Philosophy of Science (2024), **91**, 151–174
doi:10.1017/psa.2023.97

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

Epistemic Markers in the Scientific Discourse

Christophe Malaterre and Martin Léonard

Département de philosophie and Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal (UQAM), Montréal, Québec, Canada
Corresponding author: Christophe Malaterre; Email: malaterre.christophe@uqam.ca

(Received 14 July 2022; revised 21 April 2023; accepted 06 July 2023; first published online 22 August 2023)

Abstract

The central role of such epistemic concepts as *theory*, *explanation*, *model*, or *mechanism* is rarely questioned in philosophy of science. Yet, what is their actual use in the practice of science? Here, we deploy text-mining methods to investigate the usage of 61 epistemic notions in a corpus of full-text articles from the biological and biomedical sciences ($N = 73{,}771$). The influence of disciplinary context is also examined by splitting the corpus into subdisciplinary clusters. The results reveal the intricate semantic networks that these concepts actually form in the scientific discourse, not always following our intuitions, at least in some parts of science.

1. Introduction

The explication of epistemic concepts—*explanation*, *model*, *theory*, and many others—occupies a central place in the philosophy of science (Carnap 1950; Maher 2007; Justus 2012; Cappelen 2018). One only needs to think, for instance, about the numerous publications the notion of *mechanism* has triggered in the last couple of decades (Machamer et al. 2000; Bechtel and Abrahamsen 2005; Craver 2007; Glennan 2017). The rationale for this type of philosophical work is to be found in the roles these concepts play in the elaboration and formulation of scientific knowledge. Because these roles are often abstracted from specific case studies, a question arises as to their actual representativeness. The question is not new; for instance, some have criticized the lack of relevance of the early physics-inspired philosophy of science to the sciences at large and to the biological sciences in particular (Hull 1974; Ruse 1973; Grene and Depew 2004; Rosenberg and McShea 2008). Also, by often targeting epistemic concepts in relative isolation from one another, conceptual explication contributes to what can be thought of as a "silo view" of epistemic concepts, even if nuances exist (e.g., the concept of *mechanism*, for instance, has often been explicated with a view to examining its contribution to the concept of *explanation*). A most important issue

would therefore be to get a good footing about the actual usage of epistemic concepts in science: Which concepts are indeed used? In which contexts?

Using text-mining approaches, Overton examined the usage of the term *explanation* in science ($N = 781$, journal *Science*) and found the term to be more frequently used than in nonscientific texts (Overton 2013). Similarly, McPhetres and colleagues recently investigated the usage of *theory* in psychological research specifically ($N = 2{,}225$, journal *Psychological Science*) and found the term present in 53.6% of articles and referring to over 350 different named theories (McPhetres et al. 2021).[1] Research has also been done on the use of epistemic virtues (through the Journal Storage [JSTOR] Data for Research interface), highlighting the predominance of *simplicity*, *accuracy*, and *consistency* over other virtues such as *fruitfulness* or *testability*, but with differences depending on their application to *theories*, *hypotheses*, or *models* (Mizrahi 2022).

Beyond these studies that we were aware of, several questions remain, notably about the relative significance of numerous other epistemic terms, the possible influence of disciplinary context on terminological usage, and the relationships that epistemic terms entertain with one another. It is indeed quite likely that science and scientists mobilize a plurality of epistemic concepts in conjunction with one another while navigating the variegated epistemic landscapes they encounter. Hence, the objective of the present study was to investigate the actual usage of a large set of epistemic terms in scientific practice.

For feasibility reasons, we limited our domain of investigation to several of the biological and biomedical sciences. This choice was in part triggered by the availability of a corpus of full-text articles from the BioMed Central collection of journals ($N = 73{,}711$), to which we chose to apply quantitative text-mining methods. This was done with a view to assessing the relative significance of epistemic concepts and their relationships with one another. The present work builds on a previous study that examined the usage of six epistemic terms in that same corpus (*theory*, *model*, *mechanism*, *explanation*, *understanding*, and *prediction*) based on terminological co-occurrences (Malaterre and Léonard, 2023). Here, we targeted a much larger collection of terms and improved on the methods to identify sets of terms expressing the same concept.[2] We defined 61 "epistemic fields," which are collections of related terms considered to be epistemic markers of the chosen concepts. We then measured the frequency of occurrence of these epistemic fields in the corpus articles and investigated the correlation networks they formed. We also examined the influence of disciplinary context on concept. The resulting picture is one in which epistemic concepts form complex networks of semantic dependencies that sometimes align, but sometimes not, with philosophical conceptual explications.

We first describe the text-mining methods we used (sec. 2) and present our results in three stages: frequencies of epistemic fields across the entire corpus (sec. 3), correlations between epistemic fields across the entire corpus (sec. 4), and influence of disciplinary context on semantic field usage (sec. 5). Finally, the results are discussed together with future directions for research (sec. 6).

[1] We thank Maarten Derksen for this reference.

[2] For example, explanation can be expressed by the term *explanation* but also by the verb *to explain* and its multiple grammatical forms.

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

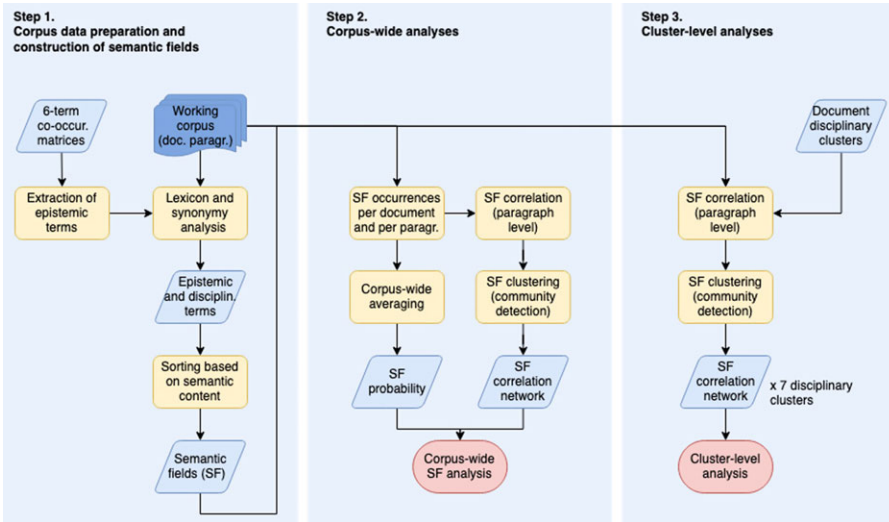

Figure 1. Research design. Three major steps of computational methods, from corpus-data preparation to semantic field correlation analyses. (Textual corpus in dark blue, data in light blue, operations in orange, analyses in red.) Figure in color online.

2. Methodology

The main intuition that drives computational textual analyses is that words are not used at random for expressing ideas but in specific patterns, whose investigation can in turn shed light on the semantic context of these words (Firth 1957). Computational approaches have turned this intuition into effective algorithms and methods that make it possible to mine word patterns so as to shed light on the semantic content of specific corpora (Aggarwal 2015). Such approaches are starting to be used in history and philosophy of science (HPS) studies (Overton 2013; Pence and Ramsey 2018; Malaterre et al. 2019; Noichl 2021; Mizrahi 2020; Khelfaoui et al. 2021). Here, we focus on terminological frequency analyses and the identification of correlation patterns between groups of terms (notably, between what we call "epistemic fields"). The research design includes three main steps (fig. 1).

The first step consisted of preparing the corpus and data for analysis. The working corpus was retrieved from Malaterre and Léonard (2023) and consisted of the full-text content of 73,771 articles from the BioMed collection that had been cleaned and lemmatized (i.e., terms had been converted to their dictionary forms, for instance, *models* to *model*) and sorted into seven disciplinary clusters (based on topic-modeling and k-means clustering). An initial major task consisted of preparing a list of "epistemic markers" or target epistemic terms that could be taken as indicators of specific epistemic concepts. We proceeded in two main stages. We built a first list of terms by analyzing the co-occurrence matrices of the six epistemic terms that had been targeted by Malaterre and Léonard (2023) (*explanation*, *understanding*, *prediction*, *model*, *theory*, *mechanism*). These matrices included co-occurrences measured over the entire corpus as well as co-occurrences measured in each of the seven disciplinary clusters. The extraction of the top-20 co-occurring terms across all matrices resulted

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

in a list of 202 unique terms, which were split into three main groups: 52 epistemic terms (e.g., *experiment*, *confirm*, *accuracy*), 52 disciplinary terms (e.g., *molecular*, *disease*, *selection*), and 98 common terms that were found to be too generic or difficult to interpret and were left aside (e.g., *number*, *result*, *change*, *system*, *difference*). Second, to get a broader coverage of epistemic concepts, we analyzed the most frequently used terms in the overall corpus (8,282 terms present in at least 1,000 documents each). In complement, a thesaurus search was run for synonyms. This resulted in the identification of 159 additional epistemic terms of potential interest. Combining the terms identified through these two stages resulted in a list of 211 target epistemic markers and 52 disciplinary terms. A second task then consisted of sorting these 263 terms according to their semantic content. For instance, *assume*, *hypothesis*, and *suppose* all share similar semantic content. This manual operation resulted in 79 "semantic fields" (noted in bold italic typeface): 61 "epistemic fields" and 18 "disciplinary fields." The 61 epistemic fields were further split into four main types depending on the different facets of knowledge they were found to highlight more specifically: 9 epistemic fields were associated with the process of discovery (e.g., ***method-approach***), 18 fields were characterized as justification (e.g., ***confirm***, ***plausible***), 25 fields concerned specific epistemic objects (e.g., ***model***, ***process***), and 9 fields were identified as epistemic virtues (e.g., ***accuracy***, ***explain***) (see table 1 for the list of semantic fields and their terms).

The second step consisted of measuring the presence of the 79 semantic fields across the entire corpus of 73,771 articles. The occurrence of each field in each document was calculated (by counting the number of terms related to each field) and averaged over all documents in the corpus.[3] This resulted in an overall occurrence map showing the relative significance of each field in the entire corpus. A subsequent analysis consisted of measuring semantic field occurrences at the paragraph level ($n = 29{,}942{,}151$ paragraphs) for each article and then calculating the Pearson correlations between semantic fields at the paragraph level.[4] The reason for moving from article level to paragraph level was to capture tighter—hence more meaningful—conjunctions between semantic fields. The resulting 79 × 79 correlation matrix was used to build a correlation network between semantic fields (limited to correlations above .1), and a Louvain community detection was carried out (Gephi implementation, with default hyperparameter values) to identify the most significant groups of semantic fields.

In a third step, the influence of disciplinary context was investigated by splitting the corpus into seven subcorpora according to article disciplinary clusters as defined by Malaterre and Léonard (2023) (see table 2). Similar analyses as those of step 2 were then carried out for each disciplinary cluster, resulting in seven semantic field correlation networks. In addition, original text excerpts were retrieved where pairs of

[3] The measure of the arithmetic means was done with custom Python code; see the Supplementary Information section.

[4] To check whether the sparsity of the semantic field × paragraph matrix affected the similarity measure, both Pearson correlation and cosine similarity were implemented. Because the results were extremely consistent with one another—especially given our objectives of understanding the most salient features of epistemic term usage in the corpus—we chose here to report the results in terms of Pearson correlation (all $ps < .001$), which is intuitively easier to interpret.

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

Table 1. The 79 Semantic Fields (Noted in Bold Italic Typeface) Sorted by Type and with Their Respective Terms

Types	Semantic fields	Terms (lemma)
disciplinary	***algorithm***	algorithm
	animal	animal, mouse, rat
	behavior	behavior, behaviour
	cancer	cancer, metastasis, tumor
	care	care, health, social
	cell	cell, stem
	clinical	clinical
	demographics	age, participant, patient, woman
	evolution	coevolution, evolution, evolutionary
	gene	code, DNA, gene, genetic, genome, miRNA, sequence
	human	human
	molecular	expression, molecular, pathway, regulate, regulation, regulatory, signal, site, substitution
	mortality-survival	mortality, survival
	pathology	disease, injury, pathogenesis, prognosis
	phylogenetics	tree
	protein	acid, amino, protein
	selection	selection
	therapy	resistance, treatment
discovery	***analysis***	analyse, analysis, analyze
	discover	discover, discovery, uncover
	experiment	experiment, experimental, experimentally, experimentation, trial
	interpretation	interpret, interpretation
	investigate	exploration, exploratory, explore, investigate, investigation
	method-approach	approach, method, methodological, methodology, protocol
	observation	detect, detectable, detection, observable, observation, observational, observe
	test	test
	underlie	underlie
justification	***alternative***	alternative
	analogy	analog, analogy, analogous, analogue
	certainty	certainty, proof, true, truth, well-established, well-known
	coherence	coherence, coherent, compatibility, compatible

(*Continued*)

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

Table 1. (*Continued*)

Types	Semantic fields	Terms (lemma)
	confirm	confirm, confirmation, confirmatory, prove, valid, validate, validation, verification, verify
	fit	fit
	inconsistency	conflict, contradict, contradiction, contradictory, incoherence, incoherent, inconsistency, inconsistent
	justification	justification, justify
	know	know, knowledge
	learn	learn, learning
	plausible	credibility, credible, plausibility, plausible, plausibly
	possible	possibility, possible, possibly, potential, potentially, putative
	probable	likelihood, likely, probable, probably
	refute	incorrect, refutation, refute, reject, rejection
	reproducibility	reproducibility, reproducible, replicability, replicable
	suggest	proposal, propose, suggest, suggestion, suggestive
	support	support, supportive
	uncertainty	doubt, doubtful, fallibility, fallible, questionable, suspicion, suspicious, uncertain, uncertainty, unclear
object	***axiom***	axiom
	cause	causal, causality, causation, cause, influence
	data	data, dataset, datum
	effect	effect
	equation	equation, formula
	estimate	estimate, estimation
	evidence	empirical, evidence, evidence-based
	framework	framework
	generalisation	generalisation, generalise, generalizability, generalizable, generalization, generalize
	hypothesis	assume, assumption, conjecture, hypothesis, hypothesise, hypothesize, hypothetical, scenario, suppose, supposition
	information	information
	law	law, lawful, lawfulness
	mathematics-statistics	correlate, correlation, Cox, distribution, error, linear, logistic, multivariate, probabilistic, probability, regression, statistic, statistical, value
	mechanism	mechanism, mechanistic
	model	model, modeling

(*Continued*)

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

Table 1. (*Continued*)

Types	Semantic fields	Terms (lemma)
	parameter-variable	parameter, variable
	pattern	pattern
	phenomena	phenomenon
	process	process
	property	property
	response	response
	role-function	function, role
	structure	structure
	theorem	theorem
	theory	theory
virtue	***accuracy***	accuracy, accurate
	clarify	clarification, clarify, elucidate, elucidation, explicate
	explain	explain, explanation, explanatory
	integration	integrate, integration, integrative
	predict	expect, expectation, forecast, foresee, predict, prediction, predictive, predictor
	simplicity	parsimonious, parsimony, simple, simplicity
	understand	understand, understanding
	unification	unification, unify
	usefulness	useful, usefulness, utility, valuable

targeted terms could be found to co-occur. These excerpts helped interpret the underlying context of semantic field correlations.

3. Occurrences of semantic fields across the entire corpus

The relative importance of the targeted semantic fields throughout the corpus can be estimated by examining the average frequency of corresponding groups of terms per article (fig. 2).

Disciplinary fields characterize the disciplinary orientations of the corpus (left-hand side of the graph in fig. 2, in beige). Unsurprisingly, they also match the disciplinary clusters of articles identified by Malaterre and Léonard (2023) (table 2). A dominant share of terms concerns research about cells, genes, and molecular biology in a broad sense, as denoted by the relative significance of the semantic fields ***cell***, ***gene***, ***protein***, and ***molecular*** (this result accords well with the presence of the article cluster CELL AND MOLECULAR BIOLOGY but also with the clusters GENETICS, DISEASE BIOLOGY, OTHERS and GENOMICS AND PHYLOGENETICS). A biomedical orientation is clearly visible

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

Table 2. The Seven Article Clusters (Noted in Small Caps Typeface) with Their Number of Articles and Their Topical Profiles

Cluster names	# docs	Top-5 topics
Bioinformatics and methods	5,113	Method-model, Network-model, Database-software, Genetic expression, Protein-domain
Cancer research	2,822	Cancer-tumor, Prognostic, Cell-oncology, Genetic expression, Genetic expression RNA
Cell and molecular biology	12,544	Cell signaling, Cell-oncology, Genetic pathway, Immunology, Cell-development
Genetics, disease biology, others	23,263	Enzyme-production, Database software, Method measurement, Hematology, Change-effect
Genomics and phylogenetics	6,504	Evolution and phylogenetics, Genetic sequence, Protein domain, Plant genetics and species, Population region
Health and clinical research	19,134	Clinical trials, Health care, Mental health, Demographics, Statistics
Public health	4,391	Health research and policy, Community health, Linguistic emphasis, Health care, Survey-report

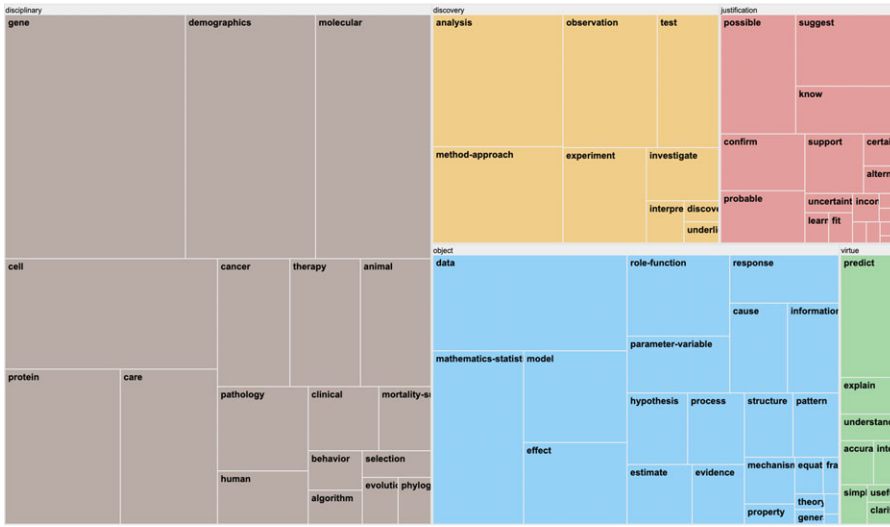

Figure 2. Average occurrence of the semantic fields per document over the complete corpus. (Semantic fields are grouped and colored by type. Surface area is proportional to the average number of occurrences of terms related to each semantic field; to give an idea of scale: ***suggest***, in the upper right-hand corner, has an average of 5.1 occurrences per article. Numerical values are available in the "Data_for_graphs" file; see Supplementary Information section.) Figure in color online.

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

through the strong presence of terms related to ***cancer***, ***therapy***, ***clinical***, and ***pathology***, as well as ***mortality-survival*** (which makes sense, given the presence of such article clusters as CANCER RESEARCH; GENETICS, DISEASE BIOLOGY, OTHERS; and HEALTH AND CLINICAL RESEARCH in the corpus). The semantic field ***demographics*** captures the presence of terms that relate to the description of patients, as is typically the case in clinical trials and public health studies; the fields ***care*** and ***behavior*** are also connected to health issues, notably mental health (these last three fields partly correspond to the clusters HEALTH AND CLINICAL RESEARCH and PUBLIC HEALTH). The fields ***animal*** and ***human*** denote research with animal models, as can be the case in cellular biology and cancer research (as in the clusters CANCER RESEARCH and CELL AND MOLECULAR BIOLOGY). The three fields ***evolution***, ***selection***, and ***phylogenetics*** characterize research in evolutionary biology and phylogenetics (as is distinctive of the cluster GENOMICS AND PHYLOGENETICS). A relatively less frequent disciplinary field is ***algorithm***, which denotes research in bioinformatics (this field relates to the article cluster BIOINFORMATICS AND METHODS).

The four types of epistemic fields are found on the right-hand side of the graph (fig. 2). Among the epistemic fields that relate to epistemic objects (in blue), several of the most significant fields denote the centrality of modeling in the practice of science. This is the case, of course, for the field ***model*** but also for ***data***, ***mathematics-statistics***, ***parameter-variable***, and ***estimate***. By contrast, theories and generalizations are quite rarely used, as shown by the reduced size of fields such as ***theory*** and ***generalization***. In between these extremes, there appears to be a moderately significant usage of hypothesis-related terms, as denoted by the medium-size field ***hypothesis***. The results also show quite a strong usage of causal notions, as revealed by the relative importance of such fields as ***cause***, ***effect***, and ***response***. Finally, specific targets of knowledge, such as mechanisms, structures, patterns, and processes, are also found in the corpus, as denoted by the presence of the corresponding epistemic fields ***mechanism***, ***structure***, ***pattern***, and ***process***.

The second-most-represented epistemic type concerns discovery-related epistemic fields (in orange, fig. 2). Terms that relate to analysis and observation are among the most frequent, as denoted by the significance of the corresponding epistemic fields ***analysis*** and ***observation***. Terms that concern experiments and tests are also quite significantly present throughout the corpus, as shown by the relative importance of ***experiment*** and ***test***. Note the relatively frequent mention of methods-related terms, as captured by the epistemic field ***method-approach***.

When it comes to modalities of justification (in red, fig. 2), one notes the dominance of weak forms of justification that consist of suggestions, possibilities, or probabilities, as denoted by the predominance of the corresponding semantic fields ***possible***, ***suggest***, ***probable***, and even ***alternative***. The second-most-significant group of justification-related epistemic fields includes somewhat stronger forms of justification, such as confirmation or corroboration, with the fields ***confirm*** and ***support***. On the other hand, negative forms of epistemic appraisal that would notably consist of refuting or establishing incompatibilities appear relatively rare, as shown by the relatively infrequent usage of terms that relate to ***inconsistency***, for instance. Similarly, very strong forms of justification—for instance, in terms of truth or certainty—are also rare, as shown notably by the reduced size of the field ***certainty***.

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

Finally, when it comes to epistemic virtues (in green, fig. 2), the dominant semantic field is ***predict***. This shows the importance of prediction in science, which is much more significant than other virtues, such as explanation or understanding (as characterized by the fields ***explain*** and ***understand***). Other virtues that are notably present, although less frequent, are accuracy and, to a lesser extent, simplicity, as denoted by their respective epistemic fields.

Note that several epistemic fields are absent from the graph, their presence in the corpus being too low (average occurrence ≤ 0.25 terms/article). Among epistemic objects, ***law*** is noticeably absent, showing that the notion of law (e.g., as in "law of nature") is clearly not mobilized in the biological and biomedical sciences of the corpus, despite the amount of conceptual work the notion has triggered in philosophy of biology, not to mention philosophy of science (for a review, see, e.g., Hamilton 2007). In terms of epistemic virtues, the field ***unification*** is also hardly present at all. This notion seems to play no effective role in the practice of science, not even in connection with explanation or understanding, which in particular seems to run contrary to unification-based accounts of explanation (e.g., Friedman 1974; Kitcher 1985). Several epistemic fields that relate to justification are also remarkably absent. This is the case for the field ***refute***, which indicates that the notion of refutation is hardly at work in the practice of science. In fact, the results tend to show that epistemic justification mostly happens in terms of possibility and probability: negative forms of justification are rarely mobilized, if at all. In short, science is more probabilistic and Bayesian than Popperian. The field ***coherence*** is also very rare across the corpus, which shows that the notion is not much used: the fact that hypotheses or models exhibit some form of coherence with other hypotheses or models and the rest of science appears to play little or no role in practice. Similar comments apply to the field ***analogy***. Despite extensive research by philosophers (for a review, see, e.g., Bartha 2019), the notion of analogy hardly seems to be mobilized, at least explicitly, in actual scientific articles. Finally, note that the field ***reproducibility*** is also barely present, if at all. Even if reproducibility and replicability should be desirable properties of scientific knowledge, the notions are not discussed in articles.

4. Correlations between semantic fields across the entire corpus

In order to analyze not just the quantitative significance of epistemic fields but also their relatedness with one another, correlation analyses were run over the whole corpus, resulting in the correlation network shown in fig. 3 (Here, semantic fields are not colored depending on their type, as was the case in fig. 2, but depending on the subnetworks they tend to form when frequently used in conjunction with one another in article paragraphs). As is immediately visible, the correlation network is dominated by three large subnetworks of semantic fields, with a small fourth one attached to the most central subnetwork.

The largest subnetwork (18 semantic fields, center position in fig. 3, in blue) includes several disciplinary fields that relate to molecular biology, genetics, and cellular biology (in decreasing order of occurrence: ***gene***, ***molecular***, ***cell***, and ***protein*** but also ***human*** and ***animal***). The epistemic field ***role-function*** is one of the fields that are most connected to these disciplinary fields, meaning that the identification of functions and roles is one major epistemic target for research on molecular processes

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

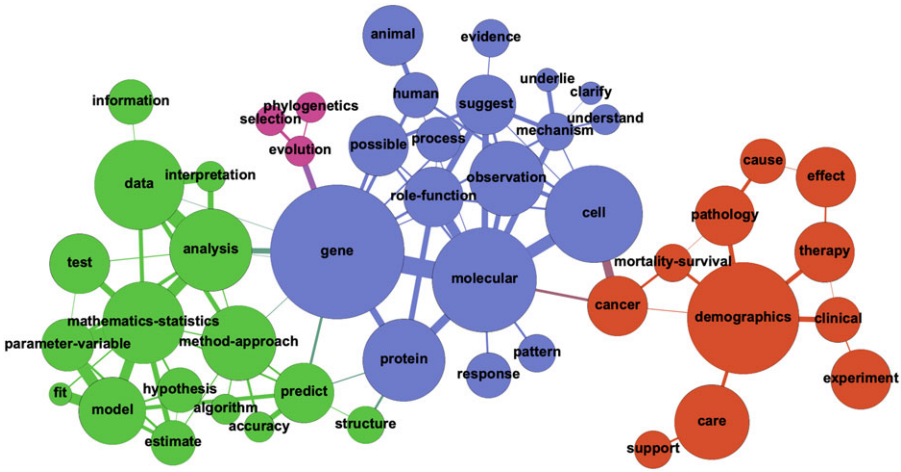

Figure 3. Correlation network of the most significant semantic fields over the whole corpus. (Correlations calculated at the paragraph level; for analysis, only correlations of >0.1 were kept; all *p*s < .001. Colors indicate correlation clusters based on Louvain community detection, node size is proportional to average occurrence in the corpus, and edge thickness is proportional to correlation strength; rendering was done with ForceAtlas2 on Gephi [Bastian et al. 2009]. To give an idea of scale: ***suggest***, in the upper-middle part of the network, has an average of 5.1 occurrences per article. Numerical values are available in the "Data_for_graphs" file; see Supplementary Information section.) Figure in color online.

or entities, notably proteins and genes (see text excerpts A10–18 in table S1, available in the Supplementary Information). ***Response*** and ***pattern*** are also frequently used in connection with ***molecular*** (A20–21). ***Role-function*** is correlated with ***process*** but also, and more significantly, with ***suggest***, which is in turn correlated with ***evidence*** and ***observation***; this tends to indicate that the role or function of molecular objects such as proteins is often, at most, suggested by observation and evidence (A30–34). This hypothetical or suggested epistemic status is also reinforced by the relatively frequent usage of ***possible*** (A35–38). Another significant epistemic object in this subnetwork—although only about half as significant as ***role-function***—is ***mechanism***. This field is strongly connected to ***molecular*** but also to ***gene*** and ***cell***, thereby corroborating the significant role that mechanisms play in molecular and cellular biology. Similarly to ***role-function***, ***mechanism*** is also significantly correlated with ***suggest***, indicating an epistemic status that is often taken to be hypothetical (A40–41). ***Mechanism*** is also frequently found in conjunction with ***underlie***, as well as with ***clarify*** and ***understand***. Mechanisms are thereby addressed by scientists as things that underlie a phenomenon of interest (e.g., an effect, a disease) and that are to be uncovered and clarified (A42–44). The correlation of ***mechanism*** with ***understand*** is ambivalent: it could mean that mechanisms provide some form of understanding, for instance, of molecular phenomena (as is defended by the new mechanicist philosophy). But it could be taken to mean that mechanisms are things that need to be understood. Most text excerpts by far corroborate the latter (A45–48). This is not to say that mechanisms never provide understanding but that they are usually rather thought of as objects that need to be understood. This finding attributes a quite different epistemic role

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

to the concept of mechanism than the one identified by the mechanicist philosophy (Machamer et al. 2000): mechanisms are rarely thought of as epistemic objects that have a form of explanatory epistemic virtue (understanding) but are more often conceived as research objectives, so to speak, in a context of discovery.

The second-largest subnetwork (in green in fig. 3, left-hand side) includes disciplinary fields that relate to data analysis and the elaboration of models: ***data***, ***analysis***, ***mathematics-statistics***, and ***model***. The field ***data*** plays a very central role (in conjunction with ***information***): it is strongly correlated with ***interpretation*** and ***analysis***, the latter being, in turn, correlated with ***mathematics-statistics*** (and with the disciplinary field ***gene***). This suggests that data, in practice, are often submitted to interpretation and analysis that are mediated by mathematical and statistical approaches (see excerpts B10–19). The epistemic field ***mathematics-statistics*** is also strongly correlated with the fields ***parameter-variable*** and ***model***, the latter also being correlated with ***fit***. This shows the importance of model fitting (B20–22), together with the strong role of statistics as the basis for both model building and model testing (as shown by the significant correlation of ***test*** with ***mathematics-statistics***) (B23–26). Interestingly, the notion of test is much more connected to the notion of model (through the use of statistical tests) than to the notions of hypothesis, theory, and experiment. But note the presence of ***hypothesis*** and its correlation with ***model*** and ***mathematics-statistics***: this is explained in particular by the practice of null-hypothesis testing in modeling (B27) but also by the testing of more general hypotheses, as would be expected in science (B28–29). The epistemic function of models is clearly one of prediction, as shown by the strong correlation between ***model*** and ***predict*** (B30–33). In this respect, accuracy is the most sought-after epistemic virtue (strong correlation between ***predict*** and ***accuracy***) (B34–36). In the disciplinary context of the present corpus, prediction notably concerns the structure of molecular entities such as proteins and others (correlation of ***predict*** with ***structure*** and ***protein***) (B37–39). On the contrary, note the absence of a significant correlation between ***model*** and ***explain*** or ***understand***. This means that neither explanation nor understanding is found to be a desirable characteristic of models, a finding that runs somewhat contrary to theses defending the explanatory role of models in science (e.g., Strevens 2008; Potochnik 2017).

The third largest subnetwork (in red in fig. 3, right-hand side) has to do with the health sciences as broadly construed. Note the central position of ***demographics*** in this cluster, which relates to the frequent usage of terms describing populations of patients. This field is most strongly correlated with ***pathology***, ***therapy***, and ***clinical***, which denotes research on public health and epidemiology, as well as clinical studies (C10–15). It is also correlated with ***care*** and ***support***, denoting research on health support and health care (C16–19) (note that *support* is not used here in the sense of epistemic justification). ***Demographics*** is correlated with ***mortality-survival*** and ***cancer***, too, denoting the presence of cancer research in the corpus (C20–23). This latter field sits at the border between the health sciences subnetwork and the molecular biology subnetwork, in particular because of its strong correlation with ***cell*** and, to a lesser extent, with ***molecular***. From an epistemic viewpoint, causal notions appear to be relatively frequent in this health sciences subnetwork, and possibly one of its epistemic characteristics, as shown by the size of ***cause*** and ***effect***, notably in conjunction with the disciplinary fields ***pathology*** and ***therapy*** (C30–33). Interestingly,

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

the field ***experiment*** is frequently used in this disciplinary context (correlation with ***clinical***) and not in the molecular biology context (as characterized by the central subnetwork, in blue in fig. 3). Text excerpts indicate that terminological usage often relates to experiments on animal models and comparison with clinical data (C40–41).

Finally, the fourth subnetwork (in pink in fig. 3) denotes research in evolutionary biology and phylogenetics. This subnetwork comprises only three disciplinary fields (***evolution***, ***selection***, and ***phylogenetics***) and no epistemic field at this level of analysis. Although conjunctions of terms pertaining to evolution and selection usually tend to characterize research on Darwinian evolution (D10–11), conjunctions of terms related to evolution and phylogenetics appear to specifically pick out taxonomical works (D12–13). In any case, most of these studies are gene based (hence the connection of this subnetwork to the field ***gene*** of the molecular biology subnetwork).

Note that the overall network comprises 48 of the 77 targeted semantic fields, meaning that 29 fields were left out. This can be explained by their relatively low level of occurrence in corpus articles and by insufficient correlation with other fields. In particular, missing fields of the type object include ***property***, ***generalization***, ***law***, and ***theory***. This indicates that corresponding epistemic objects are actually rarely mobilized in scientific practice, despite being frequently discussed in philosophy of science (e.g., Nagel 1961; Carnap 1966; Suppe 1989; Savage 1990; Hamilton 2007). Noticeable justification fields that are absent are ***analogy*** and ***coherence***, as well as ***certainty*** and ***confirm*** (as noted in sec. 3). Although it could be expected that certainty and truth would not be much present (being too strong for the biological and biomedical sciences), the results show that confirmation-related concepts are also very little used in the practice of science, which can appear somewhat counterintuitive. Finally, noticeable missing fields related to epistemic virtues include ***explain***, ***unification***, ***integration***, and ***simplicity***. Explanation therefore does not appear to be a much-discussed topic during the course of scientific investigation (nor does understanding, as discussed earlier). Similarly, unification, integration, and simplicity do not play a significant role. This contrasts quite starkly with the attention these concepts have attracted from numerous philosophers, notably since the 1950s (e.g., Hempel 1965; Friedman 1974; Kitcher 1985; Mitchell 2003; Salmon 1989; Sober 2015).

5. Effect of disciplinary context on semantic fields

To investigate whether the usage of epistemic concepts depended on disciplinary context, we made use of the relative diversity of the articles found in the corpus. Although all articles stem from the biological and biomedical sciences, broadly speaking, there are notable differences in terms of subdisciplines, ranging from molecular biology to phylogenetics and clinical research. To capture these differences, we used the article clustering conducted by Malaterre and Léonard (2023) to build semantic field correlation networks for all seven disciplinary clusters (see fig. 4). As could be expected, the relative significance of disciplinary fields reflects the disciplinary orientations of the article clusters (as depicted by their topical profiles; see table 2; for instance, as shown in fig. 4b, ***cancer*** is a dominant field in the correlation network stemming from the cluster CANCER RESEARCH). In what follows, we mostly leave aside disciplinary fields and focus on epistemic fields. The networks partly display features that were already apparent at the level of the entire corpus but

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

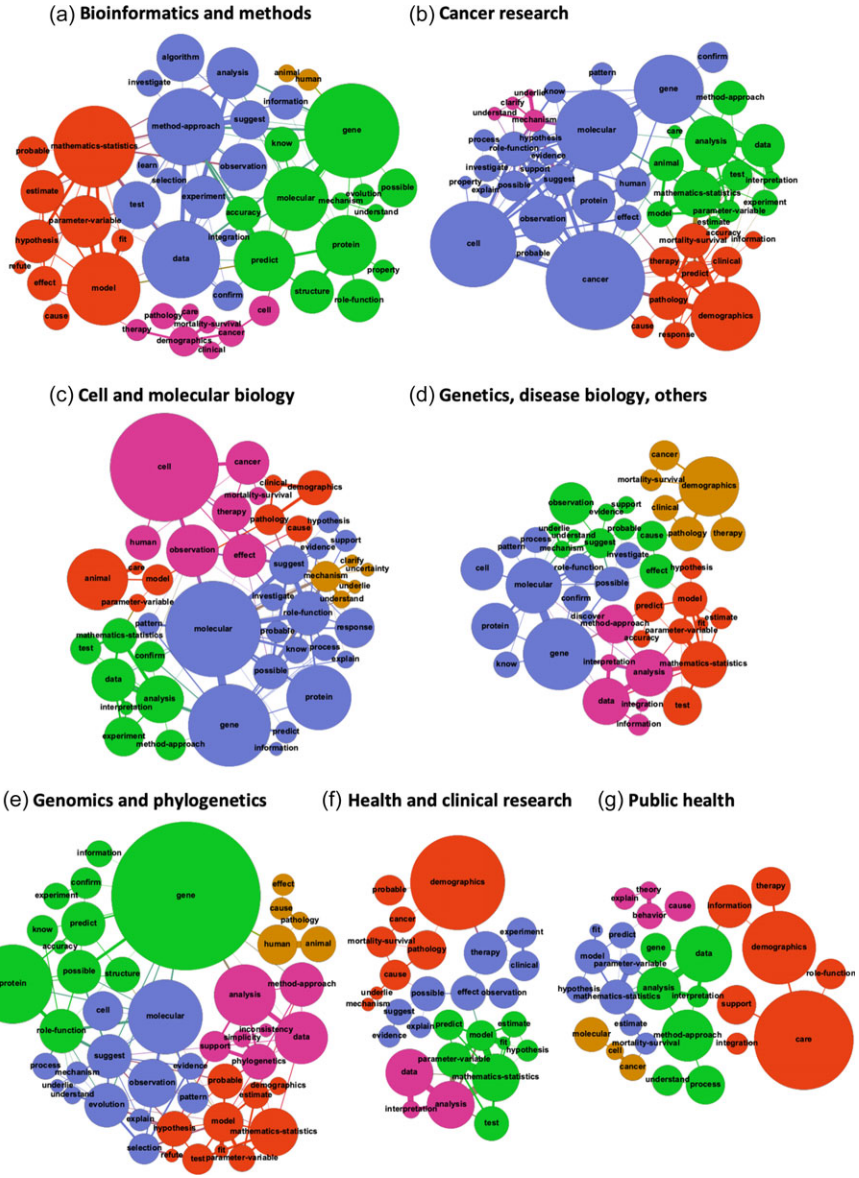

Figure 4. Correlation network of the most significant semantic fields for the seven disciplinary clusters. (Correlations calculated at the paragraph level; for analysis, only correlations > 0.1 were kept; all *p*s < .001. Node size is proportional to average occurrence in the cluster; edge thickness is proportional to correlation strength. Subnetwork identification is based on Louvain community detection; subnetwork colors depend on the size of the network, from the largest subnetwork in blue to the smallest subnetwork in brown. Rendering was done with ForceAtlas on Gephi. To give an idea of scale: ***model*** has an average of 25.6 occurrences per article in network (a) and 6.1 in network (g). Numerical values are available in the "Data_for_graphs" file; see Supplementary Information section.) Figure in color online.

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

also—and this is what matters here—specificities that depend on clusters' disciplinary orientations and their topical profiles.

Consider the correlation network for the cluster BIOINFORMATICS AND METHODS (fig. 4a): the fields ***model***, ***fit***, ***parameter***, and ***mathematics-statistics*** are all strongly correlated, showing the importance of statistical model fitting and testing (CA10–11). ***Refute***, which was absent from the corpus-level network, is here seen in conjunction with ***hypothesis***: refutation is therefore sometimes at work in science, although not very frequently, and in relation to hypotheses, notably null hypotheses, but not theories (CA12–13). Note also the usage of causal notions such as ***effect*** and ***cause*** in connection with ***hypothesis***, typically denoting that hypotheses are sometimes about observed effects and possible causes (CA14–15). Note the correlation threads between ***model***, ***predict***, ***structure***, and ***protein***: as noted earlier, models are often used to predict, and such prediction often targets the structure of molecular entities such as proteins (CA20–21). A key virtue of prediction clearly is accuracy, as denoted by the strong correlation between ***predict*** and ***accuracy*** (CA22–23), with such accuracy resulting from the use of methods (correlation with ***method-approach***) that are specifically developed for improving predictive accuracy (CA24–25). Confirmation tends to apply to prediction (correlation of ***confirm*** and ***predict***) (CA26–27, although sometimes confirmation concerns prediction methods, e.g., CA28). The relatively strong correlations between ***data***, ***method-approach***, and ***analysis*** indicate the importance of data analysis, and additional correlations with ***selection*** reveal the use of feature- and variable-selection methods and data-reduction techniques (CA30–32). Note the epistemic field ***integration*** and its correlation with ***data***, which captures the practical issue of integrating data from multiple sources (CA33–34) rather than theoretical integration (as is the case in, e.g., Mitchell [2003]). The field ***mechanism*** exhibits some correlation with ***understand*** and ***molecular***, but its role turns out to be modest (low frequency and connectivity). As noted earlier, these correlations usually do not denote any explanatory role of mechanisms but rather point to mechanisms as objects in need of understanding (CA40, but see CA41). Both ***role-function*** and ***property*** appear correlated to ***protein*** (in addition to ***structure***, as mentioned earlier), meaning that proteins are the focus of much research on their characteristics (CA50–52).

The correlation network of the cluster CANCER RESEARCH (fig. 4b) exhibits strong correlations between ***model***, ***mathematics-statistics***, and ***parameter***, as in the network of BIOINFORMATICS AND METHODS, denoting the importance of statistical model building (CB10–11). One notable specificity is the correlation of ***model*** with ***animal***, which indicates the frequent use of animal models in this domain of research (CB12–13), even when statistics are involved (CB14–15). As also noted earlier, the main epistemic virtue of models is prediction, with accuracy if possible (as shown by the correlation between ***model***, ***predict***, and ***accuracy***). Here, prediction is directed toward a specific objective: predicting survival or mortality (correlation with ***mortality-survival***) (CB16–17). The fields ***data***, ***method-approach***, ***analysis***, ***mathematics-statistics***, and ***interpretation*** are again all connected, showing that data are often statistically analyzed and interpreted (CA18–19). Defining a subnetwork of its own, ***mechanism*** is found to be mostly correlated to ***understand***, ***underlie***, and ***clarify***; these correlations actually denote that mechanisms are things to be understood, not things that provide understanding (CB20–22). In addition, discussions often concern how-possibly mechanisms, as shown by correlations with ***possible*** and ***suggest*** (CB23), and the

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

elucidation of related roles or functions (correlation with ***role-function***) (CB24). Note the correlations between ***support***, ***evidence***, and ***hypothesis***, highlighting the role of evidence in providing epistemic support, notably to hypotheses (CB30). Interestingly, ***explain*** is found to be correlated with ***possible***: often, explanations of specific phenomena are not known but offered as possible (CB31–35). This indicates the key role of how-possibly explanations, over and above how-actual explanations. There is also a correlation between ***confirm*** and ***gene***: confirmation is here predominantly tied to the presence or role of specific genes (CB36–38), as opposed to epistemic objects such as hypotheses, models, or theories (as could have been expected).

In the CELL AND MOLECULAR BIOLOGY network (fig. 4c), the fields ***model*** and ***animal*** are strongly correlated (as in the cluster CANCER RESEARCH), denoting the prevalence of models in the sense of animal models (CC10–13). Note the absence of correlation with ***predict*** in this disciplinary context; this corroborates the presence of a different type of model concept in this article cluster compared to BIOINFORMATICS AND METHODS, where models are used for prediction. This also makes sense given the correlation of ***model*** with ***pathology***, which notably captures research on animal models of specific diseases (CC14–15). As in other clusters, ***data***, ***analysis***, ***mathematics-statistics***, and ***interpretation*** are all correlated. A specificity is the correlation with ***confirm***. Data are often said to confirm other epistemic objects, for instance, other studies or other results (CC20–21), although the reverse can also be the case: sometimes results or findings are said to confirm data (CC22). As seen in other clusters, ***mechanism*** turns out to be correlated with ***underlie***, ***understand***, and ***clarify***. It is also correlated with ***role-function***, which may denote different use contexts: for instance, a system can have a role that is made possible by an underlying mechanism (CC30); elsewhere, some entity may have a role within a mechanism (CC31). In any case, much uncertainty is found in connection with mechanisms and associated roles and functions, as shown by correlations with such fields as ***possible***, ***probable***, ***uncertainty***, and ***suggest*** (CC32–35). Support can be directed toward hypotheses (correlation of ***support*** with ***hypothesis***; CC40) or toward mechanisms (correlation with ***mechanism***; CC41). Note again the correlation between ***explain*** and ***possible***; this denotes much of the uncertainty that surrounds explanations, the latter often being formulated as how-possibly explanations only (CC51–52).

As shown in the network for BIOINFORMATICS AND METHODS, the fields ***model***, ***fit***, ***parameter***, and ***mathematics-statistics*** exhibit significant correlation in the network for GENETICS, DISEASE BIOLOGY, OTHERS (fig. 4d). This shows the importance of statistical model fitting in this disciplinary context (CD10–12). The correlation of ***model*** with ***hypothesis*** maps with the confrontation of hypotheses (notably, null hypotheses) in the context of model building (CD13–14). Further correlation of ***mathematics-statistics*** with ***test*** reveals the ubiquity of statistical tests, for instance, to identify significant variables in datasets (CD15–16), sometimes associated with specific statistical models (CD17). Prediction accuracy is significant in this context, as shown earlier in the network for BIOINFORMATICS AND METHODS (correlation of ***predict*** and ***accuracy***) (CD18–19). Note that accuracy is also often discussed in association with specific methods, with some methods being known to provide more or less accuracy than others (correlation of ***accuracy*** and ***method-approach***) (CD20–21). The fields ***method-approach***, ***data***, ***analysis***, and ***mathematics-statistics*** exhibit tight correlations, notably showing the importance of data analysis by means of different statistical approaches

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

(CD22–23). The correlation of ***data*** with ***integration*** reveals the practical significance of multiple-source data integration, notably to generate and test novel hypotheses on the basis of varied sets of data (CD24–27). As shown in the network for BIOINFORMATICS AND METHODS, this specific epistemic role for integration concerns data and not models, as might have been expected. Note the presence of much uncertainty in research, with multiple correlations of the fields ***suggest***, ***possible***, and ***probable*** with such fields as ***mechanism*** or ***role-function***, indicating numerous discussions about how-possibly mechanisms and how-possibly roles or functions (CD30–33). As seen in other disciplinary networks, ***mechanism*** is also correlated with both ***underlie*** and ***understand***, typically in the sense that mechanisms that underlie phenomena of interest need to be (better) understood (CD34–35). Note also the use of causal notions such as cause and effect, mostly in connection with research on specific diseases (correlation of ***cause*** and ***pathology***) or modeling (correlation of ***effect*** and ***model***): diseases are often said to be caused by specific agents, pathogens, viruses, and so forth (CD40–41), whereas models can be used to study the effects of certain variables (CD42–43).

In GENOMICS AND PHYLOGENETICS (fig. 4e), the fields ***model***, ***parameter-variable***, ***mathematics-statistics***, and ***fit*** are correlated with one another, as seen also in BIOINFORMATICS AND METHODS and GENETICS, DISEASE BIOLOGY, OTHERS, highlighting here, too, the significance of (statistical) model fitting (CE10). ***Hypothesis*** and ***test*** are also correlated to ***model***, whereas ***refute*** is connected with ***hypothesis***: both models and hypotheses can be tested (CE11–12); sometimes hypotheses are tested with models (CE13), and at other times, hypotheses are refuted (CE14–15; note here a relatively rare occurrence of refutation). A peculiarity is the lack of a significant correlation of ***model*** with ***predict*** (as encountered in CELL AND MOLECULAR BIOLOGY). The major epistemic role of models here is to be fitted to (phylogenetic) data, not to make predictions. Prediction is more frequently used in the context of molecular structures (correlation between ***prediction***, ***protein***, ***structure***, and ***gene***) (CE20–21). Prediction accuracy is still a major concern and is affected by the use of specific methods (correlation between ***predict*** and ***accuracy***, and with ***method-approach***) (CE22–23). As can be expected, predictions can get confirmed (CE24), and experiments may play a confirmatory role (CE25) (correlations of ***predict*** with ***confirm*** and correlations of ***confirm*** with ***experiment***). Note here again that much uncertainty surrounds mechanisms and functions (CE30–31) (correlations among ***suggest***, ***possible***, ***role-function***, and ***mechanism***) and that mechanisms are epistemic objects in need of being understood (CE32–33) (correlations of ***mechanism*** with ***underlie*** and ***understand***). Interestingly, ***mechanism*** is here correlated with ***evolution***, which denotes discussions about specific mechanisms of evolution (CE34–36) (not about whether evolution itself should be considered a mechanism, as in, e.g., Barros [2008]). As for the correlation between ***explain*** and ***hypothesis***, it tends to capture instances of how-possibly explanations (CE37–39). ***Data*** is again correlated with ***analysis*** and ***method-approach***, as well as with ***phylogenetics*** (disciplinary field). In this specific context, correlation with ***simplicity*** denotes the significant use of the notion of simplicity/parsimony as an epistemic value to sort explanations, interpretations, and hypotheses (CE40–41), but also the use of specific methods that make use of parsimony, such as statistical parsimony analyses and networks (CE42–43). Also note the correlation of ***phylogenetics*** with ***support*** and ***inconsistency***, for which this disciplinary context is a rare case of

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

occurrence. This reveals the complementary use of supporting and contradicting judgments (CE44–46). In other words, justification is not just a matter of corroboration but also of noncontradiction of accepted hypotheses (or of contradiction of rival hypotheses). Finally, causal notions appear to be connected to disease research (which is still somehow present in this cluster, likely through genomics studies), as denoted by correlations between ***pathology***, ***effect***, and ***cause***.

The fields ***model***, ***parameter-variable***, ***mathematics-statistics***, and ***fit*** are correlated in the network for HEALTH AND CLINICAL RESEARCH, as seen elsewhere (fig. 4f). Modeling appears to play a significant prediction-related role (CF10–11) (correlation with ***predict***). Here, too, ***analysis***, ***data***, and ***interpretation*** are strongly correlated, showing the importance of data analysis (CF20). Uncertainty is also much present, as indicated by correlations between ***evidence*** and ***suggest***, as well as ***explain*** and ***possible***: evidence often only has a suggestive status (as opposed to being corroborative) (CF30–32), whereas explanation is mentioned in the role of how-possibly explanation (CF33–34). Causal notions are again seen very much in connection with disease and pathology (correlation of ***cause*** with ***pathology*** and ***mortality-survival***, correlation of ***effect*** with ***therapy***) (CF40–43). ***Mechanism*** plays a minor role in this disciplinary context (as opposed to the central role it had in CANCER RESEARCH; CELL AND MOLECULAR BIOLOGY; or GENETICS, DISEASE BIOLOGY, OTHERS) and is only connected to ***underlie***. Experiments appear to be most often conducted in a clinical setting (as denoted by the correlation between ***experiment*** and ***clinical***) (CF50).

The network for PUBLIC HEALTH (fig. 4g) bears some similarities with the network for HEALTH AND CLINICAL RESEARCH. The fields ***model***, ***parameter-variable***, ***mathematics-statistics***, and ***fit*** are all correlated, with correlation also to ***predict***, highlighting the predictive role of statistical models (CG10). In this cluster, behavior is one particular target of prediction, as well as what triggers the search for causes, notably causes that may influence specific behavior patterns (CG20–22) (correlations between ***prediction***, ***cause***, and ***behavior***). Yet, and most interestingly in comparison to other clusters, social psychology and behavior research appear to rely on the formulation of theories, which in turn play an explanatory role (correlation of ***behavior***, ***theory***, and ***explain***). Mentioned theories include the theory of reasoned action, the theory of planned behavior, and other diverse social cognitive theories (CG23–25), which are all frequently endowed with explanatory power (CG26–28). ***Data***, ***analysis***, ***method-approach***, and ***interpretation*** are all correlated as well, as seen elsewhere, denoting the importance of data analysis and the role of specific methods in this respect (CG30–31). Notions of support, integration, and role or function tend, in this cluster, to be used in a nonepistemic sense in connection with the notion of care—for instance, to discuss whether specific organizational support may favor the integration of different types of health-care services (CG40–43) (correlation of ***care*** with ***support***, ***integration***, and ***role-function***). Finally, note the correlation between ***process*** and ***understand***, which sometimes denotes research on the process of understanding (CG50) but more frequently on biological processes as objects that need to be understood (CG51), quite similar to the way mechanisms are also taken as targets for understanding in other clusters, such as CELL AND MOLECULAR BIOLOGY; CANCER RESEARCH; and GENETICS, DISEASE BIOLOGY, OTHERS.

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

6. Discussion and concluding remarks

Word patterns matter because they denote the presence of specific concepts of interest and the relationships these concepts have with one another. When the focus is on epistemic terms, as in this research, investigating terminological usage in scientific articles reveals the place that key epistemic concepts actually occupy in the practice of science and sheds light on the intricate connections these concepts have with one another, depending also on disciplinary context. This should be most relevant to our own work as philosophers of science. Of course, one may argue that the way philosophers talk about epistemic concepts will likely be different from the way scientists talk about these same concepts, even more so if philosophers engage in some form of normative conceptual reengineering. Yet if we think of our discipline as capturing the way knowledge is actually built in the practice of science, then philosophical reconstructions of key epistemic terms should be at least consistent with how these concepts are used in scientific publications. Although the study undertaken by Malaterre and Léonard (2023) had already noted the importance of models, data, and prediction; the secondary role of explanations; and the ambivalent correlation between mechanisms and understanding, that study was nevertheless limited to six epistemic notions and their most frequent co-occurring terms. Here, the research perimeter has been significantly enriched to encompass 61 epistemic fields (understood as sets of terms with shared semantic content) covering a vast number of epistemic objects and virtues, as well as concepts related to knowledge discovery and justification.

Among the main findings, we noted here, too, the significance of models, data, and prediction, corroborating the view that models are central epistemic objects in science (Morgan and Morrison 1999; Weisberg 2012; Frigg and Nguyen 2020). Models are found across all the biological and biomedical disciplines represented in the corpus, although perhaps somewhat less frequently in cancer research and cellular and molecular biology (where the notion of animal models tends to predominate (Leonelli and Ankeny 2012; Levy and Currie 2015) and where much effort is also deployed to identify mechanisms and the roles or functions of numerous molecular entities). The significance of models goes together with the central role played by data in day-to-day science: data are ubiquitous, as are mathematical and statistical notions of equations, parameters, variables, estimates, and so forth. In this respect, a critical activity that scientists pursue is the fitting of models to data: models are not just conceived or elaborated on the basis of observations but are finely tuned to fit the data. How this is done and how this affects the formation of knowledge deserve further investigation, as do the complex relationships that models and data entertain with one another (Krohs and Callebaut 2007; Leonelli 2019; Bokulich 2021). Another striking point is the question of data integration, which surfaces in different disciplinary contexts and is also worthy of further analysis.

Interestingly, models seem to play hardly any explanatory role at all, as our findings did not highlight any significant role played by such epistemic virtues as explanation or understanding. This contrasts quite starkly with the amount of work these concepts have triggered in philosophy of science (e.g., Strevens 2008; Potochnik 2017). The epistemic virtue that dominates—and by far—the epistemic landscape is that of prediction, much significantly as exhibited by models (especially in

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

bioinformatics and in genomics and phylogenetics) and in connection with accuracy (which triggers, in turn, specific methodological research, notably in bioinformatics). Parsimony is sometimes mobilized as well, although quite rarely, and in specific disciplinary contexts, such as bioinformatics and phylogenetics, and it would be interesting to investigate why. This finding is consistent with the relatively more frequent use of simplicity when referring to models in ecology and evolutionary biology compared to other domains of biology, as reported by Mizrahi (2022).[5] Somewhat surprising is the context in which explanation is mobilized: explanations are not offered as answers (notably, to why-questions) but as hypotheses. This is evidenced by the frequent use of conjoint terms such as *possible*, *probable*, and *suggest*. Overton interpreted the relatively high frequency of explanation-related terms (compared to nonscientific corpora) as demonstrating the significance of explanation as a goal of science (Overton 2013). Our findings relativize this interpretation: explanation in the practice of science relates much more to the notion of how-possibly explanation than to the notion of how-actual explanation.[6] This situation tends to lend support to research on how-possibly explanations (Dray 1968; Resnik 1991; Forber 2010; Bokulich 2014; Verreault-Julien 2019) while clearly changing the emphasis on the epistemic work that the concept of explanation actually plays in science, especially compared to earlier accounts of scientific explanation (e.g., Hempel 1965; Friedman 1974; Kitcher 1985; Salmon 1989).

A similar shift operates with the notion of mechanism. Whereas the new mechanistic philosophy considers mechanisms to be the most central epistemic tools for elaborating explanations in much of (molecular) biology (Machamer et al. 2000; Bechtel and Abrahamsen 2005; Craver 2007; Glennan 2017), our findings indicate that mechanisms are most often unknown objects: far from providing understanding, mechanisms require to be understood. This is most notable in molecular and cellular biology but also in phylogenetics (where there is much discussion of possible mechanisms of evolution). Our findings also indicate that mechanisms are often used in connection with specific causal notions, such as effect, role, and function, but not cause, which, in contrast, turns out to be used more frequently in domains such as disease and health research (note that in these contexts, the notion of mechanism tends to be absent, whereas process and evidence are more frequent). Such results raise questions about possible reasons for what appears to be quite specific usage patterns of some causal notions being preferable to others in given contexts.

A striking result is the very modest role played by the concept of theory. The concept is virtually absent from all of biology. This situation should encourage us to reconsider the importance given to the philosophical accounts of theory in biology,

[5] Mizrahi measured the percentage of articles in which an epistemic virtue such as simplicity would be found at least once together with theory, model, or hypothesis within a 10 + 10 word window, and for nine disciplinary subjects provided by the JSTOR Data for Research interface (Mizrahi 2022). In the present work, we measured the correlation between semantic field cooccurrence frequencies at the paragraph level.

[6] Overton had noticed that explanation was very often "used with care," as revealed by markers of possibility in retrieved groups of five terms containing *explain* (Overton 2013, 1390). Here, correlation measures make it possible to systematically identify significant relationships of semantic fields with one another, notably of ***explain*** with ***possible***, ***probable***, or ***suggest***, thereby clarifying the semantic context in which explanation and other concepts are used.

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

possibly reorienting philosophical work to the activity of theorizing (Callebaut 2013) or simply to the use of models in biology. On the other hand, the notion of theory clearly plays a role in the article cluster PUBLIC HEALTH and most notably in social cognitive research, as shown by text excerpts. The cluster exhibits an average of 1.3 occurrences of the field ***theory*** per article compared to about 0.1 for the biological clusters. This finding is consistent with those of Mizrahi, who found that psychology was a domain of science in which theory was the most used, second only to physics (Mizrahi 2022). This relatively high usage of theory also concurs with the findings of McPhetres and colleagues, who estimated that a large majority of articles in psychological research refer to theories (McPhetres et al. 2021).[7] Understanding why this is so requires further investigations: Are there specific reasons for using the concept of theory in these disciplines and not elsewhere? Are other disciplines concerned? Is the concept construed in particular ways? Interestingly, when the concept is used, correlations with ***explanation*** (complemented by text excerpts) show that it can play an explanatory role. Simply put, theories explain, which seems to corroborate Hempel's early intuitions about explanation (Hempel 1965).

Notions of confirmation and support are present more or less everywhere, as denoted by the ubiquitous presence of related fields. Yet, we highlighted the relative dominance of weak forms of justification, which consist of claims of possibility, suggestions, or identification of alternatives. This confers to the bulk of the scientific discourse a caution that is quite far from categorical assertions, be they of truth or refutation. Concerning discovery-related concepts, the notion of observation appears specific to the more biology-oriented domains; it is relatively absent in clinical and public health research. This is somewhat surprising because one would be inclined to view these domains as prone to observation (as opposed to experiment, for instance). As for experiment and test, their usage is relatively frequent in cellular and molecular biology but also in bioinformatics and methods-related research. Whether the concepts are used in similar ways or not also needs further analysis.

Several epistemic concepts are noticeably absent. The results show, for instance, that analogy, unification, and coherence are little used. It would be interesting to investigate further how such findings may be reconciled with the numerous philosophical works that have defended the role of these notions in science (e.g., Kitcher 1985; Thagard 2007; Bartha 2019). Similarly, notions of refutation and reproducibility are also quite rarely discussed, if at all. Note that the relatively low frequency of these different semantic fields does not necessarily mean that the corresponding concepts are not important in the practice of science, only that they are not explicitly mentioned in research articles. Indeed, it is possible that concepts such as unification, coherence, and reproducibility might be expressed with various complex sets of terms not easily recognizable as such. It is also

[7] According to McPhetres et al. (2021), 53.66% of articles in *Psychological Science* (2009–2019 mention at least one of 359 different named theories, whereas 15.33% explicitly claim to test predictions derived from theories. The authors interpreted the results as suggesting that most research published in this journal was not driven by theory. In comparison, our findings indicate that theory plays an even smaller role in other domains of the biological and biomedical sciences but point at domains of psychological research as exceptions where the notion of theory plays a significant role (note that the two studies do not share the same corpus; comparisons are only indicative).

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

possible that these concepts might be less discussed in research articles but more in review articles, perspectives, introductions to textbooks, or other scientific texts in which a certain degree of self-reflexivity can be found.[8] This is something that future work should investigate, comparing text-mining analyses depending on the type of scientific publication. Future work should also include iterative improvements to the list of epistemic markers we have already identified (see table 1) because these markers should be of interest to any future corpus-based approach to epistemic concepts.

Of course, another major and obvious avenue for future work would be to investigate scientific articles from a different and/or broader range of disciplines (notably, among the physical sciences), possibly covering other periods. Such research would offer additional perspectives on the usage of epistemic concepts, notably outside of the disciplines targeted here. It goes without saying that the analyses enabled by computational text-mining methods do not, and cannot, replace the precise and carefully crafted insights of conceptual explication enabled by close reading and the reconstruction of specific case studies. What these analyses make possible is to gain a broader quantitative view of the scientific discourse, depicting in somewhat broad strokes the epistemic landscape that scientists navigate, thereby providing heuristics for more focused investigations. They also establish an empirical basis for claims that may otherwise remain informal. Here, we hope to have shown that the usage of epistemic concepts in at least some parts of science does not always follow our intuitions but gives rise to a complex fabric of conceptual interdependencies worth investigating further.

Supplementary Information. A technical appendix with code and data, including table S1 and the "Data_for_graphs" file, is available on Zenodo.org (https://doi.org/10.5281/zenodo.8066574).

Acknowledgments. The authors are grateful to BioMed Central for providing access to journal articles for text-mining purposes. The authors thank the audience of the 2022 Society for Personality and Social Psychology (SPSP) Congress for comments on an earlier version of the manuscript, as well as two anonymous reviewers for *Philosophy of Science*. C. M. acknowledges funding from the Canada Foundation for Innovation (Grant 34555) and Canada Research Chairs (CRC-950-230795). M. L. acknowledges funding from the Canada Research Chair in Philosophy of the Life Sciences at UQAM.

References

Aggarwal, Charu C. 2015. *Data Mining*. New York: Springer. https://doi.org/10.1007/978-3-319-14142-8

Barros, D. Benjamin. 2008. "Natural Selection as a Mechanism." *Philosophy of Science* 75 (3):306–22.

Bartha, Paul. 2019. "Analogy and Analogical Reasoning." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University Press. https://plato.stanford.edu/archives/spr2019/entries/reasoning-analogy/

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, edited by Sharon Givon and Victor Lavrenko.

Bechtel, William, and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36 (2):421–41. https://doi.org/10.1016/j.shpsc.2005.03.010

Bokulich, Alisa. 2014. "How the Tiger Bush Got Its Stripes: 'How Possibly' vs. 'How Actually' Model Explanations." *Monist* 97 (3):321–38. https://doi.org/10.5840/monist201497321

[8] We thank an anonymous reviewer for this suggestion.

Bokulich, Alisa. 2021. "Using Models to Correct Data: Paleodiversity and the Fossil Record." *Synthese* 198 (6):5919–40. https://doi.org/10.1007/s11229-018-1820-x

Callebaut, Werner. 2013. "Naturalizing Theorizing: Beyond a Theory of Biological Theories." *Biological Theory* 7 (4):413–29. https://doi.org/10.1007/s13752-013-0122-2

Cappelen, Herman. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Carnap, Rudolf. 1966. *Philosophical Foundations of Physics*. New York: Basic Books.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Dray, William Herbert. 1968. "On Explaining How-Possibly." *Monist* 52 (3):390–407. https://doi.org/10.5840/monist196852339

Firth, John Rupert. 1957. "A Synopsis of Linguistic Theory 1930–1955." In *Studies in Linguistic Analysis, 1–32*. Oxford: Blackwell.

Forber, Patrick. 2010. "Confirmation and Explaining How Possible." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 41 (1):32–40. https://doi.org/10.1016/j.shpsc.2009.12.006

Friedman, Michael. 1974. "Explanation and Scientific Understanding." *Journal of Philosophy* 71 (1):5. https://doi.org/10.2307/2024924

Frigg, Roman, and James Nguyen. 2020. *Modelling Nature: An Opinionated Introduction to Scientific Representation*. New York: Springer. https://link.springer.com/book/10.1007/978-3-030-45153-0

Glennan, Stuart. 2017. *The New Mechanical Philosophy*. Oxford: Oxford University Press.

Grene, Marjorie and David Depew. 2004. *The Philosophy of Biology: An Episodic History*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511819018

Hamilton, Andrew. 2007. "Laws of Biology, Laws of Nature: Problems and (Dis)Solutions." *Philosophy Compass* 2 (3):592–610. https://doi.org/10.1111/j.1747-9991.2007.00087.x

Hempel, Carl G. 1965. *Aspects of Scientific Explanation*. New York: Free Press.

Hull, David L. 1974. *Philosophy of Biological Science*. Englewood Cliffs, NJ: Prentice-Hall.

Justus, James. 2012. "Carnap on Concept Determination: Methodology for Philosophy of Science." *European Journal for Philosophy of Science* 2 (2):161–79. https://doi.org/10.1007/s13194-011-0027-5

Khelfaoui, Mahdi, Yves Gingras, Mael Lemoine, and Thomas Pradeu. 2021. "The Visibility of Philosophy of Science in the Sciences, 1980–2018." *Synthese* 199 (3–4):1–31. https://doi.org/10.1007/s11229-021-03067-x

Kitcher, Philip. 1985. "Two Approaches to Explanation." *Journal of Philosophy* 82 (11):632–39.

Krohs, Ulrich, and Werner Callebaut. 2007. "Data without Models Merging with Models without Data." In *Systems Biology: Philosophical Foundations*, edited by Fred C. Boogerd, Frank J. Bruggeman, Jan-Hendrik S. Hofmeyr, and Hans V. Westerhoff, 181–213. Amsterdam: Elsevier.

Leonelli, Sabina. 2019. "What Distinguishes Data from Models?" *European Journal for Philosophy of Science* 9 (2):22. https://doi.org/10.1007/s13194-018-0246-0

Leonelli, Sabina, and Racheal A. Ankeny. 2012. "Re-Thinking Organisms: The Impact of Databases on Model Organism Biology." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1):29–36. https://doi.org/10.1016/j.shpsc.2011.10.003

Levy, Arnon, and Adrian Currie. 2015. "Model Organisms Are Not (Theoretical) Models." *British Journal for the Philosophy of Science* 66 (2):327–48. https://doi.org/10.1093/bjps/axt055

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. "Thinking about Mechanisms." *Philosophy of Science* 67 (1):1–25.

Maher, Patrick. 2007. "Explication Defended." *Studia Logica* 86 (2):331–41. https://doi.org/10.1007/s11225-007-9063-8

Malaterre, Christophe, Jean-François Chartier, and Davide Pulizzotto. 2019. "What Is This Thing Called *Philosophy of Science*? A Computational Topic-Modeling Perspective, 1934–2015." *HOPOS: The Journal of the International Society for the History of Philosophy of Science* 9 (2):215–49. https://doi.org/10.1086/704372

https://doi.org/10.1017/psa.2023.97 Published online by Cambridge University Press

Malaterre, Christophe, and Martin Léonard. 2023. "Charting the Territories of Epistemic Concepts in the Practice of Science: A Text-Mining Approach." *British Journal for the Philosophy of Science*. https://doi.org/10.1086/725656

McPhetres, Jonathon, Nihan Albayrak-Aydemir, Ana Barbosa Mendes, Elvina C. Chow, Patricio Gonzalez-Marquez, Erin Loukras, Annika Maus, Aoife O'Mahony, Christina Pomareda, Maximilian A. Primbs, Shalaine L. Sackman, Conor J. R. Smithson, and Kirill Volodko. 2021. "A Decade of Theory as Reflected in *Psychological Science* (2009–2019)." *PLoS ONE* 16 (3):e0247986. https://doi.org/10.1371/journal.pone.0247986

Mitchell, Sandra D. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

Mizrahi, Moti. 2020. "Proof, Explanation, and Justification in Mathematical Practice." *Journal for General Philosophy of Science* 51 (4):551–68. https://doi.org/10.1007/s10838-020-09521-7

Mizrahi, Moti. 2022. "Theoretical Virtues in Scientific Practice: An Empirical Study." *British Journal for the Philosophy of Science* 73 (4). https://doi.org/10.1086/714790

Morgan, Margaret S., and Mary S. Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Nagel, Ernest. 1961. *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace and World.

Noichl, Maximilian. 2021. "Modeling the Structure of Recent Philosophy." *Synthese* 198 (6):5089–100. https://doi.org/10.1007/s11229-019-02390-8

Overton, James A. 2013. "'Explain' in Scientific Discourse." *Synthese* 190 (8):1383–405. https://doi.org/10.1007/s11229-012-0109-8

Pence, Charles Hamilton, and Grant Ramsey. 2018. "How to Do Digital Philosophy of Science." *Philosophy of Science* 85 (5):930–41. https://doi.org/10.1086/699697

Potochnik, Angela. 2017. *Idealization and the Aims of Science*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226507194.001.0001

Resnik, David B. 1991. "How-Possibly Explanations in Biology." *Acta Biotheoretica* 39 (2):141–49. https://doi.org/10.1007/BF00046596

Rosenberg, Alex, and Daniel W. McShea. 2008. *Philosophy of Biology: A Contemporary Introduction*. 1st ed. Milton Park, Abingdon, Oxfordshire: Routledge. https://doi.org/10.4324/9780203926994

Ruse, Michael. 1973. *The Philosophy of Biology*. Taylors, SC: Hutchinson.

Salmon, Wesley C. 1989. *Four Decades of Scientific Explanation*. Pittsburgh, PA: University of Pittsburgh Press.

Savage, C. Wade, ed. 1990. *Scientific Theories*. Vol. 14. Minneapolis: University of Minnesota Press.

Sober, Elliott. 2015. *Ockham's Razors: A User's Manual*. Cambridge: Cambridge University Press.

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Suppe, Frederick. 1989. *The Semantic Conception of Theories and Scientific Realism*. Champaign, IL: University of Illinois Press.

Thagard, Paul. 2007. "Coherence, Truth, and the Development of Scientific Knowledge." *Philosophy of Science* 74 (1):28–47. https://doi.org/10.1086/520941

Verreault-Julien, Philippe. 2019. "How Could Models Possibly Provide How-Possibly Explanations?" *Studies in History and Philosophy of Science Part A* 73:22–33.

Weisberg, Michael. 2012. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Cite this article: Malaterre, Christophe and Martin Léonard. 2024. "Epistemic Markers in the Scientific Discourse." *Philosophy of Science* 91 (1):151–174. https://doi.org/10.1017/psa.2023.97