Development of Similarity Measures From Graph-Structured Bibliographic Metadata: An Application to Identify Scientific Convergence

IEEE Transactions on Engineering Management

Suggested Citation: Melnychuk, Tetyana et al. (2023) : Development of Similarity Measures From Graph-Structured Bibliographic Metadata: An Application to Identify Scientific Convergence, IEEE Transactions on Engineering Management, ISSN 1558-0040, IEEE, Piscataway, NJ, Iss. (early access), pp. 1-17, https://doi.org/10.1109/TEM.2023.3308008

This Version is available at:
http://hdl.handle.net/11108/582

Development of Similarity Measures From Graph-Structured Bibliographic Metadata: An Application to Identify Scientific Convergence

Tetyana Melnychuk, Lukas Galke, Eva Seidlmayer, Stefanie Bröring, Konrad U. Förstner, Klaus Tochtermann, and Carsten Schultz

***Abstract*—Scientific convergence is a phenomenon where the distance between hitherto distinct scientific fields narrows and the fields gradually overlap over time. It is creating important potential for research, development, and innovation. Although scientific convergence is crucial for the development of radically new technology, the identification of emerging scientific convergence is particularly difficult since the underlying knowledge flows are rather fuzzy and unstable in the early convergence stage. Nevertheless, novel scientific publications emerging at the intersection of different knowledge fields may reflect convergence processes. Thus, in this article, we exploit the growing number of research and digital libraries providing bibliographic metadata to propose an automated analysis of science dynamics. We utilize and adapt machine-learning methods (DeepWalk) to automatically learn a similarity measure between scientific fields from graphs constructed on bibliographic metadata. With a time-based perspective, we apply our approach to analyze the trajectories of evolving similarities between scientific fields. We validate the learned similarity measure by evaluating it within the well-explored case of cholesterol-lowering ingredients in which scientific convergence between the distinct scientific fields of nutrition and pharmaceuticals has partially taken place. Our results confirm that the similarity trajectories learned by our approach resemble the expected behavior, indicating that our approach may allow researchers and practitioners to detect and predict scientific convergence early.**

***Index Terms*—Data enrichment, machine learning, network analysis, science dynamics, scientific convergence, similarity indicator.**

Manuscript received 24 October 2022; revised 15 March 2023 and 23 June 2023; accepted 8 August 2023. This work was supported by the German Federal Ministry of Education and Research (BMBF) represented by the executing agency the German Aerospace Center (DLR) within the framework of the Research Project Q-AKTIV under Grant 01PU17013A, Grant 01PU17013B, and Grant 01PU17013C within the funding line "Quantitative Research on the Science Sector." Review of this manuscript was arranged by Department Editor G. Marzi. *(Corresponding author: Tetyana Melnychuk.)*

Tetyana Melnychuk and Carsten Schultz are with the Kiel Institute for Responsible Innovation, Kiel University, 24118 Kiel, Germany (e-mail: melnychuk@bwl.uni-kiel.de; schultz@bwl.uni-kiel.de).

Lukas Galke was with the Leibniz Information Centre for Economics (ZBW), 24105 Kiel, Germany. He is now with the Max Planck Institute for Psycholinguistics, 6525XD Nijmegen, The Netherlands (e-mail: lukas.galke@mpi.nl).

Eva Seidlmayer and Konrad U. Förstner are with the Information Centre for Life Sciences (ZB MED), 50931 Cologne, Germany (e-mail: seidlmayer@zbmed.de; foerstner@zbmed.de).

Stefanie Bröring is with the Faculty of Management and Economics, Ruhr-University Bochum, 44780 Bochum, Germany (e-mail: stefanie.broering@ruhr-uni-bochum.de).

Klaus Tochtermann is with the Leibniz Information Centre for Economics (ZBW), 24105 Kiel, Germany (e-mail: k.tochtermann@zbw-online.eu).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TEM.2023.3308008.

Digital Object Identifier 10.1109/TEM.2023.3308008

I. Introduction

IN ORDER to be prepared for future developments in technology, market, and industry, it is crucial for research organizations, funding bodies, and companies to identify scientific convergence at the front end of technology development. The present study contributes by suggesting a validated method for the early detection of scientific convergence.

Scientific convergence can be witnessed in numerous examples, such as "NanoBioInfoTech," bioinformatics, cosmeceuticals, nutraceuticals and functional foods, information and communication technology, and visual analytics [1], [2], [3], [4], [5], [6], and presents a phenomenon of growing relevance. It refers to the decreasing distance and gradual overlap of different scientific fields—as distinct fields of specialized knowledge and research—and is mirrored by intensified cross-disciplinary research [4], [7], [8]. Scientific convergence results in a blurring of the boundaries of previously distinct scientific fields and, as such, in the emergence of a novel cross-disciplinary field of science [3], [9], [10]. A combination of knowledge from distinct scientific fields can result in new technologies [11] and may lead to radical innovations and to technology, market, and industry convergence [7], [11], [12], [13], [14]. Early identification of scientific convergence helps to drive collaborations with new partners with complementary capabilities during the early phases of research and development (R&D) [7], [9], [12]. Therefore, early detection of scientific convergence creates a competitive advantage for firms and research groups.

Curran et al. [3] describe the convergence sequence as follows. First, different scientific disciplines begin to cite each other's results, and scholars start to collaborate with each other. Then, applied science starts an intensive usage of basic science results at the intersection of different scientific disciplines, resulting in the development of technology platforms [7] and leading to technology convergence. The resulting new technologies offer new product-market combinations, which trigger market convergence. Caferoglu et al. [15] propose an intermediate stage of preindustry convergence by denoting the market entrance of companies from distant industries. Finally, mergers and collaborations of different firms mark the emergence of industry

convergence [3]. If a complementary industry sector emerges, companies need to engage in strategic alliances and joint ventures that enable closing the gap of the lacking capabilities [7]. Collaborations enable firms to overcome risks and uncertainties in high technological environments through mutual learning without risk to the core business [16]. On the other hand, substitutive convergence may disrupt the existing knowledge and technology base of affected companies and might threaten the existing business model. Thus, it is pivotal for companies and research organizations to recognize at the early stage potential scientific convergences to prepare for consecutive technology, market, and industry convergence.

However, scientific convergence seems to be particularly difficult to anticipate, detect, and predict since relevant science and application fields are largely unknown [17], and scientific convergence is rather under-represented in the research on convergence [18]. In contrast to technology and industry convergence, which are identifiable by patents and alliances or collaborations at the intersection of two or more technology fields or industry sectors, scientific convergence is manifested only in scientific publications revealing new discoveries, algorithms, or methods at the interface of two scientific fields. The detection and correct interpretation of such weak signals requires constant monitoring, which is time-consuming and expensive. Moreover, these weak signals of scientific convergence are difficult to identify even by highly qualified experts. Functional fixation on the existing technologies creates biases in the technology evaluation, as researchers' existing knowledge corridors may result in "cognitive entrenchment" and prioritize familiar knowledge [19]. The dominance of field-specific research infrastructures and organizations increases the cost of processing information from other fields.

Prior research on scientific convergence offers the first promising insights into the identification of the early stages of scientific convergence [1]. Nevertheless, the incubation phase of scientific convergence is very dynamic and characterized by increasing interactions between and within distinct scientific fields [1]. Such interdisciplinary R&D may not persist, as research is more likely to stay within and emphasize disciplinary boundaries in times of crises, for instance [20]. The recent study by Hacklin et al. [4] suggests an analysis of scientific convergence through the lens of the social microlevel foundations underlying the role of the publishing behaviors of individual authors. Especially after the incubation stage [1], the intensity of social interactions in scientific communities at the boundaries of converging fields determines the scientific convergence acceleration [4]. Hence, the detection of such scientific socialization is of high relevance since it can act as a signal of scientific convergence. Thus, companies and research organizations can monitor these indicators and rely on them to develop strategies to benefit from the upcoming technology and industry convergences.

Validated instruments for measuring and predicting scientific convergence are still not fully developed [4], [14]. Therefore, we aim to develop a novel indicator based on machine learning to detect and assess scientific convergence through the lens of science dynamics' patterns. A science dynamics perspective allows us to incorporate specific factors that might foster or hinder scientific convergence and presents an opportunity to understand scientific convergence as a temporal dynamic phenomenon in an interplay with scientific divergence. Moreover, we include author-specific data in our analysis to incorporate social interactions in scientific communities, which are important for scientific convergence development [4]. Thus, we attempt to answer the call for conceptual clarity in scientific convergence research that was claimed in the review study on convergence by Sick and Bröring [10] and address open questions about convergence investigation methods [21]. Therefore, our study aims to answer the following research question: Can a machine-learning model estimate the similarity between concepts in a way that enables managers and policymakers to analyze science dynamics and detect scientific convergence?

We develop a new indicator based on a dataset rooted in the empirical context of life sciences, which we believe to be a particular subject of convergence. Life sciences also provide us with an established knowledge organization or classification system, namely, the medical subject headings (MeSH)[1] system. Within this system, concepts refer to an entire scientific field or a more fine-grained topic, depending on the level of terms in the MeSH hierarchy. We develop an indicator of scientific convergence based on the publication metadata, which we enrich with specific publication author information. We first transform the metadata into a network by connecting publications with concepts, authors, and journals and by using the concept hierarchy to connect concepts to each other. Then, we make use of the DeepWalk/node2vec algorithm [22], [23] to learn a representation for each concept. As a result, similar nodes are close to each other in the learned embedding space [23]. In this embedding space, we use the cosine distance to measure the similarity between concepts. To validate our developed indicator, we examine the existence of scientific convergence in a well-explored case of convergence, namely, cholesterol-lowering ingredients, representing a field in which "nutritional" and "pharmaceutical" science converged to some extent, forming the hybrid area of "nutraceuticals" [3], [17], [24], [25]. Finally, we apply our indicator to identify the gradual convergence process over time [10]. Our results confirm that our indicator reflects the scientific convergence within the selected use case.

In summary, our study contributes to R&D and technology management research and practice by employing machine-learning methods on bibliographic networks induced by 14 million publications to learn a similarity measure among concepts. We show how the graph representation learning algorithm DeepWalk can be incrementally trained on annual snapshots of bibliographic metadata to derive concept similarity scores. Therefore, we advance the understanding of the phenomenon of scientific convergence and enable its identification in empirical studies. We extend the existing methods to cope with longitudinal data to analyze science dynamics in general and scientific convergence in particular. We empirically validate our measurement approach on the previously well-explored convergence setting of cholesterol-lowering ingredients. The results

[1]MeSH is a hierarchical thesaurus of medical terms from the U.S. National Library of Medicine (NLM) for MEDLINE database.

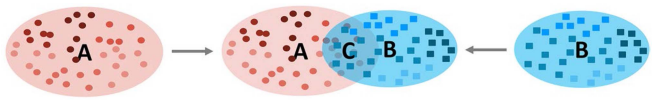

Fig. 1. Scientific fields A and B with multiple research topics within each field (represented as small colored dots and squares) move toward each other and partially merge in the overlapped field marked as C. Scientific convergence can have a substitutive nature if the new field C replaces the previously existing scientific fields A and B; or a complementary nature if the new field C coexists along with two parent fields A and B [3].

show that the similarity scores of our proposed approach reflect this convergence process. To facilitate this analysis, we compiled a dataset of 427 000 research articles in the scientific fields "pharmaceuticals" and "food and nutrition" between 2000 and 2019. To alleviate the potential bias of authors with the same name, we further enrich this corpus with disambiguated author data based on ORCIDs.

Our research, therefore, informs the growing research body focusing on the anticipation of convergence. More precisely, it extends the current approaches by not only focusing on the concept or scientific category co-occurrence in bibliographic data but also adds a novel data science approach to the current set of bibliographic methods [26], [27], [28], [29] (for a recent overview, see the article presented by Sick and Bröring [10]). Future research, as well as practitioners in the field, can profit from our source code, which is openly available. Integrating our code into a tool for the measurement of scientific convergence will allow for time-saving and inexpensive analysis of scientific convergence. The usage of the tool could enable qualified investment decisions in R&D and the search for appropriate collaboration partners between research and industry in converging scientific fields. Moreover, future research can extend our proposed instrument to measure scientific dynamics in other scientific fields, making our approach transferable to other industries.

The rest of this article is organized as follows. In Section II, we provide the theoretical background on scientific convergence. In Section III, we describe the employed methods and datasets. Our results are presented in Section IV. Finally, Section V discusses and concludes this article.

II. Theoretical Background and Related Work

In the following, we provide an overview of the theoretical background of scientific convergence. Then, we describe related work for measuring scientific convergence and, finally, discuss related work regarding data enrichment with author information.

A. Scientific Convergence in the Context of Science Dynamics

Scientific convergence renders the foundations for the consecutive technology, market, and industry convergence processes [24]. It emerges through two or more distinct scientific fields moving toward each other and the blurring and overlapping of the boundaries of these distinct fields (see Fig. 1) [3].

A system of science with distinguished scientific fields—as we are used to dealing with—is an artificial concept to structure our knowledge and delineate the communities of scientists engaging in research on similar problems, interacting with each other, and citing each other's work [30]. According to Wagner et al. [31], distinct scientific fields have a central scientific problem and a set of facts, explanations, goals, and theories related to the problem. These concepts of science have always been subject to change [30]. New scientific fields occur, and the existing ones are further differentiated, as exemplified by the emergence of computer science, nanoscience, and nanotechnology [32]. Kuhn [33] explained the role of discontinuous transformations in science as scientific paradigm shifts. Such paradigm shifts have a small effect on the established scientific communities at the beginning but may have an enormous impact on the previously existing scientific fields in later stages [33]. By investigating the role of deep learning technology in cancer imaging, Coccia [34] argues that a technological paradigm shift has an even more radical impact if such a paradigm shift influences other scientific fields, by interdisciplinary research, for instance [35]. Sun et al. [36] describe the emergence and evolution of scientific fields based on the sociocognitive interactions among researchers in respective scientific communities. They argue that disciplines arise by the splitting or merging of the existing scientific communities in collaboration networks. As such, the splitting of the researcher community, on the one hand, leads to a growing distance between hitherto close communities, resulting in a higher level of specialization and fragmentation. Coccia (2020, p. 464) defines it as "a scientific fission: the division of a scientific discipline into more research fields that evolve as autonomous entities" [37]. This process can trigger a divergence pattern of science dynamics. On the other hand, the merging of different social communities of researchers implies a combination of scientific fields that accumulate and integrate knowledge from these two or more different communities [36]. Thus, a convergence pattern results from strong social interactions in scientific communities. However, both patterns are characterized by changes in the number of publications and interactions among researchers, which is a property of science dynamics [38].

Accordingly, we study the dynamics of science and the change in distance between scientific fields. Our study, thus, focuses on the convergence of scientific fields. In a process of convergence, scientists of different scientific fields start to implement the results of their colleagues' research from other scientific fields and cite their work [3] because their own field lacks the research insights required for the continuous development of the field. Scientific convergence is characterized by the sharing of scientific methods, techniques, algorithms, and even terminology by converging fields, resulting in a converged field that becomes interdisciplinary [39]. The higher the number of disciplines and the more distant disciplines that are involved in such collaborations, the higher the integration index of interdisciplinary knowledge that can be achieved [39]. Thus, the main driver of scientific convergence is knowledge diffusion [40], and scientific convergence enriches the knowledge base of a converging scientific domain with valuable insights [17].

Most of the recent studies on convergence focus on identifying the features of scientific convergence or on describing the trajectory of scientific and technological convergence and the convergence stages [1], [4], [41]. These studies provide

valuable insights into convergence processes in science, and they reveal specific knowledge characteristics required for scientific convergence [1], [4]. In addition, within the framework of science dynamics, opposite divergence processes are of relevance, indicating stricter borders between scientific fields and a higher level of specialization [36], [37], [42]. The divergence of scientific fields can be characterized by a lower level of interaction between two distinct fields and movement apart from each other [43]. Science dynamics are characterized by the processes of growth and both convergence and divergence [38], [42], [44]. Thus, to be able to evaluate endogenous knowledge properties of scientific convergence trajectories, one should carefully analyze the interplay of convergence and divergence processes. By investigating the evolutionary processes in the entrepreneurship discipline, Grégoire et al. [45] determined that scientific convergence in this discipline undergoes several "convergence–divergence cycles." Such nonlinear patterns of scientific convergence were also identified in the case of nutraceuticals [3], [17]. Therefore, the purpose of our study is to provide a more fine-grained view of scientific convergence through the integration of a science dynamics perspective and, in particular, scientific divergence developments.

B. Approaches for Measuring Scientific Convergence

Scientific convergence emerges in scientific communities and networks [46] that publish their results [36]. Therefore, scientific publications may provide relevant indicators to detect scientific development at the embryonic stage [4], [47], [48]. One of the challenges of scientific convergence investigation lies in the analysis of a large amount of bibliographic data. Hence, machine-learning techniques, including text mining methods, and network approaches that are proven to be successful in studying the development of technology fields and are mostly based on the analysis of word co-occurrence, citations, patent classification, abstracts, claims, or full texts of patent data [49], [50], [51], [52], [53], are necessary for examining scientific convergence.

Jeong et al. [46] suggest measuring scientific convergence by scientific publications' subject categories. Although the authors demonstrate very promising results, they paid little attention to the measurement of the distance between the involved fields and to the demonstration of such interactions over time [46]. Kong et al. [41] propose applying a deep learning approach based on the graph neural networks, along with text information clustering methods, to investigate technology convergence. Kim and Sohn [26] offer a combination of machine-learning approaches, including semantic analysis, to identify and predict technology convergence based on the patent International Patent Classification (IPC) codes. Giordano et al. [52] use a name–entity–recognition technique in combination with dynamic networks for technology convergence investigation. Moreover, Hacklin et al. [4] capture knowledge flows embedded in semantic properties and microlevel behaviors of scientists in information technology and communication technology by bibliometric analysis of scientific convergence. The identification of scientific convergence can be based on the detection and measurement of technical emerging topics if these topics are shared by different scientific fields [47] or on citation patterns that reveal new fields that cite a focal field [39]. Furthermore, science overlay maps may assist in identifying scientific convergence at the broad science category level [54].

The investigation of scientific convergence as a concept of moving and merging distinct, previously separated scientific fields requires the identification and distinction of individual scientific fields. Such analysis necessitates the use of clustering techniques, such as rolling clustering, which is widely exploited for the clustering of patents and scientific publications [50], [55]. By investigating scientific convergence in bioinformatics, Zhou et al. [1] employ the fast-Newman topological clustering algorithm to capture a citation network of biology and informatics' domains and then apply a topic modeling approach to identify collaborating and citing topics within the converging field of bioinformatics. Nevertheless, some widely used topic modeling approaches based on unsupervised learning methodology have a limitation in terms of overly common and unspecific terminology, which is supposed to represent distinct topics, thus providing little technological content. Furthermore, the language of scientific publications does not clearly reveal the technical content, i.e., technology-specific word combinations that one may find in patent data [51].

In the present study, we focus on the analysis of science dynamics and the convergence of scientific fields by applying a novel machine-learning approach based on the graph representation learning, and we exploit the publications' annotations with concepts from a controlled hierarchical vocabulary (MeSH thesaurus). We define science dynamics as the change in distance between two (or more) concepts over time. These concepts represent scientific fields occupying different positions in the knowledge hierarchy. Following Sick and Bröring [10], we argue that decreasing distance over time implies that two fields of interest move toward each other, revealing a scientific convergence pattern. In contrast, an increasing distance over time means that two topics increase the degree of specialization *within two distinct topics* and both move away from each other. We denote this process as scientific divergence.

C. Examining Scientific Convergence Through the Lens of Author-Enriched Data

We use enriched bibliographic data to identify whether the composition of research groups, in terms of authors and their affiliations or journals, in which the relevant research is published, has an impact on the science dynamics in the converging fields. Research is driven by people [56], and the decisions of researchers are influenced by a wide range of social aspects [57]. Current research focuses on the ethnical privileges of researchers [58] and gender-specific aspects [59], [60], [61] to explain disparities, which also apply to other minority groups and to people with disabilities. These studies demonstrate the enormous impact that the social composition of a research group can have on scientific output. Therefore, an enrichment of the bibliometric data with social context data and rich author information is useful for measuring and understanding science dynamics [4], [48],

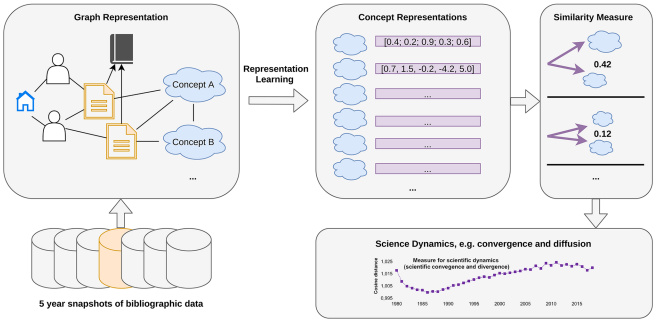

Fig. 2. Overview of our proposed approach to detect scientific convergence.

[62], [63]. Such data enrichment is particularly important due to the growing relevance of cross-disciplinary research teams [48], especially since national governmental initiatives and academia promote scientific convergence by colocating scientists from different disciplines or by providing funding for interdisciplinary research projects [64].

In summary, we seek to explore whether a machine-learning model can learn the similarity between concepts in a way that their temporal evolution unveils patterns of science dynamics. Bibliometric databases provide a unique tool to study the social and conceptual changes in science fields because they cover a long period of time [65]. In this work, we evaluate this approach at the hand of the well-known case of cholesterol-lowering ingredients, as outlined in Fig. 2.

III. Methodology and Materials

In this section, we describe our dataset (see Section III-A) along with our data enrichment strategies (see Section III-B), before we describe our employed methods (see Section III-C).

A. Dataset and Experimental Procedure

To investigate convergence processes, we select the well-explored case of convergence in the scientific field of cholesterol-lowering ingredients (phytosterols). Here, the field of "nutrition and food science" has partially converged with "pharmaceutical science" [3], [17], [25]. We replicate the findings on phytosterols with our newly developed similarity indicator. We provide a short description of the evolution of knowledge on cholesterol and the application of cholesterol-lowering ingredients in "pharmaceutical" and "food and nutrition science" in Supplementary Material A.

To investigate the convergence of two scientific fields in cholesterol research, we first defined the relevant topics for the two scientific fields of "pharmaceuticals" and "food and nutrition science." We extracted all publications from 32 "pharmaceuticals" and 41 "food and nutrition"-related scientific journals between 2000 and 2019 in five-year windows from the MEDLINE database.[2] Referring to "pharmaceuticals," we use the "Pharmacology and Pharmacy" subject category of the Web of Science (WoS) database. Similarly, we denote the "food and nutrition field" of science by referring to the "Nutrition and Dietetics" and "Food Science and Technology" categories of WoS journal-level classifications. To account for the quality of the journals' research, we use the journal impact factor since the empirical study of Saha et al. [66] confirms the reliability of this indicator for general medical journals. We rely on the WoS category normalized citation impact indicator, which is an unbiased publication impact indicator that allows comparing publications and journals of different subject categories. The value >1 denotes that a journal has at least an above-average impact, according to the WoS calculations. To be included in the analysis, a journal had to have a WoS impact factor of >3[3] and a WoS category normalized citation impact of >1. We selected 32 journals in the category "Pharmacology and Pharmacy," each having at least 140 000 citations and 41 journals in the categories "Nutrition and Dietetics" and "Food Science and Technology," each having at least 12 000 citations (i.e., according to the highest number of received citations). The latter indicator helped us to identify journals with a large number of publications if the two former quality criteria were fulfilled. This resulted in a dataset of 202 216 publications in "Pharmacology and Pharmacy" and 225 241 publications in "Nutrition and Dietetics" and "Food Science and Technology." In the same journals, we searched for publications containing the word "cholesterol" in titles or abstracts to identify publications related to the cholesterol topic. We identified 3276 publications in the category "Pharmacology and Pharmacy" and 5410 publications in the "Nutrition and Dietetics" and "Food Science and Technology" categories.

In the next step, we used the controlled vocabulary of MeSH[4] [67], [68], [69] and selected all annotations of the selected publications. We created a list of MeSH concepts relevant to each of the scientific fields "pharmaceuticals" and "food and nutrition science." We employed inverse document frequency (IDF) [70] on the MeSH concepts to select those that often occurred in one field and rarely occurred in the opposite field. To define "pharmaceuticals" and "food and nutrition science" as distinct scientific fields, we excluded MeSH concepts that often occurred in both fields. In this way, we were able to clearly separate MeSH concepts as research topics of both scientific fields and, as such, identified field-specific MeSH concepts. For this purpose, with the IDF-inspired method, we leveraged corpus characteristics to identify the most indicative corpus for each scientific field by computing the ratio of IDF scores per MeSH concept. When the ratio of IDF scores was near 1 (meaning that the concepts were equally frequent in both fields), the concepts were removed. The removed concepts also comprised

[2]MEDLINE is the database of the U.S. NLM. Description of the database: https://www.nlm.nih.gov/bsd/medline.html

[3]We relied on the median impact factors of the journals in the relevant WoS subject categories "Pharmacology and Pharmacy," "Nutrition and Dietetics," and "Food Science and Technology," which were in the range of 3–3.5 at the time of analysis.

[4]The MeSH thesaurus accounts for the breadth of scientific topics, allowing us to focus on topics with different granularities: from very broad topics, such as *Chemicals and Drugs* [D] (MeSH hierarchy level 1), to more specific topics, such as *Lipids* [D10] (MeSH hierarchy level 2), to very specific topics, such as *Cholesterol* [D10.570.938.208], or even to very fine-grained topics, such as *Cholesterol, HDL* [D10.570.938.208.270]. Hence, the MeSH hierarchy enables us to perform the science dynamics analysis at different scientific topics granularity levels.

very general MeSH concepts, such as human or animal. The resulting list was manually curated by domain experts, and 175 excessively general or irrelevant concepts (e.g., MeSH concepts, such as *patents as topics; swimming; mice, hairless; maze learning; radius; culture; climate; mass media; and discriminant analysis*) were removed from further analysis. As a result, we determined MeSH concepts for the analysis of the science dynamics of the large "pharmaceuticals" and "food and nutrition" scientific fields and for the "cholesterol-specific" subsample. First, 1008 MeSH concepts from the field of "pharmaceuticals" and 853 MeSH concepts from the "food and nutrition" field remained for investigation. Second, we extracted the MeSH concepts from cholesterol-related publications from the selected "pharmaceuticals" and "food and nutrition" scientific journals and identified 146 MeSH concepts from "pharmaceuticals" and 132 concepts from the "food and nutrition" field that we labeled "cholesterol specific."

Consequently, we intentionally excluded the MeSH concepts at the intersection of the scientific fields "pharmaceuticals" and "food and nutrition" from the analysis because the unambiguous convergence processes that use the MeSH concepts common to both scientific fields "pharmaceuticals" and "food and nutrition" could complicate the validity of the indicator we developed. As such, the convergence processes between the scientific fields "pharmaceuticals" and "food and nutrition" are rather underestimated in our results presented in Section IV.

Since we extracted MeSH concepts from the relevant "pharmaceuticals" and "food and nutrition science" journals in five-year windows, we were able to find and track the diffusion of knowledge from one scientific field into another one and into a specific application field. Referring to Petersen et al. [69], we specify an application field in our analysis according to the MeSH thesaurus domain C (Diseases) and define an application field as a scientific field denoting a demand articulation for scientific solutions from other scientific fields. As such, we examine whether our measurement approach can provide insights into knowledge diffusion processes. For this purpose, we apply our indicator (described in Section III-C) in the context of demand for disease treatment and supply of potential therapeutic agents' analysis. For the knowledge diffusion into an application fields analysis, we identified MeSH concepts that first were intensively used in one scientific field in one period but were not used or less used in another scientific field in that time period. This insight enabled us to analyze the knowledge diffusion processes that are crucial for scientific convergence [46].

We used the field-specific MeSH concepts to extract all publications that have been annotated with the identified MeSH concepts from the global dataset of 63 million scientific publications from the biomedical database ZB MED knowledge environment.[5] We use the ZB MED knowledge environment instead of the WoS database since it provides an annotation of publications with MeSH concepts, whereas WoS does not contain this information. With more than 30 million items, the ZB MED knowledge environment refers to the MEDLINE database.

[5]ZB MED knowledge environment: https://www.zbmed.de/en/research/completedprojects/zb-med-knowledge-environment

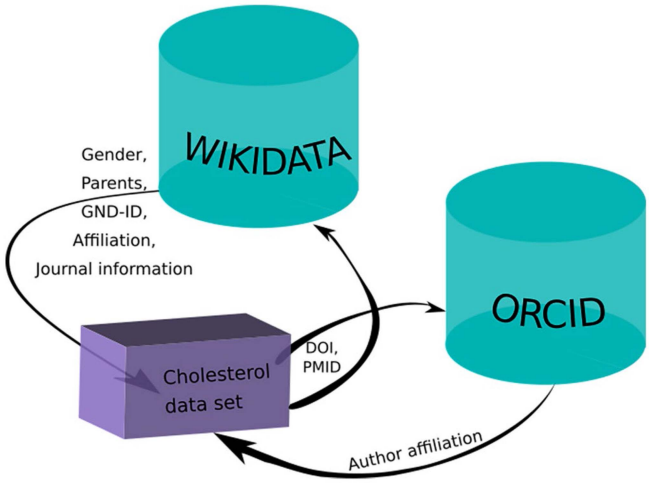

Fig. 3. Enrichment of metadata; our analysis focuses on the integration of author affiliations.

After filtering for relevant field-specific MeSH concepts, our dataset contains approximately 14 million publications.[6] The distribution of these publications over time is presented in Supplementary Material B. We created a heterogeneous information network based on the metadata (publications, concepts, authors, journal information, and MeSH hierarchy) of these publications. We employed DeepWalk [23] to learn continuous vector representations for each concept node, i.e., the MeSH concepts in each year. The details of how we set up the graph and applied the DeepWalk algorithm are described in Section III-C. To account for the latest relevant research, we used a five-year sliding window approach to compose the training sets between 1980 and 2018.[7] Due to the five-year window, the data from 1976 onward were also considered for the training.

B. Data Enrichment

As argued earlier, science dynamics can be influenced by the social composition of research groups. We focus on Wikidata and ORCID (Open Researcher and Contributor ID) for the enrichment of authors' affiliations. Wikidata is an open linked–open-data platform that, in January 2022, contained 22 574 314 scholarly articles, which represents 31.5% of all items [71], [72] (see Fig. 3). ORCID provides a persistent identifier for researchers who curate the information on scientific biography and publications themselves. The public data are available under a CC0 license. In the data sample of 2019, we detected 673 058

[6]Since these publications contained very unspecific MeSH concepts, such as *Male* (annotated in 7 689 101 publications between 1980 and 2019) or *Female* (annotated in 7 729 277 publications in 1980–2019), the number of publications used for training of the models increased to 14 million.

[7]For the extraction of MeSH concepts, we have used the time window of 2000–2019 since many journals had only a few publications before 2000. For instance, publication metadata on the *Food Chemistry* journal were available only since 2004 in the MEDLINE database. Another example is the journal *Critical Reviews* in food science and nutrition. In the period between 1980-01-01 and 1999-12-31, there are metadata on 400 publications of this journal (on average 20 publications per year), whereas in the period between 2000-01-01 and 2018-12-31, metadata on 1805 publications (on average 95 publications per year) are available.

assigned ORCID-IDs, which represent single researchers. The two databases represent two kinds of provided data: self-curated, as in the case of ORCID, and community-curated, as in Wikidata.

In our study, mainly the PubMed identifier (PMID) and DOI are present in the dataset of "food and nutrition science" and "pharmaceuticals" research papers. These are the starting points for identifying information on authors, periodicals, and institutions. While ORCID primarily contains data about authors, Wikidata lists authors and publications in a parallel manner and can function as a hub to connect publications and authors.

Using only Wikidata, we were only able to identify 4% of the authors of papers matched by PMID and DOI. A reason for this relatively small share might be the long observation period in the given sample. Older articles from the 1980s or 1990s might not be registered in Wikidata. In addition, authors from those days are included less often in the database than researchers from recent times. Another reason can be seen in the necessity of a triple registration for Wikidata, which complicates the harvesting: along with the journal article that needs to be listed, the author entry is also needed, and finally, the connection between both items also needs to be established in the database to enable retrieval. To further broaden the database, we harvested ORCID data itself. By using this strategy, we were able to allocate 14.2% of authors from ORCID to our dataset. In total, 580 283 authors of our dataset (10 366 156 authors) have ORCID. The distribution of the publications containing authors with ORCID is provided in Supplementary Material B. In our experiments, we label the model trained by using the disambiguated and enriched author data, instead of the raw string author data, as "enriched."

C. Machine-Learned Concept Similarity Indicator

We specifically employ the DeepWalk algorithm [23] to learn a similarity indicator between concepts. We have a graph $G = (V,E)$ constructed by bibliographic metadata. The set of nodes V consists of publication nodes P, concept nodes C, author nodes A, and journal nodes J. The undirected edges E may resemble the authorship relation between authors and publications, annotation relations between publications and concepts, the narrower/broader relation of concepts given by the MeSH hierarchy, and the publication's source, such as the publishing journal. Given two concepts $c_1,c_2 \in C$, the indicator should identify how similar the two concepts are. Thus, we learn a function f that maps the concepts in a vector space and apply a common distance metric d. The similarity indicator then becomes $d(f(c_1),f(c_2)) \in \mathrm{R}^+$, where $f: C \to \mathrm{R}^s$ is a trainable embedding function with s being the dimension of the embedding vectors (the embedding *size*). After training, we can compute the similarity of any two embedding vectors $f(c_1),f(c_2) \in \mathrm{R}^s$ as their cosine similarity

$$\cos_{\text{sim}}\left(f\left(c_1\right), f\left(c_2\right)\right) = \frac{f\left(c_1\right), f\left(c_2\right)}{\left|f\left(c_1\right)\right|\left|f\left(c_2\right)\right|}$$

which is the L_2-normalized dot product of embedding vectors. Subsequently, we use cosine distance $d_{\cos}: f(c_1),f(c_2) \to 1 - \cos_{\text{sim}}(f(c_1),f(c_2))$ as our distance metric. Since cosine similarity ranges between -1 and 1, cosine distance ranges between 0 and 2. The lower the cosine distance, the more similar the two concepts are.

To learn the embedding function f, we make use of the DeepWalk algorithm. DeepWalk [23] is an established approach for learning node embeddings in a graph. The Deepwalk algorithm randomly initializes a node embedding $\boldsymbol{X} \in \mathrm{R}^{n \times s}$, where n is the number of nodes and s is the embedding size. To update the node embeddings, DeepWalk samples random walks $(u_1,u_2,\ldots,u_l)$ through the graph such that $u_i \in V$ and $(u_i,u_{i+1}) \in E \ \forall i \in 1,2,\ldots,l-1$, where l is the length of the random walk. For each node on each random walk, its current embedding is used to predict neighboring nodes within a fixed context window size n_{cwnd} along the random walk. The respective node embedding is updated according to the error signal from the prediction. The embedding function f is then defined as $c_i \to \boldsymbol{X}_i$, one row of the embedding matrix, which is then used to compute the cosine distance between two concepts.

In standard DeepWalk, the error signal is computed according to the hierarchical softmax classification loss. In node2vec [22], Grover and Leskovec introduced additional hyperparameters to adjust the sampling procedure. They also explore using a negative sampling objective, as common in Word2vec [73] models. Negative sampling leads to a more efficient training procedure. With (hierarchical) softmax, the weights for all nodes not within the window size c are decreased, whereas negative sampling only draws a fixed number ns of negative examples. Because this approach is more efficient, it allows a higher throughput of random walks, which, in turn, increases the overall effectiveness [73].

For our instantiation of DeepWalk, we employ the negative sampling objective of node2vec [22] but leave the random sampling probabilities unbiased. We take inspiration from Metapath2vec [74] and use short random walks. Metapath2vec [74] is another approach specifically used for heterogeneous graphs. A metapath introduces constraints for sampling random walks. Only certain types of nodes may be sampled in the next step. We start each random walk from the concept nodes and allow only $l = 4$ steps, which would correspond to different metapaths, depending on which types of metadata we supply to the algorithm. For instance, a random walk could traverse a (meta-)path of concept-publication–author-publication or concept-publication–concept-publication. This strategy of using short random walks starting only at concept nodes was explored in the literature [75].

In the present work, we intend to learn node embeddings not only on a single graph but also on multiple graph snapshots over time. In the original iteration of DeepWalk [23], Perozzi et al. explored the possibility of incremental learning in streaming data snapshots. In our case, each snapshot consists of five years of publication data. When advancing from t to $t+1$, the new year is added, while the old year is removed, i.e., the snapshots overlap. We also explored a fully cumulative approach for taking snapshots, and the results can be found in Supplementary Material C.

In our snapshot-based setting, we run multiple epochs of training for time step t before advancing to time step $t+1$. To stabilize the training across time steps, we reuse the final trained node embeddings of time t to initialize the embeddings for $t+1$.

Thus, we avoid distorting consecutive embeddings by different random initializations, while at the same time, accounting for the changes in the graph. After training on each snapshot, we store the respective node embeddings for our analyses. For this, we normalize the concept embedding $\boldsymbol{X}_{|\text{concepts}}$, whose rows i hold concept embedding $f_t(c_i)$ in two steps. First, we center the embedding by subtracting the centroid. Then, we normalize the columns to the unit L_2-norm. We use these centered and normalized concept vectors to compute the cosine distance between any two concepts given time t.

In terms of hyperparameters, we used a walk length of $l = 4$, a context window size of $c = 4$, and an embedding size of $s = 128$. We sample 10 000 random walks per concept per year. The XL-variant was trained for 100 000 random walks of length $l = 20$ per concept per year. We optimize against $ns = 20$ negative samples with a constant learning rate of 0.025.

In summary, we incrementally train a machine-learning model on yearly snapshots of bibliographic metadata. This machine-learned concept similarity indicator yields one set of embedding vectors per concept per year. For our analysis, we center and normalize the embedding vectors. On these centered and normalized embedding vectors, we calculate the pairwise cosine similarities between different concepts to obtain year-specific similarity values.

IV. Results

In this section, we present the results of the enriched model for science dynamics in the fields of "pharmaceuticals" and "food and nutrition" and for the cholesterol-specific case. In the Supplementary Material, we supply a qualitative and quantitative evaluation, also including other model variants trained on different input modalities (Supplementary Material C). A detailed analysis with the same model trained on nonenriched data can also be found in the Supplementary Material for comparison (Supplementary Material D). With the help of our developed indicator (see Section III-C), we calculated the centered and normalized cosine distance (further termed as the normalized cosine distance) between a set of concepts belonging to two distinct scientific fields.

Fig. 4 shows the patterns of science dynamics for the concepts belonging to the fields of "pharmaceuticals" and "food and nutrition science." First, we observe a pattern of convergence between "pharmaceuticals" and "food and nutrition science" related to cholesterol research (Fig. 4, purple line with squares ■) from 1980 to 1986 since the normalized cosine distance decreased in this period. Due to increasing cosine distance from 1986 to 2007, a pattern of scientific divergence prevailed. Since 2007, science dynamics between "pharmaceuticals" and "food and nutrition science" related to cholesterol research have remained at a constant level.

We observe a similar pattern of science dynamics for the curve of "pharmaceuticals" and "food and nutrition science" in general (Fig. 4, green line with circles •). However, the cosine distance was slightly higher than that of cholesterol-specific research. The analysis of the distinct scientific fields of "pharmaceuticals" and "food and nutrition science" separately revealed a different

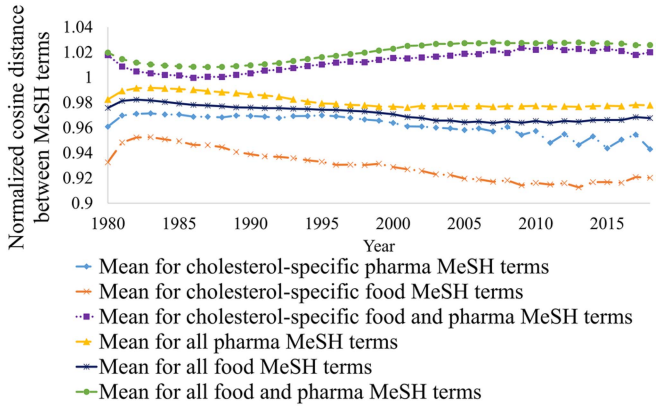

Fig. 4. Science dynamics in the fields of "pharmaceuticals" and "nutrition and food science," including cholesterol-specific science dynamics, measured on the enriched database.

picture: Up to 1982–83, a pattern of divergence prevailed for all four *distinct scientific fields* ("pharmaceuticals" and "food and nutrition science" in general and for cholesterol-specific research). However, due to the applied five-year sliding window, the data for the period 1980–1985 might be slightly biased since we had fewer publications available for training at the beginning of the period under investigation, and those publications contained very little author information (Fig. B.7 in Supplementary Material B).

Since 1983, the science dynamics in all these fields can be characterized by a convergence pattern since the cosine distance gradually decreased during this period. Such science dynamics can be explained by the properties of the evolution of scientific fields identified by Coccia [76]: a few disciplines drive the evolution of a scientific field, which indicates an accumulation and concentration of scientific production. Thus, if these disciplines concentrate primarily on research in their own scientific field, they become more specialized within the field. This specialization can be characterized by multiple linkages within a discipline and toward neighboring disciplines, which led to the decrease in cosine distance *within* a particular scientific field in our findings (see Fig. 4).

Hence, a strong specialization in specific disciplines within the "pharmaceuticals" and "food and nutrition science" fields led to moving these disciplines away from each other and, thus, to the divergence patterns of the "joint scientific field" of "pharmaceuticals" and "food and nutrition" (in general and for cholesterol-specific research). Therefore, the level of specialization determined the degree of convergence *within* the "pharmaceuticals" and "food and nutrition science" fields since the cosine distance of "food and nutrition science" for cholesterol-specific research (Fig. 4, orange line with crosses ×) was lower than that of "pharmaceuticals science" (Fig. 4, light blue line with diamonds ♦). The zigzag curve of the cholesterol-specific "pharmaceuticals" field from 2007 to 2018 could be explained by the high rates of growth in new disciplines and ambidextrous drivers of science, postulating that scientific discoveries or new technologies determine the development of a discipline [37].

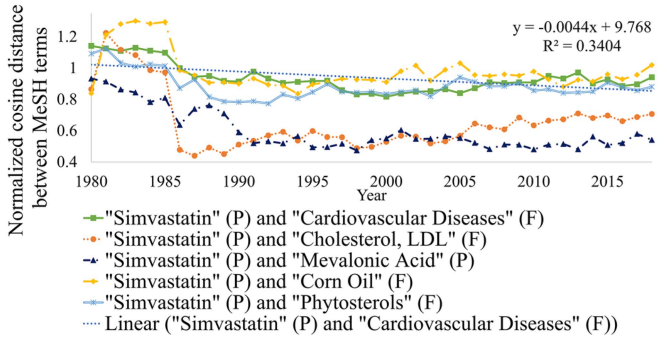

Fig. 5. Normalized cosine distances between *Simvastatin* and selected relevant MeSH concepts from the fields of "pharmaceuticals" and "food and nutrition," measured on enriched data. (F) stands for "food and nutrition" scientific field; (P) stands for "pharmaceuticals."

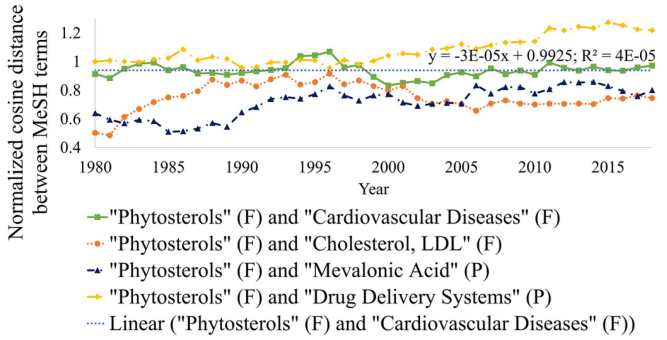

Fig. 6. Normalized cosine distances between *Phytosterols* and selected relevant MeSH concepts from the fields of "pharmaceuticals" as well as "food and nutrition," measured on enriched data. (F) stands for "food and nutrition" scientific field; (P) stands for "pharmaceuticals."

A. *Knowledge Diffusion Results*

Additionally, our indicator allows us to identify an ongoing diffusion of the cholesterol-specific concepts of each scientific field (a cholesterol-lowering drug *Simvastatin* from "pharmaceuticals" and a plant dietary cholesterol-lowering supplement *Phytosterols*[8] from "food and nutrition science") into the relevant application field (treatment and prevention of cardiovascular diseases) in the observed time frame [see Figs. 5 and 6]. We chose the *Cardiovascular Diseases* MeSH concept as an application field since it is listed in the MeSH domain C (Diseases) and, therefore, represents the demand innovation dimension according to Petersen et al. [69]. MeSH concepts *Simvastatin* and *Phytosterols* are representatives of the MeSH domain D (Chemicals and Drugs) and belong to the supply innovation dimension [69]. The trajectory of the curve of cosine distance between *Simvastatin* and *Cardiovascular Diseases* in Fig. 5 shows that the cosine distance, on average, decreased during the period of analysis. The trajectory of cosine distance between *Phytosterols* and *Cardiovascular Diseases* in Fig. 6 does not show any clear trend. Thus, an analysis of cosine distances between the cholesterol-specific concepts of each scientific field (*Simvastatin* from "pharmaceuticals" and *Phytosterols* from "food and nutrition science") and the application field concept (*Cardiovascular diseases*) in the observed time frame indicated that only the diffusion of the "pharmaceuticals" field into the application field is observable (see Figs. 5 and 6). The Mann–Kendall test (not displayed) confirmed that the slope was negative and significant only for the data of the concept pair *Simvastatin* and *Cardiovascular Diseases*.

Furthermore, the trajectories of cosine distances of *Simvastatin* and *Cholesterol,* low-density lipoprotein (*LDL),* as well as *Mevalonic Acid* (see Fig. 5) show that the cosine distance for these pairs is considerably lower than that of the pair *Simvastatin* and *Corn Oil*. This finding can be explained by the function of simvastatin as an inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase in the mevalonate pathway of cholesterol biosynthesis and its role in the reduction of LDL cholesterol [78].[9] We found that the graph of *Simvastatin* and *Corn Oil* indicates a higher cosine distance since there are still fewer findings that simvastatin (or other statins) is used in combination with corn oil (a specific source of phytosterols [77]) as a therapy for lowering cholesterol levels [79].[10]

In contrast, the cosine distance of the trajectories *Phytosterols* and *Cholesterol, LDL,* as well as *Mevalonic Acid* (see Fig. 6) is generally higher compared with those of the respective *Simvastatin* trajectories (see Fig. 5). The explanation for such differences can be the different mechanisms of action of phytosterols in the reduction of cholesterol levels via intestinal competitive absorption of phytosterols and cholesterol [77]. The analysis of the trajectory *Phytosterols* and *Drug Delivery Systems* (a concept from the field of "pharmaceuticals," belonging to the MeSH domain E (analytical, diagnostic, and therapeutic techniques, and equipment), i.e., representing technological capabilities [69]) shows that the cosine distance of this pair is much higher than that of the other pairs in Fig. 6. Due to the very low solubility of phytosterols in water and fats, it is difficult to formulate them as pharmaceutical components (drugs). Thus, drug delivery systems, as potential technological capabilities, are not relevant for the administration of phytosterols. However, food products containing phytosterols in high doses were successfully

[8]Phytosterols are plant sterols that are structurally similar to cholesterol molecules and typically have an extra methyl or ethyl group at the C-24 position or an additional double bond at the C-22 position [77]. Although the structural properties of phytosterols are similar to those of cholesterol, the mechanism of their interaction with cholesterol is different than that of statins and is based on competitive solubilization and cocrystallization with cholesterol; hydrolysis of phytosterol fatty acids that induces the replacement of cholesterol in the micelles fostering the accumulation of cholesterol in the oil phase; and competitive cholesterol/other sterols transport from the intestinal lumen to lymph [77].

[9]"Statins [...] inhibit HMG-CoA reductase, the enzyme that converts HMG-CoA into mevalonic acid, a cholesterol precursor. The statins do more than just compete with the normal substrate in the enzyme's active site. They alter the conformation of the enzyme when they bind to its active site. This prevents HMG-CoA reductase from attaining a functional structure. The change in conformation at the active site makes these drugs very effective and specific" [78, p. 380].

[10]Although there are a lot of randomized controlled trial (RCT) studies that investigated the effect of the combined therapy of high concentration of LDL with statins together with phytosterols [80], these studies usually use a combination of phytosterols that cannot be allocated to a specific plant source [81] or that was directly administered to patients in the form of vegetable oil-based spread enriched with plant stanol esters [82]. In their meta-analysis of 15 RCTs, Han et al. [80] found that "combination treatment with statins together with phytosterols significantly decreased the levels of total cholesterol by 0.30 mmol/L and LDL-cholesterol by 0.30 mmol/L, compared with statins alone" [80, p. 4].

developed and marketed [77]. For this purpose, phytosterols and phytostanols are esterified with fatty acids, enabling the use of phytosterols in margarine, cooking oils, etc. [77]. In sum, these identified patterns are consistent with the described dynamics in the field of cholesterol research, which supports the validity of our newly developed indicator.

V. Discussion and Conclusion

A. Summary of Major Results and Advantages of the Developed Indicator

1) Our indicator allows us to identify patterns of scientific convergence and divergence for the fields of "pharmaceuticals" and "food and nutrition" for cholesterol-specific research between 1980 and 2018. Since our approach uses the MeSH thesaurus to specify scientific fields based on broad publication metadata, including information on authors, it allows for the temporal investigation of scientific convergence, i.e., for a very fine-grained scientific topic analysis. The indicator thereby quantifies a contextual distance between two or more scientific fields or topics, which the science overlay mapping approach [54] and the method offered by Jeong et al. [46] lack.
2) A similar pattern of science dynamics was identified for the fields "pharmaceuticals" and "food and nutrition science" in general. However, the normalized cosine distance was higher, revealing a larger gap between the fields "pharmaceuticals" and "food and nutrition" in general compared with that of cholesterol-specific research.
3) We observed patterns of scientific convergence *within* the fields of "pharmaceuticals" and "food and nutrition science" in general and for cholesterol-specific research. We argue that different levels of specialization in research, as well as other factors, such as path dependency of critical disciplines, scientific fields' fission, and the emergence of new disciplines within scientific fields [37], [76], drive science dynamics *within* the fields "pharmaceuticals" and "food and nutrition" in general and for cholesterol-specific research.
4) Our indicator enables the investigation of knowledge diffusion from one scientific field into another one or into an application field. We identified the ongoing diffusion of cholesterol-specific MeSH concepts into their potential application fields. Our results suggest that, based on decreasing cosine distance for the MeSH concept pair *Simvastatin* of the scientific field "pharmaceuticals" and *Cardiovascular Diseases*, this specific technology diffused into the *Cardiovascular Diseases* application field. However, we found no evidence for the diffusion process of the scientific field "food and nutrition" into *Cardiovascular D*iseases.
5) The shortcomings of our developed approach include the missing inclusion of citation data; the descriptive nature of the analysis and the limited possibility of identifying determinants of the observed science dynamics; and the low coverage of author information in the dataset. Future research should tackle these limitations to further contribute to the understanding of science dynamics.

B. Discussion of Major Findings

Scientific convergence processes induce the development of new technologies and the emergence of new markets and industry sectors [12], [28], [29]. Although recent studies offer the first approaches for the identification and monitoring of scientific convergence [1], [4], a well-established, reliable measurement for the scientific convergence of multiple scientific fields, in particular, and science dynamics, in general, was still lacking. Using large-scale bibliographic metadata and a novel machine-learning approach, in this article, we were able to provide a reliable analysis of science dynamics and, in particular, scientific convergence and divergence. Thus, we expand the literature on scientific convergence by providing a novel validated measurement method. To this end, we trained a node representation learning model and applied cosine distance as an indicator for the similarity of the two concepts.

Our results provide insights into science dynamics and reveal the partial scientific convergence in the field of cholesterol research. We detected patterns of scientific convergence between the fields of "pharmaceuticals" and "food and nutrition." This is in line with prior research on the scientific convergence of "pharmaceuticals" and "food and agriculture" fields in phytosterol research [17]. We conclude that our developed indicator is an appropriate measure to investigate science dynamics and to detect scientific convergence as well as scientific divergence that can indicate evolving specialization trends within scientific fields. We, thus, contribute to the prior literature on science dynamics [33], [36], [37], [56], [76], covering specifically convergence [1], [4]. By providing empirical evidence of scientific convergence and scientific divergence processes within the framework of science dynamics, we extend the model of social dynamics and branching in scientific fields [36]. As a result, we stress the importance of investigating scientific dynamics as the interplay of convergence and divergence processes.

Furthermore, we have shown that knowledge generation is not a linear process. This insight is in line with prior findings of the authors in [3], [15], [17], [45], and [52]. We recognized that different science dynamics processes occur within and between scientific fields. We found that a convergence pattern is currently predominant *within* the scientific fields ("pharmaceuticals" or "food and nutrition"). This confirms that establishing linkages and increasing interactions between similar scientific fields, which drives convergence, may be easier to achieve [46]. Coccia [76] also claims that the path dependency of critical disciplines determines the evolution of scientific fields. These critical disciplines are the "parent" disciplines from which new disciplines emerged [76]. Thus, the strong development of particular disciplines *within* the "pharmaceuticals" and "food and nutrition" science fields determined the evolution of the "joint scientific field" of "pharmaceuticals and food and nutrition." Coccia [37] argues that the processes of specialization and fragmentation in science may dominate over time. A high level of specialization *within* one field facilitates rapid knowledge diffusion

[36]. Researchers from the same field have lower transaction costs compared with interdisciplinary research in which a lack of common language, shared norms, and research techniques, as well as a lack of background knowledge from other domains, creates higher barriers to effective interactions between distant scientific fields [83]. This is in line with the recent study on scientific convergence in bioinformatics, which identified that, in the fast-changing incubation stage of scientific convergence, intense interactions *within* a scientific field prevail over the linkages to other scientific fields [1].

Our results from the case of cholesterol research suggest the patterns of divergence prevail *between* two scientific fields with distant knowledge ("pharmaceuticals" and "food and nutrition" fields). This finding is in line with that of Curran and Leker [17], who found an increasing specialization in the field of phytosterol research over the past two decades, although the authors detected a pattern of scientific convergence in the earlier periods. Growing specialization can be an indicator of ongoing branching mechanisms that lead to increasingly fragmented and concentrated scientific fields [36]. Divergence patterns, which can also be identified with our indicator, reveal that scientific fields evolve parallel to each other or disassociate from each other, implying low interaction levels between the scientific fields in terms of collaborations or joint publications. This is in line with the authors in [38] and [44], who found that relatively new scientific fields, such as biotechnology, genomics, and nanotechnology, show a rapid growth rate and increasing specialization. Although Kim and Sohn [26] argue that a rapid growth rate of a technology field leads to its overlapping with other technology fields, such overlapping can occur with multiple fields, which are not necessarily the fields of the anticipated convergence in our study. Furthermore, our results are in line with Jeong et al. [46], suggesting that the diffusion of knowledge from the field of interest (in our case, research on simvastatin) into its application field (treatment or prevention of cardiovascular diseases) may trigger scientific convergence processes that are also similar to "convergence initiating" entities [12].

We found that the enriched information on researchers' affiliations could offer a more detailed view. This insight is in line with the study of Hacklin et al. [4] underlying the role of individual scientists in the initialization of scientific convergence processes since these scientists span boundaries between two distinct scientific fields and reuse knowledge from other scientific fields. The developed indicator of science and technology similarity includes author information and can guide scientific institutes and firms in the identification of potential collaboration partners in order to profit from convergence. However, in Supplementary Material C, we show that the choice of metadata data attributes (authors, publications sources, and concept hierarchy) has no measurable effect on the accuracy of the model as long as publications and concepts are included (see Fig. C.11). We assume that due to a relatively low percentage of author-enriched data, our current analysis was unable to reveal the entire potential of author enrichment for the science dynamics' processes. In Supplementary Material D, we provide the same analysis but on the nonenriched data, which shows similar findings. Our additional results, as presented in Supplementary Material D, reveal that even data that are not enriched with complementary author and affiliation information are an appropriate base for conducting science dynamics analysis. Since the main results presented in Section IV are similar to the findings obtained from the model trained on nonenriched data (see Supplementary Material D), our approach provides robust and reliable results in terms of measuring science dynamics.

Our method learns representations on the basis of graphs created by bibliographic metadata. As such, it does not take the raw text of the publications into account. One could argue that such approaches would only detect structure and not capture the semantic similarity between concepts. However, as the literature [75] has shown, these approaches capture the meaning of concepts better than the traditional entirely text-based approaches, such as latent semantic analysis [84] and, are on the same level, as graph neural networks [85], which take into account both text and structure. The ability of DeepWalk/node2vec to capture meaningful representations in structured data is also well supported by the literature [22], [23]. From a different angle, it is an advantage that our methods do not require any textual data and instead only use bibliographic metadata.

The MeSH thesaurus that we have exploited for the definition of scientific fields and topics result in a more stable analysis than the results of clustering algorithms, such as *K*-means in combination with the latent Dirichlet allocation topic modeling method, where the number of technology clusters may be set to 35 and the topics' distribution may be set to 1, as the very insightful study by Kong et al. [41] proposes. Our thesaurus-based method can be compared with the hierarchical system of technological knowledge structured by the IPC that allows for differentiated technology convergence analysis, for instance, conducted by Kim and Sohn [26], by using a machine-learning approach.

C. Limitations and Future Research

A limitation of our study is that we have not yet used citation data. We are aware that citations are an important factor in analyzing scientific literature [86] and scientific convergence [1]. For now, we focus on coauthorship relationships representing sociocognitive interactions of researchers, publication venues as proxies for scholar communities [36], concept hierarchies, and primarily the annotation of publications with concepts [67]. Thus, the integration of the underlined entities allowed us to develop an indicator for science dynamics that exceeds the measurements based only on individual entities. We expect that our methods would further benefit from modeling citations as another type of edge within the graph. Future research might further enrich large databases with citations. Nevertheless, it is notable that our method can still learn meaningful concept representations without resorting to citations.

Another limitation is that our indicator offers the explication of scientific trends in the scientific fields of interest only in a descriptive manner. It does not allow us to investigate the determinants of the science dynamics. This could be enhanced in future work by integrating an explainability mechanism, such as GNNExplainer [87]. This may include information about the

location and diversion of the research teams since geographical diversity may even further decelerate the convergence process [11].

We are also aware that the usage of MeSH annotations is particularly challenging because the MeSH thesaurus is updated every year. New concepts are added to the vocabulary and integrated into the tree structure of the thesaurus. Emerging concepts that might be a result of convergence are especially affected by this updating approach of MeSH. Therefore, we should pay special attention to those concepts that are new to the classification, as they may indicate that convergence has already occurred. The integration of new concepts can be preceded by another process on which the analysis focuses: concepts that have already been established in a scientific field but are suddenly used simultaneously with concepts from another context may indicate convergence and the adoption of knowledge. Nevertheless, despite the changes occurring due to the inclusion of new concepts, the MeSH concept thesaurus, in general, is still quite stable over time [69]. Consequently, we believe that it is sufficient to extract MeSH concepts from 2000 onward from the related journals to perform the analysis. Furthermore, since we deliberately excluded the MeSH concepts at the intersection of the scientific fields "pharmaceuticals" and "food and nutrition" from the analysis, our results might underestimate the convergence science dynamics because the MeSH concepts that may indicate convergence by appearing in both scientific fields were removed before the analysis. Future research can build upon our method and use all concepts appearing in both scientific fields. This task is challenging because of the time lag between the emergence of the scientific topic and its emergence as a MeSH concept in the MeSH thesaurus.[11] Therefore, we believe that new MeSH descriptors might be problematic as long as the scientific topic defined by such a MeSH descriptor has not been established. However, due to the retrospective annotation of publications with new MeSH descriptors, the lagging bias might be quite low.

Another limitation of our study is the low percentage of analyzed publications with author-enriched data. Only 5.6% of all authors of our dataset are disambiguated,[12] i.e., they have ORCID. For that reason, we could not show that author-enriched data can significantly contribute to the identification of the science dynamics processes. Moreover, due to the lack of disambiguated author data, we were unable to integrate authors' collaborations in our training dataset. Future studies can investigate science dynamics' patterns incorporating disambiguated author data. Future research may use such author-disambiguated data, e.g., to perform an entire author analysis by examining whether an author published in one scientific field (for instance, in "pharmaceuticals") or in both scientific fields ("pharmaceuticals" and "food and nutrition") and/or if authors are able to build bridges between distant scientific fields. Furthermore, additional enrichment of bibliographic metadata could improve our input data, which could lead to more specific results. Here, the networks of authors and institutions could be integrated to better understand the role of social relations in research dynamics. For this, the problem of author name disambiguation should be solved. We expect ORCID author identifiers as well as Wikidata to be even more established in the future, which will strengthen the data basis for this metadata enrichment.

Our proposed method can be applied to datasets from other domains. Controlled vocabularies, such as MeSH, are prevalent in other domains as well. To name a few, there is the Standardthesaurus Wirtschaft[13] for economics and business studies, ACM CCS[14] for computer science, AGROVOC[15] for agricultural science, and UMTHES[16] for environmental research. As other domains may have a different publication culture, further validation would be needed. We have made the first step in the well-explored case of cholesterol research in the life sciences domain. Our first results are promising, and we expect the proposed method to be transferable even to other domains. Future research should validate our developed indicator for measuring scientific dynamics in other scientific domains and use it accordingly. Based on our presented research, similar methods for understanding research dynamics as well as indicators emerging in the future should be included for an analysis of different scientific convergence processes. Future studies can elaborate on patterns of how scientific convergence evolves in comparison with divergence to better distinguish between not only these two patterns but also their particular nature and driving factors triggering these events.

High-quality author-enriched data and our indicator could help to explore whether and which companies, universities, and other research organizations are involved in the initial stage of convergence processes. This includes an analysis of those resources and capabilities that may differentiate highly influential organizations from organizations that play a less prominent role in convergence processes. Furthermore, future research could investigate whether companies are more successful than research organizations in the evolution of technology convergence and the establishment of industry convergence in the late stages of convergence processes. Moreover, our measurement approach can help to investigate the effect of scientific convergence on

[11] From our recent research, we know that, for instance, the MeSH descriptor *COVID-19* [C01.748.610.763.500] was not available at the beginning of 2020 as the coronavirus pandemic broke out, but it was added to the MeSH thesaurus on 2020-07-07 and became an established descriptor on 2021-01-01. After the inclusion of this MeSH descriptor into the MeSH thesaurus, the research articles on novel coronavirus SARS-CoV-2, which were published before 2020-07-07, were annotated with *COVID-19* MeSH descriptor, retrospectively. While the previously exploited MeSH descriptors denoting the novel virus, such as *Betacoronavirus,* were retained.

[12] An author name disambiguation problem occurs if a person can be listed in a database under several different names, or different persons may share the same name in a database. Since the techniques to perform the author name disambiguation are based on the author similarity profiles analyzing such bibliographic attributes as publication titles, journals, coauthor names, and citations [88], we were concerned that these techniques will prevent us from finding authors that may publish in different scientific fields ("pharmaceuticals" and "food and nutrition"), i.e., authors publishing in journals belonging to different disciplines or with coauthors representing different disciplines. For that reason, we used only ORCID data since authors maintain their data in this database by themselves.

[13] [Online]. Available: https://zbw.eu/stw

[14] [Online]. Available: https://dl.acm.org/ccs

[15] [Online]. Available: http://www.fao.org/agrovoc/

[16] [Online]. Available: https://sns.uba.de/umthes/de.html

the further phenomena of technology, market, or industry convergence as well as on the emergence of innovations with a high degree of innovativeness.

D. Theoretical Implications

Our developed indicator enables the detection of scientific convergence and divergence. It facilitates the identification of dynamic processes in research, pointing out that the scientific fields might undergo a number of science dynamics processes, such as scientific field fission or the emergence of new scientific fields [37]. We challenge the existing views that scientific convergence can be measured by simply counting concepts or scientific category co-occurrences. We have shown that co-occurrence counts are insufficient for the analysis of research dynamics and suggest building on a learned embedding space, where the similarity between any two concepts or scientific categories is meaningful, even if they have never co-occurred. This is especially important for the convergence of research fields, which often leads to the emergence of a new scientific field. One way to enable this is to consider publication metadata as a graph and take multihop relationships into account, as we did for our analyses. In this work, we validate the proposed indicator on the well-explored case of scientific convergence in cholesterol-lowering agents. We show that the convergence process is reflected in the similarity scores of our proposed approach. As a result, we advance the current understanding of scientific convergence and offer an indicator that allows us to empirically identify and, thus, monitor scientific convergence processes.

E. Practice and Policy Implications

We envision that the proposed approach will help academic researchers as well as practitioners and policymakers who are responsible for guiding science-based R&D programs to detect scientific convergence in the early phase. By means of our indicator, those topics can be identified that are related to each other in terms of content and whose connection is intensifying, even though they may belong to different scientific fields. On this basis, decisions can be made, for example, about the organization of scientific institutions, e.g., with the aim of promoting cooperation of distant knowledge fields in the interest of interdisciplinary research. In addition, our indicator allows for tracking scientific dynamics and, thus, identifying future-relevant topics in scientific fields. Based on the relevance of scientific convergence, scientific institutions, but also policymakers and R&D managers in commercial organizations, could invest limited resources in those fields that may benefit from intensifying convergence. The usage of the indicator could enable a focus on future-oriented topics in which universities, research organizations, or companies could initiate explorative projects and identify new R&D partners. Such projects could be grouped into programs and may inform the mission of the public and private organizations in the selected future topic. Furthermore, individual experts relevant to such topics and their organizations will be identifiable. Based on such information, firms and research organizations will be able to identify relevant cooperation partners or may aim to hire relevant experts. In sum, R&D managers can take these insights into account to consider adapting their organizations' research portfolio, putting them in a good position for coping with future trends. Thus, our approach can be integrated into a science dynamics analysis instrument that can support decision makers in developing R&D strategies for universities, research organizations, and companies.

F. Ethical Considerations

Although rich information on knowledge graphs, including the personal details of researchers, contribute to the analysis and the understanding of evolving scientific fields, it would become ethically unacceptable if a decision on science dynamics would build only upon our provided approach. The development of scientific fields as described in our study is likely related to the composition of the research group. A change in a researcher's group composition could result in a change in topics of interest. It must be assumed that globalization, the economic rise of developing countries, and gender mainstreaming partially overcome the ethnic bias that affects the research. Painting a complete picture of the development of research topics in their merging and diverging behavior, therefore, needs to include information on the social conditions. These conditions could be networks or relationships apart from direct coauthorship and citation relations but include the researchers that act within networks.

As machine-learning techniques always learn structures from the existing data, the assumptions about science dynamics calculated by our prognosis tool are necessarily characterized by the input data it is trained on. When we feed our models with information about research patterns and individuals from the past, we will reproduce the societal circumstances of the past. Therefore, we understand that feeding an algorithm with additional social information can lead to the reproduction of the existing structures and hinder change. We highly recommend including as many sources of information as possible, in addition to the provided indicator, if trying to identify science dynamics. Based on the deliberations above, we suggest not relying on a single analysis (or forecast) tool in regard to complex decisions, as in the context of R&D management and policy.

G. Reproducibility and Reuse

We provide the source code so that future research can further develop the approach, which is available at https://gitlab.com/Q-Aktiv/qgraph.

Acknowledgment

The authors would like to thank the IEEE editors and anonymous reviewers for their valuable comments and suggestions, which improved the quality of the article, and the German Network for Bioinformatics Infrastructure (de.NBI) for providing the computing infrastructure for conducting our experiments.

REFERENCES

[1] Y. Zhou, F. Dong, D. Kong, and Y. Liu, "Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies," *Technol. Forecasting Social Change*, vol. 144, pp. 205–220, Jan. 2019, doi: 10.1016/j.techfore.2019.03.014.

[2] E. Maine, V. J. Thomas, and J. Utterback, "Radical innovation from the confluence of technologies: Innovation management strategies for the emerging nanobiotechnology industry," *J. Eng. Technol. Manage.*, vol. 32, pp. 1–25, Apr./Jun. 2014, doi: 10.1016/j.jengtecman.2013.10.007.

[3] C.-S. Curran, S. Bröring, and J. Leker, "Anticipating converging industries using publicly available data," *Technol. Forecasting Social Change*, vol. 77, no. 3, pp. 385–395, Mar. 2010, doi: 10.1016/j.techfore.2009.10.002.

[4] F. Hacklin, M. W. Wallin, J. Björkdahl, and G. von Krogh, "The making of convergence: Knowledge reuse, boundary spanning, and the formation of the ICT industry," *IEEE Trans. Eng. Manage.*, vol. 70, no. 4, pp. 1518–1530, Apr. 2023, doi: 10.1109/TEM.2021.3087365.

[5] S. Bornkessel, S. Bröring, and S. W. F. Omta, "Crossing industrial boundaries at the pharma-nutrition interface in probiotics: A life cycle perspective," *PharmaNutrition*, vol. 4, no. 1, pp. 29–37, Jan. 2016, doi: 10.1016/j.phanu.2015.10.002.

[6] D. S. Ebert, A. Reinert, and F. Brian, "Visual analytics review: An early and continuing success of convergent research with impact," *Comput. Sci. Eng.*, vol. 23, no. 3, pp. 99–108, May/Jun. 2021, doi: 10.1109/MCSE.2021.3069342.

[7] N. Sick, N. Preschitschek, J. Leker, and S. Bröring, "A new framework to assess industry convergence in high technology environments," *Technovation*, vol. 84–85, pp. 48–58, Jun./Jul. 2019, doi: 10.1016/j.technovation.2018.08.001.

[8] S. Jeong, J.-C. Kim, and J. Y. Choi, "Technology convergence: What developmental stage are we in?," *Scientometrics*, vol. 104, no. 3, pp. 841–871, Sep. 2015, doi: 10.1007/s11192-015-1606-6.

[9] S. Bröring, L. M. Cloutier, and J. Leker, "The front end of innovation in an era of industry convergence: Evidence from nutraceuticals and functional foods," *R&D Manage.*, vol. 36, no. 5, pp. 487–498, Nov. 2006, doi: 10.1111/j.1467-9310.2006.00449.x.

[10] N. Sick and S. Bröring, "Exploring the research landscape of convergence from a TIM perspective: A review and research agenda," *Technol. Forecasting Social Change*, vol. 175, Feb. 2022, Art. no. 121321, doi: 10.1016/j.techfore.2021.121321.

[11] L. Ardito, A. Natalicchio, and A. M. Petruzzelli, "Evidence on the determinants of the likelihood and speed of technological convergence: A knowledge search and recombination perspective in key enabling technologies," *IEEE Trans. Eng. Manage.*, vol. 70, no. 4, pp. 1442–1455, Apr. 2023, doi: 10.1109/TEM.2021.3103878.

[12] L. J. Aaldering, J. Leker, and C. H. Song, "Uncovering the dynamics of market convergence through M&A," *Technol. Forecasting Social Change*, vol. 138, pp. 95–114, Jan. 2019, doi: 10.1016/j.techfore.2018.08.012.

[13] Y. Geum, M.-S. Kim, and S. Lee, "How industrial convergence happens: A taxonomical approach based on empirical evidences," *Technol. Forecasting Social Change*, vol. 107, pp. 112–120, Jun. 2016, doi: 10.1016/j.techfore.2016.03.020.

[14] J. Kim and S. Lee, "Forecasting and identifying multi-technology convergence based on patent data: The case of IT and BT industries in 2020," *Scientometrics*, vol. 111, no. 1, pp. 47–65, Apr. 2017, doi: 10.1007/s11192-017-2275-4.

[15] H. Caferoglu, D. Elsner, and M. G. Moehrle, "The interplay between technology and pre-industry convergence: An analysis in the technology field of smart mobility," *IEEE Trans. Eng. Manage.*, vol. 70, no. 4, pp. 1504–1517, Apr. 2023, doi: 10.1109/TEM.2021.3092211.

[16] J. Hagedoorn and G. Duysters, "External sources of innovative capabilities: The preferences for strategic alliances or mergers and acquisitions," *J. Manag. Stud.*, vol. 39, no. 2, pp. 167–188, Mar. 2002, doi: 10.1111/1467-6486.00287.

[17] C.-S. Curran and J. Leker, "Employing STN AnaVist to forecast converging industries," *Int. J. Innov. Manag.*, vol. 13, no. 4, pp. 637–664, Dec. 2009, doi: 10.1142/s1363919609002455.

[18] A. Klarin, Y. Suseno, and J. A. L. Lajom, "Systematic literature review of convergence: A systems perspective and re-evaluation of the convergence process," *IEEE Trans. Eng. Manage.*, vol. 70, no. 4, pp. 1531–1543, Apr. 2023, doi: 10.1109/TEM.2021.3126055.

[19] E. Dane, "Reconsidering the trade-off between expertise and flexibility: A cognitive entrenchment perspective," *Acad. Manage. Rev.*, vol. 35, no. 4, pp. 579–603, Oct. 2010, doi: 10.5465/amr.35.4.zok579.

[20] M. Coccia, "Evolution and structure of research fields driven by crises and environmental threats: The COVID-19 research," *Scientometrics*, vol. 126, no. 12, pp. 9405–9429, Dec. 2021, doi: 10.1007/s11192-021-04172-x.

[21] F. P. Appio, S. Bröring, N. Sick, S. Lee, and L. Mora, "Editorial deciphering convergence: Novel insights and future ideas on science, technology, and industry convergence," *IEEE Trans. Eng. Manage.*, vol. 70, no. 4, pp. 1389–1401, Apr. 2023, doi: 10.1109/TEM.2023.3242518.

[22] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. ACM SIGKDD Int. Conf. Knowl.*, 2016, pp. 855–864.

[23] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. ACM SIGKDD Int. Conf. Knowl.*, 2014, pp. 701–710.

[24] S. Bröring, *The Front End of Innovation in Converging Industries: The Case of Nutraceuticals and Functional Foods*, 1st ed. Wiesbaden, Germany: Deutscher Universitäts-Verlag, 2005.

[25] C.-S. Curran and J. Leker, "Patent indicators for monitoring convergence—Examples from NFF and ICT," *Technol. Forecasting Social Change*, vol. 78, no. 2, pp. 256–273, Feb. 2011, doi: 10.1016/j.techfore.2010.06.021.

[26] T. S. Kim and S. Y. Sohn, "Machine-learning-based deep semantic analysis approach for forecasting new technology convergence," *Technol. Forecasting Social Change*, vol. 157, Aug. 2020, Art. no. 120095, doi: 10.1016/j.techfore.2020.120095.

[27] T. Kose and I. Sakata, "Identifying technology convergence in the field of robotics research," *Technol. Forecasting Social Change*, vol. 146, pp. 751–766, Sep. 2019, doi: 10.1016/j.techfore.2018.09.005.

[28] C. H. Song, D. Elvers, and J. Leker, "Anticipation of converging technology areas—A refined approach for the identification of attractive fields of innovation," *Technol. Forecasting Social Change*, vol. 116, pp. 98–115, Mar. 2017, doi: 10.1016/j.techfore.2016.11.001.

[29] S. Bornkessel, S. Bröring, and S. W. F. Omta, "Analysing indicators of industry convergence in four probiotics innovation value chains," *J. Chain Netw. Sci.*, vol. 14, no. 3, pp. 213–229, Jan. 2014, doi: 10.3920/JCNS2014.x011.

[30] D. E. Chubin, "State of the field the conceptualization of scientific specialties," *Sociol. Quart.*, vol. 17, no. 4, pp. 448–476, Sep. 1976, doi: 10.1111/j.1533-8525.1976.tb01715.x.

[31] C. S. Wagner et al., "Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature," *J. Informetrics*, vol. 5, no. 1, pp. 14–26, Jan. 2011, doi: 10.1016/j.joi.2010.06.004.

[32] N. Battard, "Convergence and multidisciplinarity in nanotechnology: Laboratories as technological hubs," *Technovation*, vol. 32, no. 3/4, pp. 234–244, Mar. 2012, doi: 10.1016/j.technovation.2011.09.001.

[33] T. S. Kuhn, *The Structure of Scientific Revolutions*, 2nd ed. Chicago, IL, USA: Univ. Chicago Press, 1962.

[34] M. Coccia, "Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence," *Technol. Soc.*, vol. 60, Feb. 2020, Art. no. 101198, doi: 10.1016/j.techsoc.2019.101198.

[35] J. Mittelstraß, "Forschung und gesellschaft: Von theoretischer und praktischer transdisziplinarität," *Gaia*, vol. 27, no. 2, pp. 201–204, Jan. 2018, doi: 10.14512/gaia.27.2.4.

[36] X. Sun, J. Kaur, S. Milojević, A. Flammini, and F. Menczer, "Social dynamics of science," *Sci. Rep.*, vol. 3, no. 1, Jan. 2013, Art. no. 1069, doi: 10.1038/srep01069.

[37] M. Coccia, "The evolution of scientific disciplines in applied sciences: Dynamics and empirical properties of experimental physics," *Scientometrics*, vol. 124, no. 1, pp. 451–487, Jul. 2020, doi: 10.1007/s11192-020-03464-y.

[38] A. Bonaccorsi, "Search regimes and the industrial dynamics of science," *Minerva*, vol. 46, no. 3, pp. 285–315, Sep. 2008, doi: 10.1007/s11024-008-9101-3.

[39] A. L. Porter and I. Rafols, "Is science becoming more interdisciplinary? Measuring and mapping six research fields over time," *Scientometrics*, vol. 81, no. 3, pp. 719–745, Apr. 2009, doi: 10.1007/s11192-008-2197-2.

[40] D.-H. Jeong, K. Cho, S. Park, and S.-K. Hong, "Effects of knowledge diffusion on international joint research and science convergence: Multiple case studies in the fields of lithium-ion battery, fuel cell and wind power," *Technol. Forecasting Social Change*, vol. 108, pp. 15–27, Jul. 2016, doi: 10.1016/j.techfore.2016.03.017.

[41] D. Kong, J. Yang, and L. Li, "Early identification of technological convergence in numerical control machine tool: A deep learning approach," *Scientometrics*, vol. 125, no. 3, pp. 1983–2009, Dec. 2020, doi: 10.1007/s11192-020-03696-y.

[42] A. Bonaccorsi and J. Vargas, "Proliferation dynamics in new sciences," *Res. Policy*, vol. 39, no. 8, pp. 1034–1050, Oct. 2010, doi: 10.1016/j.respol.2010.05.002.

[43] N. Kim, H. Lee, W. Kim, H. Lee, and J. H. Suh, "Dynamic patterns of industry convergence: Evidence from a large amount of unstructured data," *Res. Policy*, vol. 44, no. 9, pp. 1734–1748, Nov. 2015, doi: 10.1016/j.respol.2015.02.001.

[44] G. Heimeriks and L. Leydesdorff, "Emerging search regimes: Measuring co-evolutions among research, science, and society," *Technol. Anal. Strategic Manage.*, vol. 24, no. 1, pp. 51–67, Jan. 2012, doi: 10.1080/09537325.2012.643562.

[45] D. A. Grégoire, M. X. Noël, R. Déry, and J. P. Béchard, "Is there conceptual convergence in entrepreneurship research? A co-citation analysis of frontiers of entrepreneurship research, 1981–2004," *Entrepreneurship Theory Pract.*, vol. 30, no. 3, pp. 333–373, May 2006, doi: 10.1111/j.1540-6520.2006.00124.x.

[46] D. Jeong, K. Lee, and K. Cho, "Relationships among international joint research, knowledge diffusion, and science convergence: The case of secondary batteries and fuel cells," *Asian J. Technol. Innov.*, vol. 26, no. 2, pp. 246–268, May 2018, doi: 10.1080/19761597.2018.1522961.

[47] S. F. Carley, N. C. Newman, A. L. Porter, and J. G. Garner, "An indicator of technical emergence," *Scientometrics*, vol. 115, no. 1, pp. 35–49, Apr. 2018, doi: 10.1007/s11192-018-2654-5.

[48] A. M. Petersen, M. E. Ahmed, and I. Pavlidis, "Grand challenges and emergent modes of convergence science," *Humanities Social Sci. Commun.*, vol. 8, no. 1, Aug. 2021, Art. no. 194, doi: 10.1057/s41599-021-00869-9.

[49] F. Caviggioli, "Technology fusion: Identification and analysis of the drivers of technology convergence using patent data," *Technovation*, vol. 55–56, pp. 22–32, Sep. 2016, doi: 10.1016/j.technovation.2016.04.003.

[50] H. Niemann, M. G. Moehrle, and J. Frischkorn, "Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application," *Technol. Forecasting Social Change*, vol. 115, pp. 210–220, Feb. 2017, doi: 10.1016/j.techfore.2016.10.004.

[51] S. Venugopalan and V. Rai, "Topic based classification and pattern identification in patents," *Technol. Forecasting Social Change*, vol. 94, pp. 236–250, May 2015, doi: 10.1016/j.techfore.2014.10.006.

[52] V. Giordano, F. Chiarello, N. Melluso, G. Fantoni, and A. Bonaccorsi, "Text and dynamic network analysis for measuring technological convergence: A case study on defense Patent data," *IEEE Trans. Eng. Manage.*, vol. 70, no. 4, pp. 1490–1503, Apr. 2023, doi: 10.1109/TEM.2021.3078231.

[53] C. Luan, S. Deng, A. L. Porter, and B. Song, "An approach to construct technological convergence networks across different IPC hierarchies and identify key technology fields," *IEEE Trans. Eng. Manage.*, to be published, doi: 10.1109/TEM.2021.3120709.

[54] S. Carley, A. L. Porter, I. Rafols, and L. Leydesdorff, "Visualization of disciplinary profiles: Enhanced science overlay maps," *J. Data Inf. Sci.*, vol. 2, no. 3, pp. 68–111, Aug. 2017, doi: 10.1515/jdis-2017-0015.

[55] S. P. Upham, L. Rosenkopf, and L. H. Ungar, "Innovating knowledge communities," *Scientometrics*, vol. 83, no. 2, pp. 525–554, May 2010, doi: 10.1007/s11192-009-0102-2.

[56] M. J. Mulkay, "Three models of scientific development," *Sociol. Rev.*, vol. 23, no. 3, pp. 509–526, Aug. 1975, doi: 10.1111/j.1467-954X.1975.tb02231.x.

[57] D. J. de Solla Price, *Little Science, Big Science and Beyond*. New York, NY, USA: Columbia Univ. Press, 1986.

[58] L. T. Phillips and B. S. Lowery, "Herd invisibility: The psychology of racial privilege," *Curr. Directions Psychol. Sci.*, vol. 27, no. 3, pp. 156–162, Jun. 2018, doi: 10.1177/0963721417753600.

[59] V. Larivière, C. Ni, Y. Gingras, B. Cronin, and C. Sugimoto, "Bibliometrics: Global gender disparities in science," *Nature*, vol. 504, pp. 211–213, Dec. 2013, doi: 10.1038/504211a.

[60] C. Wennerås and A. Wold, "Nepotism and sexism in peer-review," *Nature*, vol. 387, pp. 341–343, May 1997, doi: 10.1038/387341a0.

[61] J. West, J. Jacquet, M. King, S. Correll, and C. Bergstrom, "The role of gender in scholarly authorship," *PLoS One*, vol. 8, no. 7, Jul. 2013, Art. no. e66212, doi: 10.1371/journal.pone.0066212.

[62] H. Lee, P. R. Kim, and H. Zo, "Impact of cooperative R&D projects on ICT-based technology convergence," *ETRI J.*, vol. 39, no. 4, pp. 467–479, Aug. 2017, doi: 10.4218/etrij.17.0116.0874.

[63] I. Hwang, "The effect of collaborative innovation on ICT-based technological convergence: A patent-based analysis," *PLoS One*, vol. 15, no. 2, Feb. 2020, Art. no. e0228616, doi: 10.1371/journal.pone.0228616.

[64] I. Linkov, M. Wood, and M. Bates, "Scientific convergence: Dealing with the elephant in the room," *Environ. Sci. Technol.*, vol. 48, no. 18, pp. 10539–10540, Sep. 2014, doi: 10.1021/es503585u.

[65] Y. Gingras, *Bibliometrics and Research Evaluation: Uses and Abuses*. Cambridge, MA, USA: MIT Press, 2016.

[66] S. Saha, S. Saint, and D. A. Christakis, "Impact factor: A valid measure of journal quality?," *J. Med. Library Assoc.*, vol. 91, no. 1, pp. 42–46, Jan. 2003.

[67] L. Leydesdorff, D. Rotolo, and I. Rafols, "Bibliometric perspectives on medical innovation using the medical subject headings of PubMed," *J. Amer. Soc. Inf. Sci.*, vol. 63, no. 11, pp. 2239–2253, Nov. 2012, doi: 10.1002/asi.22715.

[68] D. Rotolo, D. Hicks, and B. R. Martin, "What is an emerging technology?," *Res. Policy*, vol. 44, no. 10, pp. 1827–1843, Dec. 2015, doi: 10.1016/j.respol.2015.06.006.

[69] A. M. Petersen, D. Rotolo, and L. Leydesdorff, "A triple helix model of medical innovation: Supply, demand, and technological capabilities in terms of medical subject headings," *Res. Policy*, vol. 45, no. 3, pp. 666–681, Apr. 2016, doi: 10.1016/j.respol.2015.12.004.

[70] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[71] E. Mitraka, A. Waagmeester, S. Burgstaller-Muehlbacher, L. M. Schriml, A. I. Su, and B. M. Good, "Wikidata: A platform for data integration and dissemination for the life sciences and beyond," *bioRxiv*, Nov. 2015, Art. no. 031971, doi: 10.1101/031971.

[72] Wikidata, *Wikidata:Statistics*, Jan. 2022. [Online]. Available: https://www.wikidata.org/wiki/Wikidata:Statistics

[73] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[74] Y. Dong, N. V. Chawla, and A. Swami, "Metapath2vec: Scalable representation learning for heterogeneous networks," in *Proc. ACM SIGKDD Int. Conf. Knowl.*, 2017, pp. 135–144.

[75] L. Galke et al., "Inductive learning of concept representations from library-scale bibliographic corpora," in *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft*. Bonn, Germany: Gesellschaft für Informatik, 2019, pp. 219–232.

[76] M. Coccia, "General properties of the evolution of research fields: A scientometric study of human microbiome, evolutionary robotics and astrobiology," *Scientometrics*, vol. 117, no. 2, pp. 1265–1283, Nov. 2018, doi: 10.1007/s11192-018-2902-8.

[77] S. Rozner and N. Garti, "The activity and absorption relationship of cholesterol and phytosterols," *Colloids Surfaces A, Physicochem. Eng. Aspects*, vol. 282–283, pp. 435–456, Jul. 2006, doi: 10.1016/j.colsurfa.2005.12.032.

[78] C. Stancu and A. Sima, "Statins: Mechanism of action and effects," *J. Cellular Mol. Med.*, vol. 5, no. 4, pp. 378–387, Oct. 2001, doi: 10.1111/j.1582-4934.2001.tb00172.x.

[79] M. Hoskins and T. Jacobson, "Combination use of statins and omega-3 fatty acids: An emerging therapy for combined hyperlipidemia," *Future Lipidol.*, vol. 1, no. 5, pp. 579–591, Oct. 2006, doi: 10.2217/17460875.1.5.579.

[80] S. Han et al., "Effects of plant stanol or sterol-enriched diets on lipid profiles in patients treated with statins: Systematic review and meta-analysis," *Sci. Rep.*, vol. 6, Aug. 2016, Art. no. 31337, doi: 10.1038/srep31337.

[81] D. M. T. Malina et al., "Additive effects of plant sterols supplementation in addition to different lipid-lowering regimens," *J. Clin. Lipidol.*, vol. 9, no. 4, pp. 542–552, Jul. 2015, doi: 10.1016/j.jacl.2015.04.003.

[82] M. Hallikainen, S. Kurl, M. Laakso, T. Miettinen, and H. Gylling, "Plant stanol esters lower LDL cholesterol level in statin-treated subjects with type 1 diabetes by interfering the absorption and synthesis of cholesterol," *Atherosclerosis*, vol. 217, no. 2, pp. 473–478, Apr. 2011, doi: 10.1016/j.atherosclerosis.2011.03.041.

[83] A. Yegros-Yegros, I. Rafols, and P. D'Este, "Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity," *PLoS One*, vol. 10, no. 8, Aug. 2015, Art. no. e0135095, doi: 10.1371/journal.pone.0135095.

[84] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, Sep. 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[85] F. Scarselli, M. Gori, A. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.

[86] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *J. Amer. Soc. Inf. Sci.*, vol. 24, no. 4, pp. 265–269, Jul. 1973, doi: 10.1002/asi.4630240406.

[87] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 9244–9255.

[88] D. K. Sanyal, P. K. Bhowmick, and P. P. Das, "A review of author name disambiguation techniques for the PubMed bibliographic database," *J. Inf. Sci.*, vol. 47, no. 2, pp. 227–254, 2021, doi: 10.1177/0165551519888605.

Tetyana Melnychuk received the B.Sc. and M.Sc. degrees in business chemistry in 2015 and 2018, respectively, from Kiel University, Kiel, Germany, where she is currently working toward the Ph.D. degree in technology management.

She was a Research Assistant with Kiel University, Kiel, Germany. Her research interests include open innovation, network management, scientific and technology convergence, and management of interdisciplinary research with a focus on university–industry collaborations. Her work is published in the *Journal of Product Innovation Management*.

Lukas Galke was born in Wuppertal, Germany, in 1989. He received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Kiel University, Kiel, Germany, in 2017 and 2023, respectively.

From 2017 to 2021, he was a Doctoral Researcher with Kiel University and ZBW - Leibniz Information Centre for Economics. Since 2022, he has been a Postdoctoral Researcher with the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. He is the author of more than 20 journal articles, conference proceedings papers, and workshop papers. His research interests include machine learning, natural language processing, and information retrieval.

Dr. Galke was a recipient of the Association for Computational Linguistics (ACL) Outstanding Reviewer Award in 2023. He is a member of the Gesellschaft für Informatik and ACL.

Eva Seidlmayer received the first MA degree after studying philosophy and ancient history from the University of Hannover, Hannover, Germany, and Goethe University Frankfurt, Frankfurt, Germany, in 2011, and the master's degree in library and information science from the TH Köln – University of Applied Sciences, Cologne, Germany, in 2018.

Her Ph.D. addressed the topic of contextualized particularism and fundamental universalism, offering a philosophical and historical critique of the competition and complementarity of typical relationships between thought and action. It was published by J. B. Metzler (Springer Nature) in 2018 as a monograph. Her master's thesis deals with an "Ontology of digital objects in philosophy."

Dr. Seidlmayer was a recipient of a three-year full Ph.D. scholarship by the Rosa Luxemburg Foundation as well as a fellowship in 2019/2020 through the Open Science Fellows Program run by Wikimedia Deutschland, the Volkswagen Foundation, and the Stifterverband initiative. Since December 2022, she has been in charge of the project AQUAS - Automatic Quality Assessment: NLP methods for semantic mapping of life-science texts, which is funded by the German Research Foundation (DFG). Her work as a Postdoctoral Researcher is located with ZB MED – Information Centre for Life Sciences, Cologne, and focuses on the enrichment of bibliographic metadata with quality data. As a part of that, she is developing a machine-learning-based approach that will automatically determine whether the characteristics of a text are similar to those of publications within the categories of science, popular science, or disinformation. She is a member of the DFG network "Digital Bioethics."

Stefanie Bröring studied medicine at the University of Lübeck, Germany. She received the bachelor's degree in international management from the Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands, in 2000, the master's degree in business management from the University of Münster, Germany, in 2001, the Ph.D. degree in innovation and technology management from the University of Münster, Münster, Germany, in 2005.

She holds the Chair for Entrepreneurship and Innovative Business Models and is a Professor with Ruhr-University Bochum (RUB), Bochum, Germany. She also serves as the Academic Director of the RUB Worldfactory Start-up Centre, Bochum, Germany. Her research interests span the emergence of novel (clean) technology systems, technology transfer, business model transformation, as well as the rise of cross-industry ecosystems triggered by science and technology convergence. Her work has been published in the *Journal of Product Innovation Management,* IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT*, R&D Management, Technovation, Journal of Technology Transfer, Technological Forecasting and Social Change*, and more.

Konrad U. Förstner received the master's degree in biochemistry from the University of Greifswald, Greifswald, Germany, in 2004, and the Ph.D. degree in bioinformatics from the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, in 2009.

He worked as a freelancing Software Developer after receiving his doctorate and then started postdoctoral research with Jörg Vogel with the Institute for Molecular Infection Biology and Cynthia Sharma with the Center for Infectious Diseases (ZINF), Würzburg, Germany. He first became the Head of bioinformatics with the core unit Systems Medicine of the University and the University Clinic Würzburg and then the Head of the full unit. Since May 2018, he has been a Joint Professor of data and information literacy with the TH Köln Campus, Köln, Germany, and ZB MED - Information Centre of Life Sciences, Cologne, Germany, where he leads the Data Science and Services Unit. His research covers the integration of large data corpora in the life sciences to answer a diverse set of biological questions and includes the system biological analysis of high-throughput data, automatic text analysis of biomedical literature, and the development of research software for these purposes.

Dr. Konrad U. Förstner is a member of several organizations that promote Open Science. He is the speaker of the NFDI4Microbiota consortium of the National Research Data Infrastructure in Germany and a founding member of the working group "Open Science" of the Open Knowledge Foundation Germany, and he represents the German Rectors' Conference in the working group "Digital tools: Software and Services" in the Priority Initiative "Digital Information," which is a joint initiative of the Alliance of Science Organizations in Germany. Since 2019, he has been a member of the Advisory Board of the Weizenbaum Institute for the Networked Society - the German Internet Institute, and since 2020, he has been a member of the Executive Council of The Carpentries.

Klaus Tochtermann studied computer science at Kiel University and Dortmund University, Germany. He received the master's degree in computer science and the Ph.D. degree from Dortmund University, in 1991 and 1995, respectively.

In 1996, he was a Postdoctoral Researcher with Texas A&M University. From 2004 to 2010, he was a Full Professor of Knowledge Management and Knowledge Technologies with the Graz University of Technology, Austria. Since 2010, he has been the Director of the ZBW - Leibniz Information Centre for Economics, Kiel and Hamburg and holds a professorship for Digital Information Infrastructures with Kiel University, Kiel, Germany. The ZBW is the world's largest information infrastructure for scientific information in economics, both online and offline. His research focuses on open science, digital information infrastructures, and research data management.

Mr. Tochtermann is involved in national and European committees and initiatives on the topic of open science and research data management. These include the GO FAIR initiative, which Germany launched together with The Netherlands and France to establish the FAIR principles for research data. Among others, he is also a member of the scientific senate of the National Research Data Infrastructure in Germany and a member of the Board of Directors of the European Open Science Cloud Association.

Carsten Schultz received the master's degree in business engineering and the Ph.D. degree in innovation management from the Technical University of Berlin, Berlin, Germany, in 2001 and 2006, respectively.

He has been a Professor of technology management with Kiel University, Kiel, Germany, since 2012. From 2008 to 2011, he held the assistant (junior) professorship for the Management of Service Innovations and Technology Transfer endowed by Deutsche Telekom AG with the Technical University of Berlin. In 2004–2008, he headed the Siemens Center for Knowledge Interchange, the strategic partnership between Siemens and TU Berlin. His work focuses on analyzing the challenges and success factors of innovation management, of digital business models and services, and of the increasing complexity of value creation and open innovation processes. In several research projects, his team is dedicated to the specific aspects of innovation activity in the life sciences and energy industries as well as the success factors of university–industry collaborations. He is a member of the editorial board of the *Journal of Product Innovation Management,* the journal *Creativity Innovation Management*, and the *Health Care Management Review*. He is the author of numerous publications in international journals and, among others, of the textbook Innovation Management, which was published in its 7th edition in 2022. He is committed to promoting interdisciplinary research and technology transfer as a member of the scientific board of the Center for Entrepreneurship, Kiel University, as a member of the supervisory board of Kiel Science Center, and as the Head of the Innovation Management subproject of the Mittelstand Digital Zentrum Schleswig-Holstein. He is also a member of the scientific advisory board of the German Foundation for the Chronically Ill.