RESEARCH ARTICLE

The confirmation of scientific theories using Bayesian causal networks and citation sentiments

Henry Small

SciTech Strategies Inc., Bala Cynwyd, PA 19004

Keywords: Bayes's theorem, causal networks, citation context sentiments, confirmation, history and philosophy of science, nociception

ABSTRACT

The confirmation of scientific theories is approached by combining Bayesian probabilistic methods, in particular Bayesian causal networks, and the analysis of citing sentences for highly cited papers. It is assumed that causes and their effects can be identified by linguistic methods from the citing sentences and that the cause-and-effect pairs can be equated with theories and their evidence. Further, it is proposed that citation context sentiments for "evidence" and "uncertainty" can be used to supply the required conditional probabilities for Bayesian analysis where data is drawn from citing sentences for highly cited papers from various fields. Hence, the approach combines citation and linguistic methods in a probabilistic framework and, given the small sample of papers, should be considered a feasibility study. Special attention is given to the case of nociception in medicine, and analogies are drawn with various episodes from the history of science, such as the Watson and Crick discovery of the structure of DNA and other discoveries where a striking and improbable fit between theory and evidence leads to a sense of confirmation.

an open access journal

Check for updates

Citation: Small, H. (2022). The confirmation of scientific theories using Bayesian causal networks and citation sentiments. *Quantitative Science Studies*, *3*(2), 393–419. https://doi.org/10.1162/qss_a_00189

DOI:
https://doi.org/10.1162/qss_a_00189

Peer Review:
https://publons.com/publon/10.1162/qss_a_00189

Received: 2 February 2022
Accepted: 16 March 2022

Corresponding Author:
Henry Small
hsmall@mapofscience.com

Handling Editor:
Ludo Waltman

The MIT Press

1. BACKGROUND

Scientometrics and quantitative studies of science have traditionally avoided epistemological issues such as the nature of scientific knowledge, how knowledge is discovered and confirmed, and the relationship of theory and evidence. This is despite the fact that the scientific papers we count, classify, and map are filled with arguments and descriptions dealing with theories and observations, and why we should believe one finding or theory rather than another. Clearly the field will need new tools, or to adapt old ones, to enable us to delve into this deeper level of scientific content.

This paper will discuss one possible approach: the identification of causal statements in scientific texts and the evaluation of their degree of confirmation, inspired by recent developments in causal network theory (Pearl, 2000; Pearl & Mackenzie, 2018). The concept of causality is itself the subject of much debate in philosophy from the time of Aristotle to the present (Bunge, 1963; Findler & Bickmore, 1996; Sobrino, Olivas, & Puente, 2010). Contemporary approaches to analyzing and extracting causal content from texts are increasingly focused on deep learning algorithms (Li, Li et al., 2021a; Trieu, Tran et al., 2020). Modern approaches to causal networks are based on Bayes's theorem, and we will use this framework to interpret the causal assertions found in scientific texts.

What do we mean by the statement A causes B? Because we are dealing with science, we will interpret theories, hypotheses, models, or laws as positing causal assertions that are linked to empirical findings or observations and are the effects of those causes. Thus, if a theory asserts that A causes B, and B is found to occur, this increases the probability that the theory is correct, which is a basic tenet of Bayesian philosophy of science. Of course, we know from the history of science that theories have changed radically in the past, and there is no reason to think that they will not continue to change in the future. No theory, no matter how well corroborated, is invulnerable. This means that we will not be dealing with the ultimate causes, whether A *really* causes B, or whether theory A is the final *true* explanation of B, but rather with the perception or belief that theory A is true within a particular historical context given the evidence B available at the time.

Familiar examples of changing causal explanations from the history of science are the transition from Aristotle's theory of motion to Newton's laws, Ptolemy's Earth-centered account of celestial motions to the Copernican Sun-centered account, the phlogiston theory of combustion to Lavoisier's oxygen-based theory, Newton's to Einstein's theory of gravity, and Bohr's atomic model to the Schrödinger/Heisenberg quantum mechanical theory of the atom. The Watson and Crick discovery of the structure of DNA will serve as an example of theory change in the face of confirming and disconfirming evidence.

The replacing of one theory by another is, of course, an instance of what Kuhn (1962) called a scientific revolution, although the vast majority of instances play out on a much smaller, microrevolutionary scale. The common thread in these examples is that theories act as causal constructs and effects are the observable phenomena. While the causes may change over time as one theory supersedes another, the effects are somewhat more stable, although the latter can increase in accuracy or expand dramatically when new scientific instruments are invented. The field of medicine is replete with causes and effects, such as, when a bacterium or virus is postulated as the cause of a disease. Here the bacterium or virus is the cause, or theory, and the disease is the effect or evidence. In diagnosis, the disease acts as the cause or theory and symptoms as effects or evidence. Technologies and methods might also be modeled in the same fashion, although here the mechanism or inner workings of the method plays the role of the "theory," and the end result or outcome is the "effect." Generally, the concept of "A causes B" can be viewed as a possible pathway in a complex, probabilistic network of causes and effects.

From the effect side, we know from the work of Hanson (1972) that evidence can be theory laden, and confirmation bias is always present. Of course, theories are designed to explain specific phenomena. If a theory is later found to explain or predict some other phenomenon, then our confidence in the theory is usually increased. Likewise, unexpected failure to explain some phenomenon may decrease our confidence in the theory. Effects are also subject to experimental errors, which can propagate if a chain of measurements is involved. Such seems to be the case for cold fusion, where initial experiments indicating an excess of energy output over input were interpreted as support for a nuclear fusion hypothesis ("Cold fusion," 2021). In the historical case of phlogiston, it was the neglect of weight comparisons of reactants and products, presumably irrelevant to theory, that delayed the recognition that something was being added during combustion (oxygen) rather than being lost (phlogiston) (Ihde, 1964, 57). These examples are in accord with the Bayesian framework because confirming evidence increases our confidence in a theory and disconfirming evidence decreases it.

In this paper we will deal with causal assertions at the microlevel rather than the paradigm-changing level based on a close analysis of scientific texts. Of course, collecting sufficient

textual evidence for science in earlier centuries is challenging, but given current full text resources there is no such limitation for contemporary science. Presumably, if we are seeking opinion on the status of a current theory or empirical finding, we could perform a full text search on the scientific literature or even on social media. This would generate a heterogeneous set of statements representing a broad range of opinion from experts and nonexperts.

In this paper we will constrain the process by focusing on specific highly cited papers and their citing sentences, also called *citances* (Nakov, Schwartz, & Hearst, 2004), and attempt to discern causes and effects from that more limited perspective. By restricting the data to highly cited papers and their citing sentences, we can sharpen the focus on a specific theory, and more accurately assess its degree of confirmation within a community of peers. In addition, we can expand the scope by including closely related papers drawn from a citation-based cluster. Citing sentences also reveal the degree of agreement among a community of citing authors on the core findings of the cited work (Small, 1978), and when aggregated can be represented as a network of assertions. The resulting network, it is proposed, can be interpreted as a collective model of the theory and its empirical outcomes.

The background to this effort was an analysis of a single highly cited paper on the topic of nociception (Caterina, Schumacher et al., 1997), the biological basis of the sensation of pain (Moayedi & Davis, 2013). Using a set of 763 citing sentences for this paper, it was possible to manually construct a network of assertions that linked theoretical causes with experimental effects (Small, 2021). The goal of this paper is to automate the creation of such networks as far as possible and see if they can be used to assess the degree of confirmation of the underlying theory. In the course of this work, quite unexpectedly, the senior author of the original focal paper (Caterina, Leffler et al., 2000), David Julius from the University of California, San Francisco, was awarded the 2021 Nobel Prize in Physiology or Medicine for his contributions to the field of nociception (Julius, 2021).

2. DATA

Three different data sources were used to identify highly cited papers and collect their citing sentences. At the time this research began in early 2021, no single source of citing sentences was available (see Nicholson, Mordaunt et al., 2021). First, the Centre for Science and Technology Studies (CWTS) at Leiden University provided sets of highly cited papers and their citing sentences partitioned into five algorithmically defined fields of science drawn from Elsevier's ScienceDirect database. These data were in turn drawn from full-text information of English-language scientific papers published in Elsevier journals following the procedure described in previous papers (Boyack, Van Eck et al., 2018; Larmers, Boyack et al., 2021). Using this resource, the 500 most cited papers were selected for each of five broad fields (Biomedical and Health Sciences, Life and Earth Sciences, Mathematics and Computer Science, Physical Sciences and Engineering, and Social Sciences and Humanities) in addition to their citing sentences. The cited papers cover the years 2000 to 2015, and the citing sentences are from papers published from 2000 to 2016.

A second data source was the open access subset of PubMed Central® (PMC) from the National Library of Medicine, consisting of the full text of primarily biomedical articles in XML format. The PMC includes papers that were required to be publicly available under the National Institutes of Health public access policy and other open access sources. PMC processing adds codes to the references cited by articles that allow the user to connect the reference within the text to the bibliographic information at the end of each article, and, like the Leiden ScienceDirect database, enables the retrieval of all the sentences from the full text

of covered articles that cite a given reference. SciTech Strategies downloaded the open access subset from November 2018, and imported it into a MySQL database (Small, Tseng, & Patek, 2017). The years covered are the 1990s to 2018.

The third data source used was a cluster analysis, or model, of Scopus data maintained by SciTech Strategies. The model covers Scopus data for the years 1996 to 2018 and consists of 43 million documents assigned to 104,677 clusters or research communities (Klavans, Boyack, & Murdick, 2020). Denoted as STS5, the model was created using a direct citation clustering algorithm from Leiden University (Traag, Waltman, & Van Eck, 2019).

Papers were selected from different fields using these data sources. The papers served as case studies for developing methods for extracting cause/effect (theory/evidence) relationships from their citing sentences and testing their degree of confirmation, and should not be considered as representative of the broad fields. As an initial screening, samples of citances for each paper were scanned for the presence of theoretical or experimental terms which suggested that causal connections were being made. An examination of 20 or so citances for a given cited paper reliably identified it as causal or noncausal. On this basis, roughly one-half of the papers in a sample of 500 highly cited biomedical papers were classified as causal.

While citing sentences for causal cited papers tend to be causal as well, citing sentences can also be descriptive, procedural, or programmatic and not make any theoretical assertions. Citing sentences for method papers, for example, are predominantly procedural in nature, and not causal. However, review papers, because of their role in synthesizing knowledge, can be a rich source of causal connections.

Ten papers were selected from the Elsevier/CWTS data set spread across four fields: one from Biomedical and Health Sciences, and three each from Life and Earth Sciences, Physical Sciences and Engineering, and Social Sciences and Humanities (see Table 1). These papers then served as the basis of the feasibility study. The single paper from life science, the previously mentioned paper by Caterina et al. (2000), appeared in cluster #769 from the SciTech Strategies STS5 model (denoted STS5-769). This cluster consisted of 7,971 papers and was focused on nociception. The 20 most cited Scopus papers from this cluster were also selected for analysis (see Table 2). Citing sentences for these 20 nociception papers were retrieved from the PubMed Central repository. See Table 7 for general theory statements for each of the papers.

3. METHODS

3.1. Creating Causal-Effect Pairs

One of the goals of this project was to see if pairs of words, or more precisely noun phrase pairs, could be extracted from citing sentences representing cause/effect or theory/evidence connections. This seemed feasible because the citing sentences were often restatements of the findings of the cited work, and multiple citing sentences were available.

Following the initial screening of highly cited papers for theoretical or experimental terms, there was also the need to have some indicator that the citing sentence had made a causal assertion. One way to do that is to look for general words that denote causes or effects. Examples of causal words are the verb *activated* and the noun *stimulus*. Examples of effect words are *response* and *result.* To this end, general cause and effect words were compiled by manually scanning citances for the 30 highly cited papers used in this study.

The manual selection process was augmented using machine learning in the following way taking nociception as an example. A random sample of 327 sentences was selected from

Table 1. Papers selected from Elsevier/CWTS data in four fields. The field from which the papers were selected precedes the bibliographic information on the paper. The column "number of citances" is the number of citing sentences in Elsevier's ScienceDirect database through 2016

Paper	Number of citances
Biomedical and Health Sciences	
Caterina, M. J., Leffler, A., Malmberg, A. B., Martin, W. J., Trafton, J., ... Julius, D. (2000). Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science*, *288*(5464), 306–313.	763
Life and Earth Sciences	
Mottram, D. S., Wedzicha, B. L., & Dodson, A. T. (2002). Acrylamide is formed in the Maillard reaction. *Nature*, *419*(6906), 448–449.	399
Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., ... Wardle, D. A. (2001). Ecology—biodiversity and ecosystem functioning: Current knowledge and future challenges. *Science*, *294*(5543), 804–808.	406
Alexander, M. (2000). Aging, bioavailability, and overestimation of risk from environmental pollutants. *Environmental Science & Technology*, *34*(20), 4259–4265.	395
Physical Sciences and Engineering	
Adachi, C., Baldo, M. A., Thompson, M. E., & Forrest, S. R. (2001). Nearly 100% internal phosphorescence efficiency in an organic light-emitting device. *Journal of Applied Physics*, *90*(10), 5048–5051.	560
Das, S. K., Putra, N., Thiesen, P., & Roetzel, W. (2003). Temperature dependence of thermal conductivity enhancement for nanofluids. *Journal of Heat Transfer-Transactions of the ASME*, *125*(4), 567–574.	574
Aharony, O., Gubser, S. S., Maldacena, J., Ooguri, H., & Oz, Y. (2000). Large *N* field theories, string theory and gravity. *Physics Reports—Review Section of Physics Letters*, *323*(3–4), 183–386.	480
Social Sciences and Humanities	
Berkman, L. F., Glass, T., Brissette, I., & Seeman, T. E. (2000). From social integration to health: Durkheim in the new millennium. *Social Science & Medicine*, *51*(6), 843–857.	349
Cardinal, R. N., Pennicott, D. R., Sugathapala, C. L., Robbins, T. W., & Everitt, B. J. (2001). Impulsive choice induced in rats by lesions of the nucleus accumbens core. *Science*, *292*(5526), 2499–2501.	326
Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11818–11823.	323

papers citing Caterina et al. (2000), and manually classified as causal or noncausal. The sample was divided into training and test sets, and the Scikit-learn package was used for machine learning (Pedregosa, Varoquaux et al., 2011). The algorithm finds an optimal surface in multidimensional space separating the causal and noncausal sentences where each word corresponds to an axis in the space. This is done for 10 classifiers. The median accuracy of the 10 classifiers was 73%. Three of the classifiers had an accuracy of 74%. One of these was the BernoulliNB classifier, which had an F1 of 75% based on its precision and recall scores. The coefficients of individual words for that classifier were used to select additional cause/effect words. For example, the highest coefficient words for the Bernoulli classifier included words like *induced*, *activation*, *stimuli*, and *responses*, while low coefficient words were action-oriented, like *performed* or *examined* but in general were more diverse. Eight cause/effect words appeared in the top one-half of one percent of words ranked by coefficient.

Table 2. Twenty most cited papers from the STS5-769 cluster on nociception. The selection is based on citation counts from Scopus through 2018

Caterina, M. J., Schumacher, M. A., Tominaga, M., Rosen, T. A., Levine, J. D., & Julius, D. (1997). The capsaicin receptor: A heat-activated ion channel in the pain pathway. *Nature*, *389*(6653), 816–827.
Caterina, M. J., Leffler, A., Malmberg, A. B., Martin, W. J., Trafton, J., ... Julius, D. (2000). Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science*, *288*(5464), 306–313.
Tominaga, M., Caterina, M. J., Malmberg, A. B., Rosen, T. A., Gilbert, H., ... Julius, D. (1998). The cloned capsaicin receptor integrates multiple pain-producing stimuli. *Neuron*, *21*(3), 531–543.
Clapham, D. E. (2003). TRP channels as cellular sensors. *Nature*, *426*(6966), 517–524.
Story, G. M., Peier, A. M., Reeve, A. J., Eid, S. R., Mosbacher, J., ... Patapoutian, A. (2003). ANKTM1, a TRP-like channel expressed in nociceptive neurons, is activated by cold temperatures. *Cell*, *112*(6), 819–829.
McKemy, D. D., Neuhausser, W. M., & Julius, D. (2002). Identification of a cold receptor reveals a general role for TRP channels in thermosensation. *Nature*, *416*(6876), 52–58.
Julius, D., & Basbaum, A. I. (2001). Molecular mechanisms of nociception. *Nature*, *413*(6852), 203–210.
Szallasi, A., & Blumberg, P. M. (1999). Vanilloid (Capsaicin) receptors and mechanisms. *Pharmacological Reviews*, *51*(2), 159–211.
Peier, A. M., Moqrich, A., Hergarden, A. C., Reeve, A. J., Andersson, D. A., ... Patapoutian, A. (2002). A TRP channel that senses cold stimuli and menthol. *Cell*, *108*(5), 705–715.
Holzer, P. (1991). Capsaicin: Cellular targets, mechanisms of action, and selectivity for thin sensory neurons. *Pharmacological Reviews*, *43*(2), 143–201.
Jordt, S.-E., Bautista, D. M., Chuang, H.-H., McKemy, D. D., Zygmunt, P. M., ... Julius, D. (2004). Mustard oils and cannabinoids excite sensory nerve fibres through the TRP channel ANKTM1. *Nature*, *427*(6971), 260–265.
Davis, J. B., Gray, J., Gunthorpe, M. J., Hatcher, J. P., Davey, P. T., ... Sheardown, S. A. (2000). Vanilloid receptor-1 is essential for inflammatory thermal hyperalgesia. *Nature*, *405*(6783), 183–187.
Bautista, D. M., Jordt, S.-E., Nikai, T., Tsuruda, P. R., Read, A. J., ... Julius, D. (2006). TRPA1 Mediates the inflammatory actions of environmental irritants and proalgesic agents. *Cell*, *124*(6), 1269–1282.
Venkatachalam, K., & Montell, C. (2007). TRP channels. *Annual Review of Biochemistry*, *76*, 387–417.
Bandell, M., Story, G. M., Hwang, S. W., Viswanath, V., Eid, S. R., ... Patapoutian, A. (2004). Noxious cold ion channel TRPA1 is activated by pungent compounds and bradykinin. *Neuron*, *41*(6), 849–857.
Caterina, M. J., & Julius, D. (2001). The vanilloid receptor: A molecular gateway to the pain pathway. *Annual Review of Neuroscience*, *24*, 487–517.
Caterina, M. J., Rosen, T. A., Tominaga, M., Brake, A. J., & Julius, D. (1999). A capsaicin-receptor homologue with a high threshold for noxious heat. *Nature*, *398*(6726), 436–441.
Ramsey, I. S., Delling, M., Clapham, D. E., Bautista, D. M., Jordt, S.-E., ... Julius, D. (2006). An introduction to TRP channels. *Annual Review of Physiology*, *68*, 619–647.
Nilius, B., Owsianik, G., Voets, T., Peters, J. A., Venkatachalam, K., & Montell, C. (2007). Transient receptor potential cation channels in disease. *Physiological Reviews*, *87*(1), 165–217.
Chuang, H.-H., Prescott, E. D., Kong, H., Shields, S., Jordt, S.-E., ... Julius, D. (2001). Bradykinin and nerve growth factor release the capsaicin receptor from PtdIns(4,5)P2-mediated inhibition. *Nature*, *411*(6840), 957–962.

Eventually a set of 230 cause/effect words was compiled by a combination of manual scanning and machine learning runs. Only sentences containing one or more of the cause or effect words were input to subsequent processing steps aimed at isolating the actual causal assertions. An example of a causal citance for Caterina et al. (2000) is "The transient receptor potential vanilloid 1 channel (TRPV1) is a nonselective cation channel expressed in primary sensory neurons and implicated in thermal hyperalgesia." Causal words here are *expressed* and *implicated*. A descriptive, and hence noncausal, citance is "The cell bodies of the primary afferent sensory nerves are located in the dorsal root ganglia and trigeminal ganglia."

Many of the cause/effect words were verbs, and it appeared that verbs were important separators of causes from effects within the sentences. Initially it was also noted that causes seemed to occur as subjects of the citing sentences while effects occurred in the predicates (e.g., "A causes B"). This rule, however, does not hold when the passive voice is used, for example, "B was caused by A." In either case, however, the cause and effect usually appear in different clauses separated by a verb. Thus, the output from a part-of-speech parser was input to a Python script that formed pairs of noun phrases separated by verbs, as illustrated by Figure 1.

A count is then made of the total number of identical noun phrase pairs from different verb-separated sentence segments across all citing sentences for the highly cited paper or cluster of papers. Table 3 shows the most frequent pairs for the 20 most cited papers from cluster STS5-769 on nociception.

"TRP" stands for "transient receptor potential channel," of which many subtypes have been identified that respond to a variety of agents. These subtypes are considered sensors of the cell's environment. For the Caterina et al. (2000) citances alone, the noun phrase "TRPV1" pairs with "noxious heat" 13 times. A more general wildcard search for "*TRPV1*" and "*heat*" shows that these words are paired 143 times across verb-separated segments. In causal terms we can say that the TRPV1 receptor leads to the sensation of heat, which at this point is a hypothesis in need of confirmation.

This approach is similar to the subject-predicate-object (SPO) triples method used by databases, such as the semantic MEDLINE, to facilitate search and to identify various types

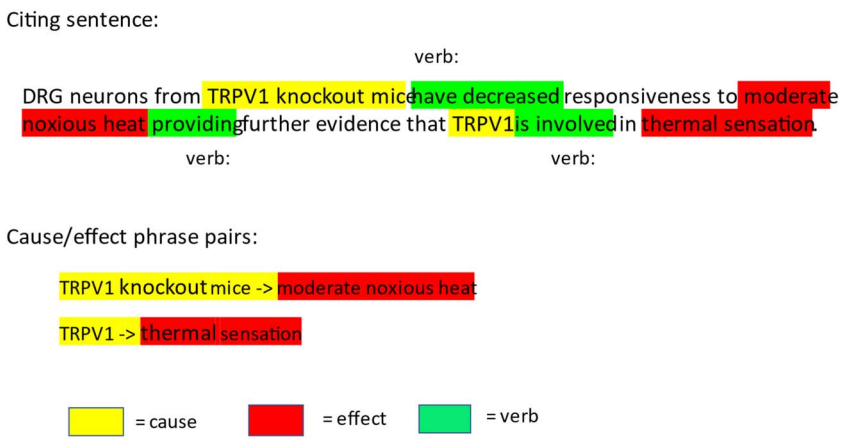


Figure 1. The creation of causal phrase pairs from a citing sentence. TRPV1 receptor phrases are highlighted in yellow, which function as causes, and "heat" words in red as effects. Verbs are highlighted in green and break the sentence into segments across which phrase pairs are formed. Two cause/effect pairs are generated in this example involving the TRPV1 receptor.

Table 3. Most frequent noun phrase pairs for the top 20 papers in the nociception cluster from SciTech Strategies, Inc. The 20 papers are listed in Table 2. Citing sentences are from the PMC database through 2018. Equivalent terms and acronyms have been unified

Cause phrase	Effect phrase	# of citances
TRPV1	capsaicin	170
TRPV1	heat	96
TRPV1	noxious heat	52
TRPV1	protons	48
capsaicin	TRPV1	43
TRPA1	allyl isothiocyanate	43
TRPM8	menthol	41
TRPV1	inflammation	35
TRPA1	noxious cold	27
TRPA1	mustard oil	24
TRPV1	pain	23
TRPV1	receptor	22
capsaicin	pain	21
TRPV1	bradykinin	20

of concept dependences in the biomedical literature (Kilicoglu, Shin et al., 2012; Rindflesch, Kilicoglu et al., 2011). The so-called "semantic predications" are available through the NLM's SemMed database and have been used by Chen and Song (2018) to map the subject and object connections involving causal type links in various ways to understand how causal connections can transform biomedical research areas. In a similar vein Li, Peng, and Du (2021b) have explored SPO triples as knowledge units in connection with the uncertainty sentiment as part of a case study of lung cancer.

The field of literature-based discovery (LBD) also uses the SPO tool to identify what Swanson (1986) called "undiscovered public knowledge," which is new knowledge somehow implicit in existing knowledge. The extensive LBD literature has recently been reviewed (Thilakaratne, Falkner, & Atapattu, 2019). The goal of LBD, however, differs from that pursued in this paper, which is not to posit new "undiscovered" knowledge but rather to identify existing causal associations in the literature and assess their degree of confirmation. A related approach to ours uses Bayesian networks among semantic predications to find novel biomedical hypotheses (Atkinson & Rivas, 2008). Their approach, however, requires that conditional probabilities be supplied by experts in the field and is not aimed at confirmation.

Another difference between the present approach and the SPO work is that phrase pairs are focused on the citing sentences for a specific highly cited paper or cluster of closely related papers and not the titles and abstracts of papers used by semantic MEDLINE. This means that we can capture the community consensus on the significance of the cited work and limit the phrase pairs to the subject matter represented by the cited paper or cluster. We can also look at causal connections across a variety of scientific fields and not be restricted to biomedicine.

As noted above, there is no guarantee that the cause will precede the effect in the sentence, and cases have been found where the cause appears in the predicate following a verb. Thus, the best policy is to look for frequently appearing noun phrases either preceding or following verbs and use other criteria to discern which is the cause and which is the effect. The rule of thumb adopted here is to take the more abstract or general entity to be the cause/theory and the more specific and concrete entity to be the effect/evidence. To give an example from one of the papers selected for analysis, if an abstract entity like the "Maillard reaction" is paired with the specific substance "acrylamide," then the chemical reaction is the cause and acrylamide, the effect. This is despite the fact that the phrase "Maillard reaction" comes after the word "acrylamide" in most sentences due to the passive voice (e.g., "acrylamide is formed by the Maillard reaction"). In this case, a total cause and effect pair frequency can be obtained by combining the forward with the backward occurrences. Later, we will differentiate these as "forward" and "backward" cases and show that the "forward" cases predominate.

3.2. The Bayesian Theory Confirmation

In his pioneering work on a computational method for evaluating theories called the *theory of explanatory coherence* (TEC), Thagard (1992) noted that the main drawback to applying Bayes's theorem to confirmation was the difficulty of specifying the conditional probabilities required for the calculation. Instead, Thagard posited a network of nodes representing statements that either cohere or conflict with one another. By passing confirming or disconfirming signals iteratively through the network, the weights for each node eventually converge to stable values for each statement.

By contrast, a Bayesian approach is based on causal relationships between a set of statements in the form of a directed, acyclic graph (DAG), where each link has, in effect, two weights associated with it, one denoting the probability that the theory agrees with the evidence and the other the probability that some other theory does. The weights are the conditional probabilities. Like the TEC process, the Bayesian network passes information back and forth among the linked statements in a series of iterations in a process called *belief propagation* until an equilibrium is reached, and new probabilities are arrived at that determine whether confirmation is achieved (Pearl & Mackenzie, 2018, p. 112). This process has been implemented in the Bnlearn package running in R (Nagarajan, Scutari, & Lebre, 2013), and later will be applied to a network of causal relationships in the field of nociception.

Bayesian confirmation theory was proposed by Carnap in the 1950s and was developed by philosophers of science beginning in the 1970s. It is based on a subjective interpretation probability in contrast to a frequentist one where countable events set the probabilities (Pearl, 2000). In either the subjective or frequentist interpretation, probabilities vary between 0 and 1, where 1 indicates complete certainty. For example, the probability of a theory T being true, such as quantum mechanics or the Watson/Crick double helix for DNA, is a matter of subjective opinion, whether individual or collective, and is called the *prior probability*, denoted as $P(T)$. The fundamental assumption of Bayesian confirmation is that T and E are logically independent, that the prediction of the theory does not affect or influence the acquisition of the evidence, and vice versa. Thus, the joint probability of T and E, $P(T \& E)$ represents the agreement of theory with evidence.

The notation $P(E|T)$ is the probability of observing E given that theory T is true. This has the character of a deduction of E from T, going from the general to the specific. The inverse, $P(T|E)$, is the probability of theory T being true, given that evidence E is observed, has the character of an induction going from the specific to the general. $P(T|E)$ is called the *posterior probability*,

the probability of the theory conditional on the evidence E, which indicates confirmation if it is greater than the prior probability $P(T)$. In this case we apply Bayes's rule and update the prior probability for the theory $P(T)$ to the value of the posterior probability $P(T|E)$, awaiting the arrival of further evidence either confirming or disconfirming the theory. The deductive step $T \rightarrow E$ requires time and effort on the part of the scientist whereas the inductive step $E \rightarrow T$ does not, which means that realizing T agrees with E is delayed even if E is old.

Bayes's theorem can be written as:

$$P(T|E) = P(T) * P(E|T)/P(E)$$

which follows from the definition of $P(T|E)$ as $P(T \,\&\, E)/P(E)$, and $P(E|T)$ as $P(E \,\&\, T)/P(T)$.

An extension of this formula using a theorem in probability theory called "total probability" is:

$$P(T|E) = P(T) * P(E|T)/(P(T) * P(E|T) + P(\sim T) * P(E|\sim T))$$

where $\sim T$ is "not T" or "anything other than T" and $P(\sim T) + P(T) = 1$.

In the context of theory and evidence, the $\sim T$ indicates any possible theory other than T that might explain E such as an alternative or competing theory. "Total probability" states that any probability, say $P(E)$, can be expressed as the sum of all possible mutually exclusive theories T_i, that is, the sum of $P(E|T_i) * P(T_i)$ over i, or equivalently the sum of all joint probabilities $P(E \,\&\, T_i)$ (Pearl, 2000, p. 4).

The conditional probability $P(E|T)$ expresses how well the theory T fits the evidence E, and $P(E|\sim T)$ how well an alternative theory fits the evidence E. The ratio of these two quantities is called the *likelihood ratio* and determines whether the hypothesis is confirmed or disconfirmed (Howson & Urbach, 2006, p. 21; Pearl, 2000, p. 7). It follows from Bayes's theorem that if $P(E|T)$ is greater than $P(E|\sim T)$, $P(E|T)$ must be greater than the prior probability $P(T)$. This indicates that the hypothesis is confirmed. Conversely, if $P(E|T)$ is less than $P(E|\sim T)$, the theory is disconfirmed and $P(E|T)$ is less than $P(T)$. If $P(E|T) = P(E|\sim T)$ then the theory is neither confirmed nor disconfirmed, and the posterior probability $P(T|E)$ equals the prior probability $P(T)$, which means that taking the evidence E into account does not change the probability of the theory. These relationships can be illustrated graphically by plotting the three probabilities $P(T|E)$, $P(E|T)$, and $P(E|\sim T)$ as a three-dimensional surface for a given value of $P(T)$ (Small, 2020). Note that $P(E|\sim T)$ is the probability of a false positive assuming T is true.

It is obvious that most scientists do not follow such a formal mathematical procedure when formulating or testing their theories (Glymour, 1980; Kuhn, 1977). However, it is possible that many scientists intuitively apply two principles of the Bayesian approach in the conduct of their research: first, when they assess the fit between a theory and the evidence, that is, the ability of the theory to explain or predict the evidence, and second, when they assess whether an alternative theory can explain the evidence equally well or better. Hence, the Bayesian apparatus does suggest some simple rules of thumb for evaluating theories.

As an historical example, consider James Watson's realization that the DNA bases fit together in a unique way. By playing with cardboard cut-outs of the four bases (adenine, thymine, guanine, and cytosine), he saw that the pattern of hydrogen bonding fit together neatly for A linking to T and G linking to C (Olby, 1974; Watson, 1968). This unique pattern also explained the Chargaff rules of base ratios, as well as the observed symmetry from X-ray diffraction by Rosalind Franklin (Schindler, 2008). Thus, at least three increments of confirmation (stereochemistry, X-ray symmetry, and base ratios) gave a boost to the theory, increasing its $P(E|T)$. At the same time, Watson's previous model of DNA, where bases were bonded like-to-

like (Watson, 1968, p. 185), an alternative model, could not explain these findings, thus decreasing $P(E|{\sim}T)$. Hence, the autobiographical and historical accounts of Watson and Crick's work are consistent with a Bayesian framework, although they do not show that Bayesian precepts actually governed the actions of the participants.

3.3. Estimating Probabilities Using Sentiment Analysis

It is not immediately obvious how bibliometric methods can be adapted to a Bayesian model. One approach is to use autobiographical accounts of discoveries such as Watson's to look for events that increment or decrement confidence in a theory or competing theory. Linus Pauling's competing theory of a triple helix structure for DNA was rejected by Watson because the structure could not be acidic, which contradicted experimental evidence. This reduced Watson's confidence in the model. However, we have no way of knowing how much the probability of the model was reduced. Nor does the Bayesian theory give us any guidance on what counts as evidence. For example, "accuracy" is just one of the five criteria of theory choice discussed by Kuhn (1977). Another very different approach is to survey the opinion of peers on the model. This can be done in retrospective studies by analyzing a large sample of contemporary texts, for example, by a sentiment analysis of citation contexts. Presumably, the community would be using their own subjective criteria when citing the theory, which may or may not match those used by the discoverers.

The quantity $P({\sim}T)$, the prior probability of "not T," seems amenable to an analysis of uncertainty. By searching for the number of sentences jointly mentioning the theory (or causal entity) and uncertainty terms, we get a measure of the uncertainty of T. Dividing this quantity by the number of sentences containing T gives a number between 0 and 1. This provides a probability measure of uncertainty for T or certainty for ${\sim}T$. We obtain a quantity proportional to the prior probability of the theory $P(T)$ by subtracting $P({\sim}T)$ from one because $P(T) = 1 - P({\sim}T)$.

A similar approach might be taken to indirectly estimating $P(E|{\sim}T)$ because we are looking for instances of support for an alternative to T, namely ${\sim}T$, as an explanation of E which implies a weakening of T. We do this by searching for sentences containing both theory T and evidence E (i.e., both cause and effect) in conjunction with uncertainty terms. In this instance, the uncertainty terms weaken the theory and there is no need to subtract from one. To estimate $P(E|T)$ we need to find sentences where support is provided for the theory-evidence or cause-effect combination. In this case, we use a vocabulary of words indicating that supporting evidence is provided and search for them in conjunction with the theory-evidence pair. The number of such sentences divided by the total number of sentences with the theory-evidence pair gives a rate of support for the theory by the evidence.

It is important to recognize the approximate and indirect nature of these estimates of conditional probabilities. In the case of $P(E|T)$ we are assuming that the appearance of words denoting supporting evidence for a hypothesis boosts the probability that T leads to E. In the case of $P(E|{\sim}T)$ we are assuming that the appearance of uncertainty words in a sentence involving the theory increases the probability that some other theory (${\sim}T$) explains the evidence without, however, saying what that other theory is. We will discuss the limitations of this approach in the discussion section. No doubt the existence of viable competing or alternative theories increases the uncertainty of the theory under consideration (Chen & Song, 2018), but there may be other reasons for this lack of confidence and by itself it does not imply support for an alternative theory.

Another difficulty with using uncertainty and support terms to estimate probabilities is due to the inherent differences in the rate of occurrence of these words for different topics. For

example, in most cases examined, the "supporting evidence" term occurrences exceed the "uncertainty" occurrences. This may simply express a "confirmation bias" or tendency to use supporting words in citation contexts, as pointed out by Greenberg (2009). Large-scale studies, such as Nicholson et al. (2021), based on deep learning showed an even larger imbalance between "supporting" and "contrasting" citances, although they appear not to have taken "uncertainty" terms into account. There also may be inherent differences between topics in the rates of sentiment words that could lead to biases in comparing topics. A simple solution to compensate for such differences is to make the theory-evidence rates relative to a baseline specific to the topic in question. To do this we divide the rates derived from the cause-effect sentences by a baseline rate obtained from a broader sample of sentences that includes the sentences under analysis. For example, if the sentences are contained as a subset of a broader topic, we can divide by the "support" and "uncertainty" rates computed from the broader topic. Such baseline rates have been computed using all citing sentences for individual highly cited papers or, alternatively, for a cluster of closely related papers on the topic.

As an example, suppose the theory-evidence or cause-effect terms occur in 615 sentences in a data set consisting of 4,752 sentences. Of the 615 sentences, 79 (12.8%) contain uncertainty terms, while 123 (20%) sentences contain supporting evidence terms. The corresponding rates for the broader baseline sample of 4,752 sentences are 20.3% and 24.7%. Dividing by the baseline rates gives 0.63 for uncertainty and 0.81 for supporting evidence. Because we are equating uncertainty with $P(E|{\sim}T)$ and supporting evidence with $P(E|T)$, these values give a likelihood ratio greater than 1 and the theory is confirmed.

3.4. Compiling Sentiment Word Sets

We have relied on the presence of specific cue or signal words to classify the citing sentences. Three types of sentiment word sets have been compiled: words denoting causes and effects, words expressing supporting evidence, and words expressing uncertainty. For uncertainty words, important prior work has been carried out by Chen and Song (2018) and by Chen, Song, and Heo (2018). They use a seed set of uncertainty words from Hyland (2004) including hedging terms and expand the set by the word2vec method (Mikolov, Sutskever et al., 2013). In one of their studies, they use predications from semantic MEDLINE involving causal predications such as "HIV CAUSES Aids." When they combine these data with the presence of uncertainty words they can show the time evolution of certainty or uncertainty for the claim over a period of years. They point out that predications are much enhanced by the inclusion of uncertainty.

The approach taken here involves manual coding of random samples of sentences for each of these sentiments, coding each sentence as having the sentiment or not having it. The sentences coded as having the sentiment were tokenized and word counts generated. The resulting ranked lists were scanned for possible cue words for the sentiment. The cue words selected were as independent as possible of subject matter or technical meaning. Lists compiled by other authors were also consulted to see what cue words were used in their studies. For example, the recent paper on identifying "disagreement" citances (Larmers et al., 2021) was used to augment the uncertainty word set as it seemed likely that disagreement contributes to the lack of certainty of an assertion.

Machine learning was also used to aid in the compilation of cue words, as described previously for the cause/effect sentiment, by dividing the coded random samples of sentences into training and test sets. The output from machine learning includes the accuracy of the various

classifiers and the coefficients for individual words for a given classifier that define the optimal surface in multidimensional word space. Because these coefficients are higher for words that occur in sentences classified having a particular sentiment (assuming the sentence is coded 1 for presence of the sentiment, and 0 for its absence), scanning the list of words having the highest coefficients can also reveal potential cue words for the sentiment.

The precision and recall of a given word can be computed by matching the manually coded sentences with the sentences retrieved by the sentiment word. For example, the cause/effect cue word "stimuli" retrieved 30 sentences that contained the word, of which 25 were coded causal and five noncausal. Thus, the precision for this word in retrieving causal sentences is 25/30, or 83%, based on this sampling. Recall for this single word is 25/254, or 10%, although recall is expected to be low for single words.

A similar exercise was undertaken for compiling and testing "uncertainty" sentiment words. A small set of 25 uncertainty words was compiled and tested against 300 randomly selected sentences from the fields of life science, biological science, physical science, and social science. These sentences were coded independently by two coders as uncertain or certain. Matching the set of 25 prospective uncertainty words (using wildcard searches to retrieve variants) and comparing the hits to the manually coded sentences gave an overall precision of 75% and a recall of 56% for the aggregate of 300 sentences from the four fields combining the results from both coders. The relatively low recall statistic indicates that the 25 uncertainty words were inadequate for retrieving all the sentences that had been coded as uncertain. Using Cohen's Kappa (Cohen, 1968), only a moderate interrater reliability of 0.43 was found for the two coders. Nevertheless, the precision computed for individual words revealed a core of reliable uncertainty words (Table 4).

The compilation of words for the "supporting evidence" sentiment followed a similar course. This sentiment was designed to capture sentences that seek or claim evidence supporting the cause/effect assertion. Thus, words that indicate support, such as *demonstrate, show*, or *measure*, are included, as are words denoting actions to find evidence such as *study, observe*, and *experiment.* Ten of these cue words were tested on the same set of 300 sentences from four fields using the two coders as described above. In this case overall precision and recall improved to 90% and 79% respectively. Again, overall recall was lower than precision, indicating that not all cases of "supporting evidence" were retrieved. The precision and recall for eight individual words are shown in Table 5.

Table 4. Uncertainty words with the highest recall and precision, based on a random sample of 300 sentences from four fields. Sentences were coded independently by two coders

Word	Wildcard search	Precision (%)	Recall (%)
however	*however*	91.6	23.6
may	*may*	76.9	21.5
could	*could*	78.6	11.0
although	*although*	100.0	12.9
appears	*appear*	75.0	12.9
suggests	*suggest*	75.0	6.5
failed	*fail*	75.0	9.7

Table 5. Supporting evidence words with the highest recall and precision, based on the same sample of sentences used for Table 4.

Word	Wild card search	Precision (%)	Recall (%)
report	*report*	92.0	33.7
observe	*observ*	90.0	11.0
experiment	*experiment*	85.4	20.3
study	*stud*	89.3	29.1
demonstrate	*demonstrat*	90.0	5.2
found	*found*	83.3	5.8
show	*show*	97.0	19.2
measure	*measur*	97.2	20.3

4. RESULTS

4.1. Computing Confirmation for Individual Causal Pairs by the Likelihood Ratio

Each of the highly cited papers in Table 1 and corresponding citing sentences were represented by frequently occurring cause-and-effect phrase pairs. As described previously, these pairs are generated by combining noun phrases separated by verbs across the citing sentences containing causal words and ranking the phrase pairs by frequency. This results in a list with a few frequently encountered pairs at the top of the list and a long tail of less frequently occurring pairs. First, we will focus on the most frequent phrase pair for each paper and present a typical citing sentence for each.

Table 6 shows the principal causal phrase pair for each highly cited paper, the number of instances of the phrase pairs in verb-separated segments of the citing sentences, and the

Table 6. Principal causal phrase pairs for highly cited papers in Table 1. The first column gives the primary author and year of the paper. The third column gives the cause and effect separated by an arrow →. The verb-separated count column shows the forward (F) and backward (B) occurrences

Highly cited paper	Field	Principal causal phrase pair	Verb-separated count	Distinct sentences with both phrases
Caterina (2000)	Life sci	TRPV1 → heat	108 F + 35 B	99
Mottram (2002)	Biological sci	Maillard reaction → acrylamide	35 F + 100 B	152
Loreau (2001)	Biological sci	biodiversity → ecosystem	69 F + 26 B	100
Alexander (2000)	Biological sci	time → bioavailability	7 F + 23 B	46
Adachi (2001)	Physical sci	excitons → quantum efficiency	40 F + 25 B	56
Das (2003)	Physical sci	nanofluid → thermal conductivity	134 F + 156 B	283
Aharony (2000)	Physical sci	ads/cft → boundary	19 F + 2 B	30
Berkman (2000)	Social sci	social network → health	22 F + 10 B	50
Cardinal (2001)	Social sci	brain lesions → impulsivity	74 F + 24 B	71
Blood (2001)	Social sci	music → reward	40 F + 34 B	76

Table 7. Typical citing sentences and theory statements for the principal causal pairs in Table 6. The first column gives the primary author and year of the paper from Table 1. The second column contains a typical citing sentence in quotes, and in the following row a summary statement of the theory

Highly cited paper	Typical citing sentence for the causal pair/statement of theory
Caterina (2000)	"Temperature gating is an important feature of TRPV1, critical for the somatosensory response to noxious heat."
Theory	There are a variety of genetically expressed molecular receptors on neurons responsible for the sensation of heat and other environmental stimuli.
Mottram (2002)	"The major mechanistic pathway for the formation of acrylamide in foods so far established is via the Maillard reaction."
Theory	The Maillard reaction mechanism accounts for acrylamide formation in high-starch foods during cooking at high temperatures.
Loreau (2001)	"Many studies were focused on so called biodiversity effects, i.e., the way in which diversity affects ecosystem function and services."
Theory	Plant diversity is crucial for maintaining the function and stability of ecosystems.
Alexander (2000)	"Bioavailability and toxicity of organic chemicals in soil can change over time."
Theory	The aging of contaminated sediment and soil reduces bioavailability of pollutants to microorganisms due to sequestration.
Adachi (2001)	"Due to the ability to harvest both singlet and triplet excitons, phosphorescent organic light emitting devices can have 100% internal quantum efficiency."
Theory	The internal quantum efficiency of the OLED devices can be greatly enhanced approaching 100%.
Das (2003)	"From the investigations in the past decade, nanofluids were found to exhibit significantly higher thermal properties, in particular, thermal conductivity, than those of base fluids."
Theory	In a nanofluid, thermal conductivity enhancement can be explained based on the stochastic or Brownian motion of the nanoparticles.
Aharony (2000)	"The AdS/CFT correspondence asserts there is an equivalence between a gravitational theory in the bulk and a conformal field theory in the boundary."
Theory	The anti-de Sitter/conformal field theory conjecture postulates a duality between field theories and Type IIB string theory in various geometries.
Berkman (2000)	"Structural and functional characteristics of social networks influence health via several other pathways."
Theory	Social support theory deals with the various sources of positive or protective influences associated with an individual's social relationship and network.

Table 7. (*continued*)

Highly cited paper	Typical citing sentence for the causal pair/statement of theory
Cardinal (2001)	"In animal studies, lesions in the ventral striatum or in specific regions within the orbitofrontal cortex have been shown to increase impulsivity."
Theory	The nucleus accumbens is involved in codifying and computing the value of future rewards and therefore acts as a driving force to perform goal-directed actions.
Blood (2001)	"Music activates brain regions involved in reward and emotion and can provoke intensely pleasurable responses in these areas."
Theory	Chills that occur in response to preferred music are partly mediated by reward-associated brain regions, which are similarly activated by sex and addictive drugs.

distinct number of sentences containing the phrase pair. In determining these counts, the cause-and-effect phrases were searched using wildcards so that variants could be retrieved. For example, for Cardinal (2001) in Table 1 the search was for "*brain lesion*" and "*impulsiv*". The counts for verb separated phrases are divided between the cause coming before effect (F = forward) and after the effect (B = backwards). The sum of F + B can be less or greater than the distinct sentence counts (given in the last column) because a pair can repeat within a sentence, which makes the count higher, or not be separated by a verb, which makes the count lower.

In 7 of 10 cases, the forward count exceeds the backward count, meaning the cause usually precedes the effect in the sentences. In most cases, the causal direction is clear, even if the effect precedes the cause, such as in the case of acrylamide caused by the Maillard reaction. The main exception is the theoretical physics paper Aharony et al. (2000) on string theory, where the causal direction is not clear. In this case both the cause and effect ("ads/cft" → "boundary") are theoretical constructs that are mathematically related. Whether our analysis can apply to such cases remains to be seen.

Table 7 gives examples of citing sentences illustrating the principal causal phrases in Table 6. Instances of effects preceding causes in the sentences are Mottram et al. (2002) and Alexander et al. (2000). Table 7 also gives a one-sentence summary of the theory that underlies the causal phrases in Table 6. These summaries are manually constructed by scanning a large sample of citing sentences for each paper. The summaries enable the specific causal connections in Table 6 to be seen in the context of a more general theory. For example, TRPV1 is just one type of receptor for pain perception.

The aim of the analysis is to compute a likelihood ratio $P(E|T)/P(E|{\sim}T)$, as defined in Section 3.2, for each of the cause/effect relations in Table 6 that determines whether the causal connection is confirmed by sentiment analysis. Hence, we are dealing with simple causal patterns $A \rightarrow B$, disregarding other factors that might impinge on either B or A or other effects that might flow from them. The approach is to approximate the conditional probabilities $P(E|T)$ and $P(E|{\sim}T)$ by computing the "supporting evidence" and "uncertainty" sentiments respectively.

The data for this calculation are shown in Table 8. Each paper is represented by two rows, the first of which is data on the subset of citing sentences containing the cause-effect or theory-evidence phrase pair, and the second is data on all the citing sentences for the highly cited paper which serves as the baseline for the phrase pair. We start with the number of citing sentences containing the phrase pair shown in the column headed "Total citances." The next

Table 8. Computing confirmation based on citing sentence sentiments for the 10 highly cited papers. Each paper is represented by two rows: The first row is data on the subset of citing sentences containing the causal phrase pair and the second row is data on all citing sentences for the individual highly cited paper which serves as the baseline for the phrase pair. The column labeled "Norm evid wrt paper baseline" divides the "Percent evidence" for the causal pair by the "Percent evidence" for the paper in the following row. The "Confirm" column is "Yes" if the "Norm evid wrt paper baseline" exceeds the "Norm uncert wrt paper baseline" and "No" if it does not

Paper	Causal pair	Total citances	Evidence citances	Percent evidence	Uncertain citances	Percent uncertain	Norm evid. wrt paper baseline	Norm uncert. wrt paper baseline	Confirm
Caterina (2000)	TRPV1 → heat	99	43	43.4	36	36.4	1.07	1.32	No
	paper baseline	411	167	40.6	113	27.5			
Mottram (2002)	maillard → acrylamide	152	46	30.3	20	13.2	0.97	0.71	Yes
	paper baseline	399	125	31.3	74	18.5			
Loreau (2001)	biodiversity → ecosystem	100	22	22.0	30	30.0	0.82	0.81	Yes
	paper baseline	406	109	26.8	151	37.2			
Alexander (2000)	time → bioavailability	46	15	33.6	12	26.1	1.29	0.69	Yes
	paper baseline	395	100	25.3	150	38.0			
Adachi (2001)	excitons → quantum efficiency	56	7	12.5	6	10.7	0.51	0.97	No
	paper baseline	560	137	24.5	62	11.1			

Table 8. (*continued*)

Paper	Causal pair	Total citances	Evidence citances	Percent evidence	Uncertain citances	Percent uncertain	Norm evid. wrt paper baseline	Norm uncert. wrt paper baseline	Confirm
Das (2003)	nanofluid → thermal conductivity	283	196	69.3	29	10.2	1.11	0.86	Yes
	paper baseline	598	373	62.4	71	11.9			
Aharony (2000)	ads/cft → boundary	30	5	16.7	1	3.3	1.00	0.21	Yes
	paper baseline	480	80	16.7	77	16.0			
Berkman (2000)	social network → health	50	10	20.0	15	30.0	0.92	0.87	Yes
	paper baseline	349	76	21.8	120	34.4			
Cardinal (2001)	lesions → impulsive	71	39	54.9	37	52.1	1.09	1.15	No
	paper baseline	326	164	50.3	148	45.4			
Blood (2001)	music → reward	76	50	65.8	19	25.0	1.04	0.94	Yes
	paper baseline	323	205	63.5	86	26.6			

column, labeled "Evidence citances," is a count of the sentences containing the "supporting evidence" sentiment words, followed by its percentage of the total citances.

The count for the "Uncertain citances" and "Percent uncertain" follow. The columns labeled "Norm evid wrt paper baseline" and "Norm uncert wrt paper baseline" are the "Evidence" and "Uncertainty" percentages for the causal pair divided by the corresponding percentages for the paper as a whole given in the row immediately below it labeled "Paper baseline." Hence, the total citances for the paper serve as a reference baseline for the specific causal pair derived from it. This preserves the topic focus as well as compensating for any over- or underuse of specific sentiment words in the topic.

The relative magnitudes of these two normalized percentages determine the likelihood ratio under the assumptions we are using on the interpretations of the sentiments. If the normalized supporting evidence sentiment is greater than the normalized uncertainty, the causal pair is confirmed. This is indicated by a "Yes" or "No" in the last column labeled "Confirm." In Table 8 it is interesting to note that in eight of 10 cases the evidence sentiment outweighs the uncertainty, but following normalization, five of 10 cases show a reversal of sentiments where the dominant sentiment prior to normalization is reversed after normalization.

We also note that three of the 10 causal relations are disconfirmed because the uncertainty outweighs the evidence, including "TRPV1 → heat" from the Caterina et al. (2000) paper. However, another prominent causal link for Caterina et al. (2000), not shown in Table 6, namely "TRPV1 → capsaicin" (the sensation of capsaicin) is confirmed, so confirmation can vary from link to link within a given paper. The explanation of why "TRPV1 → heat" is disconfirmed is more subtle. It turns out that the response of the receptor depends on the temperature of the stimuli as made clear by the following citance: "Even though there is no doubt that TRPV1 mediates thermal pain, the presence of additional heat sensors was suggested due to the fact that TRPV1 knock-out mice still exhibited residual nociceptive behaviors to noxious thermal stimuli." In other words, suppressing the receptor did not eliminate the sensation of extreme or noxious heat. We will see later on (in Table 8) that when compared to a cluster of papers on nociception, this distinction between moderate and noxious heat is diminished and the causal link is confirmed. Hence, confirmation can also depend on the scope of the corpus.

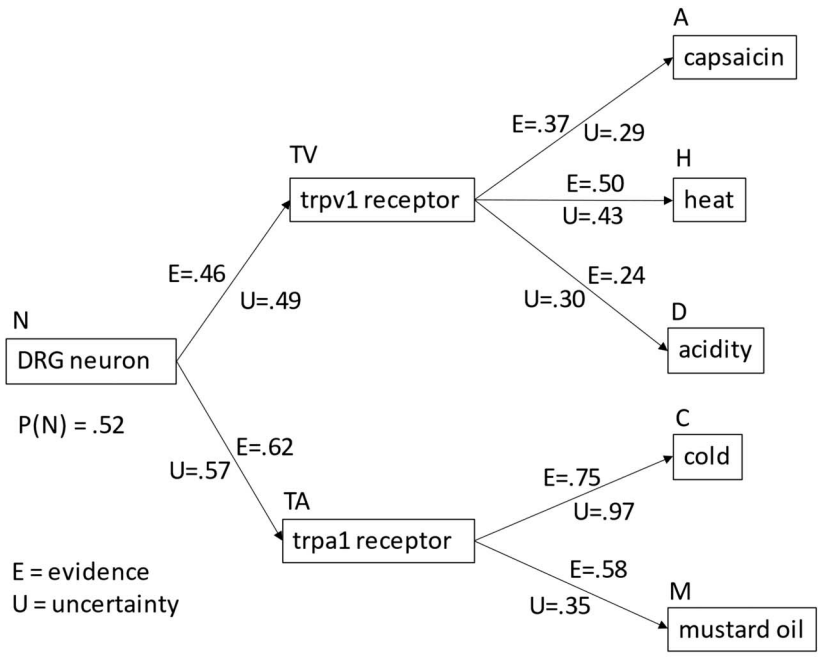
4.2. Computing Confirmation for a Network

Each of the cause/effect assertions in Table 6 can be considered a simple one link networks A → B which have an exact solution using Bayes's theorem. However, when multiple causal links are connected in a network, an exact solution is not possible, and an algorithm is required that iteratively exchanges information between nodes until the network converges to a stable solution.

A network was created by merging the citances for the top 20 papers from the nociception cluster from the SciTech Strategies model. Noun phrase pairs were created as described above for the combined citances. Table 3 showed that TRPV1 and TRPA1 receptors were involved in multiple prominent causal assertions, leading to the sensations of heat, cold, acidity, capsaicin, mustard oil, and other agents. Citances also revealed that the two receptors had a common origin in neurons, as indicated by the following citance:

> "The TRPA1 channel is found in a subset of rat DRG neurons in which it is co-expressed with the TRPV1, but not the TRPM8 channel."

This led to a linking together of seven causal assertions to form the directed acyclic graph (DAG) in Figure 2. The causal network involved eight nodes, starting with a "neuron" node on



$$P(N)*P(TV|N)*P(TA|N)*P(A|TV)*P(H|TV)*P(D|TV)*P(C|TA)*P(M|TA) = .54$$

Figure 2. The causal network for seven nociception links and eight nodes, starting with a "neuron" node on the left and progressing to the sensations evoked on the right via two receptor types. Nodes are labeled with upper case letters. Each link is coded by two condition probabilities, *E* and *U*, derived from evidence and uncertainty sentiments. The joint probability distribution expression based on the "chain rule" for the network is shown below the network, as is the final $P(T|E)$ value of 0.54 which is an average of 20 runs using Bnlearn software using the "logic sampling" option.

the left, and progressing to the sensations evoked on the right via two receptor types: TRPV1 and TRPA1. In contrast to the simple $A \rightarrow B$ pattern, here an effect can act as a cause leading to another effect, creating causal chains. In Figure 2 we also give the formula for so-called "joint probability distribution" for the network, which is a product of conditional probabilities for every link in the network following the "chain rule" of probabilities. The first term in this expression is the prior probability of the initial node $P(N)$ where *N* stands for neuron. Following terms are conditional probabilities each of which corresponds to an arrow in the network of the form $P(\text{effect} \mid \text{cause})$.

Our aim is to compute the probability that the network is confirmed as a representation of a theory of nociception based on the sentiments of the citing authors. Thus, we need to compute, as before, two conditional probabilities for each link in addition to the prior probability for the initial node in Figure 2 and input these into the Bnlearn software. Table 9 shows how these numbers were calculated. As a baseline we use the cumulated citances for the cluster, rather than the citances for individual papers, as in Table 8. This baseline is shown in the second row of Table 9. Beginning in the fourth row we give data for each separate link in the network computed in the same manner as in Table 8 except that the columns headed "Norm evid wrt cluster" and "Norm uncert wrt cluster" show the sentiment rates divided by the cluster baseline. The columns headed "rescale" divide each normalized value by a constant (= 2.2) so that their values will fall between 0 and 1, as required by probabilities. The scaled values are labeled as *E* for evidence and *U* for uncertainty on Figure 2 and are the values input into the

Table 9. Computing confirmation based on citing sentence sentiments for the network of Figure 2. The second row in the table labeled "Cluster baseline" contains sentiment counts for the aggregate citances for the top 20 papers in the cluster listed in Table 2. Beginning in the fourth row, each link of the network of Figure 2 is listed. The columns labeled "Norm evid wrt cluster" and "Norm uncert wrt cluster" divide the "Percent evidence citances" and the "Percent uncertain citances" by the values of the respective cluster baselines in the second row. The two "Rescale" columns divide the normalized evidence and uncertainty percentages by a constant of 2.2 so that the normalized values fall within the 0–1 interval required by probabilities. The last row in the table shows the computation of the prior probability for the leftmost node in the network of Figure 2, $P(N)$. This is based on the uncertainty of "neuron" citances, normalized and rescaled as above, and subtracted from 1 to get a certainty value

	Total citances	Evidence citances	Percent evidence citations	Uncertain citances	Percent uncertain citances	Norm evid wrt cluster	Norm uncert wrt cluster	Rescale evid wrt cluster (E)	Rescale uncert wrt cluster (U)	Confirm
Cluster baseline	4,752	1,173	24.7	964	20.3					
				Causal pair						
DRG neuron → TRPV1	428	106	24.8	93	21.7	1.00	1.07	0.46	0.49	No
TRPV1 → capsaicin	615	123	20.0	79	12.8	0.81	0.63	0.37	0.29	Yes
TRPV1 → heat	687	186	27.1	133	19.4	1.10	0.95	0.50	0.43	Yes
TRPV1 → acid	151	20	13.2	20	13.2	0.54	0.65	0.24	0.30	No
DRG neuron → TRPA1	247	83	33.6	63	25.5	1.36	1.26	0.62	0.57	Yes
TRPA1 → cold	252	103	40.9	109	43.3	1.66	2.13	0.75	0.97	No
TRPA1 → mustard oil	172	54	31.4	27	15.7	1.27	0.77	0.58	0.35	Yes
DRG neuron (1-prior prob)	846			181	21.4		1.05		0.48	

software. It was found that confirmation was not sensitive to the value of the scaling constant and $P(T)$ and $P(T|E)$ were both shifted up or down proportionally.

The last row in Table 9 shows how the prior probability of $P(N)$ is computed. As discussed previously, we base this on the uncertainty sentiment which is computed for citances containing the terms "DRG [or trigeminal] neuron." The prior is also subject to the same normalization and rescaling applied to the conditional probabilities. The final number 0.48 must, however, be subtracted from 1 to convert it to a probability of certainty rather than one of uncertainty, hence the value of 0.52 = (1 − 0.48) in Figure 2.

The last column in Table 9 shows that four of the individual links were confirming based on the likelihood ratio. Running the full network using the Bnlearn software gives a probability of 0.54 (an average of 20 separate runs using the "logic sampling" option), which thus narrowly confirms the network with respect to the prior of 0.52. Similar to the individual links in Table 8, in five of seven links in Table 9 the evidence outweighs the uncertainty and the links are confirmed. Only one of the seven links changes the dominant sentiment after normalization. One of the two disconfirmed links in Figure 2 is the "TRPA1 receptor" leading to the sensation of "cold." Examining the citances for this link we find statements like "noxious cold activation of TRPA1 is somewhat controversial," which perhaps explains why this link is not confirmed. However, the two disconfirming links were not strong enough to disconfirm the full network.

5. DISCUSSION

The next step in this research is to automate the formation of as many causal networks as possible using the cumulative citances for a cluster of papers. This involves linking up as many causal word phrase pairs as possible given some threshold or limit on pair frequency. Two main problems remain to be solved. First, we need a systematic criterion for differentiating which member of the pair is the cause/theory and which is the effect/evidence. Second, when computing sentiments, we need to normalize the different presentations of cause-and-effect phrases which we have done here based on wildcard searching. But the synonym problem remains to be addressed. A possible solution to the first problem is to take the more uncertain entity of the pair as the cause or theory and the more certain entity of the pair as the effect or evidence.

Regarding the measuring of sentiments, there is also the need to expand and sharpen the lists of evidence and uncertainty cue words. The list of terms denoting evidence was a mix of words indicating the effort to obtain evidence, such as *study* or *experiment,* in addition to words indicating that supporting evidence was found, such as *determined* or *shown.* The uncertainty words represented only a small sample of possible ways of expressing this sentiment (Chen & Song, 2018). The normalization procedure of dividing the evidence and uncertainty rates for cause-effect pairs by paper or cluster baselines may, to some extent, compensate for the incompleteness of the cue word sets, but results at this stage must be considered tentative. A related problem is misclassification. The lower precision rates for some cue words mean that misclassifications will inevitably occur. Another issue is failure to classify, which is indicated by low recall rates, particularly for uncertainty words. This calls for the broadening of the uncertainty cue word set.

A question yet to be examined is whether confirmation changes over time, as Chen and Song have shown for the uncertainty of predications. For some papers we have 18 years of citing sentences, which could be subdivided by citing years to see if the confirmation status of a particular cause/effect relation changed from period to period. No doubt slicing the time periods too narrowly would lead to random fluctuations in the ratio of evidence and uncertainty sentiments. Such a community-based confirmation measure should be more stable than

an individual participant's perception, which in real time might fluctuate from day to day as new evidence comes to light.

Another fundamental question relates to how we have used the uncertainty of the theory as a proxy for the probability of an alternative theory explaining the evidence $P(E|{\sim}T)$, assuming, in effect, that uncertainty is due to the existence of alternative or competing theories. This makes confirmation a balancing act of supporting evidence versus uncertainty. However, it is important to develop a more direct way of estimating the probability of an alternative theory.

Some perspective is offered by the history of science. In most research programs, the DNA history included, investigators move from one theory to another sometimes over a series of years (Small, 1971). These can be denoted as T_1, T_2, T_3, ..., and so on. In the case of DNA, the Pauling triple helix might be T_1 and Watson's like-with-like base pairing model T_2, with T_3 their final published model. According to Crick, the debate about whether their model for DNA was correct continued for nearly 25 years, with a number of alternative models suggested and rejected (Crick, 1988, 73). From a Bayesian perspective, each theory must be evaluated on its own merits based on its fit with evidence. But precursor theories can serve as alternative or competing theories, which are needed for Bayes's theorem to work. $P(E|{\sim}T)$ is, in fact, the sum of all mutually exclusive alternative theories, published or unpublished, which can have varying degrees of fit with the evidence. This argues for a nonzero floor or minimum $P(E|{\sim}T)$ even if T_1 is merely an uninformed initial hunch.

In the case of nociception, David Julius in his Nobel lecture (2021) briefly alludes to a competing theory that the capsaicin receptor, rather than being a specific molecular entity that acted as an ion channel, was due to integrating capsaicin into the cell membrane to form an ion channel that functioned nonselectively. This set off what he referred to as the "Holy Grail" of pain research: the search for the molecular capsaicin receptor. Michael Caterina in Julius's lab succeeded in cloning genes from neurons and those genes stimulated fibroblast cell cultures to express the receptor and respond to capsaicin (Caterina et al., 1997). Julius describes this as a "Eureka moment."

A 1995 paper describing a competing hypothesis that capsaicin had created the receptor was found in the STS5-769 direct citation cluster. In addition, this paper was cited in the 1997 discovery paper (Caterina et al., 1997) as a previously "proposed model," and by examining its citances we could perhaps assess its degree of support or uncertainty. This suggests that a good way to find competing theories is to look at the references made by the discovery team itself, as social norms call for citing competing theories. Obviously, this approach works only when the competing theory corresponds to a published paper.

Many writers on science have concluded that discovery in science is spurred by chance occurrences or serendipity. For example, Francis Crick claimed that Watson's discovery of base pairing in DNA was due in part to luck (Crick, 1988, p. 65). Similarly, Hall (1954, p. 125) stated that Kepler accidentally noticed that an ellipse fitted the orbit of Mars using Tycho's observations and Koestler (1964, p. 112) attributed Pasteur's discovery of vaccination for chicken cholera in part to chance. The discovery process may be initiated by a novel observation (some chickens did not get cholera), an inconsistency in theory (Einstein's theory of relativity), or even a dream (Kekulé's structure of benzene). Whatever inspires the hypothesis, once it is generated a long process of critical evaluation begins. The evaluation can spur new experiments, or modifications of the theory. The discoverer may only reluctantly ask whether there are competing theories due to his or her interests in priority. Whether we take the point of view of the individual scientist or the collective view of a community, the evaluation needs to look for positive and negative evidence as well as alternative explanations.

The question of time slicing raises an interesting question if we view the discovery and confirmation process as a series of random events. This contrasts with the empiricist notion that discovery is a systematic process of working backwards from the evidence to the theory (Losee, 1972, p. 103; Popper, 1962; Schindler, 2008). Reading Watson's account of the discovery of the structure of DNA, we see almost day-to-day swings in confidence as Watson and Crick are buffeted by incoming evidence and theoretical insights favoring one model or another. For example, Linus Pauling's triple helix model is rejected (Watson, 1968, p. 160). Watson's own like-to-like base pairing model was rejected because he had used the wrong tautomeric form for two of the bases, and Crick also objected that it would violate the Chargaff rules (Olby, 1974, p. 412). The final model of two right-handed helices with unique base pairings between them satisfied all the objections and fit with the available evidence so well that Watson proclaimed: "a structure this pretty just had to exist" (Watson, 1968, p. 205). In Bayesian terms we could ascribe this feeling to a large jump in $P(E|T)$ leading to a similar jump in $P(T|E)$ versus the prior $P(T)$ where T is the double helix. Likewise, the ups and downs of the other models could be interpreted as incremental changes in probabilities $P(E|T)$ or $P(E|\sim T)$ depending on the evidence at hand. The day-to-day swings in confidence experienced by Watson and Crick are analogous to the precarious balance of supporting evidence and uncertainty proposed in this paper as expressed by the likelihood ratio.

Whether such a qualitative application of Bayes's theorem is possible based on historical examples is beyond the scope of this paper. If we are correct, then Eureka or "aha" moments are indicators of shifts in the prior vis-à-vis posterior probabilities of a theory. We further assume that these moments will continue to occur randomly during the extended process of confirmation, including disappointing moments of disconfirmation. The personal and subjective point of view of Watson contrasts with the method used in this paper based on citing sentences from a community of peers. The latter is by contrast a delayed, retrospective reaction. In the long run we might expect a convergence of opinion between the subjective view of the discoverer and the collective perspectives of the community. But given the different interests of these parties, it would not be surprising to see differences. A discoverer who expends considerable effort to support the validity of a knowledge claim would be expected to take a more sanguine view of the evidence than a peer group with competing interests in an alternative theory.

6. CONCLUSIONS

This paper proposes a network model of confirmation in science based on cause-and-effect linkages interpreted as theory and evidence connections. The model is a hybrid citation and language approach that draws on citing sentences for single papers or clusters of papers. This combines the capability of citation-based clustering methods to defined specialty areas with the in-depth conceptual-level detail afforded by textual and linguistic methods to identify cause-effect linkages. The present paper points to the possibility of using Bayes's rule to understand the process of confirmation.

The use of citation context sentiments for computing conditional probabilities is attempted for the first time, but issues remain, particularly regarding the evaluation of competing theories. This problem might be resolved if competing theories have been published and their citances analyzed, reducing confirmation to a comparison of sentiments for competing published theories.

It is interesting that Kuhn argued against the Bayesian approach to theory choice, because he maintained that scientists in historical contexts used a variety of subjective criteria (Kuhn, 1977; Salmon, 1990). For example, he argued that a phlogiston theorist might prefer their theory over the oxygen theory because it explained the "similarity" of metals, all of which

contained phlogiston. At the same time, there was widespread acceptance of oxygen's explanation of weight gain of calxes. On the other hand, an oxygen theorist might argue that the similarity of metals was due to the absence of oxygen. A Bayesian might say that these divergent criteria would have simply offset one another and at worst delayed the decision in favor of the oxygen theory until further evidence emerged.

The "no miracles" argument, attributed to the realist philosopher Hilary Putnam (1975, p. 73), says that the striking agreement between theory and evidence sometimes achieved in modern science would not be possible unless the underlying theory was true (Howson & Urbach, 2006, p. 26). The Bayesian, on the other hand, would point to the improbability of a close fit between theory and evidence and the resulting higher probability of the theory being true given the evidence, but no possibility of absolute truth as long as there are alternative theories. Arthur Koestler in his classic book *The Act of Creation* (1964) talks about the "Eureka" moment when two seemingly unrelated events come together for which he coins the term "bisociate"—the transition from thinking something is unlikely to seeing that it works. Such moments occur when theory closely fits with evidence, for example, when James Watson lines up the molecular models of the DNA base pairs, or when Caterina and Julius clone the capsaicin receptor.

Assuming "Eureka" moments occur randomly during the course of theory testing means that conditional probabilities are incremented or decremented as the scientific community critically examines and refines the theory's and its competitor's fit with the evidence. Thus, a theory's confirmation status will remain in flux for an extended period. Clearly, a community and citation-based assessment, as we have outlined here, filtered through cool scientific prose, lacks the emotional impact of the "Eureka" or "aha" moment. A challenge for future research is to show how the force of a sudden change in a theory's probability, such as a discovery, is communicated to the community and reflected in citing sentences.

ACKNOWLEDGMENTS

I would like to thank Nees Van Eck of CWTS and Kevin Boyack of SciTech Strategies, Inc. for providing citation context and cluster data, Mike Patek of SciTech Strategies for programming, and Harriet Noble for assistance in citation sentiment coding. Two anonymous referees provided detailed comments which were very helpful.

COMPETING INTERESTS

The author has no competing interests.

FUNDING INFORMATION

No funding has been received for this research.

DATA AVAILABILITY

Data are available from the author.

REFERENCES

Atkinson, J., & Rivas, A. (2008). Discovering novel causal patterns from biomedical natural-language texts using Bayesian nets. *IEEE Transactions on Information Technology in Biomedicine*, *12*(6), 714–722. https://doi.org/10.1109/TITB.2008.920793, PubMed: 19000950

Boyack, K. W., Van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73. https://doi.org/10.1016/j.joi.2017.11.005

Bunge, M. (1963). *Causality: The place of the causal principle in modern science*. Cleveland: Meridian Books.

Caterina, M. J., Leffler, A., Malmberg, A. B., Martin, W. J., Trafton, J., ... Julius, D. (2000). Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science*, *288*(5464),

306–313. https://doi.org/10.1126/science.288.5464.306, PubMed: 10764638
Caterina, M. J., Schumacher, M. A., Tominaga, M., Rosen, T. A., Levine, J. D., & Julius, D. (1997). The capsaicin receptor: A heat-activated ion channel in the pain pathway. *Nature*, *389*(6653), 816–827. https://doi.org/10.1038/39807, PubMed: 9349813
Chen, C., & Song, M. (2018). *Representing scientific knowledge: The role of uncertainty*. London: Springer. https://doi.org/10.1007/978-3-319-62543-0
Chen, C., Song, M., & Heo, G. E. (2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics*, *12*(1), 158–180. https://doi.org/10.1016/j.joi.2017.12.004
Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220. https://doi.org/10.1037/h0026256, PubMed: 19673146
Cold fusion. (2021, December 10). In *Wikipedia*. https://en.wikipedia.org/wiki/Cold_fusion.
Crick, F. (1988). *This mad pursuit: A personal view of scientific discovery*. New York: Basic Books.
Findler, N. V., & Bickmore, T. (1996). On the concept of causality and a causal modeling system for scientific and engineering domains, CAMUS. *Applied Artificial Intelligence*, *10*(5), 455–487. https://doi.org/10.1080/088395196118506
Glymour, C. (1980). *Theory and evidence*. Princeton, NJ: Princeton University Press.
Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *British Medical Journal*, *339*, b2680. https://doi.org/10.1136/bmj.b2680, PubMed: 19622839
Hall, A. R. (1954). *The scientific revolution 1500–1800: The formation of the modern scientific attitude* (2nd edn). Boston: Beacon Press.
Hanson, N. R. (1972). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge: Cambridge University Press.
Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach*. Chicago: Open Court Publishing Co.
Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor: The University of Michigan Press.
Ihde, A. J. (1964). *The development of modern chemistry* (Chapter 3). New York: Harper & Row.
Julius, D. (2021, December 7). *From peppers to peppermints: Insights into thermosensation and pain*. https://www.nobelprize.org/prizes/medicine/2021/julius/lecture/
Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics Applications Note*, *28*(23), 3158–3160. https://doi.org/10.1093/bioinformatics/bts591, PubMed: 23044550
Klavans, R., Boyack, K. W., & Murdick, D. A. (2020). A novel approach to predicting exceptional growth in research. *PLOS ONE*, *15*(9), e0239177. https://doi.org/10.1371/journal.pone.0239177, PubMed: 32931500
Koestler, A. (1964). *The act of creation*. London: Penguin.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Kuhn, T. S. (1977). Objectivity, value judgment and theory choice. In *The essential tension* (pp. 320–339). Chicago: University of Chicago Press.
Larmers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., ... Murray, D. (2021). Investigating disagreement in the scientific literature. *eLife*, *10*, e72737. https://doi.org/10.7554/eLife.72737, PubMed: 34951588
Li, Z., Li, Q., Zou, X., & Ren, J. (2021a). Causal extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing*, *423*, 207–219. https://doi.org/10.1016/j.neucom.2020.08.078
Li, X., Peng, S., & Du, J. (2021b). Towards medical knowmetrics: Representing and computing medical knowledge using semantic predications as the knowledge unit and the uncertainty as the knowledge context. *Scientometrics*, *126*, 6225–6251. https://doi.org/10.1007/s11192-021-03880-8, PubMed: 33612884
Losee, J. (1972). *A historical introduction to the philosophy of science*. London: Oxford University Press.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 3111–3119).
Moayedi, M., & Davis, K. D. (2013). Theories of pain: from specificity to gate control. *Neurophysiology*, *109*(1), 5–12. https://doi.org/10.1152/jn.00457.2012, PubMed: 23034364
Nagarajan, R., Scutari, M., & Lebre, S. (2013). *Bayesian networks in R with applications in systems biology*. New York: Springer. https://doi.org/10.1007/978-1-4614-6446-4
Nakov, P., Schwartz, A., & Hearst, M. (2004). Citances: Citation sentences for semantic analysis of bioscience text. *SIGIR Workshop of Search and Discovery on Bioinformatics*.
Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., ... Rife, S. C. (2021). scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, *2*(3), 882–898. https://doi.org/10.1162/qss_a_00146
Olby, R. (1974). *The path to the double helix*. Seattle: University of Washington Press.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Pearl, J., & Mackenzie, D. (2018). *The book of why*. New York: Basic Books.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge* (Chapter 8). New York: Basic Books.
Putnam, H. (1975). *Collected papers: Mathematics, matter and method* (Vol. 1). Cambridge: Cambridge University Press.
Rindflesch, T. C., Kilicoglu, H., Fiszman, M., Rosemblat, G., & Shin, D. (2011) Semantic MEDLINE: An advanced information management application for biomedicine. *Information Services & Use*, *31*, 15–21. https://doi.org/10.3233/ISU-2011-0627
Salmon, W. C. (1990). *Rationality and objectivity in science, or, Tom Kuhn meets Tom Bayes*. University of Minnesota Press, Minneapolis. Retrieved from the University of Minnesota Digital Conservancy: https://hdl.handle.net/11299/185726
Schindler, S. (2008). Model, theory and evidence in the discovery of the DNA structure. *British Journal for the Philosophy of Science*, *59*(4), 619–658. https://doi.org/10.1093/bjps/axn030
Small, H. (1971). *The helium atom in the old quantum theory* (doctoral dissertation). University of Wisconsin, ProQuest #7125217.
Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, *8*, 327–340. https://doi.org/10.1177/030631277800800305
Small, H. (2020). Past as prologue: Approaches to the study of confirmation in science. *Quantitative Science Studies*, *1*(3), 1025–1040. https://doi.org/10.1162/qss_a_00063

Small, H. (2021). From citing sentences to causal networks: The causality index. In W. Glanzel, S. Heefer, P.-S. Chi, & R. Rousseau (Eds.), *Proceedings of the 18th Conference on Scientometrics and Informetrics: ISSI2021* (pp. 1039–1044).

Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, *11*, 46–62. https://doi.org/10.1016/j.joi.2016.11.001

Sobrino, A., Olivas, J. A., & Puente, C. (2010). Causality and imperfect causality from texts: A frame for causality in social sciences. *International Conference on Fuzzy Systems* (pp. 1–8). Barcelona: IEEE. https://doi.org/10.1109/FUZZY.2010.5584863

Swanson, D. R. (1986). Undiscovered public knowledge. *Library Quarterly*, *56*(2), 103–118. https://doi.org/10.1086/601720

Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9780691186672

Thilakaratne, M., Falkner, K., & Atapattu, T. (2019). A systematic review on literature-based discovery: General overview, methodology, & statistical analysis. *ACM Computing Surveys*, *52*(6), Article 129. https://doi.org/10.1145/3365756

Traag, V. A., Waltman, L., & Van Eck, N.-J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*, 5233. https://doi.org/10.1038/s41598-019-41695-z, PubMed: 30914743

Trieu, H.-L., Tran, T. T., Duong, K. N. A., Nguyen, A., Miwa, M., & Ananiadou, S. (2020). DeepEventMine: End-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, *36*(19), 4910–4917. https://doi.org/10.1093/bioinformatics/btaa540, PubMed: 33141147

Watson, J. D. (1968). *The double helix: A personal account of the discovery of the structure of DNA*. New York: Atheneum. https://doi.org/10.1063/1.3035117