation	Academic age	Prod. (co-auth.)	Prod. (solo-auth.)
Any	20	50	2000–2009 2015–2019	**+1**$^{+0.14}_{-0.14}$	+0.04$^{+0.1}_{-0.1}$	−0.03$^{+0.2}_{-0.2}$	+0.04$^{+0.2}_{-0.2}$	**−0.21**$^{+0.12}_{-0.12}$	**−0.28**$^{+0.15}_{-0.14}$	−0.02$^{+0.1}_{-0.1}$
1st/last	20	50	2000–2009 2015–2019	**+1.1**$^{+0.18}_{-0.18}$	+0.01$^{+0.1}_{-0.1}$	−0.12$^{+0.2}_{-0.2}$	−0.05$^{+0.3}_{-0.3}$	−0.15$^{+0.16}_{-0.16}$	−0.12$^{+0.19}_{-0.19}$	−0.01$^{+0.1}_{-0.1}$
Any	20	50	2000–2004 2005–2009	**+1**$^{+0.15}_{-0.14}$	+0.04$^{+0.1}_{-0.1}$	**−0.18**$^{+0.16}_{-0.16}$	−0.07$^{+0.3}_{-0.3}$	**−0.18**$^{+0.13}_{-0.13}$	**−0.25**$^{+0.15}_{-0.15}$	+0.04$^{+0.1}_{-0.1}$
Any	20	50	2010–2014 2015–2019	**+0.89**$^{+0.12}_{-0.11}$	+0.04$^{+0.09}_{-0.1}$	+0.1$^{+0.1}_{-0.1}$	**−0.22**$^{+0.22}_{-0.22}$	−0.04$^{+0.1}_{-0.1}$	**−0.34**$^{+0.13}_{-0.13}$	**−0.09**$^{+0.09}_{-0.09}$
Any	20	50	2000–2009 2010–2019	**+0.88**$^{+0.093}_{-0.09}$	**+0.16**$^{+0.077}_{-0.076}$	−0.01$^{+0.1}_{-0.1}$	+0.07$^{+0.2}_{-0.2}$	−0.06$^{+0.09}_{-0.08}$	**−0.32**$^{+0.11}_{-0.11}$	**−0.09**$^{+0.07}_{-0.07}$
Any	20	50	2000–2009 2015–2019	**+0.91**$^{+0.13}_{-0.13}$	−0.05$^{+0.1}_{-0.1}$	−0.07$^{+0.2}_{-0.2}$	+0.001$^{+0.2}_{-0.2}$	**−0.16**$^{+0.12}_{-0.12}$	**−0.28**$^{+0.14}_{-0.14}$	−0.002$^{+0.1}_{-0.1}$
Any	15	50	2000–2009 2015–2019	**+1.1**$^{+0.14}_{-0.14}$	−0.003$^{+0.1}_{-0.1}$	−0.15$^{+0.15}_{-0.15}$	−0.19$^{+0.23}_{-0.23}$	−0.05$^{+0.1}_{-0.1}$	**−0.21**$^{+0.14}_{-0.14}$	−0.07$^{+0.1}_{-0.1}$
Any	25	50	2000–2009 2015–2019	**+1.1**$^{+0.16}_{-0.16}$	−0.07$^{+0.1}_{-0.1}$	**+0.17**$^{+0.17}_{-0.17}$	−0.06$^{+0.3}_{-0.3}$	**−0.21**$^{+0.14}_{-0.14}$	**−0.39**$^{+0.15}_{-0.16}$	+0.09$^{+0.1}_{-0.1}$
Any	15	150	2000–2009 2015–2019	**+1.3**$^{+0.15}_{-0.14}$	+0.05$^{+0.1}_{-0.1}$	−0.11$^{+0.15}_{-0.15}$	+0.04$^{+0.2}_{-0.2}$	−0.06$^{+0.1}_{-0.1}$	**−0.23**$^{+0.15}_{-0.15}$	+0.04$^{+0.1}_{-0.1}$
Any	20	150	2000–2009 2015–2019	**+1.1**$^{+0.14}_{-0.14}$	+0.09$^{+0.1}_{-0.1}$	−0.09$^{+0.2}_{-0.2}$	+0.09$^{+0.2}_{-0.2}$	−0.11$^{+0.12}_{-0.12}$	**−0.21**$^{+0.15}_{-0.15}$	−0.06$^{+0.1}_{-0.1}$
Any	25	150	2000–2009 2015–2019	**+1.1**$^{+0.17}_{-0.17}$	+0.1$^{+0.1}_{-0.1}$	+0.16$^{+0.18}_{-0.18}$	+0.24$^{+0.27}_{-0.27}$	**−0.14**$^{+0.14}_{-0.14}$	**−0.44**$^{+0.16}_{-0.16}$	−0.05$^{+0.1}_{-0.1}$



(a) Effect of intellectual capital ($\gamma_{kk'}$). (b) Effect of social capital ($\delta_{kk'}$).

Figure 19 Effect of intellectual and social capital on transfers across research areas. Rows represent research areas of origin and columns represent target areas. Effect sizes are expressed in log-odds (log $[\theta_{akk'}/(1-\theta_{akk'})]$, where $\theta_{akk'}$ is the fraction of attention to k redirected to k') per unit of intellectual or social capital in the target research area k'. Effects that are not significant (at the 95% credible level) are displayed in white for purposes of clarity

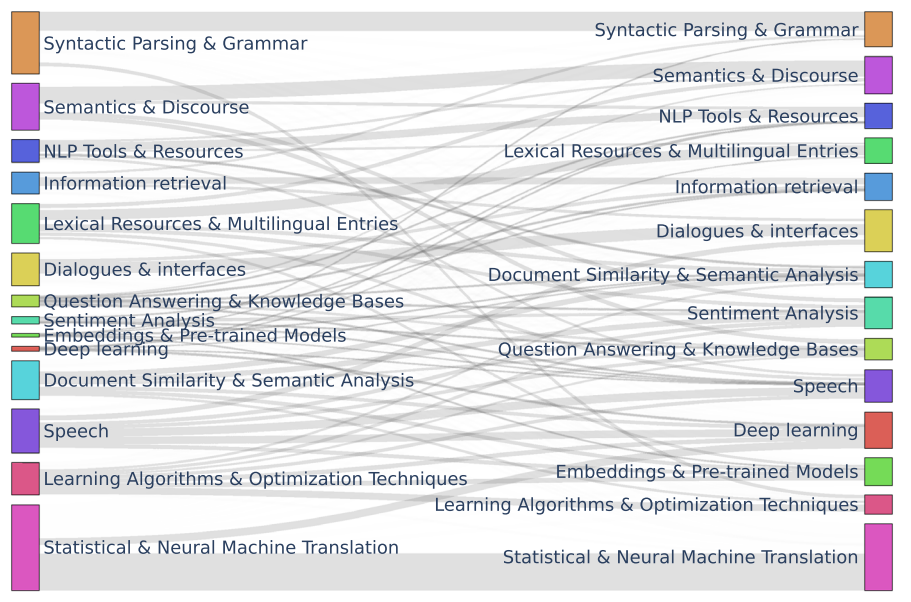


Figure 20 Aggregate transfers of attention across research areas in the ACL anthology corpus of Computational Linguistics, between 2002-2011 (to the left) and 2012-2022 (to the right). Widths of flows are proportional to $\sum_a X_{ak}\theta_{akk'}$. Transfers less frequent than expected by chance alone are transparent

A.8 Additional robustness checks

The robustness of the results of the comparative analysis is assessed by varying different parameters:

- The papers included in each authors' portfolio (any paper versus first-authored and last-authored papers only).
- The amount of topics in the topic model (K_0).
- The amount of dimensions for the word embeddings (L).
- The temporal segmentation for the early and late research portfolios.

A.9 Trajectory model parameters evaluated on a different time period

To further assess the robustness of the findings, the effect of intellectual and social capital on individual trajectories is measured on a different temporal segmentation (2000-2004 to 2005-2009). We make similar findings: the concentration of intellectual capital in one area promotes either commitment to this research area (or transfers in related areas). Social capital, on the other hand, matters increasingly as cognitive distance increases.

A.10 Replication corpus

For purposes of testing and replication, certain analyses have been reproduced on the ACL anthology corpus of Computational Linguistics research.

The transfers of attention are shown in Fig. 20. Compared to the high-energy physics corpus, it features significant disruptions (e.g., the emergence of new topics, such as "deep learning", "sentiment analysis" and "embeddings & pre-trained models").

Acronyms

APS, American Physical Society; HEP, High-Energy Physics; LHC, Large Hadron Collider; OT, Optimal Transport; PACS, Physics and Astronomy Classification Scheme®.

Acknowledgements

I thank Thomas Heinze, Radin Dardashti, and Elisa Omodei for useful discussions; Marc Santolini for rich feedback; Georges Ricci for raising my interest in Optimal Transport; Marco Schirone for providing useful references; the LATTICE lab in Montrouge for the opportunity to discuss this work; and Elizabeth Zanghi for corrections. I finally thank the reviewers for their helpful comments and suggestions.

Author contributions

The sole author did all of the work.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work has received support from the G-Research Grant for Early-Career Researchers.

Data Availability

The data and the code for the paper can be found at the following address: https://github.com/lucasgautheron/specialization-adaptation/

Declarations

Competing interests

The authors declare no competing interests.

Received: 21 July 2024 Accepted: 19 December 2024 Published online: 08 January 2025

References

1. Kuhn TS (1977) The essential tension: tradition and innovation in scientific research. In: The essential tension. University of Chicago Press, Chicago, pp 225–239. https://doi.org/10.7208/9780226217239-010
2. March JG (1991) Exploration and exploitation in organizational learning. Organ Sci 2(1):71–87. https://doi.org/10.1287/orsc.2.1.71
3. Foster JG, Rzhetsky A, Evans JA (2015) Tradition and innovation in scientists' research strategies. Am Sociol Rev 80(5):875–908. https://doi.org/10.1177/0003122415601618
4. Jia T, Wang D, Szymanski BK (2017) Quantifying patterns of research-interest evolution. Nat Hum Behav 1:4. https://doi.org/10.1038/s41562-017-0078
5. Aleta A, Meloni S, Perra N, Moreno Y (2019) Explore with caution: mapping the evolution of scientific interest in physics. EPJ Data Sci 8:27. https://doi.org/10.1140/epjds/s13688-019-0205-9
6. Zeng A, Shen Z, Zhou J, Fan Y, Di Z, Wang Y, Stanley HE, Havlin S (2019) Increasing trend of scientists to switch between topics. Nat Commun 10:1. https://doi.org/10.1038/s41467-019-11401-8
7. Tripodi G, Chiaromonte F, Lillo F (2020) Knowledge and social relatedness shape research portfolio diversification. Sci Rep 10:1. https://doi.org/10.1038/s41598-020-71009-7
8. Singh CK, Tupikina L, Lécuyer F, Starnini M, Santolini M (2024) Charting mobility patterns in the scientific knowledge landscape. EPJ Data Sci 13:12. https://doi.org/10.1140/epjds/s13688-024-00451-8
9. Liu F, Zhang S, Xia H (2024) Science as exploration in a knowledge landscape: tracing hotspots or seeking opportunity? EPJ Data Sci 13(1):27
10. Fujimura JH (1988) The molecular biological bandwagon in cancer research: where social worlds meet. Soc Probl 35(3):261–283. https://doi.org/10.2307/800622

11. Galesic M, et al (2023) Beyond collective intelligence: collective adaptation. J R Soc Interface 20(200). https://doi.org/10.1098/rsif.2022.0736
12. Monge G (1781) Mémoire sur la théorie des déblais et des remblais. Mem Math Phys Acad R Sci 666–704
13. Kantorovich L (1942) On the translocation of masses. C R (Dokl) Acad Sci URSS 37:199–201
14. Peyré G, Cuturi M (2019) Computational optimal transport: with applications to data science. Found Trends Mach Learn 11(5–6):355–607. https://doi.org/10.1561/2200000073
15. Abbasi A, Wigand RT, Hossain L (2014) Measuring social capital through network analysis and its influence on individual performance. Libr Inf Sci Res 36(1):66–73. https://doi.org/10.1016/j.lisr.2013.08.001
16. Wu J, O'Connor C, Smaldino PE (2023) The cultural evolution of science. In: The Oxford handbook of cultural evolution. Oxford University Press, London. https://doi.org/10.1093/oxfordhb/9780198869252.013.78
17. Chiu W-T, Wang P, Shafto P (2022) Discrete probabilistic inverse optimal transport. In: Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S (eds) Proceedings of the 39th international conference on machine learning. Proceedings of machine learning research, vol 162. PMLR, pp 3925–3946
18. Gieryn TF (1978) Problem retention and problem change in science. Sociol Inq 48(3–4):96–115. https://doi.org/10.1111/j.1475-682x.1978.tb00820.x
19. Battiston F, Musciotto F, Wang D, Barabási A-L, Szell M, Sinatra R (2019) Taking census of physics. Nat Rev Phys 1(1):89–97. https://doi.org/10.1038/s42254-018-0005-3
20. Bourdieu P (1980) Le capital social. Actes Rech Sci Soc 31(1):2–3
21. Bourdieu P (1986) The forms of capital. In: Richardson J (ed) Handbook of theory and research for the sociology of education, Greenwood, pp 241–258
22. Mulkay M (1974) Conceptual displacement and migration in science: a prefatory paper. Soc Stud Sci 4(3):205–234. https://doi.org/10.1177/030631277400400301
23. Schon DA (1963) Displacement of concepts, 1st edn. Routledge, London. https://doi.org/10.4324/9781315014111
24. Mahoney J, Thelen K (2009) A theory of gradual institutional change. In: Explaining institutional change: ambiguity, agency, and power. Cambridge University Press, Cambridge, pp 1–37. https://doi.org/10.1017/CBO9780511806414.003
25. Hallonsten O, Heinze T (2013) From particle physics to photon science: multi-dimensional and multi-level renewal at DESY and SLAC. Sci Public Policy 40(5):591–603. https://doi.org/10.1093/scipol/sct009
26. Hallonsten O, Heinze T (2015) Formation and expansion of a new organizational field in experimental science. Sci Public Policy 42(6):841–854
27. Heinze T, Hallonsten O, Heinecke S (2017) Turning the ship: the transformation of DESY, 1993–2009. Phys Perspect 19(4):424–451. https://doi.org/10.1007/s00016-017-0209-4
28. Heinze T, Münch R (2012) Intellektuelle Erneuerung der Forschung durch institutionellen Wandel. Institutionelle Erneuerungsfähigkeit der Forschung. VS Verlag für Sozialwissenschaften, pp 15–38. https://doi.org/10.1007/978-3-531-94274-2_2
29. Smaldino PE, Moser C, Pérez Velilla A, Werling M (2023) Maintaining transient diversity is a general principle for improving collective problem solving. Perspect Psychol Sci 19(2):454–464. https://doi.org/10.1177/17456916231180100
30. Schimmelpfennig R, Razek L, Schnell E, Muthukrishna M (2021) Paradox of diversity in the collective brain. Philos Trans - R Soc, Biol Sci 377:1843. https://doi.org/10.1098/rstb.2020.0316
31. Muthukrishna M, Henrich J (2016) Innovation in the collective brain. Philos Trans - R Soc, Biol Sci 371(1690):20150192. https://doi.org/10.1098/rstb.2015.0192
32. Page S (2008) The difference: how the power of diversity creates better groups, firms, schools, and societies-new edition. Princeton University Press, Princeton
33. Acerbi A, Ghirlanda S, Enquist M (2012) Old and young individuals' role in cultural change. J Artif Soc Soc Simul 15:4. https://doi.org/10.18564/jasss
34. Lykken J, Spiropulu M (2014) Supersymmetry and the crisis in physics. Sci Am 310(5):34–39
35. Roser T, et al (2023) On the feasibility of future colliders: report of the Snowmass'21 implementation task force. J Instrum 18(05):P05018. https://doi.org/10.1088/1748-0221/18/05/p05018
36. Harlander R, Martinez J-P, Schiemann G (2023) The end of the particle era? Eur Phys J H 48:1. https://doi.org/10.1140/epjh/s13129-023-00053-4
37. Kosyakov B (2023) A farewell to particles. https://doi.org/10.48550/arXiv.2305.09692
38. Shifman M (2020) Musings on the current status of HEP. Mod Phys Lett A 35(07):2030003. https://doi.org/10.1142/s0217732320300037
39. Abbott B, Abbott R, Abbott TD, Abernathy M, Acernese F, Ackley K, Adams C (2016) Observation of gravitational waves from a binary black hole merger. Phys Rev Lett 116:6. https://doi.org/10.1103/physrevlett.116.061102
40. Gautheron L, Omodei E (2023) How research programs come apart: the example of supersymmetry and the disunity of physics. Quant Sci Stud 4(3):671–699. https://doi.org/10.1162/qss_a_00262
41. Moskovic M (2021) The INSPIRE REST API. https://doi.org/10.5281/ZENODO.5788550
42. Perović S, Radovanović S, Sikimić V, Berber A (2016) Optimal research team composition: data envelopment analysis of Fermilab experiments. Scientometrics 108(1):83–111. https://doi.org/10.1007/s11192-016-1947-9
43. Chall C, King M, Mättig P, Stöltzner M (2019) From a boson to the standard model Higgs: a case study in confirmation and model dynamics. Synthese 198:3779–3811. https://doi.org/10.1007/s11229-019-02216-7
44. Strumia A (2021) Gender issues in fundamental physics: a bibliometric analysis. Quant Sci Stud 2(1):225–253. https://doi.org/10.1162/qss_a_00114
45. Sikimić V, Radovanović S (2022) Machine learning in scientific grant review: algorithmically predicting project efficiency in high energy physics. Eur J Philos Sci 12:3. https://doi.org/10.1007/s13194-022-00478-6
46. Ioannidis JPA, Boyack KW, Klavans R (2014) Estimates of the continuously publishing core in the scientific workforce. PLoS ONE 9(7):e101698. Ed. by L. A. N. Amaral. https://doi.org/10.1371/journal.pone.0101698
47. Dieng AB, Ruiz FJR, Blei DM (2020) Topic modeling in embedding spaces. Trans Assoc Comput Linguist 8:439–453. https://doi.org/10.1162/tacl_a_00325
48. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

49. Omodei E (2014) Modeling the socio-semantic dynamics of scientific communities. Thesis, Ecole Normale Supérieure
50. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Bengio Y, LeCun Y (eds) 1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, workshop track proceedings.
51. Schirone M (2023) Field, capital, and habitus: the impact of Pierre bourdieu on bibliometrics. Quant Sci Stud 4(1):186–208. https://doi.org/10.1162/qss_a_00232
52. Roth C, Cointet J-P (2010) Social and semantic coevolution in knowledge networks. Soc Netw 32(1):16–29. https://doi.org/10.1016/j.socnet.2009.04.005
53. Burt RS (2007) Brokerage and closure: an introduction to social capital. Oxford University Press, London. 296 pp.
54. Newman ME (2004) Who is the best connected scientist? A study of scientific coauthorship networks. In: Complex networks. Springer, Berlin, pp 337–370. https://doi.org/10.1007/978-3-540-44485-5_16
55. Jost L (2006) Entropy and diversity. Oikos 113(2):363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x
56. Muzellec B, Nock R, Patrini G, Nielsen F (2017) Tsallis regularized optimal transport and ecological inference. In: Proceedings of the AAAI conference on artificial intelligence, vol 31.
57. Li R, Ye X, Zhou H, Zha H (2019) Learning to match via inverse optimal transport. J Mach Learn Res 20
58. Martens NCM, King M (2023) Doing more with less: dark matter & modified gravity. In: Philosophy of astrophysics, Springer, Berlin, pp 91–107. https://doi.org/10.1007/978-3-031-26618-8_6
59. Martens NC, Sahuquillo MÁC, Scholz E, Lehmkuhl D, Krämer M (2022) Integrating dark matter, modified gravity, and the humanities
60. Galichon A (2018) Optimal transport methods in economics. Princeton University Press, Princeton
61. Dupuy A, Galichon A (2014) Personality traits and the marriage market. J Polit Econ 122(6):1271–1319
62. Galison P (1987) How experiments end. University of Chicago Press, Chicago
63. Waltman L (2012) An empirical analysis of the use of alphabetical authorship in scientific publishing. J Informetr 6(4):700–711
64. Venturini S, Sikdar S, Rinaldi F, Tudisco F, Fortunato S (2024) Collaboration and topic switches in science. Sci Rep 14:1. https://doi.org/10.1038/s41598-024-51606-6
65. Kummerfeld E, Zollman KJS (2016) Conservatism and the scientific state of nature. Br J Philos Sci 67(4):1057–1076. https://doi.org/10.1093/bjps/axv013
66. Smith CL (2015) Genesis of the large hadron collider. Philos Trans R Soc A, Math Phys Eng Sci 373(2032):20140037. https://doi.org/10.1098/rsta.2014.0037
67. Singh CK, Tupikina L, Lécuyer F, Starnini M, Santolini M (2023) Charting mobility patterns in the scientific knowledge landscape. https://doi.org/10.48550/arXiv.2302.13054
68. Stirling A (2007) A general framework for analysing diversity in science, technology and society. J R Soc Interface 4(15):707–719. https://doi.org/10.1098/rsif.2007.0213
69. Porter AL, Cohen AS, Roessner JD, Perreault M (2007) Measuring researcher interdisciplinarity. Scientometrics 72(1):117–147. https://doi.org/10.1007/s11192-007-1700-5
70. Leahey E, Beckman CM, Stanko TL (2016) Prominent but less productive. Adm Sci Q 62(1):105–139. https://doi.org/10.1177/0001839216665364
71. Rohatgi S (2022) ACL Anthology Corpus with Full Text. Github
72. Skilling J, Gull SF (1991) Bayesian maximum entropy image reconstruction. Lecture notes-monograph series, pp 341–367
73. MacKay DJC, Peto LCB (1995) A hierarchical Dirichlet language model. Nat Lang Eng 1(3):289–308. https://doi.org/10.1017/s1351324900000218

Publisher's Note

Chapter 3

Dilemmas and trade-offs in the diffusion of conventions

Dilemmas and trade-offs in the diffusion of conventions

Lucas Gautheron[1,2]
lucas.gautheron@gmail.com

April 13, 2025

Abstract

Outside ideal settings, conventions are shaped by heterogeneous competing processes that can challenge the emergence of norms. In order to acknowledge this complexity, this paper develops a generalized account of conventions and identifies three trade-offs involved in their diffusion: (I) the trade-off between the imperatives of social, sequential, and contextual consistency that individuals balance when choosing between conventions; (II) the competition between local (bottom-up) and global (top-down) coordination, depending on whether individuals coordinate their behavior via interactions throughout a social network or external factors transcending the network; and (III) the balance between decision optimality (e.g., collective satisfaction) and decision costs when collectives with conflicting preferences choose a convention. A broadly applicable statistical physics framework for exploring these trade-offs is developed and applied to a sign convention in physics. The method can infer the structure of the underlying coordination game, the networks of social interactions involved, and the processes through which conflicts are resolved. This shows that the purpose of conventions may exceed coordination, and that individual preferences towards conventions are concurrently shaped by cultural factors and multiple social networks. Finally, this work emphasizes the role of leadership in the resolution of conflicts.

Keywords: conventions; collective cognition; cultural evolution; Ising model; inverse problems; simulation-based inference.

1 Introduction

Since David Lewis [1], conventions (including linguistic norms, technological or manufacturing standards, and many other social norms) are primarily conceived as solutions to coordination problems [2]. Yet, the attitude of individuals towards conventions involves

[1]Interdisciplinary Centre for Science and Technology Studies (IZWT), University of Wuppertal, Germany
[2]Département d'Études Cognitives, École Normale Supérieure, Paris, France

a multitude of factors beyond social coordination, resulting in tensions that may disrupt the emergence of a universal norm. To acknowledge this complexity, this paper develops a generalized statistical physics account of conventions and identifies three trade-offs involved in their diffusion and the resolution of conflicts in the absence of consensus. The first trade-off is the balance between i) social consistency (driven by coordination with peers), ii) sequential consistency (driven by the cost of switching from one practice to another), and iii) contextual consistency (driven by a need for consistency with other choices or cultural traits) (§1.2). The second trade-off involves the balance between *local* (or bottom-up) versus *global* (or top-down) coordination, depending on whether individual preferences are formed endogenously through local interactions on a social network, or by factors transcending the network structure (or both, in possibly contradicting ways) (§1.3). Finally, the last trade-off is the balance between decision costs and the optimality of the outcome in the resolution of conflicts (§1.4). To explore these trade-offs, we apply a statistical physics framework to behavioral data about a sign convention in physics. This statistical framework allows us to retrieve information about individuals' decision-making, the structure of the underlying coordination problem, and the multiple infrastructures (whether social or cultural) involved in the propagation of a convention.

First, we show that scientists' attitude is driven by sequential consistency, as they tend to maintain a preferred choice in their solo-authored publications independently of the target research area (§2.1). Then, we show that scientists' preferences are correlated – albeit imperfectly – with those of their co-authors, which means that some level of social coordination is achieved (§2.2). In order to explain how, the relative contribution of local coordination (via dyadic interactions with peers) and global coordination (i.e. via shared culture) is measured by solving an inverse Ising problem over the authors' collaboration and citation network. This shows that both local and global processes contribute to coordinating scientists' preferences. Third, we assess the plausibility of three mechanisms of preference-formation according to their ability to explain the observed magnitudes of local and global coordination, and find slightly more evidence for a model of cultural transmission involving the imitation of peers (§2.3). Finally, we infer the process through which scientists resolve conflicts about which convention to use in collaborations (§2.4). We find evidence that the last author's preference most often prevails, thus highlighting the role of seniority and power in the resolution of conflicts. Taken together, these results indicate that decision-making processes related to conventions involve multiple and sometimes conflicting factors.

1.1 Background

While formal models of the diffusion of conventions provide rich insights by focusing on one or a few key features of the phenomena of interest, they may also leave out crucial aspects of reality by stripping away too much of its complexity [3], or by neglecting the interactions between phenomena studied in isolation. For instance, [4, 5] demonstrated the importance of accurately representing the topological features of complex networks

(including their small-world, scale-free or clustering properties) for modeling and simulating the propagation of conventions. Similarly, while controlled experiments can uncover certain aspects of conventions in idealized settings [6–10], they may conceal the fact that complex heterogeneous processes and multiple social infrastructures can drive or prevent the emergence of conventions in naturalistic situations [11]. Fortunately, the advent of large online communities has opened up opportunities to investigate the diffusion of real norms and conventions in complex networks [12–15]. Such data-driven approaches, however, have barely extended to the study of scientific conventions[1]. Yet, "conventionalism" can be traced back to Poincaré and his account of the epistemic status of the axioms of geometry [16]. In fact, conventions are ubiquitous in science [17], including statistical practices (e.g. statistical significance thresholds [18]), measurement strategies [19], and unit systems. By exploring a scientific convention, this work highlights the interactions between multiple phenomena involved in the diffusion of conventions that prior works have addressed separately or ignored, and provides cues for understanding how conventions can fail to develop into universal norms, in naturalistic settings.

Let us introduce the convention emphasized in the present paper. In relativistic physical theories (such as general relativity and quantum field theory), the "metric tensor" is a mathematical object that represents the metric properties of space-time. Broadly speaking, the metric tensor can take either of the two following forms:

$$\begin{pmatrix} +1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \text{ or } \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix}$$

The first choice $(+,-,-,-)$ is known as the mostly minus convention while the second choice, $(-,+,+,+)$ is referred to as the mostly plus convention. These choices are physically equivalent and lead to identical predictions. However, depending on which choice one makes, certain quantities arising in calculations will take either positive or negative values. Interestingly, there is no norm and both conventions are used.

1.2 The trade-off between social, sequential, and contextual consistency

While Lewis' account of conventions is focused on their social dimension, earlier accounts provide different perspectives: the holist account of conventionalism, for instance, contends that we may choose freely between distinct but collectively coherent systems of beliefs [16]. Below, these two perspectives on conventions are unified into a single notion of collective consistency, formalized using elements of game theory and statistical physics. This reveals that conventions involve multiple dimensions that can compete with each other.

[1] with the exception of [15], which investigates LaTeX macros naming conventions in scientific papers.

	$x_j =$	$x_j =$
$x_i =$	$(1, 1)$	$(0, 0)$
$x_i =$	$(0, 0)$	$(1, 1)$

(a) **Social consistency**. Alice and Bob are better off if they agree on either or .

	$x_{t+1} =$	$x_{t+1} =$
$x_t =$	1	0
$x_t =$	0	1

(b) **Sequential consistency**. Alice is better off if she consistently chooses or .

	$y =$	$y =$
$x =$	1	0
$x =$	0	1

(c) **Contextual consistency**. Alice is better of if she chooses either or .

Table 1: Collective consistency as coordination games involving Alice ( ) and Bob ( ), or Alice alone. Each table represents a payoff matrix associated with a "collective" choice.

Social consistency and coordination costs Conventions are mainly conceived as solutions to coordination problems [1], which arise when individuals would benefit from acting in a mutually consistent way, but struggle to do so – maybe, for instance, because they lack the information necessary for achieving joint-action [1, 2]. Conventions can solve coordination problems by providing individuals with expectations about how others will behave in a given setting, a paradigmatic example being left-hand versus right-hand traffic. In absence of universal conventions, individuals experience *coordination costs* in their interactions. When interactions involve two people at a time, coordination costs can be represented by a payoff matrix that defines the utility (i.e. the rewards) $u_{i,j}(x_i, x_j)$ for agents i and j as a function of x_i and x_j, their respective strategies (Table 1a) (for clarity, we consider binary conventions labeled by $x \in \{-1, +1\}$). Additionally, coordination costs are specified by a network structure capturing the frequency of interactions w_{ij} between any pair (i, j) of agents. In naturalistic scenarios, given observations of individuals' strategies, one may want to retrieve the structure of the underlying game or to identify the relevant social network(s). Fortunately, coordination games such as 2 can be mapped onto models from statistical mechanics such as the Ising model [20, 21], which, as we show, enables empirical explorations of conventions. To this end, one constructs a "potential" $U(x_1, \ldots, x_N)$ [22] (a collective utility), which is a function of the joint strategy of every individual $1 \leq i \leq N$ that varies by $\sum_j w_{ij}[u_i(x'_i, x_j) - u_i(x_i, x_j)]$ as any agent i unilaterally changes their strategy from x_i to x'_i. Under a simple evolutionary rule[2], the probability of a particular combination of individual strategies is:

$$P(x_1, \ldots, x_N) = \frac{1}{Z} e^{\beta U(x_1, \ldots, x_N)} \tag{1}$$

Where Z is a normalization constant and $\beta \geq 0$ controls the degree of rationality – and efficiency – of the agents [21]. In statistical physics, (1) is the Boltzmann distribution; U is (up to a minus sign) the energy potential of a particular configuration, and β is the inverse temperature[3]. This probabilistic framework enables the retrieval of information about the coefficients of the payoff matrices (u_i, u_j) or the network structure (w_{ij}) from observations of individuals' strategies, as shown in §2.2. In the case of Table 1a, this

[2]For "potential" games, the "logit" rule and the Glauber dynamics lead to the above Boltzmann distribution [22, 23].

[3]Often, β may be omitted without loss of generality through proper rescaling of U.

gives the Ising model [24]:

$$P(x_1, \ldots, x_N) = \frac{1}{Z} e^{\frac{\beta}{2} \sum_{ij} w_{ij} x_i x_j} \tag{2}$$

Sequential consistency and switching costs In addition to addressing coordination problems, conventions enable individuals to settle on one choice once and for all, in a way that facilitates future moves. Consider keyboard layouts (e.g. qwerty). While there exists many such layouts, we benefit from settling on a single one, even if our choice is arbitrary and different from our peers'. In that respect, certain conventions can serve a purely internal purpose of consistency, as if individuals "played" a coordination game with themselves, such that their payoffs depend on the mutual coherence of their actions. To model sequential consistency, let x_{it} be the convention employed by agent i at time $t \in \{1, \ldots, T\}$. A simple model of the utility of a sequence of choices for an isolated individual is a Markov model $U(x_{i1}, \ldots, x_{iT}) = \sum_{t=1}^{T-1} u(x_{i,t}, x_{i,t+1})$, where $u(x_t, x_{t+1})$ is the payoff matrix associated with the transition from x_t to x_{t+1} (Table 1b). In such a model, agents experience costs every time they switch from one convention to another. Alternatively, sequential consistency may reflect lasting preferences with memory effects due to complex long-range interactions between individual actions. Instead, one might consider the effective model $U(x_{i1}, \ldots, x_{iT}) = \sum_{t=1}^{T} u_i^{x_{it}}$ where u_i^x designates the utility associated with choice x for agent i. Again, we may assume that the probability of a particular sequence takes the form $P(x_{i1}, \ldots, x_{iT}) \propto e^{\beta U(x_{i1}, \ldots, x_{iT})}$, where β is, as before, a measure of efficiency.

Contextual consistency and maladaptation costs Some conventions are less conventional than others [25, 26]: certain choices can be *maladaptive* and less likely to be adopted. However, which conventions are more or less adaptive may depend on the context. Unit systems are a good example: while light-years might be a convenient unit of length for astronomers, engineers may reasonably prefer meters. Maladaptation costs indicate an inconsistency between a convention and other interacting choices or cultural traits. This can be thought of in terms of a cultural fitness landscape [27], where $U(x_{i,1}, \ldots, x_{i,C})$ describes the fitness of a configuration of C traits $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,C}) \in \{\pm 1\}^C$. It is possible that the choice between, say, $x_{i,1} = -1$ or $+1$ is "conventional", in that there is no universally superior choice across the landscape (i.e. $\mathbb{E}_{\boldsymbol{x} \sim p}[U(\boldsymbol{x}) | x_1 = -1] \simeq \mathbb{E}_{\boldsymbol{x} \sim p}[U(\boldsymbol{x}) | x_1 = +1]$, where p is the joint probability distribution over all traits), even though certain regions in the landscape may locally favor a specific choice for x_1[4]. The cultural landscape can be modeled using the same building blocks as for social and sequential consistency, by considering pairwise interactions between cultural traits (i.e. epistasis [28]), cf. Table 1c. Then, assuming the co-evolution of traits follows the Glauber dynamics or the logit rule, their distribution converges to an Ising model.

[4]This is obvious in the context of language. The mapping between objects and symbols is highly conventional; however, for a given pre-existing language, the choice of how to name a new object can be constrained by preceding linguistic infrastructure.

Therefore, cultural landscapes involving multiple traits and conventions can sometimes be reconstructed empirically (although approximately) by solving an inverse Ising problem, following [27]. In certain cases, this can reveal a plurality of collectively consistent systems of choices: A demonstration is proposed in S4.2, using collections of naming conventions in a scientific typesetting language, following upon [15]. When the position of agent i in the landscape can be considered fixed, except for a trait k, the relative reward for their choice $x_{i,k} \in \{\pm 1\}$ reduces to $U(x_{i,1}, .., x_{i,k}, .., x_{i,C}) - U(x_{i,1}, .., -x_{i,k}, .., x_{i,C}) \equiv B_i x_k^i = \pm B_i$.

Broadly speaking, conventionality arises when behavior is determined by "collective" rather than individual constraints – in other words, the marginal probability of a particular outcome $p(x_i)$ is weakly constrained; only the joint probability of all outcomes $p(x_1, \ldots, x_n)$ is, due to synergistic interactions between individuals, cultural traits, or consecutive choices. For example, in the case of sequential consistency, the first move does not matter, as long as the *entire* sequence of actions is collectively consistent. Contextual consistency is also a collective constraint, since it assumes there is no way to universally reject a particular choice independently from other choices[5]. Interestingly, all three imperatives can be modeled using the same fundamental game-theoretic building blocks (Table 1), given that two-person coordination games, and the Ising model, provide a simple account of the interactions between more-or-less conventional traits.

In the most general case, all three factors can be involved in conventions, albeit to varying extents. For the metric signature, coordination costs are plausible (it should be easier to collaborate with scientists who will systematically agree to using your favorite convention, and it is easier to copy results if they are systematically derived with the same convention). Switching costs are seemingly plausible, as working with different metric signatures implies keeping track of which sign certain quantities must take according to which convention is used. Finally, context-dependent maladaptation costs might be involved too. For instance, for problems that involve "proper time" calculations, the mostly minus metric is advantageous, since then proper time is equal to the pseudo-distance between events rather than minus the pseudo-distance. The diffusion of conventions involving these three imperatives in conjunction involves the co-evolution of $(x_{ict}) \in \{\pm 1\}^{N \times C \times T}$. It can be simulated by flipping traits ($x_{i,c,t+1} = -x_{i,c,t}$) with probability $p = \min[1, \exp(\beta \Delta u_{ic})]$ where Δu_{ic} is the variation in reward associated with switching from x_{ict} to $-x_{ict}$. Traits with comparatively large switching costs are less likely to flip and can be considered fixed. The ability of such a model to explain physicists' attitude towards the metric signature is evaluated in §2.3.

When individuals behave consistently and uniformly, it is difficult to infer which factor was determinant in the adoption of a norm. However, when variations are observed, these can be leveraged to infer the underlying processes. In §2.1, we start by evaluating the importance of sequential and contextual consistency in the case of the metric signature. It will be shown that both matter, but sequential consistency matters more, such that individuals tend to stick to their favorite convention across different contexts. Physi-

[5]See epistemological holism, according to which beliefs are constrained collectively rather than individually [29].

cists therefore have *preferences* towards a metric signature, and we may ask how these preferences are formed. In §4, we also examine the relative contribution of social and contextual consistency in the formation of scientists' preferences. When the two compete with each other, individuals play an asymmetric game (Table 2) which parameters (J, the contribution of social consistency) and B (contextual consistency) can be measured empirically using an Ising model. Interestingly, these two parameters simultaneously encode a universal competition between local and global coordination.

Table 2: Generic payoff matrix of a two-player two-action coordination game. Cells indicate $(u_i(x_i, x_j), u_j(x_i, x_j))$, the rewards of i and j as a function of their joint strategy. J measures the synergistic benefit of coordination, and (B_i, B_j) measures the "preferences" of i and j, due to (for instance) their positions in the cultural landscape.

	$x_j =$	$x_j =$
$x_i =$	$(+J - B_i, +J - B_j)$	$(-J - B_i, -J + B_j)$
$x_i =$ 	$(-J + B_i, -J - B_j)$	$(+J + B_i, +J + B_j)$

1.3 Local and global processes in the diffusion of conventions

The emergence of social norms is the byproduct of both "local", "dyadic" processes and pre-existing "broader population-level infrastructure" [2], including social networks or central authorities [30]. In particular, we propose a distinction between *local* and *global* processes of coordination. "Local" coordination refers to bottom-up coordination via local interactions on a network (e.g. by the imitation of peers [31], or strategic adjustment to their behavior), as opposed to "global", top-down processes resulting from external factors transcending the network structure, including institutions, "central authorities" [30], but also any pre-established cultural traits or common knowledge shared within different groups. Global coordination can arise when individuals share the understanding that one option is intrinsically superior ($\text{sign}(B_i) = \text{sign}(B_j)$ and $|B_i|, |B_j \gg 1$ in Table 2). In scientific communities, local processes may propagate over a co-authorship network, while global factors may include a shared "disciplinary matrix" [32][6].

Figure 1 illustrates how local and global processes may generate different patterns of coordination. In this particular example, local coordination fails to produce consensus as the network is stuck into a Nash equilibrium. Occasionally, "global" processes may solve this type of failure. Alternatively, local and global forces may push in opposite directions and complicate the emergence of a norm [33] – for instance, if different groups with incompatible inclinations come into contact. Figure 1 also shows that the Ising model can correctly infer the actual coordination process for each toy example.

In §2.2, using an Ising model, we measure the contribution of local (J) and global ($\boldsymbol{B}$) mechanisms to the formation of physicists' preferences. We find evidence for both

[6]This distinction between local and global differs from that suggested in [15], which opposes dynamics in the diffusion of conventions at the microscopic and macroscopic (or aggregate) levels.

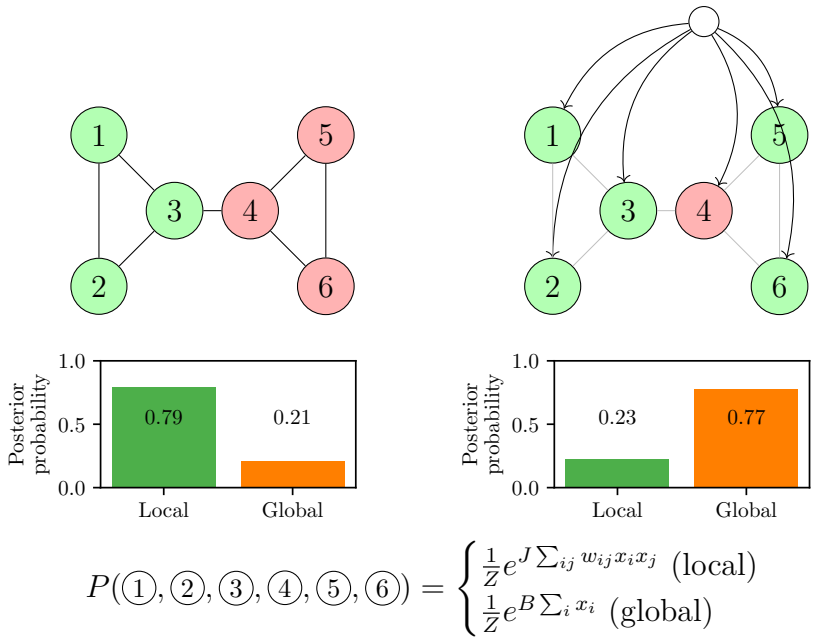


Figure 1: **Left**. Local coordination: nodes align to their neighbors through pairwise interactions. They may get stuck in a Nash equilibrium. **Right**. Global coordination: nodes are coordinated by a common cause transcending the graph structure (with some possible noise). Local and global processes generally predict different patterns of coordination, which means their contribution can be inferred from behavioral data. In each of these toy examples of local and global coordination, the Ising model correctly identifies the most likely process.

local and global effects in the case of the metric signature, while the latter seem to predominate. Moreover, as will be shown in §2.3, this Ising model approach can serve as a basis for comparing the plausibility of more realistic mechanisms of preference-formation, according to whether they generate local or global coordination patterns.

1.4 Optimality versus decision costs in the resolution of conflicts

In the absence of norms, how can individuals with conflicting preferences achieve coordination? In scientific collaborations, authors must sometimes overcome such conflicts. They must then operate a trade-off between "optimality" (e.g. the maximization of their collective satisfaction), and "decision costs". Indeed, co-authors can seek to maximize their collective satisfaction by making a collective decision, through deliberation of bargaining. However, this can be cumbersome: not all decisions deserve to be put under the whole collective's scrutiny, and it might be easier to let a leader decide, potentially at the expense of collective agreement. It is indeed well known that power and leadership can mitigate decision and coordination costs [34, 35]. In §2.4, we infer the mechanisms via which physicists resolve conflicts in co-authored papers. We find some evidence that leadership also plays a role in the resolution of conflicting preferences towards the metric signature, resulting in suboptimal decisions given that individual preferences are only partially aggregated within collaborations. While dictatorial strategies of conflict-resolution

fail to represent individual preferences, they may occasionally yield superior outcomes by producing more mutually coherent collections of decisions than e.g. majority voting [36, p. 23].

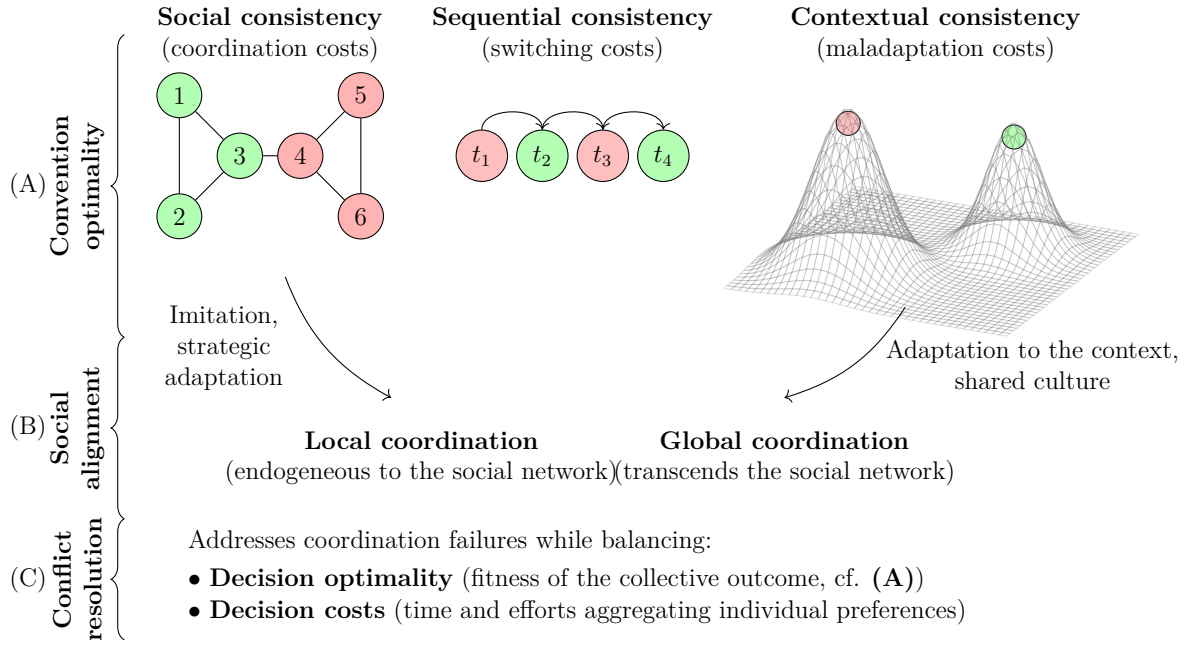


Figure 2: Three trade-offs affecting conventions and their relationships.

We have identified three trade-offs affecting conventions. Figure 2 summarises these trade-offs and highlights their interactions. In what follow, we show how statistical physics and inverse problems can provide empirical evidence for these trade-offs.

1.5 Data

Literature in high-energy physics is collected from the Inspire HEP database, which includes various metadata (authorship, institutional affiliations, etc.). When available, the LaTeX source of each paper is retrieved from arXiv. 22500 papers from four categories (Phenomenology-HEP, Theory-HEP, General Relativity & Quantum Cosmology, and Astrophysics) are successfully classified into either metric signature (± 1) using a small set of regular expressions (see S4.1).

2 Results

2.1 Beyond coordination: the role of sequential and contextual consistency

We have postulated that in addition to social coordination, individuals' attitude towards conventions may also be influenced by imperatives of sequential and contextual

consistency. If sequential consistency matters, individuals should tend to use the same convention throughout their own works. By contrast, if individuals behave differently across research areas, we may infer that they value contextual consistency.

Below, we measure the importance of sequential and contextual consistency in scientists' behavior. We consider only solo-authored papers, for which the choice of metric purely reflects the sole author's choice. In order to capture the imperatives of sequential and contextual consistency, we assume that the probability that an author i uses the $+1$ sign convention in a paper d is:

$$P(x_d = +1|i, c) = \text{logit}^{-1}(\theta_i + b_c) = \frac{e^{\frac{1}{2}(\theta_i + b_c)}}{e^{\frac{1}{2}(\theta_i + b_c)} + e^{-\frac{1}{2}(\theta_i + b_c)}} \tag{3}$$

where θ_i is a latent parameter that encodes author i's preference ($\theta_i > 0$ implying a preference for the $+1$ convention) and b_c is a latent parameter that encodes the bias associated with context c (the category of literature to which the paper belongs[7]). In our account of conventions, θ_i is the mean-field effect of sequential consistency, and b_c is the mean-field effect of cultural traits interacting with the choice of the metric signature, given their distribution in a research area c. We assume that θ_i is drawn from a mixture of two distributions ($\theta_i = \pm\mu$), such that the model may capture the existence of two populations with a preference for each metric. We also assume that $b_c \sim \mathcal{N}(0, 1)$[8]. If $|\mu|$ is typically large, and larger than $|b|$, this would imply that scientists have preferences that generally exceed the influence of the context. As shown in Figure 3a, we find that scientists *do* have preferences that they tend to maintain across contexts, although there is some evidence that they occasionally adapt to the target research area. While we interpret such deviations from an author's preference as adaptation to the subject matter, they could indicate adaptation to the audience of the paper, in pursuit of social consistency (code-switching).

Figure 8b [9] confirms that authors tend to generally stick to the same metric in their works and that the prevalence of each preference varies depending on the authors' primary research area. This shows that authors manage the tension between sequential and contextual consistency by developing preferences adapted to their cultural environment.

2.2 Local versus global coordination: an Ising model approach

If scientists' attitude towards the sign convention was dictated by social consistency, then, their preferences should be aligned with their social environment. While there exists no

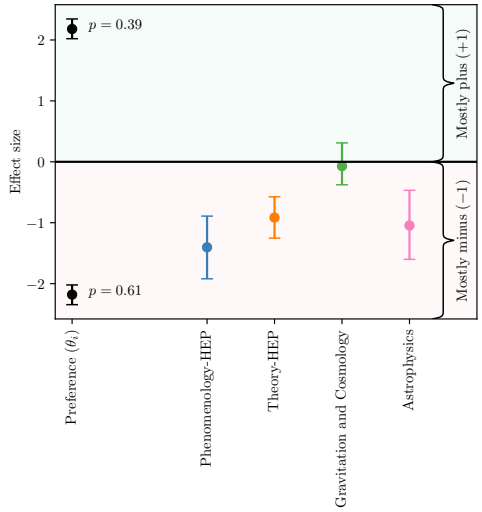
[7] In case a paper belongs to multiple categories, we average b_c over all these categories.

[8] We assume that:

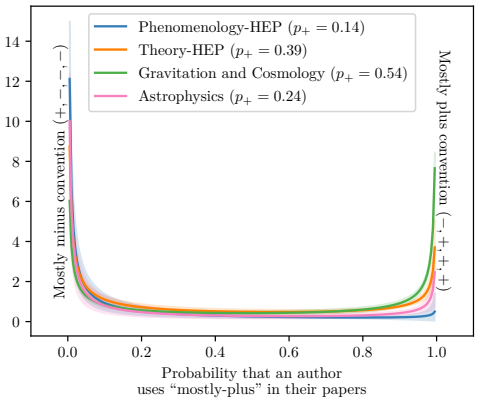
$$\theta_i = \begin{cases} +\mu & \text{with probability } p_{C_i} \\ -\mu & \text{with probability } 1 - p_{C_i} \end{cases}$$

where C_i is the primary research area of author i and $\mu \sim \text{Exponential}(1)$. The ability of this item-response model to reconstruct the latent parameters μ and b is tested with simulated data assuming no effect of sequential consistency, i.e. $\theta_i = 0$ for every author (S4.3, Figure 9).

[9] Given N_i, the amount of solo-authored papers by an author i with an explicit choice of metric signature, and k_i the amount of those using the $+1$ convention, we assume that $k_i \sim \text{Binomial}(N_i, p_i)$, with $p_i \sim \text{Beta}(\alpha_{C_i}, \beta_{C_i})$ and $\alpha_c, \beta_c \sim \text{Exponential}(1)$.



(a) Effect of sequential consistency (i.e. preferences, in black), and context (in color), on the choice of a convention in solo-authored papers. p indicates the prevalence of each preference (± 1).



(b) Distribution of authors' probability of using the mostly plus metric signature, according to their primary subject category. Shaded areas represent 95% credible intervals (CI). Distributions are generally bimodal, with two peaks at 0 and 1, which imply that authors tend to use always one or the other signature but rarely a mix of both in solo-authored papers.

Figure 3: **Importance of sequential and contextual consistency in scientists' behavior.**

universal norm at the level of the entire field, it could still be the case that scientists are at least behaving in a way consistent with their own collaborators. To establish whether this is the case, we explore the co-authorship graph (Figure 4), where each node i on the graph (each author) possesses a favorite convention $x_i \in \{\pm 1\}$ (as measured from their solo-authored publications). The weights of the edges (w_{ij}) encode the strength of the relationship between co-authors i and j[10]. We may then measure the average alignment between co-authors, $\langle x_i x_j \rangle = \sum_{i,j} w_{ij} x_i x_j / \sum_{i,j} w_{ij}$. We find $\langle x_i x_j \rangle = +0.32$, which is significantly more than would be expected by chance alone ($P < 10^{-4}$)[11]: despite the absence of universal norm, scientists' preferences are positively correlated with those of their collaborators.

How did such partial alignment emerge? Coordination among physicists may be achieved either locally (via short-range interactions between scientists), or globally, via shared culture. To delineate these two possibilities, we model physicists' preferences with an Ising model, with parameters J and $\boldsymbol{B}$, such that the probability $P(x_1, \ldots, x_n | J, \boldsymbol{B})$ of observing a particular configuration $x_1, \ldots, x_n$ is:

[10] We use $w_{ij} = \sum_{d | \{i,j\} \subset A_d} \frac{1}{|A_d| - 1}$, where A_d is the set of co-authors of publication d, following [37].

[11] We compare the observed value of $\langle x_i x_j \rangle$ to what would be expected if authors chose one or the other convention at random, with probabilities equal to the frequency of each convention. This null model predicts $\mathbb{E}[\langle x_i x_j \rangle] = 0.10$, far below the observed value.

$$P(x_1, \ldots, x_n | J, \boldsymbol{B}) = \frac{1}{Z(J, \boldsymbol{B})} e^{U(x_1, \ldots, x_n, J, \boldsymbol{B})} \tag{4}$$

$$\text{with } U = \underbrace{\sum_{i,j} J w_{ij} x_i x_j}_{\substack{\text{local} \\ \text{coordination}}} + \underbrace{\sum_i B_{C_i} x_i}_{\substack{\text{global} \\ \text{coordination}}} \tag{5}$$

Where C_i is the primary research area of i. J captures the effect of *local* coordination via pairwise interactions on the graph. $\boldsymbol{B} = (B_c)$ captures the *global* effect of each research area: their effect is global in that they equally affect all individuals within a group regardless of their position in the network. The Ising model follows naturally from eq. (1), §1.2 in coordination games. The $\boldsymbol{B}$ term introduces an asymmetry between authors from different research areas[12].

If $J > 0$, the potential U is higher in configurations in which nodes share the orientation of their neighbors. Such systems may undergo phase transitions towards configurations in which individual nodes spontaneously align over large distances. Although originated from spin physics, the Ising model provides a concise description of the emergence of collective behavior at large [24, 38].

In our case, we would like to infer the posterior distribution $P(J, B | x_1, \ldots, x_n)$ given $(x_1, \ldots, x_n)$. However, this distribution is computationally intractable, and we use the pseudo-likelihood approximation [39] which is accurate, efficient, and robust to missing data as we show in S4.4. The results are shown in Table 3. The inverse Ising approach reveals that research areas have large global effects, and that local coordination of co-authors has a small but statistically significant effect. However, this convention may propagate locally via channels others than collaborations, including citations (Figure 5). We account for this possibility by introducing an additional local contribution $J^{\text{cit}} \sum_j w_{ij}^{\text{cit}} x_j$ in the pseudo-likelihood ((6)), induced by the authors' citation graph G^{cit} which captures "who cites who"[13]. The weights w_{ij}^{cit} of the edges of G^{cit} measure the frequency of citations of j by i, given $w_{ij}^{\text{cit}} = \sum_{d,d' | i \in A_d, j \in A_{d'}, i \neq j} \frac{c_{dd'}}{|A_d||A_{d'}|}$ with $c_{dd'} = 1$ if d cites d' and 0 otherwise. After adding this contribution to (6), we find that both J and J^{cit} are significantly positive; that is, both co-authors and publications seem to carry an influence[14].

To assess which of local or global coordination dominate, we evaluate the fraction of authors for which local contributions in (6) exceed the global effect of $\boldsymbol{B}$. We find that local effects exceed and reverse global effects for 7% of the sample of 2 277 authors ($\text{CI}_{95\%} = [3\%\text{–}15\%]$). In addition, we find that the inclusion of local effects only

[12]Unlike Table 2, we assume that the effect of the asymmetry between research areas does not scale linearly with each node's degree centrality ($k_i = \sum_j w_{ij}$). Instead, each strategy is associated with a constant payoff $r_i = B_{C_i} x_i$ regardless of the interactions involving i [21]

[13]The pseudo-likelihood approach can directly accommodate asymmetric interactions in directed networks.

[14]That J remains positive after accounting for citations suggests that correlations between co-authors' preferences may not be explained solely by correlations in their research.

Figure 4: **Metric signature preferences in the co-author network**. Each node is an author. Edges represent co-authorship relationships between authors. Nodes' colors indicate authors' preferences (pink for -1, green for $+1$). Only the largest connected component is shown.

marginally improves the model's predictive accuracy, from an average of 67.7% (only considering global effects) to 70.2%. Therefore, local processes play a smaller role.

Measurements of J and $\boldsymbol{B}$ may be confounded by hidden structures. For instance, while $\boldsymbol{B}$ was assumed to be uniform within each of the four research areas, it may vary across subtopics within each research areas. If their effect is omitted, this might inflate the estimate of J. To assess this possibility, a linear Support Vector Machine classifier was trained to predict the metric signature from sentence embeddings of scientific abstracts. The accuracy on a test-set was 73%, slightly above the accuracy of a classifier relying only on the four categories (70%); in other words, categories contain most of the relevant contextual information. Conversely, the effect of each research area may reflect unmodeled social structures. Therefore, the Ising model is an effective parameterization, and the values of J and $\boldsymbol{B}$ may vary depending on the networks and scales under consideration. Finally, when the social structure is entirely correlated with another interacting trait, their relative effect cannot be teased apart.

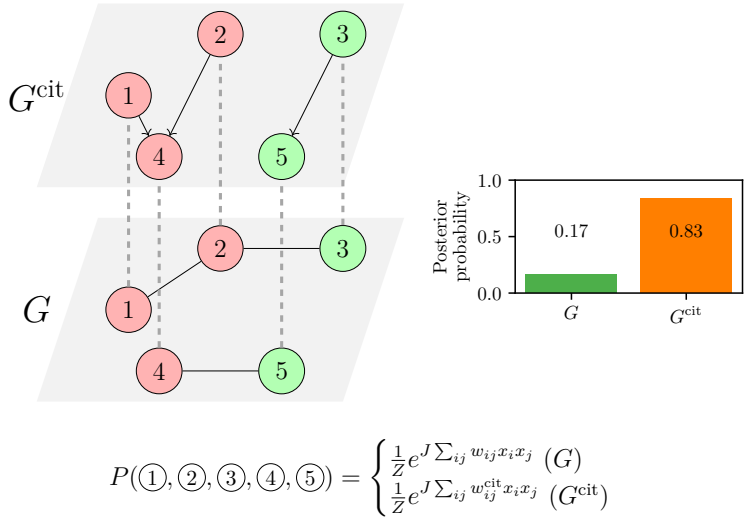


Figure 5: **Illustration of local coordination in multilayered social networks**. Nodes can be connected through different kinds of relationships (for instance, authors can be related via collaborations (G) or citations (G^{cit})). In this diagram, patterns of coordination are better explained by the directed graph at the top (G^{cit}). The Ising model correctly identifies the relevant social structure.

2.3 Inferring mechanisms of preference formation

The Ising model is certainly not a realistic description of how individuals form preferences. Nevertheless, idealized models from statistical physics can provide clues about the actual process. Below, we assess the relative plausibility of three hypothetical mechanisms according to their ability account for the observed values of J and $\boldsymbol{B}$.

The first proposed mechanism (M_1) is an agent-based model in which scientists operate a trade-off between social consistency (driven by coordination costs), sequential consistency (driven by switching costs), and contextual consistency (driven by maladap-

tation costs, i.e. incompatibility with their research area). In this model, the network is initialized in a random state; then, at every step of the simulation, scientists follow a best response strategy, by evaluating whether they would be better off changing their preference or not, given the magnitude of each of these costs, their probability of publishing in each research area, and their collaborators' preferences[15] (in that scenario, coordination is channeled by co-authorship and not citations). The second mechanism considered (M_2) is a global process of cultural transmission whereby scientists adopt a convention at the start of their career with a probability that depends on their primary research area, and on the time at which their career started. Such process is meant to capture the transmission of conventions via cultural artefacts such as textbooks (S4.7). Finally, the third mechanism considered (M_3) is a process of local cultural transmission, in which scientists copy the preference of their first co-author[16].

Figure 6 shows that each model predicts different patterns for J and $\boldsymbol{B}$. In particular, since it explicitly implements coordination costs (which are themselves driven by local interactions), the "strategic agent" model can predict large values of J. The model of cultural transmission via imitation predicts slightly higher values of J than global cultural transmission, but generally smaller values of B. Because of these distinctive patterns, we can compare each model's ability to account for the data using simulation-based inference [40]. As shown in Figure 6, the results seem to rule out purely global cultural transmission which fails to explain the magnitude of local coordination. There is slightly more evidence of partial local cultural transmission model.

2.4 Inferring mechanisms of conflict resolution

Coordination failures give rise to conflicts. Given that physicists' preferences are not perfectly aligned to those of their collaborators, they must occasionally resolve disagreements about which metric signature to use as they co-author a paper. We stressed that the resolution of conflicts in such scenarios implied a trade-off between optimality and decision costs: while some decisions may be superior to others, the cost of arguing and properly aggregating each author's input may exceed the benefits.

Below, we consider multiple preference aggregation strategies and estimate their prevalence given data about the metric signature selected in co-authored papers. As we will show, this provides indirect information about how authors navigate this trade-off in the case of the metric signature. We leverage papers with an identified metric signature $S \in \{\pm 1\}$ for which all authors' preferences $(x_i, \ldots, x_n) \in \{\pm 1\}^n$ were measured independently from single-authored papers. For many of these papers (182 papers with two authors, 28 papers with three authors, and 4 papers with four authors), authors have conflicting preferences. Since different processes of preference-aggregation occasionally predict different outcomes given $(x_i, \ldots, x_n) \in \{\pm 1\}^n$, we may infer their relative

[15] See S4.6 for a more precise description.

[16] The preference of scientists with no "parent" in the graph is drawn according to the same global process as in the global cultural transmission model (M_2), such that the process M_3 includes both local and global mechanisms. In total, in this model, 10% of authors form a preference by imitation.

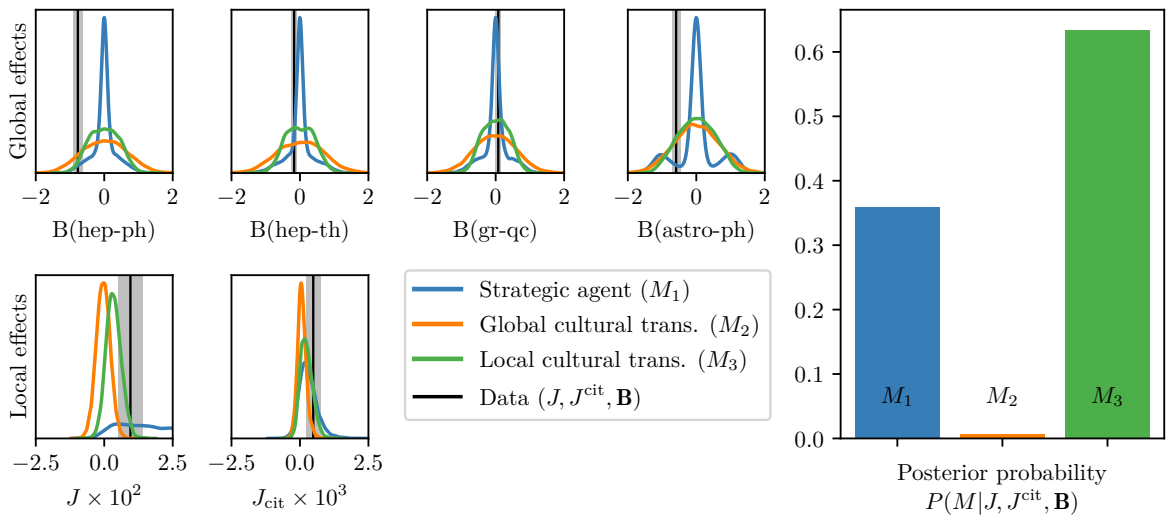


Figure 6: **Left plot:** marginal posterior distribution of summary statistics for each model (shown in colors), compared to the summary statistics derived from the data (indicated by black bars). Gray bars represent the 95% posterior credible interval of each parameter given the data. **Right plot:** posterior probability of each model given the observed parameters of the Ising model.

likelihood from the data.

In contrast to [15], which explored conventions for LaTex macros, we consider strategies of conflict-resolution suggested by the literature on judgment aggregation [41, 42]. First, we consider "dictatorial" strategies, whereby a specific author (the first author, the last author, or any other one) imposes their favorite convention (which, again, is independently measured from their solo-authored publications). Dictatorial strategies dismiss all information about other authors or the research context, such that the resulting decision is potentially suboptimal. We also consider a "majoritarian" process, whereby the majority preference is selected, thus maximizing collective satisfaction. These two strategies (dictatorial and majoritarian) are probably the most classic examples in social choice theory and in the preference and judgment aggregation literature [41, 42]. It is also tempting to consider the achievement of consensus through deliberation, another popular example. However, it seems difficult to infer whether a decision was reached from deliberation based solely on the observed outcome and each individual's initial preference. Instead, we consider a "random" process, equivalent to a coin-flip (in fact, in the two-author case, a coin-flip is presumably equivalent to deliberation, if both authors are equally influential in the deliberation). Finally, we include a "conventional" process, whereby the signature most frequent in a given context is retained, irrespective of the authors' preferences. We then estimate the prevalence prevalence π_k of each preference aggregation strategy $A_k \in \{A_1, \dots\}$, given that $P(S|x_1, \dots, x_n) = \sum_k P(S|x_1, \dots, x_n, A_k)P(A_k)$, and $A_k \sim \text{Categorical}(\pi_k)$.

Results are shown in Figure 7, given a flat Dirichlet prior on π_k. Due to the sample size, error bars are quite wide. Nevertheless, dictatorial strategies prevail ($\pi_{\text{dictatorial}} > 0.73$ at

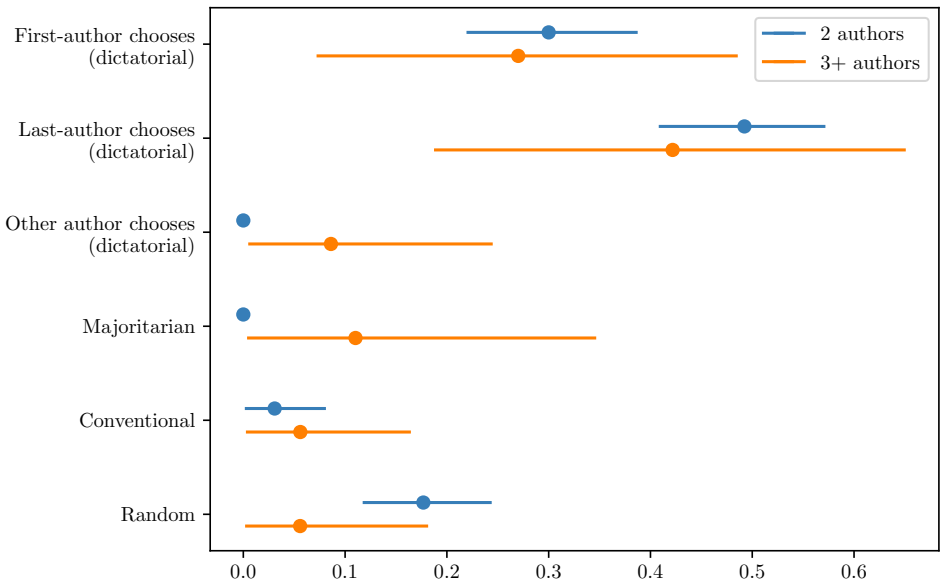


Figure 7: Prevalence of aggregation strategies. Error bars indicate 95% credible intervals. The dominant strategy seems to be that the last author dictates the metric convention.

the 95% credible level for the two-author case and $\pi_{\text{dictatorial}} > 0.57$ for the three+-author case – which is almost always three authors), even in the 3+ authors case (for which majority vote is possible): inequalities in the status of authors can help expedite judgment aggregation. More interestingly, in the two-author configuration, there is conclusive evidence that the first author is less likely to choose the metric signature compared to the last author ($P(\pi_{\text{first-author}} > \pi_{\text{last-author}}) = 0.008$). For 3+ authors, the data leans towards this direction as well ($P(\pi_{\text{first-author}} > \pi_{\text{last-author}}) = 0.222$); moreover, middle-authors seem less likely to dictate the final choice. The last author (who is generally in a leadership position) therefore seems to enjoy more influence, even though the first author carries a greater share of work (in principle) and would benefit from using their favorite metric signature[17]. This convergences with [15], which finds that "fights" for visible conventions in scientific papers are more often won by experienced authors. This emphasizes the role of leadership in the resolution of conflicts, and suggests that for this particular convention, "optimality" (whether in the sense of promoting collective agreement, or the first author's satisfaction) is sacrificed.

3 Discussion

This paper introduced a statistical physics account of conventions capturing the heterogeneous competing processes influencing their adoption. This account acknowledges that individuals attitude' towards conventions is dictated not only by an imperative of social

[17]Authorship norms are known to vary across fields [43]. To verify that these interpretation hold in fundamental physics, we evaluated the probabilities that the first-author or the last-author are strictly older than the other co-authors. We found an association between last-authorship and seniority (see S4.9 for more details).

coordination, but also by purely internal imperatives of sequential consistency (the need to avoid switching back-and-forth between different choices) and contextual consistency (the need to adopt mutually coherent systems of choices, as in epistemological holism [16]). Broadly speaking, conventionality arises when a collection of choices or traits is constrained only collectively, due to synergistic interactions between individual choices. Interestingly, all three imperatives (social, sequential, and contextual consistency) can be modeled by coordination games on some underlying graph structure, or, alternatively, by an Ising model on a graph. This connection with the Ising model can be leveraged to extract empirical information about real-life conventions. As we showed, it can infer the structure of the underlying social coordination game and the social networks involved, or reconstruct the underlying cultural landscape, in the case of systems of conventions.

Additionally, the framework predicts that social coordination can arise out of either *local* (or bottom-up) coordination, driven by dyadic interactions on a network, or from *global* (top-down) coordination, relying on shared culture and knowledge or institutions transcending the social infrastructure. The contribution of these two processes in real-life conventions can also be measured with an Ising model approach. Additionally, different models of preference-formation (imitation, adaptation, cultural artifacts, etc.) predict different magnitudes of local and global coordination, such that the magnitude of these two channels of coordination can help determine the actual underlying process in real conventions. In the case of the metric signature, using simulation-based inference, we found slightly more evidence in favor of cultural transmission of preferences via the imitation of a peer, a process that can explain a small but non-vanishing magnitude of local coordination. In scientific communities, it may explain which aspects of epistemic cultures belong to a "disciplinary matrix" [32] (the set of practices and values that scientists adopt as part of the process of acquiring and conforming to a disciplinary identity) and which aspects emerge more spontaneously and locally. More generally, we show how the Ising approach provides a relatively model-independent way of discriminating local (i.e. emergent and endogenous) from global (exogenous) collective synchronization using behavioral network data.

Finally, it was argued that the resolution of conflicts between multiple conventions implies a trade-off between the optimality of the outcome (e.g., the degree of collective satisfaction) and decision costs (i.e. the cost of reaching a decision). We illustrated this trade-off using the example of the metric signature. We inferred the prevalence of various preference-aggregation strategies in co-authored papers, and found more evidence for "dictatorial" strategies. Specifically, we found that the last-author's preference has a higher chance of prevailing, leading to suboptimal outcomes. Therefore, leadership and seniority play a role in addressing coordination problems in the absence of norm.

Overall, this work provides an array of tools for understanding either the lack of norm or the persistence of inferior norms and practices in a wide range of contexts. The proposed framework can be generalized in several ways. For instance, while the Ising model presupposes pairwise interactions (between, say, individuals, or cultural traits), complex systems often involve higher-order interactions. In particular, scientists frequently in-

teract in collaborations involving more than three authors. Such complex interactions, whether they involve multiple individuals or cultural traits, can be encoded by hypergraphs [21]. In our framework, this leads to a generalized Ising model [44]. Moreover, although this paper this paper limits itself to a binary convention, the approach can be extended to conventions involving more than two alternatives, as in the Potts model [45]. Finally, in contrast to [15], this paper has not paid much attention to temporal dynamics, due to the temporal sparsity of the data. Nevertheless, exploring such dynamics would provide more information about the underlying processes of transmission, or about how sequential consistency plays out over time.

Acknowledgements Many thanks to Radin Dardashti, Thomas Heinze, Olivier Morin, and Cailin O'Connor for their feedback. I am also grateful to Jeffrey Barrett, Sean Carroll, John Miller, Scott Page, and Brian Skyrms for inspiring discussions. Finally, I would like to thank Robert D. Hawkins for further suggestions.

Code and data The code and data for this paper is available at `https://gin.g-node.org/lucasgautheron/dilemmas-conventions`.

Funding The author acknowledges funding from the DFG Research Training Group 2696 "Transformations of Science and Technology since 1800".

Competing interests The author declares a personal inclination towards the $(+,-,-,-)$ metric signature.

References

[1] D. Lewis. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press., 1969.

[2] R. X. Hawkins, N. D. Goodman, and R. L. Goldstone. "The emergence of social norms and conventions". In: *Trends in cognitive sciences* 23.2 (2019), pp. 158–169.

[3] C. Elsenbroich and N. Gilbert. "Modelling Norms". In: *Modelling Norms*. Springer Netherlands, 2013, pp. 143–149. DOI: 10.1007/978-94-007-7052-2_10.

[4] J. Delgado. "Emergence of social conventions in complex networks". In: *Artificial Intelligence* 141.1–2 (2002), pp. 171–185. DOI: 10.1016/s0004-3702(02)00262-x.

[5] J. M. Pujol, J. Delgado, R. Sangüesa, and A. Flache. "The role of clustering on the emergence of efficient social conventions". In: *Proceedings of the 19th international joint conference on Artificial intelligence*. 2005, pp. 965–970.

[6] F. Guala and L. Mittone. "How history and convention create norms: An experimental study". In: *Journal of Economic Psychology* 31.4 (2010), pp. 749–756. DOI: 10.1016/j.joep.2010.05.009.

[7] D. Centola and A. Baronchelli. "The spontaneous emergence of conventions: An experimental study of cultural evolution". In: *Proceedings of the National Academy of Sciences* 112.7 (2015), pp. 1989–1994. DOI: 10.1073/pnas.1418838112.

[8] R. X. D. Hawkins and R. L. Goldstone. "The Formation of Social Conventions in Real-Time Environments". In: *PLOS ONE* 11.3 (2016). Ed. by C. T. Bauch, e0151670. DOI: 10.1371/journal.pone.0151670.

[9] A. Formaux, D. Paleressompoulle, J. Fagot, and N. Claidière. "The experimental emergence of convention in a non-human primate". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1843 (2021). DOI: 10.1098/rstb.2020.0310.

[10] R. D. Hawkins, M. Franke, M. C. Frank, A. E. Goldberg, K. Smith, T. L. Griffiths, and N. D. Goodman. "From partners to populations: A hierarchical Bayesian account of coordination and convention." In: *Psychological Review* 130.4 (2023), p. 977.

[11] V. Boyce, R. D. Hawkins, N. D. Goodman, and M. C. Frank. "Interaction structure constrains the emergence of conventions in group communication". In: *Proceedings of the National Academy of Sciences* 121.28 (2024). DOI: 10.1073/pnas.2403888121.

[12] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. "No country for old members: user lifecycle and linguistic change in online communities". In: *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013. DOI: 10.1145/2488388.2488416.

[13] F. Kooti, H. Yang, M. Cha, K. Gummadi, and W. Mason. "The emergence of conventions in online social networks". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 6. 1. 2012, pp. 194–201.

[14] B. Heaberlin and S. DeDeo. "The Evolution of Wikipedia's Norm Network". In: *Future Internet* 8.2 (2016), p. 14. DOI: 10.3390/fi8020014.

[15] R. Rotabi, C. Danescu-Niculescu-Mizil, and J. Kleinberg. "Competition and Selection Among Conventions". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. International World Wide Web Conferences Steering Committee, 2017, pp. 1361–1370. DOI: 10.1145/3038912.3052652.

[16] Y. Ben-Menahem. *Conventionalism: From Poincaré to Quine*. Cambridge University Press, 2006.

[17] J. Wu, C. O'Connor, and P. E. Smaldino. "The Cultural Evolution of Science". In: *The Oxford Handbook of Cultural Evolution*. Oxford University Press, 2023. DOI: 10.1093/oxfordhb/9780198869252.013.78.

[18] T. Wilholt. "Bias and values in scientific research". In: *Studies in History and Philosophy of Science Part A* 40.1 (2009), pp. 92–101.

[19] P. E. Smaldino and C. O'Connor. "Interdisciplinarity can aid the spread of better methods between scientific communities". In: *Collective Intelligence* 1.2 (2022), p. 263391372211318. DOI: 10.1177/26339137221131816.

[20] A. Correia, L. Leestmaker, H. Stoof, and J. Broere. "Asymmetric games on networks: Towards an Ising-model representation". In: *Physica A: Statistical Mechanics and its Applications* 593 (2022), p. 126972. DOI: 10.1016/j.physa.2022.126972.

[21] F. Zimmaro, S. Galam, and M. A. Javarone. "Asymmetric games on networks: Mapping to Ising models and bounded rationality". In: *Chaos, Solitons & Fractals* 181 (2024), p. 114666. DOI: 10.1016/j.chaos.2024.114666.

[22] G. Szabó and I. Borsos. "Evolutionary potential games on lattices". In: *Physics Reports* 624 (2016), pp. 1–60. DOI: 10.1016/j.physrep.2016.02.006.

[23] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki. "Statistical physics of human cooperation". In: *Physics Reports* 687 (2017), pp. 1–51. DOI: 10.1016/j.physrep.2017.05.004.

[24] M. W. Macy, B. K. Szymanski, and J. A. Hołyst. "The Ising model celebrates a century of interdisciplinary contributions". In: *npj Complexity* 1.1 (2024). DOI: 10.1038/s44260-024-00012-0.

[25] C. O'Connor. "Measuring Conventionality". In: *Australasian Journal of Philosophy* 99.3 (2020), pp. 579–596. DOI: 10.1080/00048402.2020.1781220.

[26] L. Gasparri. "Inherent and probabilistic naturalness". In: *Philosophical Studies* 181.2–3 (2023), pp. 369–385. DOI: 10.1007/s11098-023-02070-x.

[27] V. M. Poulsen and S. DeDeo. "Inferring Cultural Landscapes with the Inverse Ising Model". In: *Entropy* 25.2 (2023), p. 264. DOI: 10.3390/e25020264.

[28] I. Pascual, J. Aguirre, S. Manrubia, and J. A. Cuesta. "Epistasis between cultural traits causes paradigm shifts in cultural evolution". In: *Royal Society Open Science* 7.2 (2020), p. 191813. DOI: 10.1098/rsos.191813.

[29] W. V. O. Quine and J. S. Ullian. *The web of belief*. Vol. 2. Random House New York, 1978.

[30] H. P. Young. "The economics of convention". In: *Journal of economic perspectives* 10.2 (1996), pp. 105–122.

[31] R. Moore. "Imitation and conventional communication". In: *Biology & Philosophy* 28.3 (2012), pp. 481–500. DOI: 10.1007/s10539-012-9349-8.

[32] T. S. Kuhn. *The Structure of Scientific Revolutions*. 2nd edition, with postscript. Chicago: University of Chicago Press, 1970.

[33] E. Lee, J. Lee, and J. Lee. "Reconsideration of the winner-take-all hypothesis: Complex networks and local bias". In: *Management science* 52.12 (2006), pp. 1838–1848.

[34] O. E. Williamson. *Markets and Hierarchies, Analysis and Antitrust Implications: A study in the economics of internal organization*. New York: Free Press, 1975.

[35] R. Calvert. "Leadership and its basis in problems of social coordination". In: *International Political Science Review* 13.1 (1992), pp. 7–24. DOI: 10.1177/019251219201300102.

[36] F. Dietrich and C. List. "Arrow's theorem in judgment aggregation". In: *Social Choice and Welfare* 29.1 (2006), pp. 19–33. DOI: 10.1007/s00355-006-0196-x.

[37] M. E. Newman. "Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks". In: *Complex Networks*. Springer, 2004, pp. 337–370. DOI: 10.1007/978-3-540-44485-5_16.

[38] M. Galesic and D. Stein. "Statistical physics models of belief dynamics: Theory and empirical tests". In: *Physica A: Statistical Mechanics and its Applications* 519 (2019), pp. 275–294. DOI: 10.1016/j.physa.2018.12.011.

[39] H. C. Nguyen, R. Zecchina, and J. Berg. "Inverse statistical problems: from the inverse Ising problem to data science". In: *Advances in Physics* 66.3 (2017), pp. 197–261. DOI: 10.1080/00018732.2017.1341604.

[40] K. Cranmer, J. Brehmer, and G. Louppe. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062. DOI: 10.1073/pnas.1912789117.

[41] K. J. Arrow. *Social Choice and Individual Values*. John Wiley & Sons, 1951.

[42] C. List and P. Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, 2011. DOI: 10.1093/acprof:oso/9780199591565.001.0001.

[43] L. Waltman. "An empirical analysis of the use of alphabetical authorship in scientific publishing". In: *Journal of Informetrics* 6.4 (2012), pp. 700–711.

[44] T. Robiglio, L. Di Gaetano, A. Altieri, G. Petri, and F. Battiston. *Higher-order Ising model on hypergraphs*. 2024. DOI: 10.48550/ARXIV.2411.19618.

[45] F.-Y. Wu. "The potts model". In: *Reviews of modern physics* 54.1 (1982), p. 235.

[46] S. T. Radev, M. D'Alessandro, U. K. Mertens, A. Voss, U. Köthe, and P.-C. Bürkner. "Amortized bayesian model comparison with evidential deep learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 34.8 (2021), pp. 4903–4917.

[47] S. T. Radev, M. Schmitt, L. Schumacher, L. Elsemüller, V. Pratz, Y. Schälte, U. Köthe, and P.-C. Bürkner. *BayesFlow: Amortized Bayesian Workflows With Neural Networks*. 2023.

[48] M. Galesic, H. Olsson, J. Dalege, T. van der Does, and D. L. Stein. "Integrating social and cognitive aspects of belief dynamics: towards a unifying framework". In: *Journal of The Royal Society Interface* 18.176 (2021). DOI: 10.1098/rsif.2020.0857.

4 Supplementary materials

4.1 Regular expressions for determining the metric signature

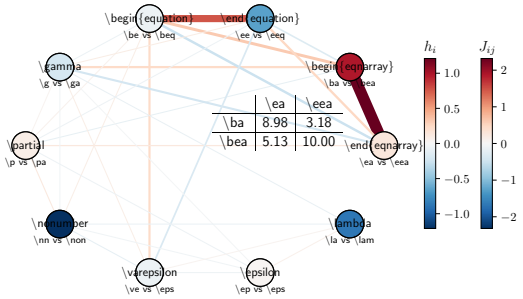
The following case-insensitive regular expressions have been used to detect occurences of the mostly minus signature:

- `(([,\s\{\}]*)(\+|1)([,\s\{\}1]*)){1}(([,\s\{\}]*)\-([,\s\{\}1]*)){3,}`
- `(mostly[-\s]*minus|west[-\s]*coast)`
- `g_\{(00|tt)\}[\s]*=[\s]*[+]?[\s]*1`
- `\\Box(\^(\{2\}|2))?[\s]*\+[\s]*m\^(\{2\}|2)`

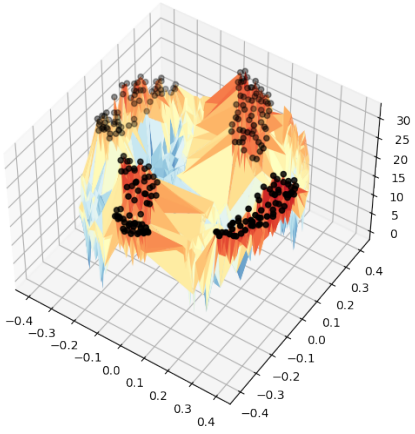
Symmetric expressions are conversely employed for detecting instances of the mostly plus metric signature.

4.2 Reconstructing cultural landscapes of conventions

This section illustrates the ability of the Ising model to reconstruct cultural fitness landscapes of conventions. To this end, following [15], we explore ten naming conventions involving LaTeX macros abbreviating the name of frequently used LaTeX commands. For instance, instead of writing `\begin{equation}` at the beginning of each equation, authors often use an abbreviated name (e.g. `\be`) by defining a custom macro (e.g. `\newcommand{\be}{\begin{equation}}`). The choice of abbreviated name is conventional and in principle at the author's discretion. For instance, certain authors prefer "`\beq`" over `\be`. We collect occurrences of the two most frequent abbreviated names for each of ten such conventions, labeled by -1 (for the shorter version, for instance `\be`) and $+1$ (for the longer version, for instance `\beq`). Following our framework, we assume that the fitness of a combination of choices is given by the Ising model, $U(x_1, \ldots, x_{10}) = \sum_{ij} J_{ij} x_i x_j + \sum_i h_i x_i$. Assuming $J_{ij}, h_i \sim \mathcal{N}(0, 1)$ and, using data from 77 000 papers in which at least one convention appears, we solve an inverse Ising problem to recover (J_{ij}) and (h_i). The results are shown in Figure 8. This reveals strong interactions between two pairs of choices: the abbreviations of (`\begin{equation}`, `\end{equation}`), and those of (`\begin{eqnarray}`, `\end{eqnarray}`).



(a) Ising model representation of a cultural landscape involving ten binary conventions, following [27]. Each convention represents a choice between two options to shorten the name of a LaTeX command. Edges represent interactions between conventions. Thick edges designate traits that must be mutually consistent. Node colors indicate the intrinsic advantage of a specific choice (blue indicates that the shorter abbreviation is favored, red indicates a general preference for the longer abbreviation). Interactions between two traits can be seen as coordination games. The inverse Ising problem recovers the coefficients of the underlying payoff matrices.



(b) Three-dimensional representation of the cultural landscape. Each of the $2^{10} = 1024$ potential combination of traits is mapped onto two dimensions using multidimensional scaling. The height and color of the landscape indicates the fitness of each combination. There are four large peaks (in red). The black dots represent the configurations in which the abbreviations for the pairs (`\begin{equation}`, `\end{equation}`) & (`\begin{eqnarray}`, `\end{eqnarray}`) are consistent.

Figure 8: **Cultural landscape of common abbreviations of LaTeX commands.**.

4.3 Sequential versus contextual consistency: model assessment

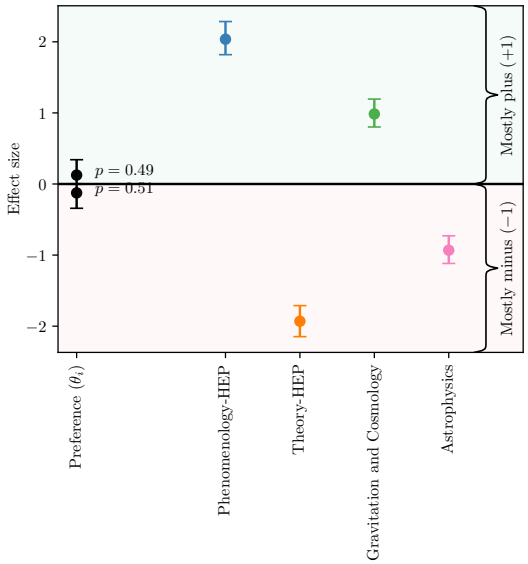


Figure 9: The analysis in 2.1 is re-iterated with simulated data instead of actual data. The simulation assumes that $\theta_i = 0$ for all authors (i.e. there is no effect of consistency), while each research area has a significant effect. The inference correctly finds that $|\theta|$ is nearly zero and correctly identifies the ground truth size of the effect of each research area (+2, -2, +1, and -1 respectively).

4.4 Inverse ising problem and the pseudo-likelihood approach

The pseudo-likelihood method [39] transforms the inverse Ising problem into a tractable logistic regression, based on the likelihood of observing each individual spin conditional on the others, i.e.:

$$\prod_i P(x_i = +1 | \{x_{j \neq i}\}) = \prod_i \frac{e^{+J \sum_j w_{ij} x_j + B_{C_i}}}{e^{+J \sum_j w_{ij} x_j + B_{C_i}} + e^{-J \sum_j w_{ij} x_j - B_{C_i}}} \tag{6}$$

Using simulated configurations of G, we demonstrate that the pseudo-likelihood approach provides reliable estimates of J and $\boldsymbol{B}$, if all x_j are observed, and for $J \leq 10^{-2}$ (Figure 10). In the case that a value x_j is unknown, due to a lack of paper solo-authored by j with an identified metric signature, then author j is omitted from the sums in (6). This is equivalent to imputing $x_j = 0$[18]. We find that this approach is able to recover reliable information about the true value of J (Appendix 4.4, Figure 10). However, we may fear that the imputation of missing data (equivalently interpretable as the removal of unobserved nodes from the network) introduces bias in our inference [27]. A proper handling of unknown authors' preferences would require marginalizing eq. (6) over the

[18]This imputation strategy is also equivalent to restricting the inference procedure to a sub-graph of the co-authorship graph, including only the nodes and edges involving the 2 277 authors whose preference could be identified in at least one solo-authored paper.

2^m possible combinations of the m underlying unobserved signatures[19]. Unfortunately, the amount of missing data makes this impossible. However, this issue is not necessarily critical if, ultimately, we are less interested in recovering the exact values of J and $\boldsymbol{B}$ than in using the estimates as summary statistics for the purpose of comparing multiple models of the formation of individual preferences. Then, as long as each model predicts distinct patterns for the best-fit values of J and $\boldsymbol{B}$, the procedure remains useful. In any case, simulations show that the measured value of J is very correlated with the true value, even when nodes with missing data are masked during the inference process (cf. Appendix 4.4, Figure 10). Finally, missing data could be a feature rather than a bug; they might manifest the fact that certain authors make no explicit use of a specific metric signature, in which case it is reasonable to assume that they may not exert any influence over their co-authors' preferences.

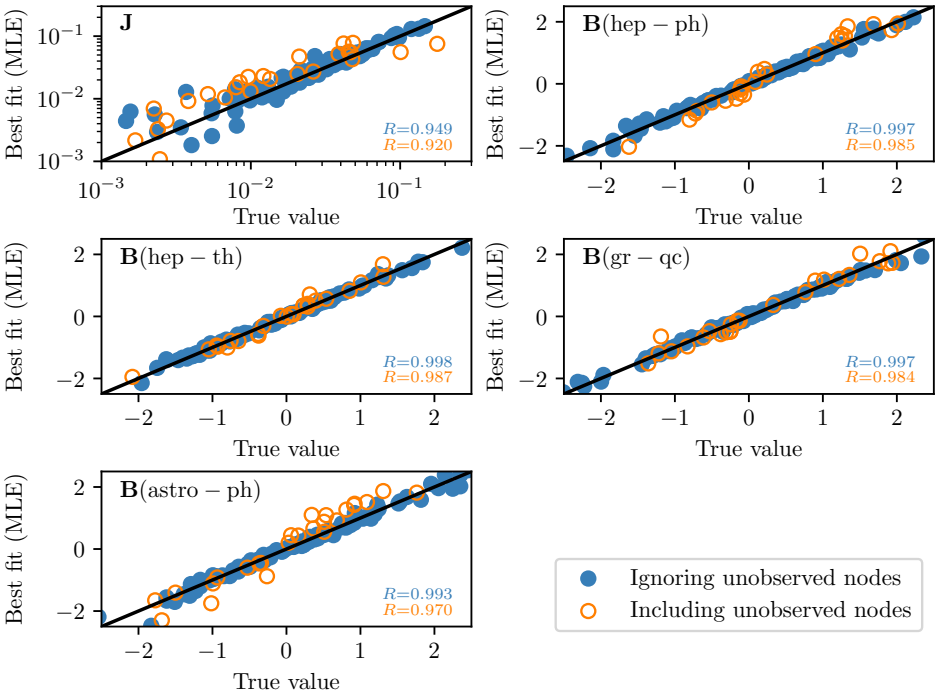


Figure 10: Robustness of the pseudo-likelihood approach for measuring J and $\boldsymbol{B}$. "True" values of J and $\boldsymbol{B}$ are drawn at random [$J \sim \text{Exponential}(1/J^*)$, $\boldsymbol{B} \sim \mathcal{N}(0,1)$]. Node configurations (x_i) are drawn at random according to the Ising model for each values of J and $\boldsymbol{B}$, using Gibbs sampling, either i) removing or ii) including nodes corresponding to authors whose preference is not observed in the data. Finally, the maximum likelihood estimates (MLE) J^{MLE} and $\boldsymbol{B}^{\text{MLE}}$ are recovered with the pseudo-likelihood approach, for each configuration (x_i), imputing $x_i = 0$ for authors whose preference was not observed in our data. The best-fit values are in reasonably good agreement with the true values over the simulated range, although they are much less accurate in the case where unobserved authors are included in the Gibbs sampling process.

[19] An alternative would be Gibbs sampling, which may handle missing data without marginalization, though it turned out to perform worse than HMC in the present case.

Parameter	Effect size	$CI_{95\%}$	Effect size	$CI_{95\%}$
J	+0.013	[+0.009, +0.017]	+0.0095	[+0.0052, +0.014]
J^{cit}	-	-	+0.00049	[+0.00023, +0.00075]
$B(\text{hep} - \text{ph})$	-0.86	[-0.99, -0.73]	-0.77	[-0.91, -0.64]
$B(\text{hep} - \text{th})$	-0.22	[-0.29, -0.15]	-0.17	[-0.24, -0.095]
$B(\text{gr} - \text{qc})$	+0.075	[-0.0069, +0.16]	+0.076	[-0.0066, +0.16]
$B(\text{astro})$	-0.6	[-0.74, -0.47]	-0.59	[-0.73, -0.46]

Table 3: Parameters of the Ising model.

4.5 Models of preference formation

Three models are considered: a strategic agent model (M_1), a global-transmission model (M_2), and a local transmission model (M_3). Many samples are drawn according to each generative process M_1, M_2, M_3. For each sample, we infer the parameters of the Ising model ($\boldsymbol{B}$, J and J^{cit}) – ignoring the authors whose actual preference is unknown, in order to preserve the compatibility with the values of $\boldsymbol{B}$, J and J^{cit} inferred from the actual data). Since each model generates slightly different patterns for these parameters (Figure 6), these can be used as summary statistics for estimating their relative plausibility given the observed data, $P(M|J, J^{\text{cit}}, \boldsymbol{B})$. For this task, we use simulation-based inference [40] with BayesFlow [46, 47]. This procedure allows to perform Bayesian inference when one lacks an analytical expression for the likelihood $P(D|M)$, and all that can be done is drawing samples by simulating the generative process M. This technique is especially useful for making inferences about models defined by complex programs, such as agent-based models. When the data is highly dimensional (as in the present case), this approach requires "summary statistics" [40]. Interestingly, the parameters of the Ising model can serve this role. Figure 12 confirms that the procedure exhibits some ability to discriminate the three models.

4.6 Strategic agent model

The "strategic agent" model proceeds as follow:

1. The parameters of the model are drawn at random:
   - $c_b \sim \mathcal{N}(0, 1)$, defined for each research area b, is the (dis)advantage of the +1 convention in b. The cost of using a convention x in context b is $\max(0, -xc_b)$.
   - $c_c \sim \text{Exponential}(\langle d_i \rangle)$ represent the magnitude of coordination costs, where $\langle d_i \rangle$ is the average degree-centrality of authors in the co-authorship graph. The mean is thus set such that $\langle c_c \rangle \langle d_i \rangle = 1$.
   - The cost of switching from one convention to another is fixed ($c_s = 1$)[20].

[20]This breaks a degeneracy of the model due to scale-invariance (if all costs were rescaled by a certain quantity, agents' behavior would remain identical).

2. At $t = 0$, the network is initialized in a random state: $x_{i,t=0}$ is set to either -1 or $+1$ with equal probabilities.

3. At $t + 1$, each agent compares their payoff in two scenarios: i) they switch their preference ($x_{i,t+1} = -x_i$) or ii) they maintain it ($x_{i,t+1} = x_i$). The difference in payoffs is:

$$\Delta = -c_s - c_c \sum_j w_{ij} \left(\max(0, x_{j,t} x_{i,t}) - \max(0, -x_{j,t} x_{i,t})\right) - \sum_b p_{ib} \left(\max(0, x_{i,t} c_b) - \max(0, -x_{i,t+1} c_b)\right) \tag{7}$$

   Where p_{ib} is the probability that i publishes in research area b. If $\Delta > 0$, i switches their preference. The cost of switching (c_s) introduces an asymmetry in Δ and has the effect of a conservative bias.

4. The process is repeated 50 times. The amount of steps reflects a compromise between performance and convergence.

This best-response strategy model is similar to common logit-response approaches to belief dynamics such as [48], in the limit $\beta \to +\infty$ (see eq. 1.6).

4.7 Global transmission model

For the global transmission model, we assumed that the probability of adopting a specific convention depends on both time and the author's primary research area. The time-dependence was captured by a random walk. The rate of change in the random walk was obtained by fitting the model to data on reference books for which approximate patterns of citations throughout time could be measured. We manually determined the metric convention used in each of these references. These gave us a measure of the prevalence of each convention in the citations of reference textbooks' throughout time. Unfortunately, this measure itself was too imperfect to reflect the actual probability that a scientist adopts a convention from a specific textbooks. Nevertheless, we used the rate of variation of this measure with time in our random walk model.

4.8 Distribution of summary statistics across models

Conditioning the outcome of simulations on high-dimensional data D to evaluate $P(\cdot|D)$ is difficult because the probability of generating exactly D becomes virtually zero. One should therefore condition on summary statistics T living in a lower dimensional space. Ideally, the mapping $f : D \mapsto T$ should be chosen in a way that maximizes our ability to tell apart the hypotheses that we seek to discriminate. In our case, $f : (x_1, \ldots, x_n) \mapsto J, \boldsymbol{B}$ may not be optimal in that specific sense, but it has *some* discriminating power (see Figure 12) and has the merit of interpretability. A trivially better summary statistic for assessing the plausibility of, say, the model of local cultural transmission would be, for instance, the average rate of agreement between each author's and their first co-author (whose preference they should have imitated, according to the model).

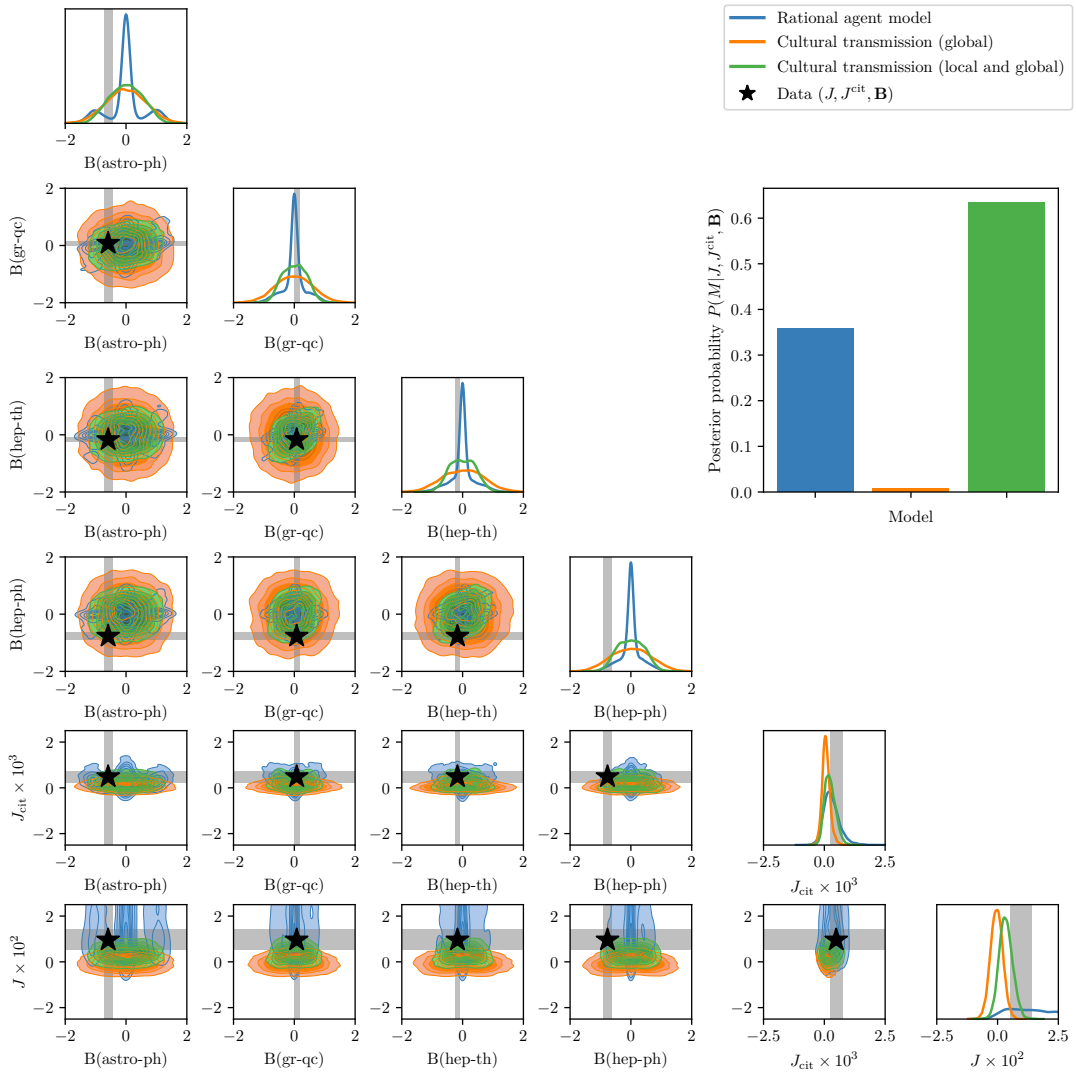


Figure 11: **Bottom-left pair plot:** distribution of summary statistics for each model (shown in colors), compared to the summary statistics derived from the data (shown as black stars). Plots on the diagonal show the marginal posterior distribution of each summary statistics for each model (gray bars represent the 95% posterior credible interval of each parameter given the data). **Top-right bar plot:** posterior probability of each model given the observed parameters of the Ising model.

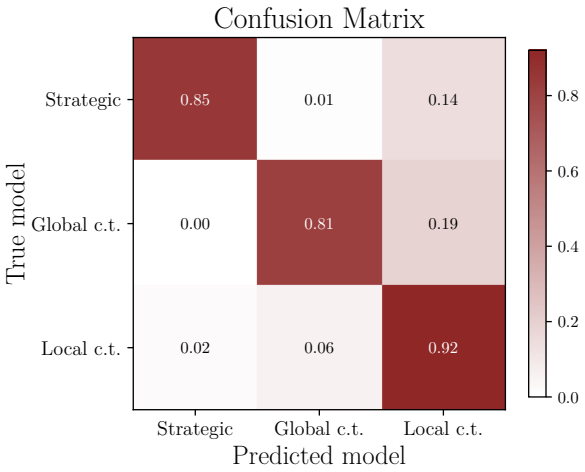


Figure 12: Reliability of the simulation-based model comparison approach. The confusion matrix represents the probability that a sample drawn from the horizontal model is attributed to the vertical model.

4.9 Authorship norms

We investigated authorship norms in fundamental physics (excluding experimental physics, which are not considered in this paper and have very unusual norms). We found that the author-list of 79% of two-author papers are alphabetically ordered. Given that for n authors, there is a $1/(n!)$ chance that any ordering is equal to the alphabetical order, this implies that 56% of two-author papers author-lists are *intentionally* ordered [43]. This number goes down to 45% for four-author publications. Therefore, despite a high prevalence of alphabetical ordering in fundamental physics compared to other disciplines (as found by [43]), in about half of the publications the ordering of authors is meaningful.

Most importantly, we found evidence that last-authorship is associated with seniority: in 54% of two-author papers, the last author has an academic age strictly higher than the first author; in comparison, in only 40% of cases, the first-author has strictly higher seniority compared to the last-author. In the three-author case, the last author has the strictly highest seniority in 29% of cases, versus 17% for the first-author.