Scientometrics (2026) 131:6045–6096
https://doi.org/10.1007/s11192-026-05736-5

From green thesis to green invention: tracing the contribution of PhDs to green technology transfer of german universities

Guido Buenstorf[1] · **Alana Hindiyeh**[1] · **Johannes Koenig**[1,2] · **Liam Paul Neumann**[1] · **Burcu Ozgun**[1,3]

Received: 24 September 2025 / Accepted: 29 June 2026 / Published online: 21 July 2026
© The Author(s) 2026

Abstract
As a driver of innovation, university research plays an important role in the transition to a more sustainable future. Yet the role of PhD researchers in fostering green innovation remains underexplored. This study examines how PhD graduates contribute to green technology transfer by linking their academic and inventive outputs. Focusing on German PhD recipients in two fields of research (physics, and electrical engineering and computer science), we construct a new record linkage connecting PhD theses, scientific publications, patents, and patent-to-publication citations. To consistently identify research output related to environmental sustainability across document types, we use SPECTER2.0 embeddings to construct a semantic similarity–based measure of sustainability in PhD theses and patents, utilizing a curated corpus of publications. Our findings offer insights into how PhDs contribute to green innovation, both directly as inventors and indirectly through citations of their academic research in patents. We find that environmental sustainability-oriented PhDs are more frequently associated with green patenting, and that indirect contributions via citations received in patents are much more prevalent than direct contributions via patenting. We discuss and address potential limitations stemming from the absence of a direct benchmark through comparisons to established classification schemes, robustness checks, and qualitative analyses.

Keywords Green innovation · Green patents · PhD · Green technology transfer · Germany · Thesis-publication linkage

✉ Alana Hindiyeh
alana.hindiyeh@uni-kassel.de

[1] INCHER and Institute of Economics, University of Kassel, Kassel, Germany

[2] IAB Institute for Employment Research, Saarbruecken, Germany

[3] Department of Human Geography and Spatial Planning, Utrecht University, Utrecht, Netherlands

Introduction

Universities are central actors in generating and transferring knowledge that drives technological and social innovation (Mansfield, 1995). Their role has become particularly salient in the context of sustainability transitions, where addressing challenges such as climate change requires not only new knowledge but also its effective dissemination across society. In this regard, knowledge transfer through commercialization and academic engagement (Perkmann et al., 2021, 2013), based on collaborations with industry, government, and civil society (Di Maria et al., 2019; Stephens et al., 2008; Trencher et al., 2013), is of crucial importance.

While the importance of universities for sustainable development and for technological and social innovations has long been recognized, the sustainability dimension has not received much prior attention in the technology transfer literature (Chams et al., 2024; Yang, 2024). In particular, limited evidence exists on the extent to which PhD training contributes to the production and dissemination of knowledge relevant to green innovation. PhD programs generate cutting-edge research that can drive technological and societal transformation (Matthies et al., 2022; Stephens et al., 2008), while also shaping future experts who contribute to sustainability transitions. Their impact materializes through both direct and indirect channels: some PhD researchers become inventors who translate research into new products and processes, whereas others influence innovation indirectly by producing knowledge that flows to firms and societal actors (Breschi & Lissoni, 2001) incorporating it into sustainable solutions. These contributions extend across multiple pathways, including academic publications, industry collaborations, and interdisciplinary research networks. Yet, despite the recognized importance of academic patenting (Breschi et al., 2005), academic engagement (Perkmann et al., 2021, 2013), and labor mobility-driven knowledge transfer (Buenstorf & Heinisch, 2020; Kaiser et al., 2018), the specific role of PhD recipients in green innovation remains underexplored. We explore this role utilizing data on PhD graduates' own patent applications as well as citations their publications receive in patent documents (Narin et al., 1997).

A further challenge is that research relevant to environmental sustainability is not consistently delineated across datasets. In this paper, we adopt a deliberately focused definition that centers on environmental sustainability, rather than on broader social or economic dimensions of the concept. Specifically, we conceptualize sustainability in terms of research and innovation activities that contribute to environmental protection and climate-change mitigation, including the development of environmentally related and low-carbon technologies.[1] Building on this definition, we adopt a broad perspective in identifying relevant research outputs: we consider not only work that explicitly addresses sustainability, but also work that is oriented toward it. This wide view allows us to capture the full range of contributions PhDs make to green inventions. Throughout the paper, we refer to such work and to the researchers who engage in it as sustainability-oriented.

Building on a large-scale data linkage, this paper examines the contributions of PhD recipients to green innovation, focusing on both their academic and inventive outputs. We address three research questions: (1) How can text-based methods be used to identify research outputs relevant for environmental sustainability? (2) To what extent, and in what

[1] Other dimensions of sustainability, such as social sustainability, economic sustainability, or issues related to equity and inclusion, are beyond the scope of the present analysis.

ways, do PhD recipients contribute to green inventive activities? (3) Which characteristics of PhD training and scholarly output are associated with direct and indirect contributions to green innovation?

We address these questions in the empirical context of Germany, the largest producer of PhDs in the European Union (Oecd, 2018) and a global leader in patenting activity (European Patent Office, 2024). These characteristics make Germany an ideal setting for analyzing the role of PhDs in green innovation. Our analysis covers PhD theses, associated publications, and patents from PhD recipients who obtained their degree between 2000 and 2020 in two STEM fields with both high patenting propensity and strong relevance for sustainability: physics and electrical engineering/computer science (EECS). We trace PhD recipients' research and inventive activity by linking theses to subsequent publications, patents, and patents citing their work, and assess the sustainability orientation of both their PhD research and downstream inventions. To this end, we develop an approach that leverages SPECTER2.0 embeddings applied to publications and inventions by PhDs, offering a valuable complement to traditional expert-based patent classifications commonly used in the literature.

With this analysis, we bridge two largely separate streams of research: work on sustainability-oriented scholarship and work on knowledge and technology transfer between universities and other societal actors. Conceptually, the study advances understanding of how early-career researchers contribute to sustainability-oriented knowledge transfer through both direct inventorship and indirect channels of research utilization. Methodologically, it introduces a text-based approach to identifying sustainability-oriented research outputs, extending and complementing established expert-based classifications such as the Green Inventory of the International Patent Classification (IPC) by the World Intellectual Property Organization (WIPO), and the Y02/Y04 classification scheme developed by the European Patent Office (EPO). Together, these contributions provide descriptive evidence on the role of PhD recipients in green innovation and offer tools for systematically analyzing sustainability-oriented knowledge transfer.

The remainder of this paper is structured as follows. Section "Academic and inventive contributions of PhDs to green innovation" reviews the relevant literature and establishes the motivation for the research questions. Section "Data and methodology" details the datasets, explains the linkage between PhD recipients, their publications, and patents, and outlines the text-based approach used to assess sustainability orientation. Section "Analysis and findings" presents the empirical findings, offering insights into the contributions of PhDs to green innovation. Finally, Sect. "Conclusions" summarizes the key findings and discusses the implications for future research.

Academic and inventive contributions of PhDs to green innovation

Tackling climate change and other global environmental challenges requires a fundamental shift that transitions established socio-technical systems toward more sustainable modes of production and consumption (Markard et al., 2012). Such transitions to a greener economy are inherently tied to technological change (Stern & Valero, 2021). The sustainability transitions literature emphasizes that these transformations are driven not only by incremental technological improvements, but by the development and diffusion of new knowledge, skills, and capabilities that enable novel technological trajectories (Schot & Steinmueller, 2018). Because scientific advances play a central role in enabling innovation, academic research,

and higher education; institutions are well-positioned to contribute to addressing the climate crisis (Di Maria et al., 2019; Stephens et al., 2008; Trencher et al., 2013). Traditionally, universities have pursued missions encompassing education, research, and—more recently—knowledge transfer and entrepreneurship (Clark, 1998; Etzkowitz & Leydesdorff, 2000). Within this framework, universities serve not only as producers of explicit knowledge but also as training grounds for highly specialized human capital capable of translating scientific advances into practical applications. The knowledge they produce has been shown to significantly influence industrial innovation and technological change (Cohen et al., 2002; Jaffe, 1989; Mansfield, 1995). Given the complexity of sustainability challenges, such scientific knowledge is especially critical, as innovation in this domain is typically more dependent on knowledge-intensive competencies than in other fields (De Marchi, 2012; Schot & Steinmueller, 2018). Green innovation often involves higher uncertainty, stronger regulatory interaction, and greater interdisciplinarity (De Marchi, 2012; Popp, 2012; Porter & van der Linde, 1995; Schot & Steinmueller, 2018), increasing the importance of advanced scientific training and problem-solving capacity (Nooteboom et al., 2007; Spender & Grant, 1996). Recent research on science-industry relations further emphasizes that such knowledge-intensive innovation processes are increasingly shaped by interactive and collaborative forms of knowledge production rather than linear models of technology transfer (Cantner et al., 2024; Yström et al., 2025). Therefore, gaining deeper insight into how academic research contributes to the resolution of environmental challenges through innovation is timely and essential.

A key element of this contribution lies in the knowledge generated through PhD research. PhD candidates conduct independent research and acquire in-depth knowledge of their field of research. From a human capital and knowledge-based perspective, PhD graduates represent carriers of both codified and tacit knowledge, enabling the recombination of scientific insights with applied technological contexts (Auriol et al., 2016; Buenstorf & Heinisch, 2020; Spender & Grant, 1996). Accordingly, PhDs are frequently conceptualized as key micro-level agents in knowledge transfer processes linking academic research to technological development (Perkmann et al., 2021, 2013). PhD training has also been emphasized by national science policies as playing an integral role in shaping early-career researchers' knowledge transfer across academic and non-academic domains (Gaughan & Robin, 2004; Shibayama, 2019). Reflecting these policy expectations, a growing body of empirical work has begun to examine PhD contributions to innovation beyond academic settings (Buenstorf & Heinisch, 2020; Kaiser et al., 2018; Toole & Czarnitzki, 2010; Zucker et al., 2002). While PhD graduates are expected to drive advances in research, systematic evidence on the nature and scope of their contributions to innovation outside academia remains limited and uneven (Auriol et al., 2016). Kaiser et al. (2018) find that hiring individuals with university research experience can enhance a firm's innovative output; however, their study primarily focuses on the mobility aspect of researchers, with limited exploration of the specific roles PhD recipients play in the innovation process. Buenstorf and Heinisch (2020) show that PhD graduates are key contributors to private-sector innovation and that hiring PhDs may enable firms to diversify into new fields of technology. These studies suggest that PhD recipients influence innovation outcomes beyond academia, but provide little insight into the substantive orientation of the knowledge being transferred.

More recent studies provide insight into the mechanisms through which PhD researchers contribute to knowledge transfer within collaborative research environments. Plantec et al. (2023) show that early-career researchers engaged in university-industry collaborative PhD projects act as boundary-spanning agents whose research orientation and engagement influence project performance. Related work highlights that science-industry collaborations are characterized by interdependencies between scientific and technological outcomes, with

research orientations emerging endogenously from collaborative knowledge creation processes (Cabanes et al., 2024; Cantner et al., 2024). At the same time, academic engagement with external partners can feed back into scientific knowledge production itself: Nast et al. (2025) find a positive association between academic engagement and scientific impact, suggesting that interaction with non-academic contexts can serve as a source for novel insights for research. While this literature substantially advances our understanding of PhD contributions to innovation and engagement more broadly, it remains largely technology-neutral. In particular, there is still little direct empirical evidence on whether, and how, PhD recipients contribute specifically to sustainability-oriented or green innovation. This points to a gap in the literature concerning the role of PhDs in shaping green innovation outcomes.

Empirically identifying the contribution of PhDs to green technological change is challenging. Traditional measures of innovation, such as R&D expenditures or the number of active researchers, often do not adequately capture the nuances of green innovation, since aggregated data may overlook specific sustainable practices (Haščič & Migotto, 2015). Furthermore, even when data on sustainable activities are available, it can be inconsistent or difficult to compare between different studies (Popp, 2012). Consequently, both patent and bibliometric data have been identified as potentially useful alternatives for analyzing green innovation (Haščič & Migotto, 2015; Popp, 2012). Our study aims to take advantage of patent and bibliometric data by exploiting machine learning techniques to develop record linkage algorithms.

Furthermore, most studies have explored these topics through the lens of sustainable development, often relying on standardized definitions or alignment with the Sustainable Development Goals (SDG) (Adams et al., 2016; Beynaghi et al., 2016; Celone et al., 2022). While such approaches provide valuable benchmarks, they may obscure emerging or interdisciplinary forms of sustainability-oriented research that do not neatly align with predefined taxonomies. Our study differs by adopting an approach that identifies sustainability-oriented content without adhering to preset definitions. Utilizing SPECTER2.0, a pre-trained language model fine-tuned for scientific texts, presents a promising avenue for enhancing our understanding of the contributions of academic research to sustainability.

By aggregating sustainability-oriented publications into a unified embedding, we can compare each thesis individually to an overarching sustainability corpus embedding. This approach operationalizes thematic proximity and cognitive relatedness (Nooteboom et al., 2007; Rafols, 2014) between PhD research and sustainability-oriented knowledge, concepts that capture the extent to which knowledge bases overlap and can be effectively recombined in processes of innovation and technological change (Schot & Steinmueller, 2018). It not only captures thematic coherence across fields but also mitigates the noise associated with variances in individual publication topics. The emerging use of SPECTER2.0 in related studies (Caravale et al., 2023; Duede et al., 2024) positions it as a valuable tool to analyze academic contributions to green innovation. Caravale et al. (2023) leverage this model to generate embeddings that capture the semantic relationships in digital archaeology texts, producing a more nuanced and detailed classification of digital archaeology research. Duede et al. (2024) use SPECTER2.0 embeddings to calculate the semantic similarity between AI-engaged and non-AI-engaged academic publication titles and abstracts across different disciplines compared to AI-concentrated research in Computer Science.[2] Furthermore, Ji et al. (2024) emphasizes the effectiveness of SPECTER2.0 embeddings

[2] At the time of writing, this article is only available as a preprint, and to the best of the authors' knowledge, no other research using the SPECTER2.0 embedding model in a similar comparative method has yet been published.

in managing large datasets, showcasing its utility in investigating the academic and inventive contributions of PhDs to green innovation. Building on this emerging methodological literature, our study applies SPECTER2.0 with thematic proximity measures to examine how PhD research aligns with sustainability-oriented knowledge and how such alignment translates into academic and inventive outputs.

Our study contributes to the literature in two main ways. First, it examines the direct and indirect contributions of PhD recipients to sustainability-oriented innovation, along with the determinants of these contributions. By explicitly focusing on PhDs as carriers of advanced scientific knowledge, the study extends existing research on academic knowledge transfer and green innovation. Second, it offers a methodological advancement by proposing a approach to linking PhD recipients with their publications and patents, thereby facilitating the identification of knowledge transfer activities.

Data and methodology

This section outlines the methodological approach for tracing the contribution of PhDs to green technology transfer of German universities. Linking theses to the journal articles and patents that stem from them is complicated by homonymy (different researchers with identical names) and synonymy (individuals appearing under multiple name variants). Section "Record linkage for PhD recipients, publications and patents" therefore introduces an automated record-linkage methodology that overcomes these obstacles, connecting theses, publications, and patents with minimal reliance on human annotators. The resulting pipeline scales efficiently to large datasets and can be readily repurposed for other publication author and inventor matching tasks. Later, in Sect "Identifying the sustainability-orientation of theses, publications, andpatents", we specify the criteria and procedures applied to the linked records to identify theses, publications, and patents that are explicitly sustainability-oriented.

Record linkage for PhD recipients, publications and patents

This subsection provides a detailed description of the automated record-linkage methodology that links PhD recipients to their publications and patents, and its overall workflow is summarized in Fig. 1.

Identifying PhDs and their theses

We identify PhD recipients from German universities through the German National Library (DNB) catalog, which archives nearly all German university theses, both physical and electronic. This librarian-curated collection—Europe's largest repository of online theses—is considered highly reliable and has been used in previous studies (Buechele et al., 2026; Buenstorf & Geissler, 2014; Buenstorf et al., 2023, 2025; König, 2022). The raw DNB dataset contains more than 1.9 million records.

We focus on PhD theses submitted between 2000 and 2020 in two STEM fields with high patenting propensity and strong sustainability relevance: physics and EECS.[3] To pinpoint theses in the focal fields, we augmented the DNB subject codes with faculty and degree titles, improving classification precision as detailed in Appendix A. We ultimately identify 44,720 relevant theses—29,561 in physics and 15,159 in EECS.

Identifying publications of the PhDs

We apply a four-stage process to link the authors of these PhD theses with their Scopus-listed publications,[4] which combines classification based on high-quality matching criteria with machine learning techniques. While our final analysis focuses on physics and on EECS, the linkage of thesis authors with their Scopus-listed publications was performed for all subjects. Our approach thus leverages specific characteristics of population-level data to establish reliable linkages between two databases.

Data Preprocessing

The linkage is based on all publications listed in Scopus between 1996 to 2022 and PhD theses across all fields submitted between 2000 and 2020. To prepare the data, the names of thesis authors and Scopus authors were standardized. All names were converted to lowercase and stripped of punctuation,[5] For each author, tokens were then created from the last name and the initials of the first name.[6]

Linking PhD and Publication Databases

After author names were cleaned in Scopus and DNB, both databases were linked using the initials and last name tokens. The analysis was limited to individuals with fewer than 300 potential matching partners,[7] and only Scopus authors with publications where at least one author has a German affiliation were considered. The final linked dataset includes 484,004 PhD recipients matched to 1,096,177 unique Scopus authors.

For the matching process executed by machine learning algorithms, synthetic training and test datasets were created based on a probabilistic approach. The foundation consisted of unique matches, defined as cases in which a single PhD author listed in the DNB catalog (1975–2021) shared a first and last name combination with a single Scopus author. Additionally, to qualify as a unique match, the Scopus author had to have at least one publication with a co-author listing an affiliation in Germany. This resulted in approximately 100,000 unique cases (true positives). From this, a sample of exactly 50,000 unique matches was

[3] The electrical engineering and computer science discipline is known as *Informatik/Elektrotechnik* in German.

[4] We use Scopus data (version 04/2023). The dataset access is provided by the Competence Network for Bibliometrics (Schmidt et al., 2024). For more information about the center, see: https://bibliometrie.info/en/.

[5] The German language contains several umlauts not represented in the Latin alphabet. A transliteration was performed, converting umlauts such as ö ü, and ä into oe, ue, and ae, and ß into ss. Other characters were converted to their Latin variants.

[6] The use of first-name initials rather than full first names reflects limitations in the publications metadata: particularly in the early 2000s, but also in later years; depending on journal policy, only author initials are recorded. Using initials consistently avoids systematically excluding publications for which full first names are unavailable and ensures comparability across publication years.

[7] An evaluation revealed that only 1.4% of PhD candidates were linked to more than 300 potential Scopus authors, which accounted for 31.81% of all potential matches.

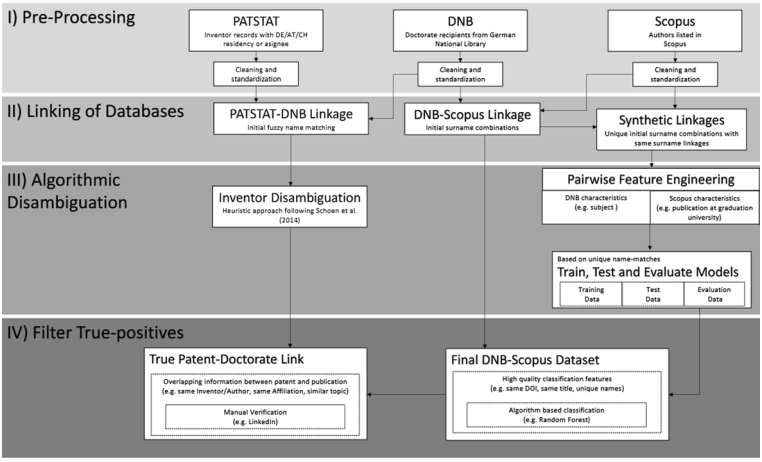


Fig. 1 Overview of the data processing and record linkage procedure

drawn. True negatives were assigned based on cases where the last name matched, but the initials differed. Additionally, careful consideration was given to ensure that the training dataset maintained the same distribution of multiple linkages as the full original database.[8]

The advantage of using a synthetic training database lies in its high precision in generating a sufficiently large training and testing set while avoiding human decision-making errors. The curated true positives and the artificially generated true negatives should be highly precise, because both the DNB thesis database and the Scopus publication records provide near-complete population coverage. Consequently, if a unique name token appears in both datasets—one linked to a thesis in Germany and the other to a publication with a German affiliation—it is highly probable that it belongs to the same individual. This methodology also accounts for shifting publication patterns across time and disciplines, as the algorithm can leverage extensive datasets to detect unique trends. In contrast, manually curated training sets tend to remain small because record-level verification is highly labor-intensive. At the same time, such datasets can reflect misclassifications or other inaccuracies, which in turn may reinforce existing social biases. The chosen approach mitigates these biases by assuming that the classification characteristics of true negatives are independent of the initial letter of the first name.[9]

[8] Original Database: min=0; median=3; max=300; Training Database: min=0; median=3; max=297.

[9] At the same time, synthetic training approaches may also introduce other potential sources of bias, although these biases may differ from those arising from manually curated training datasets. In particular, the constructed true positive and true negative linkages may not fully reflect complex real-world matching constellations. Systematic distortions could emerge if the distinctive name combinations used to construct the training data are correlated with individual characteristics such as international mobility. Similarly, potential bias may arise if characteristics associated with scientific careers systematically correlate with surname and initial combinations used in the linkage procedure. In such cases, the synthetic training data may not fully represent the underlying population structure. Consequently, linkage results based on synthetic training data should be critically evaluated for plausibility and complemented by additional robustness checks such as manual annotation exercises.

Training the algorithm

In the next step, variables were created to serve as input for the classification task. These variables are categorized based on whether they provide high precision in linking PhDs to their corresponding Scopus authors or, when considered alone, exhibit low precision and thus serve as input for the machine learning algorithm that recognizes relevant patterns in the training data.[10]

High-quality matching variables include:

- DOI/ISSN of the thesis matches the DOI/ISSN of a publication linked to a specific Scopus author.
- ORCID ID of the thesis author matches the ORCID ID in Scopus.
- Standardized title of the thesis corresponds to the title of a publication by the linked Scopus author.
- Initial name combination is unique among thesis authors in Germany and among Scopus authors having at least one publication with a German affiliation.

Lower-quality matching variables include:

- The first publication in Scopus was released within 4 years prior to and 2 years after graduation/publication of the thesis.
- The Scopus author has at least one affiliation with the university where the PhD was granted.
- A co-author from the Scopus author's affiliation is at the same university as the individual's PhD-granting university.
- Year of graduation/publication of the thesis.
- A dummy variable that takes the value of 1 if the Scopus author has/had an affiliation at a German university.
- The Scopus author publishes in the same disciplinary group to which the PhD is classified (humanities and arts, human sciences, economics, law, and social sciences, natural sciences with mathematics, engineering sciences, medicine, agricultural, forestry, and nutritional sciences, and sports).
- A SCOPUS co-author shares the same surname and initial as a committee member of the PhD audit committee.[11]

The artificial training dataset was first split into equal-sized training and evaluation sets. The training set was further divided, with one half used for training the algorithm and the other for testing. The model with the highest F1 score was selected for each approach (Random Forest, AdaBoost). Since algorithm selection also requires validation, the best-performing model's tuned parameters (based on the F1 criterion) were re-estimated on the training dataset and then evaluated on the evaluation dataset.

[10] While incorporating additional matching variables could in principle increase precision, our approach is constrained by the information consistently available in PhD dissertations across universities and over time. To ensure coverage over two decades and all German universities, we therefore rely on variables that are available in a standardized form across the full sample.

[11] Advisor information was sourced from the full text of online theses or was previously collected by the DNB.

For the classification task, Random Forest and AdaBoost algorithms were employed.[12] Prior work indicates that these algorithms yield superior classification performance in similar tasks (Heinisch et al., 2020).[13]

Taking supervisor information into account, the best AdaBoost algorithm—selected based on the F1 criterion—achieved recall, precision, and F1 scores of 0.8647, 0.9480, and 0.9045. The best Random Forest algorithm achieved recall, precision, and F1 scores of 0.8690, 0.9559, and 0.9103. When advisor information was not included, recall, precision, and F1 scores for Random Forest were 0.7460, 0.8986, and 0.8152, and for AdaBoost, they were 0.7337, 0.9006, and 0.8086.

Filtering true positives

A multi-stage procedure was then chosen for the final classification task. This procedure accounts for the varying quality of matching variables available for different segments of the data. Initially, the highest quality matching criteria were applied. For cases that could not be uniquely identified using these criteria, classification is based on the previously described machine learning approach. The sequence in which the procedure was followed is as follows:

- DOI/ISSN of the thesis matches the DOI/ISSN of publications in Scopus (44 pairs linked).
- Scopus author shares the same ORCID as the thesis author (239 pairs linked).
- The title of the thesis corresponds to a publication by the Scopus author (1844 pairs linked).
- Unique name surname combination (105,111 pairs linked).
- Linkage based on Random Forest algorithm first with supervisor information, then without (107,110 and 108,662 pairs linked).
- Linkage based on AdaBoost for samples with and without supervisor information (3151 and 4127 pairs linked).

For the full dataset covering all disciplines, 330,288 PhD recipients from graduation cohorts 2000 to 2020 were clearly identified from a total of 484,004 individuals for whom at least one author shares the same initials and last name as the PhD recipient.

Final evaluation

In a final step, we applied the algorithm to a manually annotated sample for all PhD recipients in STEM fields. This application provides insights not only into how effectively the synthetic training and test datasets has functioned, but also into how well the combination of high-quality matching criteria and machine learning algorithms performs for this classification task. We exclude the top 5% of individuals with the most matches due to data handling considerations. Otherwise the dataset would have become too comprehensive for three rounds of manual annotation.[14] To achieve high precision and recall of the manually annotated sample, the dataset was initially independently classified by two student assistants. In cases where the two student assistants arrived at different conclusions, a third

[12] For AdaBoost, the ada package (Culp et al., 2007) was used, while for Random Forest, the randomForest package (Liaw, 2002) was applied.

[13] Tuning parameters for the Random Forest algorithm include the number of trees to grow, varying between 20, 50, 100, 200, and 500, and the number of variables randomly sampled as candidates at each split, varying between 2, 5, 7, 10, and 15. For AdaBoost, the number of boosting iterations (50, 100, 250, 500) and the shrinkage parameter for boosting (0.01, 0.2, 0.5, 0.9, 1) were varied.

[14] Our sample contains 1344 PhD thesis authors linked to 8116 Scopus authors.

manual annotation was conducted by a senior researcher. A comparison of this sample of STEM graduates with the results from the classification task indicates that our classification approach achieved an accuracy of 96.75%, a precision of 91.35%, a recall of 82.39%, and F1 score of 86.64%. The rest of the analysis is based on 29,053 and 14,992 PhD recipients in physics and EECS, respectively.

Identifying PhDs' contribution to inventions

We now turn to the contribution to invention made by PhDs through two key channels: (1) the extent to which their research influences subsequent innovation, and (2) their direct participation in the invention process. To assess the first channel, we analyze non-patent literature (NPL) citations in patent documents, i.e. references that patents make to scientific research (Narin et al., 1997; Roach & Cohen, 2013). By evaluating the frequency and context in which patents cite these PhDs' publications, we gain insights into the practical impact of their academic work on technological progress. This impact may be based on various pathways through which knowledge "spills over" to firms, allowing them to absorb it in their innovation processes. For the second channel, we track the direct involvement of PhDs in patenting activities, which allows us to measure their practical contributions to innovation. Since both channels rely on patent data, we use the Worldwide Patent Statistical Database (PATSTAT 2023 Spring Edition) maintained by the European Patent Office.[15]

Identifying patent citations to PhD research

To identify patents that cite the research of the identified PhDs, we use the Patent-to-Paper Citations dataset by Marx and Fuegi (2020, 2022),[16] which captures both front-page and in-text citations USPTO and EPO patents make to scientific literature. This dataset covers citations between patents and publications through the end of 2023. Using the list of PhD publications from the Scopus match (see Sect. "Identifying Publications of the PhDs"), we search for these publications within the Patent-to-Paper Citations dataset. Since this dataset indexes papers using OpenAlex IDs, an intermediary step is required. As OpenAlex is an open-source bibliographic database, we use its API—via the *openalexR* (Aria et al., 2024) package—to retrieve the corresponding OpenAlex IDs for the PhDs' publications, using DOIs provided by Scopus. We successfully matched 77.5% of Scopus publications with DOIs to OpenAlex records. This matching enables us to identify the patents that cite the PhDs' research. Additional patent information is obtained through PATSTAT.

Identifying PhD-invented patents

We investigate all patenting activity PhD recipients may have conducted between 1995 and 2023. The start date precedes the PhD sample period by 5 years, so patents filed before PhD completion can also be captured. We restrict our search to priority applications to avoid double-counting patents filed in multiple jurisdictions and to better capture the original invention date (de Rassenfosse et al., 2013).

For several reasons, matching researchers with their patents is a challenging task. Among the issues that must be handled is that patent databases such as PATSTAT do not include reliable identifiers for inventors. Researchers with multiple patents therefore need to be matched

[15] The PATSTAT 2023 Spring Edition contains information on patents published up to the end of January (European Patent Office, 2023).

[16] The Patent-to-Paper Citations dataset is part of the Reliance on Science project and can be freely accessed at: https://relianceonscience.org/.

with each of their patents. In addition, the extent and presentation of information included in patent documents differs between patent offices. The literature proposes a wide array of techniques to deal with these challenges (Huberty et al., 2014; Lissoni et al., 2006; Maraut & Martínez, 2014; Meyer, 2006; Pezzoni et al., 2014; Schmoch et al., 2012). Since our context is close to that of Schoen et al. (2014) who linked German professors to inventors in PATSTAT, we primarily build on their approach, adapting it with minor modifications.

First, we retrieve all priority patent applications filed worldwide between 1995 and 2023, provided at least one inventor or applicant is from Germany, Switzerland, or Austria. Such geographic restrictions are common practice (Breschi et al., 2005; Meyer, 2006) to reduce false positives and computational complexity. However, they also limit our ability to fully capture relevant patenting activity of PhD recipients if it occurred abroad without co-inventors or applicants from the target countries.

This initial retrieval yields 3,067,351 inventor records linked to 1,298,242 priority patent applications. Next, we filter out irrelevant inventor records by calculating a similarity score between the inventor and PhD names using the 2-g algorithm. This score is derived by breaking names into consecutive pairs of characters (bigrams) and calculating the proportion of shared bigrams between the two names. To improve filtering, inventor names are cleaned and standardized by removing or replacing corrupted characters and extracting academic titles, address information, or applicant details, if present. We also generate alternative spelling variations for each name, by reversing first and last names, adapting German linguistic elements in different ways,[17] and abbreviating first or middle names while keeping the first given name in full. Using these variations, we select the highest similarity score for each inventor–PhD pair. Applying the 0.9 similarity threshold proposed by Schoen et al. (2014), we reduce the pool to 206,315 inventor records.

To identify cases where multiple inventor records correspond to the same individual, Schoen et al. (2014) creates validated patterns that determine when different inventor records refer to the same person based on applicant information, technical classifications related to the invention, and inventor-specific factors like name rarity and address information. Their approach distinguishes between missing and contradictory information to improve accuracy and applies transitive logic.[18] We examine these patterns across all filtered inventor-PhD pairs from the previous step. Using this methodology, we identify 101,939 disambiguated inventors.

Finally, to establish a definitive match between PhDs and disambiguated inventors, we leverage the link between PhDs and Scopus publication data (see Sect. "Identifying publications of the PhDs") to identify overlapping information patterns between PhD publications and inventor patent applications. We restrict this search to cases where publication and patent application dates fall within a 3-year window of each other. We classify matching patterns as either strong or weak based on their reliability in identifying a true match. This approach deviates from Schoen et al. (2014)'s methodology due to differences in available data on the academics being matched. We investigate the presence of the following patterns:

[17] For example, both "a" and "ae" may be suitable adaptations for the umlaut "ä".

[18] If inventor records A&B and B&C match, then A&C must also belong to the same individual.

- The same name appears as both a co-author in the PhD's publications and a co-inventor on the inventor's patents. To identify these overlaps, we apply the same fuzzy name-matching methodology used in the initial inventor filtering step.[19]
- The same organization appears as both the PhD's affiliation in their publications and the applicant on the inventor's patents. To identify these cases, we manually code instances where the two entity names share at least one non-generic word.[20] This method outperforms fuzzy or edit-distance-based similarity measures, as it better handles the inconsistencies in unstructured company and university names across our data sources.[21]
- The technological field covered by the PhD's publications and the inventor's patent is identical. For this purpose, we utilize the concordance table developed by Schmoch et al. (2012), which maps the Scopus research areas to the technological field classifications of patents. This information is available on the journal level.
- A PhD appears both as an inventor of a patent and as an author of a publication cited by the patent. We leverage the Patent-to-Paper Citations dataset (see section "Identifying PhDs' contribution to inventions") and apply the previously used fuzzy name matching methodology, to identify instances where the inventor of a citing patent and the PhDs who authored cited publications are the same individual.

A PhD–inventor match is accepted only when it meets at least one strong and one weak criterion. Multiple instances of the same pattern can occur for a single inventor-PhD pair. For example, if two co-authors of a PhD holder also appear as co-inventors, the presence of two strong matching criteria would establish a definitive match.[22] Of the 101,939 disambiguated inventors, we match 8691 (8.53 %) to 4347 PhDs.[23]

Table 1 reports the total number of thesis authors identified in physics and EECS, as well as those identified as Scopus authors, recipients of citations in patents, and inventors.

[19] To minimize false positives, we exclude co-author information from publications with more than 15 authors, as this exceeds the outlier threshold (the 75th percentile plus 1.5 times the interquartile range). Additionally, we manually review all non-exact matches and remove false positives.

[20] Generic words, such as "university" or "corporation," are excluded. A full list is provided in the Appendix B.

[21] During manual coding, we do not distinguish between parent companies and subsidiaries. However, for large publicly funded German research organizations (e.g., Fraunhofer Society), which employ many researchers, we use a more detailed approach. Matches at the subdivision level (e.g., Fraunhofer Institute for Solar Energy Systems) are classified as strong, while broader-level matches are weak. If both entries provide conflicting subdivision information (e.g., Solar Energy Systems vs. Integrated Circuits), it is not considered a match.

[22] To evaluate our matching criteria, we manually reviewed a random sample of definitive matches (n = 100), barely not matched pairs (n = 50), and all cases where a patent predates PhD completion by over 7 years (n = 147). Using online sources (e.g., LinkedIn, university websites), we found no false positives among definitive matches. Six false positives emerged in the barely not matched group, and eight in the time-lagged cases, five of which are caused by incorrect Scopus linkages. All false positives were removed from the final dataset.

[23] Sometimes multiple disambiguated inventors are matched to a single PhD. This occurs because the algorithm by Schoen et al. (2014) does not identify that these disambiguated inventors actually refer to the same individual, due to algorithm's conservative nature and or missing patent information.

Table 1 PhD–publication–patent linkage summary

	No. of PhDs	No. scopus matched	No. PhDs cited by patents	No. inventors
Physics	29,053	22,437	8,006	2,152
EECS	14,992	10,653	5,396	2,195

Identifying the sustainability-orientation of theses, publications, and patents

Identifying sustainability-oriented publications and building reference corpora

The remaining ingredient in our empirical analysis is to identify sustainability-oriented researchers and green inventors with a PhD in Germany. To this end, we construct sustainability-oriented text corpora for our two focal disciplines using a three-step approach designed to minimize reliance on domain-specific expert judgment. First, we identify a small set of journals whose stated aims and scopes are explicitly focused on environmental sustainability. Second, we analyze the most frequent words and phrases in the keywords and titles of articles published in these journals. Third, we use these keywords to search the titles and keyword lists of articles in other high-quality journals within each discipline, allowing us to capture all relevant sustainability-oriented research outputs.

Crucially, the journals identified in the first step are used solely to derive an initial and intentionally conservative set of sustainability-oriented keywords, rather than to delimit the full universe of sustainability-oriented research. To this end, we rely on the explicit aims and scopes of selected journals rather than on expert-driven assessments of individual technical terms. In both disciplines, we therefore draw on two highly visible and explicitly sustainability-focused interdisciplinary journals: Nature Sustainability and Nature Climate Change. While these journals are classified broadly in standard bibliographic databases, they are strongly interdisciplinary in practice and regularly publish contributions by engineers and physicists. To ensure disciplinary relevance, we restrict keyword extraction to articles (co-)authored by researchers affiliated with departments corresponding to our focal fields. In addition, for EECS we include IEEE Transactions on Sustainable Energy, a discipline-specific journal with an explicit focus on environmental sustainability.[24] In total, our database for the first step consists of 1911 articles from EECS, and 378 articles from physics. In the second step of the process, the 100 most common key phrases (words, bigrams, and trigrams) are extracted from the titles and keyword lists of the articles identified in the first step. From these key phrases, we exclude any term that does not explicitly signal an orientation toward sustainability, thereby adopting a deliberately conservative keyword selection strategy. While, in principle, even a small set of highly salient seed terms (e.g., sustainability or climate change) would be sufficient for SPECTER embeddings to recover a much broader set of semantically related concepts, we deliberately go beyond this minimum. Our intentionally conservative filtering ensures that, even if some relevant terms are omitted or excluded due to non-expert judgment, the dominant signal captured by the embeddings robustly reflects sustainability relevance. The complete lists of keywords for each discipline are provided in Appendix D. The complete lists for each discipline are

[24] No comparable sustainability-focused disciplinary journal was identified for physics. Accordingly, for physics, the keyword derivation relies exclusively on the two interdisciplinary journals mentioned above, and we only consider articles co-authored by researchers from Physics departments.

provided in Appendix D. As the third and last step, we search for the keywords identified among all high-quality field journals.[25]

This process yields two discipline-specific corpora of sustainability-oriented publications. The collection of abstracts of these articles provides the reference frame against which we evaluate the sustainability orientation of PhD theses and patents. Figure 2 displays the most frequent keywords in each corpus, with no post-processing or filtering applied. As expected, terms that mirror our original search strings dominate the word clouds, yet a rich set of additional sustainability-oriented expressions also emerges. Examining these clouds not only highlights the leading sustainability themes in each field but also points to the sub-disciplines and methodological lenses that shape the debate.

Identifying sustainability-oriented theses

Drawing on the record-linkage procedure described in Sect "Identifying sustainability-oriented publications and building reference corpora", we identify theses whose authors qualify as sustainability-oriented researchers. A sustainability-oriented publication is attributed to a researcher's PhD trajectory if it appears during the PhD period or within 3 years of the defense. Publications emerging after this 3-year window are treated as post-PhD output, reflecting work conducted later in the researcher's career rather than during the PhD. For thesis authors who are not previously identified as sustainability-oriented researchers from their publications, we assess the thematic similarity between their theses and sustainability-oriented publications. The central idea of this step is to compare the textual content of theses with a reference corpus of sustainability-oriented publications and to identify theses that are unusually similar to this corpus.

To implement this comparison, we collect the abstracts of the PhD theses identified in 3.1.1. Although the DNB catalogue offers rich thesis metadata, it does not include a summary or abstract for the theses. We therefore gathered the abstracts in two ways: (1) harvesting text from external URLs archived by the DNB and (2) scraping abstracts from the online PDF versions of the theses. These methods produced abstracts for 12,767 Physics theses and 4833 theses in EECS.[26] Because most thesis titles and abstracts are in German, we machine-translated them into (American) English using the DeepL API.[27]

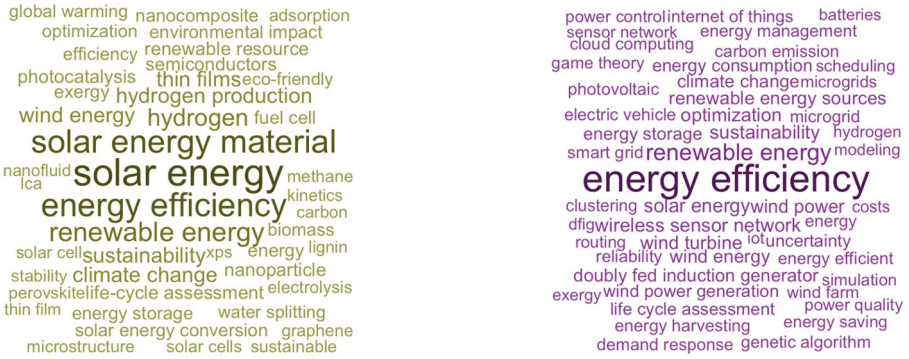
Representing Theses and Publications in a Common Semantic Space

To quantify thematic similarity in a systematic and scalable way, we represent both theses and publications as numerical vectors that capture their semantic content. Specifically, we generate document-level embeddings from thesis and publication titles and abstracts

[25] We only consider journals in Scimago Journal and Country Rank (ScimagoJR, https://www.scimagojr.com) that have a Q1 or Q2 ranking. In both journal lists, we exclude journals that also publish in unrelated fields, such as humanities. Although the journal sets are field-specific, they also encompass interdisciplinary journals in which the focal field is represented. The full list of journals containing sustainability-oriented articles is reported in Appendix C.

[26] Discrepancies between thesis counts and available abstracts reflect collection constraints: incomplete metadata, OCR and formatting issues hampering PDF scraping, and differences across historical archiving differences. Comparing year and institution level distributions shows the abstract subset generally mirrors the full dataset but underrepresents 2000–2005 and displays slight institutional skew, likely tied to early digital dissemination.

[27] DeepL SE is a German company offering a machine translation system that is shown to be more accurate than other available tools for German-English translations (Avramidis et al., 2020). The DeepL API provides programmatic access to DeepL's machine translation technology. We integrated the DeepL API into R using the *deeplr*- package (Zumbach & Bauer, 2021).



(a) Physics (b) EECS

Fig. 2 Wordclouds of keywords of the identified publications

using a pre-trained neural language model (LM) designed for scientific documents (Cohan et al., 2020). Document-level embeddings transform entire documents into fixed-size vectors such that texts with similar semantic meaning are located close to each other in a high-dimensional semantic space.

We use SPECTER2.0 (Singh et al., 2022), a state-of-the-art transformer-based neural network algorithm specifically trained on scientific literature. Intuitively, the model learns which documents are related by observing patterns in citations and textual content, and it places semantically similar documents closer together in the embedding space. Compared to its predecessor SPECTER (Cohan et al., 2020), SPECTER2.0 (Singh et al., 2022) is trained on a larger corpus from the Semantic Scholar (A12, 2024) database, which features 6 million document triplets spanning 23 fields of study.

At a technical level, SPECTER2.0 generates representative embeddings of scientific documents from the titles and abstracts of papers. These document-level embeddings leverage multiple signals of inter-document relatability, including citations, co-authorship information, and publisher metadata, ensuring that similar articles are positioned close to each other in vector space. In other words, the model is trained to reflect similarity based on citation and content patterns. SPECTER2.0 relies on a triplet margin loss function to optimize the relative distance between documents. This helps ensure that semantically similar documents are located closer together in the embedding space than dissimilar ones. The triplet margin loss function is defined as:

$$\mathcal{L}(P^Q, P^+, P^-) = \max\left(0, d(P^Q, P^+) - d(P^Q, P^-) + \text{margin}\right)$$

where:

- P^Q is the query or anchor paper (e.g. a thesis),
- P^+ is a positive paper (e.g. a related publication),
- P^- is a negative paper (semantically dissimilar to the query),
- $d(x, y)$ represents the Euclidean norm distance between samples x and y,
- margin is a hyperparameter that defines the minimum distance between positive and negative pairs. The value is 1, based on preliminary experiments by Cohan et al. (2020).

By minimizing this loss during training, the model learns to embed documents in such a way that similar pairs are close and dissimilar pairs are farther apart. While this training procedure is not re-estimated in this study, it ensures that the resulting embeddings provide a meaningful measure of semantic proximity for downstream similarity analysis.

Measuring similarity to sustainability-oriented research

Following embedding generation, we compute cosine similarity between document vectors, a standard measure of semantic similarity that captures the angle between two vectors in the embedding space (Gómez & Vázquez, 2022). Higher cosine similarity values indicate stronger thematic alignment between a thesis and sustainability-oriented research.

Rather than comparing individual theses to individual publications, we construct a single reference embedding by aggregating the titles and abstracts of all sustainability-oriented publications previously identified. Each thesis embedding is then compared to this reference corpus. This approach avoids noisy one-to-one comparisons and captures the overall semantic proximity of a thesis to the sustainability-oriented corpus, by discipline.

To identify theses that are unusually semantically similar to the sustainability-oriented corpus, we treat the distribution of cosine similarity scores across all theses as an empirical reference distribution. We compute z-scores for each thesis, measuring how many standard deviations its similarity score lies from the mean. Using a two-tailed significance level of 0.05, we determine the critical z-value cutoff ($\sim \pm 1.96$).[28] Theses with z-scores exceeding $+1.96$ are considered statistically significantly more similar to the sustainability-oriented corpus than expected by chance. These theses are classified as exhibiting strong semantic proximity to sustainability-oriented research. For analytical purposes, their authors are treated as sustainability-oriented PhD researchers alongside those identified via publication-based matching. These cases are added to the final count of sustainability-oriented PhD recipients.

Identifying green patents

The process of identifying green patents mirrors the processes described in Sect. "Identifying sustainability-oriented theses". However, instead of comparing the similarity of each thesis title and abstract with the sustainability corpus using SPECTER2.0 embeddings, we use all of the patent titles and abstracts that are matched with a PhD holder (see Sect. "Identifying green patents"). Similar to before, we select patents whose similarity scores fall within the upper 5% tail of the empirical reference distribution, corresponding to a z-score above $+1.96$, indicating significant semantic alignment with the sustainability-oriented corpus. This identifies patents with pronounced semantic similarity to sustainability-oriented publications.

[28] While theses below this cutoff may also engage with sustainability-oriented topics, the z-score threshold provides a transparent and statistically grounded rule to identify theses that are unusually close to the sustainability-oriented corpus relative to the full distribution. Alternative thresholds could be adopted in more or less conservative classification schemes, depending on the research question and analytical context. Substantively, the $z > 1.96$ cutoff approach identifies the ~ 2.5% of documents whose semantic content is most strongly oriented toward the sustainability reference corpus relative to all other theses in the same discipline. The threshold is therefore a relative one, and the measure is inherently comparative and sensitive to the composition of the reference corpus and the underlying population. Appendix F shows that our regression results are robust to alternative thresholds (1 and 10%).

Table 2 summarizes, for each field, the total number of PhD graduates and the subset classified as sustainability-oriented on the basis of their publications, theses, and patents, using the previously described SPECTER-based approach.

Comparing our patent classification approach to existing metrics

Although our methodology is based on an unsupervised approach and utilizes an embedding model specifically trained on publication data, evaluating its performance remains essential. A key challenge, however, is the absence of a gold standard dataset that reliably classifies theses and patents as sustainability-oriented. Sustainability is inherently complex and difficult to categorize in binary terms, particularly in technical documents where expert judgment is required. Given its broad scope and the need for specialized expertise, constructing such a dataset is beyond our current capabilities.

Recognizing these challenges, to evaluate the success of our approach, we compare our classification with established sustainability classification schemes such as the Green Inventory of the International Patent Classification (IPC) by the World Intellectual Property Organization (WIPO), and the Y02/Y04 classification scheme developed by the European Patent Office (EPO), which specifically aims to identify technologies that mitigate climate change (Favot et al., 2023). Each measure has strengths and limitations, as results of patent analyses may critically depend on the specific measure employed (Bruns & Kalthaus, 2020). Our proposed method positions itself as a complement to the existing green patent identification techniques and may thus help enhance the robustness of patent analyses.

Figure 3 compares the overlap between SPECTER2.0 and the two established rule-based classification schemes for both PhD-invented patents and patents citing PhDs. The figure shows that none of the three approaches fully subsumes the others and that each method identifies a substantial number of patents not captured by the alternative classifications. These differences reflect the distinct principles on which the classification schemes are built. Y02/Y04 and the IPC Green Inventory rely on expert-defined, rule-based taxonomies, though they differ in scope and emphasis. Y02/Y04 focuses on a relatively specific set of climate-mitigation technologies as defined by the European Patent Office, whereas the IPC Green Inventory aggregates a broader range of IPC classes identified by WIPO as environmentally relevant. Conceptually, one might expect Y02/Y04 to be largely a subset of the IPC Green Inventory given its narrower focus, yet in practice this is not the case, reflecting inconsistencies between the inclusion criteria used by the two classification systems. SPECTER2.0, by contrast, identifies patents based on semantic similarity between patent texts and sustainability-oriented scientific publications. As a result, it captures patents whose textual content is closely related to sustainability research. At the same time, although we benchmark our approach against two established expert-based patent classification systems, we do not provide the formal gold standard hand-coded validation sample for our specific set of parents. Therefore, it cannot be conclusively determined whether all additional patents identified by the semantic approach represent substantively sustainability-oriented inventions, or whether some reflect more peripheral or ambiguous connections. To address this limitation, we conduct an exploratory analysis to further assess the plausibility of the semantic identification strategy. Specifically, we base our SPECTER2.0 classifications on a similarity threshold corresponding to the 95th percentile of similarity to the sustainability corpus. We compare patents above this threshold with those in the lowest 5% of similarity scores, examining differences in the most frequently occurring terms in titles across our two fields (see Figs. 5–8 and Table 12 in Appendix E.). We argue

Table 2 Breakdown of identified sustainability-oriented PhDs, by discipline

	Physics	EECS
Total number of PhDs	29,053	14,992
Identified as sustainable author	780	702
Via Publication Data	302	600
Via SPECTER2.0	478	102
Identified as green inventor	161	187
Identified as being cited by green invention	3710	2648

that the observed discrepancies in frequently used terms between patents in the 95th and 5th percentiles are consistent with the notion that the SPECTER2.0 classification predominantly captures patents that are more closely related to sustainability-oriented concepts. While this evidence does not constitute definitive validation, it provides indicative support for the plausibility of the semantic classification approach.

Analysis and findings

Building on the linkage framework introduced in Sect. "Data and methodology", we assess the extent to which PhD recipients are linked to sustainability-oriented inventive activity through direct and indirect channels, as well as the characteristics of their PhD training environments and research portfolios. As our methodology relies on multiple record-linkage procedures with imperfect recall, the reported figures should be interpreted as lower bound estimates of the overall contribution of PhDs to sustainability-oriented innovation.

Tables 3 and 4 present the patent-level findings on PhD recipients' contributions to inventive activity, directly via their own patents and indirectly via patent citations to their publications, respectively. Alongside our text-based sustainability assessment, we report results based on the two alternative sustainability-relevance metrics, IPC Green Inventory and Y02/Y04, introduced in Sect. "Comparing our patent classification approach to existing metrics".[29] The first finding is that indirect contributions via patent citations are about four times more common than direct inventorship: while around 2100 PhDs in each field appear as inventors (corresponding to roughly 10% of physics PhDs and 20% of EECS PhDs), more than 8000 physics PhDs (36%) and over 5000 EECS PhDs (51%) are cited in patents. On a per-capita basis, physics PhDs are associated with an average of 1.6 patent citations, compared to 4.5 for EECS PhDs, which is consistent with stronger spillover effects in more applied fields.

Moreover, in both channels, sustainability-oriented patents comprise a noteworthy share of inventive activity linked to PhD research, though the composition and magnitude of this share varies across classification approaches. In both physics and EECS, rule-based schemes (Y02/Y04 and IPC Green Inventory) identify larger absolute numbers of green patents than the SPECTER2.0-based measure, both among citing patents and among patents invented by PhD recipients. At the same time, the semantic approach consistently flags a distinct subset of patents that are not captured by expert-based classifications. This

[29] All patent counts reflect unique simple patent families.

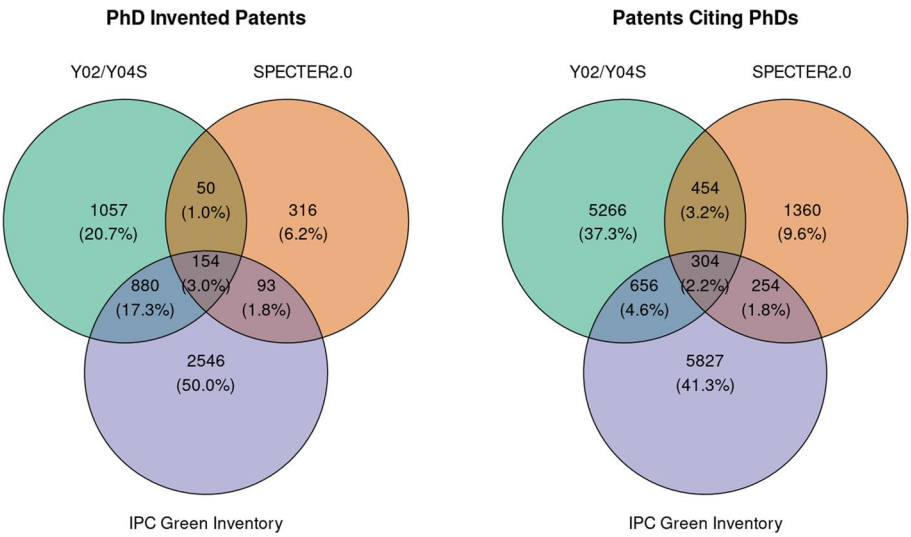


Fig. 3 Greenness classification of patents using SPECTER2.0 vs. other classification

pattern is evident both in patent citations to PhD publications and in PhD-invented patents, where each classifier highlights partially overlapping but non-identical sets of sustainability-oriented inventions (see Fig. 3 in Sect. "Comparing our patent classification approach to existing metrics" for a visualization of this pattern). Among all PhDs, between 0.8 and 5.8% hold at least one green patent depending on the classification scheme used, highlighting the sensitivity of descriptive conclusions to how sustainability is operationalized. Taken together, these results suggest that PhD-trained researchers appear more prominently connected to green innovation through the spillovers of their scientific work than through their own patents, while also highlighting the importance of using multiple, complementary classification strategies to avoid overlooking sustainability-relevant inventive activity that falls outside established expert-based taxonomies.

We next examine descriptive associations between PhDs' direct and indirect contributions to green invention using regression models. The key explanatory variables in these models indicate whether a scholar engaged in sustainability-oriented research during the PhD period (Sus. PhD) and in the post-PhD period (Sus. Post PhD).[30] In addition, we include an indicator for whether a PhD had at least one publication with an industry affiliation prior to PhD completion (Ind. Affil. pre PhD), capturing early exposure to industry-linked research environments. Our baseline specification includes a field dummy coded 1 for PhDs in electrical engineering and computer science (EECS) and 0 for physics. To account for institutional differences within the German higher education landscape, we add

[30] We consider a PhD to engage in sustainability-oriented research during their PhD if they either have a thesis which was classified by SPECTER2.0 as sustainability-oriented (see Sect. "Identifying sustainability-oriented theses") or if they publish sustainability-oriented publications (see Sect. "Identifying sustainability-oriented publications and building reference corpora") during their PhD or within 3 years after having completed their PhD. Likewise, a PhD is considered to be engaging in sustainability-oriented research after their PhD if they publish a sustainability-oriented publication more than 3 years after their PhD completion.

Table 3 Patent citations of PhD publications

	Physics	EECS	Physics	EECS
	Absolute No.		Per PhD	
Citing patents	36,095	48,186	1.609	4.531
Of those SPECTER Green	991	1668	0.044	0.157
Of those Y02/Y04	4552	2651	0.203	0.249
Of those Green Inv	3020	4292	0.135	0.404
	Absolute No.		Share of PhDs	
Cited PhDs	8006	5396	0.357	0.507
Cited by SPECTER Green Patent	1001	1017	0.045	0.096
Cited by Y02/Y04 Patent	2056	1193	0.092	0.112
Cted by Green Inv. Patent	1765	1868	0.079	0.176

Table 4 PhD invented Patents

	Physics	EECS	Physics	EECS
	Absolute No.		Per PhD	
Patents	9535	13,064	0.425	1.228
Of those SPECTER Green	257	431	0.011	0.041
Of those Y02/Y04	1025	1171	0.046	0.110
Of those Green Inv	1963	1795	0.087	0.169
	Absolute No.		Share of PhDs	
Patenting PhDs	2152	2195	0.096	0.206
With a SPECTER Green Patent	169	165	0.008	0.016
wWth a Y02/Y04 Patent	394	413	0.018	0.039
With Green Inv. Patent	664	612	0.030	0.058

dummy variables for membership in the TU9 (Germany's long-established technical universities with a strong engineering focus) and the U15 group (research-intensive comprehensive universities that host medical schools). Depending on the specification, we also control for a scholar's cumulative publication output (Log. No. of Publ.),[31] direct inventive output (No. of Patents), and the volume of knowledge spillovers their publications generate for patents (No. of citing Patents). In all models we include fixed effects for the PhD-completion year to capture time trends and account for the shorter citation and patenting windows available to more recent cohorts. Table 5 presents summary descriptive statistics for all variables used across the regression specifications, based on the full non-subsetted dataset.

In the first set of regressions, reported in Table 6, we focus on the indirect contribution of PhDs to inventive activity, exploring factors associated with the likelihood that PhDs' publications are cited by patents. We begin by investigating how engagement with

[31] We log-transformed the number of publications due to its highly skewed distribution, as some PhDs have an exceptionally high number of publications exceeding 1000.

sustainability-oriented research during and after a PhD is associated with the general transferability of PhDs' research to an inventive context (Column 1 and 2). Columns 3–8 narrow the analysis to PhDs' whose publications are cited at least once and test whether those citations originate from green patents. To ensure robustness, we apply all three alternative greenness classifications; for the SPECTER-based classification, we provide further robustness checks by rerunning the regression models utilizing different threshold levels (1 and 10%) in Appendix Section F.[32] Across all specifications, we first focus on sustainability engagement during the PhD period alone (Columns 1, 3, 5, 7), since this stage is central to shaping a researcher's long-term focus in terms of both methods and topics. We then add an indicator for post-PhD sustainability engagement, but only for cohorts with sufficient observation time after graduation; this necessitates a reduced sample in Columns 2, 4, 6, 8. This stepwise approach allows us to distinguish patterns associated with PhD work from subsequent career developments while maintaining comparability across alternative measures of green invention.

The findings suggest that engagement in sustainability-oriented research during the PhD is negatively associated with the odds of a graduate's work being cited by patents overall (Columns 1–2), but is positively associated with the odds of being cited by a sustainability-oriented patent among those with at least one citation (Columns 3–8). The estimated coefficients are sizable and consistently significant at the 1% level. Engagement with sustainability-oriented research after the PhD is not associated with the likelihood of being cited by a patent (Column 2) at a statistically significant level, but increases the odds of being cited by a sustainability-oriented patent among those PhDs whose work is cited in a patent (Column 4, 6, 8), with a statistical significance of at least 5% across all specifications. Conditional on controlling for other factors, gender does not appear to have a statistically discernible association with the likelihood of a PhD's research being cited by patents. EECS PhDs are more likely than physics PhDs to have their research cited by patents, reflecting the more applied focus of the field. However, among those with patent citations, field differences vary with the classification scheme: the EECS field either has a statistically negative (Y02/YO4), or positive relationship (Specter Green & Green Inv.) on being cited by a sustainability-oriented patent. One might be tempted to attribute these patterns to the relevance of specific underlying technologies, with physics research playing a greater role in climate change mitigation technologies, while EECS research contributes more broadly to general environmental technologies. However, results differ across both outcomes (see Tables 6 vs. 7). PhDs who had at least one industry-affiliated publication prior to completing their degree are more likely to be cited by a patent, highlighting the role of early exposure to industry-linked research. However for those PhDs who get cited having a industry-affiliated publication prior to completing their degree does not predict having a green sustainability-oriented patent citation.

In Table 7 we investigate how PhDs' involvement with sustainability-oriented research during and after their PhD is associated with their likelihood of patenting (Columns 1 & 2). We then restrict our analysis to the PhDs who patent and investigate how a PhD's involvement with sustainability-oriented research during and after their PhD is associated with the likelihood of having a sustainability-oriented patent (Column 3–8). As in Table 6, we estimate specifications that include only the indicator for sustainability engagement during the

[32] Among the regression models with different threshold levels, our variables of interest remain robust, while the other control variables remain mostly stable.

Table 5 Descriptive statistics (N = 32,043)

Variable	Mean	SD	Min	Max
Has Citing Patent	0.405	0.491	0	1.000
Has Citing Specter Green Patent	0.061	0.239	0	1.000
Has Citing Y02/Y04 Patent	0.098	0.297	0	1.000
Has Citing Green Inv. Patent	0.110	0.312	0	1.000
Has Patent	0.131	0.337	0	1.000
Has Specter Green Patent	0.010	0.099	0	1.000
Has Y02/Y04 Patent	0.024	0.153	0	1.000
Has Green Inv. Patent	0.038	0.191	0	1.000
Sus. PhD	0.038	0.191	0	1.000
Sus. Post PhD	0.009	0.095	0	1.000
Gender	0.841	0.366	0	1.000
EECS	0.324	0.468	0	1.000
Log No. of Publ	2.456	1.311	0	7.434
No. of Patents	0.750	5.029	0	495.000
Ind. Affil. pre PhD	0.058	0.235	0	1.000
TU9	0.302	0.459	0	1.000
German U15	0.318	0.466	0	1.000
PhD completion year	2010.844	6.386	2000	2021.000

PhD (Columns 1, 3, 5, 7), as well as specifications that also include the indicator denoting post-PhD sustainability-oriented research (Columns 2, 4, 6, 8).

The estimates do not suggest that involvement in sustainability-related research during the PhD is associated with the likelihood of holding a patent (Columns 1–2). However, among PhDs who do have a patent, a sustainability orientation in research significantly increases the probability that at least one of their patents is sustainability-oriented (Columns 3–8); the corresponding coefficients are all statistically significant at the 1% level. Sustainability-oriented research after the PhD predicts both the likelihood of having a patent (Column 2) and the likelihood of having a sustainability-oriented patent among patent holders (Column 4, 6, 8) with a statistical significance of 1%. Gender also predicts patenting activity: male PhDs are more likely to hold a patent (Columns 1–2), with coefficients significant at the 1% level. These findings are based on regressions controlling for academic productivity differences. However, gender is not significantly associated with the likelihood of holding a sustainability-oriented patent among those with patents (Columns 3–8), controlling for the overall number of patents. EECS PhDs are more likely than physics PhDs to hold a patent, again reflecting the more applied orientation of the field. Similarly, PhDs who had at least one industry-affiliated publication prior to completing their degree are more likely to hold a patent, reinforcing the role of early exposure to industry-linked research. Among those who do hold patents, however, prior industry affiliation does not increase the likelihood of having a sustainability-oriented patent. The statistically significant and positive relationship between the number of publications and the likelihood of having a patent should be interpreted in light of our patent-matching methodology, as it relies on identifying overlaps between publication and patent information (see Sect. "Identifying PhDs' contribution to inventions").

Table 6 Likelihood of publishing PhDs being cited by patents

	Model specification:							
	Any patent		Specter green		Y02/Y04		Green Inv.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sus. PhD	− 0.665***	− 0.740***	1.613***	1.528***	1.423***	1.289***	0.580***	0.491***
	(0.068)	(0.076)	(0.110)	(0.124)	(0.110)	(0.122)	(0.110)	(0.120)
Sus. Post PhD		0.166		1.178***		1.273***		0.539**
		(0.139)		(0.167)		(0.159)		(0.164)
Gender	0.067	0.065	0.054	0.021	− 0.045	− 0.037	0.067	0.036
	(0.037)	(0.039)	(0.082)	(0.085)	(0.064)	(0.066)	(0.064)	(0.066)
EECS	0.758***	0.719***	0.422***	0.404***	− 0.384***	− 0.439***	0.627***	0.640***
	(0.028)	(0.030)	(0.059)	(0.060)	(0.051)	(0.053)	(0.047)	(0.048)
No. of patents	0.090***	0.085***	0.008	0.007	0.006*	0.006*	0.004	0.004
	(0.013)	(0.013)	(0.006)	(0.006)	(0.003)	(0.003)	(0.003)	(0.003)
No. of citing Patents			0.024***	0.024***	0.037***	0.037***	0.037***	0.036***
			(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Log No. of Publ	0.704***	0.730***	0.196***	0.173***	0.119***	0.103***	0.190***	0.190***
	(0.012)	(0.013)	(0.022)	(0.023)	(0.019)	(0.019)	(0.019)	(0.019)
Ind. Affil. pre PhD	0.580***	0.596***	0.027	0.073	− 0.159	− 0.102	− 0.012	− 0.010
	(0.057)	(0.061)	(0.094)	(0.097)	(0.084)	(0.087)	(0.072)	(0.076)
TU9	0.092**	0.062	0.136*	0.140*	0.184***	0.193***	0.018	0.013
	(0.030)	(0.032)	(0.060)	(0.062)	(0.051)	(0.052)	(0.048)	(0.050)
German U15	− 0.211***	− 0.218***	0.014	0.025	− 0.095	− 0.106	0.053	0.068
	(0.031)	(0.033)	(0.068)	(0.070)	(0.057)	(0.058)	(0.054)	(0.055)
Constant	− 2.318***	− 2.349***	− 2.770***	− 2.687***	− 1.612***	− 1.575***	− 2.363***	− 2.346***
	(0.071)	(0.073)	(0.142)	(0.144)	(0.117)	(0.119)	(0.123)	(0.124)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 6 (continued)

	Model specification:							
	Any patent		Specter green		Y02/Y04		Green Inv.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pseudo R^2	0.169	0.168	0.109	0.115	0.102	0.106	0.102	0.104
Observations	32,043	27,989	12,961	11,968	12,961	11,968	12,961	11,968

Note: In column (1) and (2) we estimate the likelihood of a publishing PhD being cited by at least one patent. For column (3)–(8) we restrict the sample to those PhDs who are cited by at least one patent and estimate those PhD's likelihood of being cited by at least one patent which fulfills one of our three sustainability-oriented classification criteria. For specifications which contain Sus. Post PhD variable, the utilized sample had to be further restricted to PhD's who completed their PhD prior to the year of 2019. For all of these we do not consider instances of self-citations (cited publication and citing patent by the same PhD). The table lists the log odds ratios of these estimations with robust standard errors in parentheses and a pseudo R^2 value as proposed by McFadden (1974). All specifications include year fixed effects for the PhD completion year. *p<0.05; **p<0.01; ***p<0.001

Table 7 Likelihood of publishing PhDs inventing patents

	Model specification:							
	Any patent		Specter green		Y02/Y04		Green Inv.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sus. PhD	0.041	− 0.036	1.774***	1.722***	1.922***	1.852***	1.009***	0.941***
	(0.090)	(0.099)	(0.200)	(0.216)	(0.173)	(0.196)	(0.172)	(0.186)
Sus. post PhD		0.604***		0.788**		1.449***		0.825***
		(0.140)		(0.298)		(0.233)		(0.211)
Gender	0.714***	0.712***	0.361	0.349	− 0.238	− 0.207	− 0.020	− 0.005
	(0.063)	(0.065)	(0.280)	(0.286)	(0.160)	(0.166)	(0.137)	(0.142)
EECS	0.726***	0.705***	0.031	− 0.030	− 0.172	− 0.255**	− 0.240**	− 0.265**
	(0.038)	(0.040)	(0.136)	(0.139)	(0.095)	(0.099)	(0.080)	(0.083)
No. of patents			0.037***	0.038***	0.060***	0.060***	0.083***	0.082***
			(0.006)	(0.006)	(0.006)	(0.006)	(0.008)	(0.008)
Log No. of Publ	0.385***	0.382***	0.087	0.048	0.026	− 0.035	− 0.050	− 0.070*
	(0.012)	(0.012)	(0.052)	(0.055)	(0.039)	(0.041)	(0.032)	(0.034)
Ind. Affil. pre PhD	1.467***	1.449***	− 0.144	− 0.116	0.161	0.157	− 0.015	0.018
	(0.056)	(0.059)	(0.170)	(0.174)	(0.111)	(0.115)	(0.097)	(0.101)
TU9	0.139***	0.147***	− 0.256	− 0.244	0.046	0.053	− 0.024	− 0.026
	(0.039)	(0.041)	(0.143)	(0.147)	(0.094)	(0.097)	(0.081)	(0.084)
German U15	− 0.429***	− 0.421***	0.152	0.145	− 0.352**	− 0.338**	0.128	0.111
	(0.047)	(0.049)	(0.164)	(0.168)	(0.126)	(0.130)	(0.097)	(0.100)
Constant	− 3.451***	− 3.443***	− 3.253***	− 3.127***	− 1.810***	− 1.668***	− 1.187***	− 1.146***
	(0.096)	(0.098)	(0.390)	(0.394)	(0.249)	(0.254)	(0.212)	(0.217)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.123	0.117	0.089	0.091	0.109	0.119	0.092	0.095
Observations	32,043	27,989	4,241	3,961	4,241	3,961	4,241	3,961

Note: In column (1) and (2) we estimate the likelihood of a publishing PhD having invented at least one patent. For column (3)–(8) we restrict the sample to those PhDs who have invented a patent and estimate those PhDs' likelihood of inventing at least one patent which fulfills one of our three sustainability-oriented classification criteria. For specifications which contain Sus. Post PhD variable, the utilized sample had to be further restricted to PhD's who completed their PhD prior to the year of 2019. The table lists the log odds ratios of these estimations with robust standard errors in parentheses and a pseudo R^2 value as proposed by McFadden (1974). All specifications include year fixed effects for the PhD completion year. *p<0.05; **p<0.01; ***p<0.001

Conclusions

This study examined how PhD recipients contribute to sustainability-oriented innovation and which individual and organizational factors are associated with these contributions by using data on two decades of PhD graduates in physics and EECS from German universities. Our paper addressed three research questions: (1) How can text-based methods be used to identify research outputs relevant for environmental sustainability? (2) To what extent, and in what ways, do PhD recipients contribute to green inventive activities? (3) Which characteristics of PhD training and scholarly output are associated with direct and indirect contributions to green innovation?

Methodologically, our contribution lies primarily in the large-scale record-linkage procedures that connect PhD theses, publications, and patents across heterogeneous databases. While this approach inevitably entails some uncertainty, especially in cases of incomplete metadata, the combination of high-quality identifiers with machine-learning-based disambiguation techniques provides a consistent framework for tracing PhD contributions across academic and inventive domains. Building on these linkages, we use SPECTER2.0 semantic embeddings to identify sustainability-oriented research and green inventions. Rather than uniformly expanding the set of green inventions, the embedding-based approach applies a content-driven classification that partially overlaps with, but does not fully coincide with, Y02/Y04 classes or the IPC Green Inventory. These patterns suggest that semantic embeddings offer a complementary perspective on sustainability-oriented inventive activity, consistent with recent bibliometric work employing transformer-based representations (Beltagy et al., 2019). We do not propose the embedding-based measure as a replacement for expert-curated classifications, which carry the legitimacy of domain expertise and codified rules. Rather, the value of the semantic approach lies in providing a flexible, complementary tool that seems to flag sustainability-oriented inventive activity falling outside established taxonomies. Our study contributes to ongoing methodological debates on the use of semantic embeddings for large-scale classification and the development of science–technology linkage procedures, while providing a basis for further methodological refinement.

PhD graduates contribute to green inventive activities both directly through patenting and indirectly through producing knowledge that informs innovation through various pathways in innovation systems, often associated with knowledge spillovers(Breschi & Lissoni, 2001). As indicated by patents citing PhD graduates' scholarly work, indirect contributions clearly dominate. While roughly 10% of physics PhDs and 20% of EECS PhDs appear as inventors, more than one-third of physics PhDs and one-half of EECS PhDs are cited in patents. On average, physics PhDs are linked to 1.6 patent citations and EECS PhDs to 4.5, a pattern consistent with stronger spillover effects in more applied fields and with the broader literature emphasizing the role of academic science as an upstream input to innovation (Jaffe, 1989; Mansfield, 1995). Moreover, in both channels, a substantial share of all patents is sustainability-oriented, though the extent and composition of this share strongly depends on the classification approach employed. A small share of citing patents–around 3% in physics and 3.5% in EECS–are classified as green under the SPECTER2.0 measure, and only a small fraction of patenting PhDs hold at least one SPECTER2.0-classified green patent. These findings underscore the dual role of PhD research as both a source of inventions, and even more strongly, as a source for knowledge linked to further inventive activity in green technologies (Perkmann et al., 2013).

With respect to the characteristics of PhD training and scholarly output, our findings suggest that sustainability engagement during the PhD does not increase the overall likelihood of contributing to patents, but is associated with a different direction of knowledge transfer. While a focus on sustainability-oriented topics during the PhD program is negatively associated with the probability of receiving patent citations in our analysis, it is positively associated with the likelihood that the citing patents are sustainability-oriented. On the inventive side, sustainability-oriented PhD research is not significantly associated with the likelihood of patenting in general; however, among those who patent, sustainability-oriented PhD research is positively associated with patenting in green technologies. Post-PhD engagement is associated with a similar pattern. Other factors, such as field differences, gender, industry-affiliated research, and institutional environment, are associated with the volume of inventive activity, but the sustainability focus of PhD research appears relevant for whether observed contributions are aligned with green innovation. The positive association between prior industry-affiliated publications and subsequent patenting is consistent with the view that early exposure to applied research environments shapes PhDs' capacity to act as boundary spanners between academic and industrial knowledge bases (Buenstorf & Heinisch, 2020; Plantec et al., 2023). This reinforces the role of doctoral training not only as a means towards an academic credential, but also as a structural element of innovation systems.

Our results highlight PhD research as a key channel through which academic sustainability research is connected to practical applications. Encouraging PhD students to engage with sustainability issues in their research, and supporting them through mentoring, topic selection, and collaboration opportunities with industry and applied research groups, may be associated with greater sustainability-oriented knowledge transfer, although our descriptive design does not allow causal conclusions on this point. Our findings moreover indicate that the observed spillovers of sustainability-oriented PhD research are noteworthy: while only 0.8–5.8% of PhDs in our sample hold at least one green patent (depending on the classification scheme), between 4.5 and 17.6% are cited in at least one green patent. Exclusively focusing on PhDs' direct transfer activities, e.g., through patents, misses an important part of the overall role PhDs appear to play in these observed knowledge flows. Accordingly, conventional indicator systems of transfer outcomes may need to be expanded in order to provide a full account. At the same time, university managers and policymakers will often be interested in transfer activities that entail benefits to the respective institution or jurisdiction. Citations that sustainability-oriented research receives from the "outside" may inform them about under-utilized "inside" transfer potential. Our results, in addition, highlight that university managers and policymakers seeking to maximize the sustainability impact of research need to differentiate the rate of transfer activities from their direction. As shown by the above findings on patent citations to PhD research, a reduced overall citation rate of sustainability-focused research may be compensated by an increased citation rate in green patents. Finally, the differences between the various types of universities in our sample suggest that institutional environments matter. Targeted support in patent-prone fields, as well as organizational cultures that value sustainability-oriented research, may be associated with stronger links between fundamental research and technologies. In this sense, universities and policymakers have scope to change the conditions under which PhD research is more closely aligned with climate and environmental objectives.

This study sheds new light on how sustainability-oriented PhDs contribute to technology transfer, but several limitations remain. First, our sustainability classification relies on document-level embeddings constructed from titles and abstracts rather than full texts, which may miss additional nuances or hidden sustainability links. Moreover, the identification of

sustainability-oriented research through semantic similarity depends on the composition of the reference corpus as well as threshold choices and embedding aggregation strategies. In addition, the performance of SPECTER2.0 embeddings may vary across disciplines, publication types, and time periods due to differences in language use and training data. While we assess robustness by comparing our classification to existing schemes (IPC and Y02/Y04) and by examining the frequency of sustainability-related terms (see Sect. "Comparing our patent classification approach to existing metrics" and Appendix E), we are unable to directly benchmark its performance. The absence of a gold-standard annotation-based validation sample remains a key limitation of this approach; however, it would require domain-specific expertise across a wide range of technical fields that lies beyond the scope of this study. Second, our findings should be interpreted as lower-bound estimates of the contribution of PhDs to sustainability-oriented innovation and technology transfer. The analysis relies on multiple matching and linkage procedures across theses, publications, patents, and citations, each of which is characterized by imperfect recall. As a result, some relevant linkages are likely to remain undetected, implying that the observed contribution of PhDs may understate their true role. Third, the scope of the paper is restricted in several ways. Our study focuses on German PhDs in physics and electrical engineering and computer science (EECS), and our patent sample is limited to inventors and applicants based in the DACH (Germany, Austria, and Switzerland) region. While Germany provides a particularly informative case given the strong movement of doctoral graduates into private-sector R&D (Eurostat, 2024), extending the framework to other countries, disciplines, and research systems would provide valuable evidence on the broader generalizability of the findings. Finally, we observe PhD inventiveness only when publication and patent records overlap, leaving out individuals who neither patented nor published during their PhD studies and later patented in areas disconnected from their earlier academic work. Addressing these limitations in future research would further refine the methodology and provide a more comprehensive understanding of how PhDs contribute to the transfer of sustainability-oriented knowledge and technologies from universities.

Appendices

A: Classification of thesis subjects

To address potential inconsistencies in the DNB's subject classifications, we utilize data extracted from the cover pages of thesis PDFs, including degree titles, faculty and department names, and keywords, to refine the classifications.[33] For instance, when a thesis is classified in DDC under multiple focus fields (e.g., physics and electrical engineering), we use additional collected information to determine its primary classification. If the cover page indicates that the awarded degree is Dr.-Ing. (granted to engineers in Germany), the thesis is categorized under engineering. Conversely, if the degree is Dr. rer. nat. (awarded in the natural sciences), it is classified under physics. See Fig. 4 for further details.

[33] The DNB catalog is organized using the Dewey Decimal Classification (DDC) subject categories. Between 2003 and 2004, the library altered these categories, which resulted in inconsistencies in some thesis classifications that potentially affect our fields of interest.

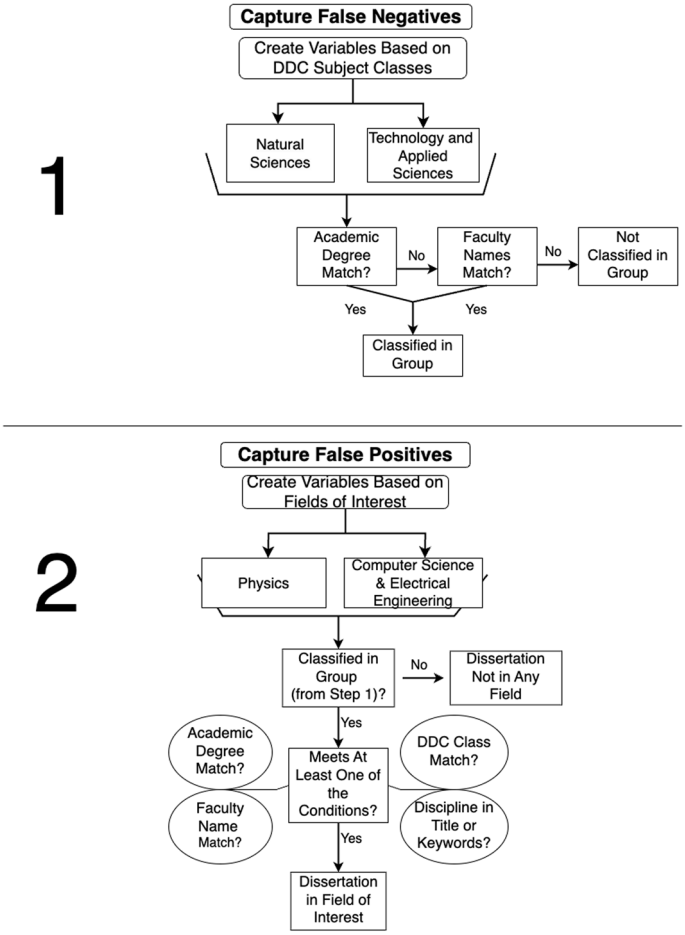


Fig. 4 Approach for thesis discipline classification

B: List of generic words

ag	gmbh	universitaet	university
co	kg	corporation	company
gesellschaft	technology	institute	limited
group	ltd	college	ev
llc	inc	inc.	holding
china	germany	industry	industrial
power	aktiengesellschaft	science	gmbh
science institute	science technology	research institute	new technology
chemie	research	university technology	co ltd
engineer*	state	switzerland	equipment
energy	electric*	and	und

C: List of journals containing sustainability-oriented publications

See Tables 8 and 9

Table 8 Journals in EECS

acm computing surveys	ieee journal of the electron devices society	journal of computer information systems
acm sigbed review	ieee journal on emerging and selected topics in circuits	journal of computer-aided molecular design
acm sigplan notices	ieee journal on exploratory solid-state computational	journal of computer-mediated communication
acm transactions on computer systems	ieee journal on selected topics in signal processing	journal of computing in civil engineering
acm transactions on computer-human	ieee micro	journal of control and decision
acm transactions on cyber-physical systems	ieee microwave and wireless components letters	journal of design research
acm transactions on graphics	ieee multimedia	journal of digital imaging
acm transactions on intelligent	ieee nanotechnology magazine	journal of dynamic systems, measurement
acm transactions on internet technology	ieee network	journal of educational computing research
acm transactions on multimedia	ieee pervasive computing	journal of electroceramics
acm transactions on reconfigurable	ieee photonics journal	journal of electromagnetic engineering and science
acm transactions on sensor networks	ieee photonics technology letters	journal of electronic commerce research
acm transactions on software engineering	ieee power and energy magazine	journal of electronic packaging, transactions
acm transactions on the web	ieee power electronics magazine	journal of energy storage
acs applied energy materials	ieee robotics and automation letters	journal of field robotics
acs photonics	ieee robotics and automation magazine	journal of grid computing
acta informatica	ieee sensors journal	journal of guidance, control, and dynamics
ad hoc networks	ieee signal processing letters	journal of information security and applications
advanced engineering informatics	ieee signal processing magazine	journal of information technology in construction
advanced robotics	ieee software	journal of infrared, millimeter, and terahertz
advances in computational design	ieee solid-state circuits letters	journal of intelligent and robotic systems: theory and
advances in engineering software	ieee systems journal	journal of intelligent information systems
advances in multimedia	ieee transactions on aerospace and electronic systems	journal of intelligent manufacturing
advances in nano research	ieee transactions on antennas and propagation	journal of intelligent transportation systems: technology

Table 8 (continued)

advances in natural sciences: nanoscience	ieee transactions on applied superconductivity	journal of internet services and applications
aeu—international journal of electronics	ieee transactions on automatic control	journal of king saud university computer and
ai magazine	ieee transactions on automation science and engineering	journal of machine engineering
algorithmica	ieee transactions on biomedical circuits and systems	journal of machine learning research
animal biotelemetry	ieee transactions on broadcasting	journal of manufacturing science and engineering
annales des telecommunications	ieee transactions on circuits and systems for video	journal of manufacturing systems
annals of gis	ieee transactions on circuits and systems i	journal of materials processing technology
annual reviews in control	ieee transactions on cloud computing	journal of mechanical design, transactions of the asme
applied clinical informatics	ieee transactions on cognitive and developmental	journal of medical systems
applied computing and informatics	ieee transactions on cognitive communications and	journal of microelectromechanical systems
applied intelligence	ieee transactions on communications	journal of network and computer applications
applied nanoscience (switzerland)	ieee transactions on components, packaging and	journal of optical communications and networking
applied optics	ieee transactions on computational social systems	journal of organizational and end user computing
applied soft computing journal	ieee transactions on computer-aided design of	journal of organizational computing and electronic
archives of computational methods in	ieee transactions on computers	journal of parallel and distributed computing
astronomy and computing	ieee transactions on consumer electronics	journal of pathology informatics
automatica	ieee transactions on control of network systems	journal of power sources
autonomous agents and multi-agent systems	ieee transactions on control systems technology	journal of process control
autonomous robots	ieee transactions on cybernetics	journal of sensors
beilstein journal of nanotechnology	ieee transactions on dependable and secure computing	journal of sensors and sensor systems
big earth data	ieee transactions on dielectrics and electrical insulation	journal of solid state electrochemistry
biological cybernetics	ieee transactions on electromagnetic compatibility	journal of systems and software
biomedical signal processing and control	ieee transactions on electron devices	journal of systems architecture
bmc bioinformatics	ieee transactions on emerging topics in computational	journal of the acm
briefings in bioinformatics	ieee transactions on emerging topics in computing	journal of the association for information systems
business and information systems	ieee transactions on energy conversion	journal of the energy institute
cad computer aided design	ieee transactions on evolutionary computation	journal of the franklin institute
calphad: computer coupling of phase	ieee transactions on fuzzy systems	journal of the optical society of america a

Table 8 (continued)

canadian journal of electrical and	ieee transactions on geoscience and remote sensing	journal of the society for information display
chemometrics and intelligent laboratory	ieee transactions on green communications and	journal of theoretical and computational acoustics
china communications	ieee transactions on image processing	journal of universal computer science
chinese optics letters	ieee transactions on industrial electronics	journal of visual communication and
communications of the acm	ieee transactions on industrial informatics	journal of visualization
communications of the association for	ieee transactions on industry applications	knowledge and information systems
computational and structural biotechnology	ieee transactions on information forensics and security	lighting research and technology
computational geosciences	ieee transactions on instrumentation and measurement	machine learning
computational management science	ieee transactions on intelligent transportation systems	materials horizons
computational materials science	ieee transactions on intelligent vehicles	mathematical structures in computer
computational mechanics	ieee transactions on knowledge and data engineering	mathematics and computers in simulation
computational methods and function theory	ieee transactions on magnetics	mechanical systems and signal processing
computational statistics and data analysis	ieee transactions on microwave theory and techniques	mechanism and machine theory
computational toxicology	ieee transactions on mobile computing	mechatronics
computer	ieee transactions on multi-scale computing systems	memetic computing
computer applications in engineering education	ieee transactions on multimedia	microelectronic engineering
computer assisted language learning	ieee transactions on nanobioscience	microsystems and nanoengineering
computer communication review	ieee transactions on nanotechnology	mobile networks and applications
computer communications	ieee transactions on network and service management	modelling and simulation in materials
computer graphics forum	ieee transactions on network science and engineering	molecular systems biology
computer journal	ieee transactions on neural networks and learning systems	msystems
computer law and security review	ieee transactions on neural systems and rehabilitation	multibody system dynamics
computer methods and programs in biomedicine	ieee transactions on nuclear science	multimedia tools and applications
computer methods in applied mechanics	ieee transactions on parallel and distributed systems	multiscale modeling and simulation
computer networks	ieee transactions on pattern analysis and machine	nano energy
computer physics communications	ieee transactions on power delivery	nano research
computer science review	ieee transactions on power electronics	nano-micro letters
computer standards and interfaces	ieee transactions on power systems	nanophotonics

Table 8 (continued)

computer-aided civil and infrastructure	ieee transactions on reliability	nanotechnology
computers and chemical engineering	ieee transactions on robotics	natural computing
computers and composition	ieee transactions on semiconductor manufacturing	nature climate change
computers and education	ieee transactions on services computing	nature machine intelligence
computers and electrical engineering	ieee transactions on signal and information processing	nature nanotechnology
computers and electronics in agriculture	ieee transactions on signal processing	nature sustainability
computers and fluids	ieee transactions on smart grid	netnomics: economic research and
computers and geosciences	ieee transactions on software engineering	networks
computers and geotechnics	ieee transactions on sustainable computing	networks and spatial economics
computers and industrial engineering	ieee transactions on sustainable energy	neural computing and applications
computers and mathematics with applications	ieee transactions on systems, man and cybernetics part a	neural networks
computers and operations research	ieee transactions on systems, man and cybernetics: systems	neurocomputing
computers and security	ieee transactions on ultrasonics, ferroelectrics, and	nonlinear analysis: hybrid systems
computers and structures	ieee transactions on vehicular technology	nonlinear dynamics
computers in biology and medicine	ieee transactions on very large scale integration	nonlinear engineering
computers in human behavior	ieee transactions on visualization and computer graphics	npj computational materials
computers in industry	ieee transactions on wireless communications	numerical heat transfer, part b
computers in the schools	ieee wireless communications	ocean modelling
computers, materials and continua	ieee wireless communications letters	optical fiber technology
computing and visualization in science	ieee/acm transactions on networking	optical materials
computing in science and engineering	ieee/asme transactions on mechatronics	optical materials: x
constraints	ieee/caa journal of automatica sinica	optical switching and networking
construction innovation	iet control theory and applications	optics and laser technology
control engineering practice	iet electric power applications	optics and lasers in engineering
cybernetics and information technologies	iet electrical systems in transportation	optics and photonics news
data mining and knowledge discovery	iet generation, transmission and distribution	optics communications
design automation for embedded systems	iet image processing	optik
diamond and related materials	iet microwaves, antennas and propagation	optimization and engineering

Table 8 (continued)

diangong...transactions of china electro	iet nanodielectrics	optimization methods and software
dianli... power system protection and control	iet power electronics	organic electronics
dianli... automation of electric power systems	iet radar, sonar and navigation	osa continuum
discrete event dynamic systems: theory and	iet smart grid	pattern recognition
discrete optimization	iete technical review	pattern recognition letters
displays	infor	peer-to-peer networking and applications
distributed computing	information and software technology	peerj computer science
ecological informatics	information fusion	pervasive and mobile computing
electric power systems research	information processing in agriculture	photonics and nanostructures
electronic commerce research	information systems	physica b: condensed matter
electronics letters	information systems and e-business management	physica status solidi (a) applications
empirical software engineering	information systems frontiers	physical communication
energy	information systems journal	plos computational biology
energy and buildings	informs journal on computing	polymer reviews
engineering	integrated computer-aided engineering	proceedings of the ieee
engineering applications of artificial..	intelligent buildings international	progress in computational fluid dynamics
engineering applications of computational..	international journal for numerical methods in fluids	progress in photovoltaics: research..
engineering science and technology	international journal of adaptive control and..	protection and control of modern..
engineering with computers	international journal of advanced manufacturing..	queue
entertainment computing	international journal of ambient computing..	remote sensing letters
environmental modelling and software	international journal of applied engineering research	requirements engineering
epj data science	international journal of approximate reasoning	robotica
eurasip journal on information security	international journal of artificial intelligence	robotics and autonomous systems
experimental heat transfer	international journal of child-computer interaction	science china information sciences
expert systems with applications	international journal of computational..	science robotics
finite elements in analysis and design	international journal of computational methods	security and communication networks
flexible and printed electronics	international journal of computer integrated..	semantic web
flow measurement and instrumentation	international journal of computer mathematics	semiconductor science and technology

Table 8 (continued)

foundations and trends in communications..	international journal of computer-supported..	sensing and bio-sensing research
frontiers in ict	international journal of control	sensors and actuators a: physical
frontiers in neuroinformatics	international journal of control, automation..	sensors and actuators b: chemical
frontiers in neurorobotics	international journal of data science and analytics	siam journal on computing
frontiers in robotics and ai	international journal of digital earth	sigmod record
frontiers of information technology and..	international journal of electrical power and..	signal processing
frontiers of optoelectronics	international journal of fuzzy system applications	signal processing: image communication
future generation computer systems	international journal of fuzzy systems	signal, image and video processing
fuzzy optimization and decision making	international journal of human computer studies	simulation modelling practice and theory
fuzzy sets and systems	international journal of human-computer interaction	slas technology
gaodianya jishu/high voltage engineering	international journal of information technology..	smart materials and structures
gigascience	international journal of intelligent systems	smart structures and systems
high voltage	international journal of machine learning and..	social science computer review
human-centric computing and information..	international journal of modelling and simulation	soft computing
iaeng international journal of computer science	international journal of neural systems	soft robotics
ibm journal of research and development	international journal of precision engineering and..	software—practice and experience
ict express	international journal of robotics research	softwarex
ieee access	international journal of robust and nonlinear control	solar rrl
ieee antennas and propagation magazine	international journal of satellite communications..	solid-state electronics
ieee antennas and wireless propagation letters	international journal of simulation modelling	statistics and computing
ieee circuits and systems magazine	international journal of social robotics	structural and multidisciplinary..
ieee cloud computing	international journal of systems science	studies in informatics and control
ieee communications letters	international journal of unconventional computing	superconductor science and technology
ieee communications magazine	international journal of ventilation	superlattices and microstructures
ieee communications surveys and tutorials	international journal of web and grid services	sustainable computing: informatics..
ieee computational intelligence magazine	international review of aerospace engineering	sustainable energy, grids and networks
ieee consumer electronics magazine	ipsj transactions on computer vision and..	swarm and evolutionary computation

Table 8 (continued)

ieee control systems	isa transactions	swarm intelligence
ieee electrical insulation magazine	it professional	systems and control letters
ieee electrification magazine	journal of aerospace information systems	systems engineering
ieee electron device letters	journal of ambient intelligence and humanized..	telecommunication systems
ieee embedded systems letters	journal of ambient intelligence and smart..	transactions on embedded computing..
ieee geoscience and remote sensing letters	journal of biomedical informatics	transactions on emerging telecommunications..
ieee geoscience and remote sensing magazine	journal of biomedical semantics	universal access in the information society
ieee industrial electronics magazine	journal of building performance simulation	virtual and physical prototyping
ieee intelligent systems	journal of chemical theory and computation	virtual reality
ieee intelligent transportation systems magazine	journal of cloud computing	visualization in engineering
ieee internet computing	journal of combinatorial optimization	vldb journal
ieee journal of biomedical and health informatics	journal of communications and networks	wiley interdisciplinary reviews: computational..
ieee journal of emerging and selected topics in power..	journal of computational design and..	wiley interdisciplinary reviews: data mining..
ieee journal of oceanic engineering	journal of computational physics	wireless communications and mobile computing
ieee journal of photovoltaics	journal of computational physics: x	wireless networks
ieee journal of quantum electronics	journal of computational science	zhongguo proceedings of the chinese society..
ieee journal of selected topics in quantum electronics	journal of computer and system sciences	
ieee journal of solid-state circuits	journal of computer assisted learning	

Table 9 Journals in Physics

2d materials	journal of applied physics	new journal of physics
acoustics australia	journal of biomedical optics	nonlinear processes in geophysics
acs nano	journal of biophotonics	npg asia materials
acs photonics	journal of chemical physics	npj 2d materials and applications
acs sensors	journal of chemical thermodynamics	nuclear engineering and design
acta crystallographica section b	journal of computational physics	nuclear fusion
adsorption science and technology	journal of computational physics: x	nuclear instruments and methods in physics..
advanced engineering materials	journal of crystal growth	nuclear technology
advanced functional materials	journal of dynamic systems, measurement..	numerical heat transfer; part a: applications
advanced optical materials	journal of electroceramics	optica
advanced science	journal of electron spectroscopy and..	optical fiber technology
american journal of physics	journal of energetic materials	optical materials
animal biotelemetry	journal of engineering thermophysics	optical materials: x
annalen der physik	journal of enhanced heat transfer	optics and laser technology
applied acoustics	journal of fluid mechanics	optics and lasers in engineering
applied nanoscience (switzerland)	journal of heat transfer	optics and photonics news
applied optics	journal of hydrodynamics	optics communications
applied physics b: lasers and optics	journal of infrared, millimeter, and terahertz..	optics express
applied physics express	journal of instrumentation	optics letters
applied physics letters	journal of lightwave technology	optik
applied physics reviews	journal of low temperature physics	organic electronics
applied radiation and isotopes	journal of magnetic resonance	osa continuum
applied surface science	journal of magnetism and magnetic materials	particle and particle systems characterization
astronomical journal	journal of materials research	particuology
astronomy and astrophysics	journal of mathematical physics	philosophical magazine
astronomy and computing	journal of molecular liquids	philosophical transactions of the royal society a
astroparticle physics	journal of non-crystalline solids	photochemistry and photobiology
astrophysical journal	journal of non-equilibrium thermodynamics	photonics and nanostructures—fundamentals and..

Table 9 (continued)

astrophysical journal letters	journal of nuclear materials	photonics research
beilstein journal of nanotechnology	journal of nuclear science and technology	physica a: statistical mechanics and its applications
building acoustics	journal of optics (united kingdom)	physica b: condensed matter
celestial mechanics and dynamical..	journal of photochemistry and photobiology a	physica d: nonlinear phenomena
chaos	journal of photochemistry and photobiology b	physica e: low-dimensional systems..
chaos, solitons and fractals	journal of photonics for energy	physica status solidi (a) applications and..
chemical physics letters	journal of physics a: mathematical and..	physica status solidi (b) basic research
chemphyschem	journal of physics and chemistry of solids	physica status solidi—rapid research letters
chinese journal of physics	journal of physics b: atomic, molecular and..	physical chemistry chemical physics
chinese optics letters	journal of physics condensed matter	physical mesomechanics
colloids and surfaces a	journal of physics d: applied physics	physical review a
colloids and surfaces b	journal of polymer science, part b	physical review b
combustion and flame	journal of quantitative spectroscopy and..	physical review c
combustion science and technology	journal of rheology	physical review d
combustion theory and modelling	journal of sensors	physical review e
communications physics	journal of solid state chemistry	physical review letters
composite interfaces	journal of solid state elektrochemistry	physical review materials
comptes rendus physique	journal of sound and vibration	physics letters, section a: general, atomic..
computational materials science	journal of statistical physics	physics of fluids
computer methods in applied..	journal of supercritical fluids	physics of plasmas
computer optics	journal of the electrochemical society	physics of the earth and planetary interiors
computer physics communications	journal of the energy institute	physics today
contemporary physics	journal of the mechanics and physics of solids	planetary and space science
continuum mechanics and..	journal of the optical society of america a	plasma chemistry and plasma processing
contributions to plasma physics	journal of the optical society of america b	plasma physics and controlled fusion
cryogenics	journal of the physical society of japan	plasma processes and polymers
crystal growth and design	journal of the society for information display	plasma sources science and technology
crystengcomm	journal of thermal analysis and calorimetry	polymer degradation and stability

Table 9 (continued)

current applied physics	journal of thermal science	powder metallurgy
diamond and related materials	journal of thermal spray technology	pramana—journal of physics
european journal of mechanics, b/fluids	journal of thermal stresses	probabilistic engineering mechanics
european journal of physics	journal of turbulence	proceedings of the japan academy series b
european physical journal plus	journal of vacuum science and technology a	proceedings of the royal society a
european physical journal: special topics	journal of vibration and acoustics	progress in crystal growth and..
european polymer journal	journal of visualization	progress in photovoltaics: research..
europhysics letters	langmuir	publications of the astronomical society of japan
experimental astronomy	laser and photonics reviews	radiation physics and chemistry
experimental heat transfer	laser physics letters	rare metals
fire safety journal	leukos—journal of illuminating	reports on progress in physics
flow measurement and instrumentation	light: science and applications	results in physics
flow, turbulence and combustion	macromolecular chemistry and physics	review of scientific instruments
fluid dynamics research	materials at high temperatures	reviews of modern physics
fluid phase equilibria	materials characterization	rheologica acta
frontiers in physics	materials chemistry and physics	romanian journal of physics
fusion science and technology	materials letters	romanian reports in physics
geophysical and astrophysical fluid dynamics	materials research bulletin	science china: physics, mechanics..
granular matter	materials science and engineering a	scripta materialia
hardwarex	materials science and engineering b	semiconductor science and technology
heat and mass transfer	materials science and engineering c	sensors and actuators a: physical
heat transfer engineering	materials science and technology	sensors and actuators b: chemical
icarus	materials science in semiconductor..	shock waves
ieee antennas and propagation magazine	materials technology	smart materials and structures
ieee geoscience and remote sensing magazine	materials today	soft matter
ieee journal of photovoltaics	materials today nano	solar physics
ieee journal of quantum electronics	materials today physics	solar rrl
ieee journal of selected topics in quantum..	materials transactions	solid state ionics

Table 9 (continued)

ieee microwave and wireless components letters	measurement science and technology	solid state nuclear magnetic resonance
ieee photonics journal	meccanica	solid state sciences
ieee photonics technology letters	mechanics based design of structures..	solid-state electronics
ieee sensors journal	mechanics of materials	space science reviews
ieee transactions on antennas and propagation	mechanics research communications	spectrochimica acta—part a
ieee transactions on applied superconductivity	metallurgical and materials transactions a	steel research international
ieee transactions on electromagnetic..	metallurgical and materials transactions b	superconductor science and technology
ieee transactions on instrumentation and..	metals and materials international	superlattices and microstructures
ieee transactions on microwave theory and..	methods and applications in fluorescence	surface and coatings technology
ieee transactions on nuclear science	microelectronic engineering	surface and interface analysis
ieee transactions on semiconductor..	microporous and mesoporous materials	surfaces and interfaces
ieee transactions on ultrasonics, ferroelectrics..	microscopy and microanalysis	synchrotron radiation news
ieee wireless communications letters	microsystems and nanoengineering	synthetic metals
iet nanodielectrics	modelling and simulation in materials..	theoretical and applied fracture mechanics
infrared physics and technology	monthly notices of the royal astronomical..	theoretical and computational fluid dynamics
international communications in heat and..	mrs bulletin	thermochimica acta
international journal of heat and fluid flow	nano letters	thin solid films
international journal of heat and mass..	nano research	transactions of nonferrous metals society..
international journal of hydrogen energy	nano-structures and nano-objects	transactions of the institute of measurement..
international journal of mechanical sciences	nanophotonics	tribology international
international journal of multiphase flow	nanoscale advances	tribology letters
international journal of radiation oncology	nanoscale research letters	tribology transactions
international journal of solids and structures	nature astronomy	ultrasonics
international journal of thermal sciences	nature climate change	ultrasonics sonochemistry
ionics	nature communications	vacuum
isa transactions	nature materials	wave motion
journal of adhesion	nature nanotechnology	waves in random and complex media
journal of applied clinical medical physics	nature photonics	wear
journal of applied fluid mechanics	nature sustainability	

D: List of search terms

See Tables 10 and 11.

Table 10 List of keywords, EECS

Trigram	Bigram	Word
Closed-loop supply chain	Energy efficien*	Sustainab*
Life cycle assessment	Climate change	Renewable
Green supply chain	Climate impact	Energy-aware
Carbon dioxide emission*	Environmental impact	
	Environmental damage	
	Green energy	
	Green software	
	Wave energy	
	Wind energy	
	Wind power	
	Solar energy	
	Geothermal energy	
	Greenhouse gas	
	Emission reduction	
	Carbon footprint	
	Evironmental remediation	

Table 11 List of keywords, physics

Trigram	Bigram	Word
Carbon dioxide emission	Climate change	Sustainab*
Sea surface temperature	Global warming	Renewable
Sea ice loss	Ocean warming	Greenhouse
Sea level rise	Climate sensitivity	
Life cycle assessment	CO_2 emission*	
	Climate extremes	
	Temperature variability	
	Environmentally friendly	
	Environmentally benign	
	Environmental impact	
	Green energy	
	Carbon footprint	
	Greenhouse gas	
	Energy efficien*	

E: Top and bottom 5% of patent title distribution

The following Figs. 5, 6, 7, and 8 show which words are relatively more common in patent titles from the bottom 5% vs the top 5% similarity score distributions (by field), using a log-odds style ratio. Table 12 show a random selection of patent titles from the bottom 5% vs the top 5% similarity score distributions (by field).

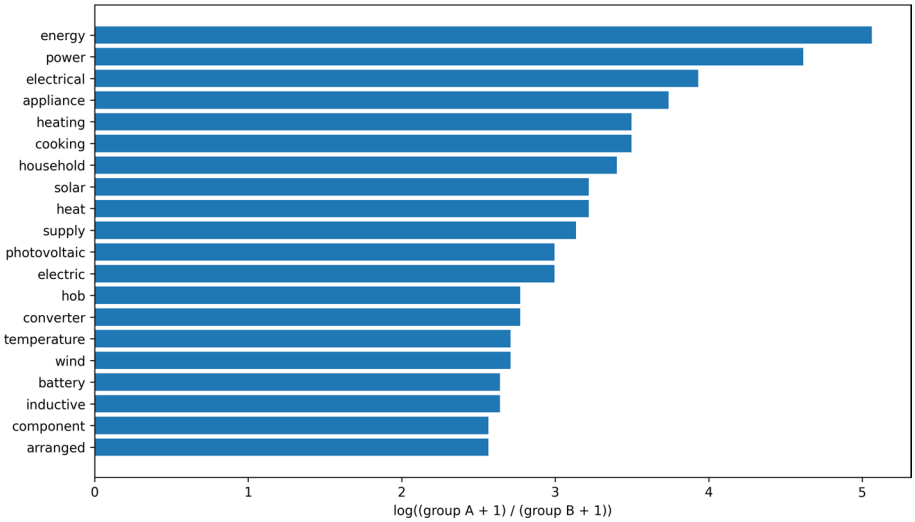


Fig. 5 EECS: Terms more common in top 5% vs bottom 5%

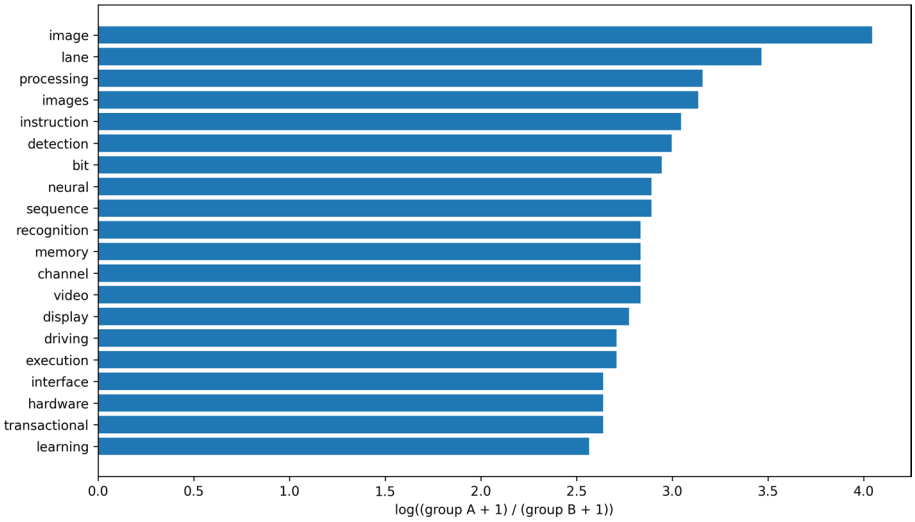


Fig. 6 EECS: Terms more common in bottom 5% vs top 5%

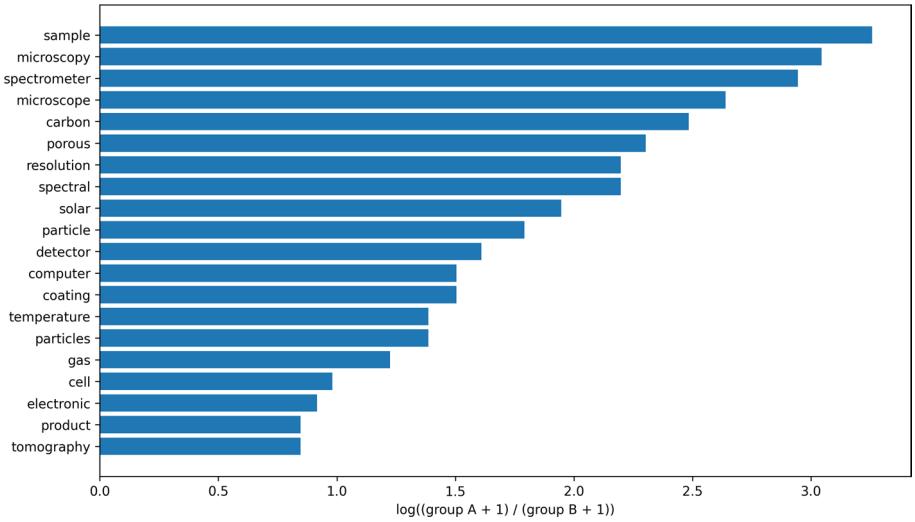


Fig. 7 Physics: Terms more common in top 5% vs bottom 5%

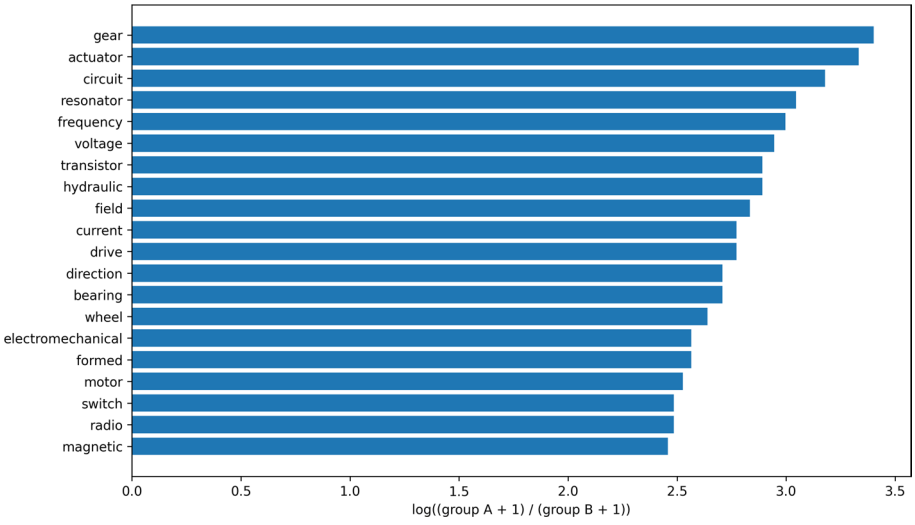


Fig. 8 Physics: Terms more common in bottom 5% vs top 5%

Table 12 Illustrative patent titles from the bottom and top tails of the similarity distribution

Field	Group	Patent title
EECS	Bottom 5%	Meta-data inputs to front end processing for automatic speech recognition
EECS	Bottom 5%	Evaluation method for blood vessel system of patient by computer, involves taking three-dimensional spatially resolved volume image of blood vessel system of patient by computer
EECS	Bottom 5%	An intelligent graph-based expert searching system
EECS	Bottom 5%	Object segmentation based on detected object-specific visual cues
EECS	Bottom 5%	Intrinsic authentication of program code
Physics	Bottom 5%	Semiconductor devices with adhesion-promoting structures and process for their manufacture
Physics	Bottom 5%	Semiconductor device comprising transistors with deformed channel regions and method of manufacturing the same
Physics	Bottom 5%	Fluid actuator for generating a pulsed outlet flow in the flow around an aerodynamic body, a blow-out device with such a fluid actuator and such an aerodynamic body
Physics	Bottom 5%	Varactor diode for LC-series circuit, has space-charge zone whose width is larger than half of finger-shaped/comb-shaped structure thickness when specific depletion voltage is applied to form charge carrier free regions
Physics	Bottom 5%	Method and device for electrically contacting components in a semiconductor wafer
EECS	Top 5%	Method for controlling heat/cold exchange between several energy systems and control platform
EECS	Top 5%	Heating body i.e. radiation heating body, for household appliance i.e. cooktop, for preparing foods, has heating resistor arranged in heating body plate and heat reflector shields formed around heating body at outer side of body plate
EECS	Top 5%	Method and control device for controlling the actual temperature of a part of a building
EECS	Top 5%	Method for controlling heat exchanges and associated control device
EECS	Top 5%	Method for the model-based determination of at least one design parameter of a thermal energy storage device and electronic computing device
Physics	Top 5%	Photovoltaic semiconductor device for converting radiant power into electric power, method for its manufacture and its use
Physics	Top 5%	System and method for determining the spectral sensitivity of a spectrograph
Physics	Top 5%	Hematite pigments

Table 12 (continued)

Field	Group	Patent title
Physics	Top 5%	Electric heating device for cover of opening in component, particularly bumper of vehicle, is arranged at or in component and electric supply line of heating device is arranged near heating device at or in component
Physics	Top 5%	Disconnect assembly for active cooling of packaged electronics

F: Alternative thresholds for SPECTER based classification

See Tables 13 and 14.

Table 13 Specter Threshold Robustness: Likelihood of Publishing PhDs being Cited by Patents

	Dependent variable					
	Specter 95th (Baseline)		Specter 90th		Specter 99th	
	(1)	(2)	(3)	(4)	(5)	(6)
Sus. PhD	1.613***	1.528***	1.675***	1.593***	2.126***	1.993***
	(0.110)	(0.124)	(0.106)	(0.118)	(0.131)	(0.144)
Sus. Post PhD		1.178***		1.124***		1.137***
		(0.167)		(0.167)		(0.202)
Gender	0.054	0.021	− 0.035	− 0.073	0.169	0.114
	(0.082)	(0.085)	(0.067)	(0.070)	(0.149)	(0.151)
EECS	0.422***	0.404***	0.258***	0.245***	1.090***	0.992***
	(0.059)	(0.060)	(0.050)	(0.052)	(0.101)	(0.103)
No. of patents	0.008	0.007	0.008	0.008	0.006	0.006
	(0.006)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)
No. of citing patents	0.024***	0.024***	0.040***	0.040***	0.001	0.001
	(0.002)	(0.002)	(0.003)	(0.003)	(0.001)	(0.001)
Log no. of Publ	0.196***	0.173***	0.198***	0.179***	0.346***	0.307***
	(0.022)	(0.023)	(0.019)	(0.020)	(0.037)	(0.037)
Pre PhD Ind. Affil	0.027	0.073	0.010	0.023	0.110	0.157
	(0.094)	(0.097)	(0.083)	(0.087)	(0.134)	(0.137)
TU9	0.136*	0.140*	0.057	0.042	0.218*	0.195
	(0.060)	(0.062)	(0.051)	(0.053)	(0.097)	(0.100)
German U15	0.014	0.025	0.069	0.055	0.074	0.078
	(0.068)	(0.070)	(0.057)	(0.058)	(0.124)	(0.125)
Constant	− 2.770***	− 2.687***	− 2.190***	− 2.109***	− 5.021***	− 4.821***
	(0.142)	(0.144)	(0.123)	(0.125)	(0.268)	(0.265)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.109	0.115	0.123	0.128	0.114	0.113
Observations	12,961	11,968	12,961	11,968	12,961	11,968

Note: *p<0.05; **p<0.01; ***p<0.001

Table 14 Specter Threshold Robustness: Likelihood of Publishing PhDs Inventing Patents

	Dependent variable:					
	Specter 95th (Baseline)		Specter 90th		Specter 99th	
	(1)	(2)	(3)	(4)	(5)	(6)
Sus. PhD	1.774***	1.722***	0.931***	0.958***	1.471***	1.423***
	(0.200)	(0.216)	(0.173)	(0.191)	(0.227)	(0.243)
Sus. Post PhD		0.788**		0.612**		0.786**
		(0.298)		(0.224)		(0.283)
Gender	0.361	0.349	0.071	0.041	0.189	0.206
	(0.280)	(0.286)	(0.162)	(0.166)	(0.315)	(0.321)
EECS	0.031	− 0.030	0.700***	0.667***	1.063***	1.014***
	(0.136)	(0.139)	(0.088)	(0.089)	(0.163)	(0.165)
No. of Patents	0.037***	0.038***	0.006	0.006	0.002	0.002
	(0.006)	(0.006)	(0.006)	(0.006)	(0.003)	(0.003)
Log No. of Publ	0.087	0.048	0.866***	0.846***	0.758***	0.705***
	(0.052)	(0.055)	(0.041)	(0.042)	(0.067)	(0.068)
Pre PhD Ind. Affil	− 0.144	− 0.116	0.231*	0.237*	0.318	0.331
	(0.170)	(0.174)	(0.104)	(0.107)	(0.169)	(0.172)
TU9	− 0.256	− 0.244	− 0.114	− 0.137	− 0.134	− 0.173
	(0.143)	(0.147)	(0.087)	(0.089)	(0.153)	(0.158)
German U15	0.152	0.145	0.185	0.169	− 0.066	− 0.030
	(0.164)	(0.168)	(0.104)	(0.105)	(0.202)	(0.205)
Constant	− 3.253***	− 3.127***	− 4.076***	− 3.970***	− 6.218***	− 6.054***
	(0.390)	(0.394)	(0.254)	(0.258)	(0.471)	(0.471)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.089	0.091	0.159	0.157	0.15	0.151
Observations	4,241	3,961	4,241	3,961	4,241	3,961

Note: *p<0.05; **p<0.01; ***p<0.001

Funding Open Access funding enabled and organized by Projekt DEAL. Funding by Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) under grant 16WIT001 is gratefully acknowledged. Use of publication data from Scopus was enabled by the German Competence Network for Bibliometrics funded by BMBFTR under grant numbers 16WIK2101C and 16WIK2101A.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

A12. (2024). Semantic scholar. Retrieved 30 July 2024, from https://www.semanticscholar.org/.

Adams, R., Jeanrenaud, S., Bessant, J., Denyer, D., & Overy, P. (2016). Sustainability-oriented innovation: A systematic review. *International Journal of Management Reviews, 18*(2), 180–205.

Aria, M., Le, T., Cuccurullo, C., Belfiore, A., & Choe, J. (2024). openalexr: An r-tool for collecting bibliometric data from openalex. *The R Journal, 15*, 167–180. https://doi.org/10.32614/RJ-2023-089

Auriol, L., Misu, T. M., & Galindo-Rueda, F. (2016). Doctorate holders' labor market and mobility: The academic career as the first choice. In L. Gokhberg, N. Shmatko, & L. Auriol (Eds.), *The science and technology labor force: the value of doctorate holders and development of professional careers* (pp. 51–74). Springer International Publishing.

Avramidis, E., Macketanz, V., Strohriegel, U., Burchardt, A., & Möller, S. (2020). Fine-grained linguistic evaluation for state-of-the-art machine translation. In Proceedings of the fifth conference on machine translation (pp. 346–356). Association for Computational Linguistics.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In K. Inui, J. Jiang, V. Ng, and X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (pp. 3615–3620), Hong Kong, China. Association for Computational Linguistics.

Beynaghi, A., Trencher, G., Moztarzadeh, F., Mozafari, M., Maknoon, R., & Leal Filho, W. (2016). Future sustainability scenarios for universities: Moving beyond the United Nations Decade of Education for Sustainable Development. *Journal of Cleaner Production, 112*, 3464–3478.

Breschi, S., & Lissoni, F. (2001). Knowledge spillovers and local innovation systems: A critical survey. *Industrial and Corporate Change, 10*(4), 975–1005.

Breschi, S., Lissoni, F., & Montobbio, F. (2005). From publishing to patenting: Do productive scientists turn into academic inventors? *Revue d'Économie Industrielle, 110*(1), 75–102.

Bruns, S. B., & Kalthaus, M. (2020). Flexibility in the selection of patent counts: Implications for p-hacking and evidence-based policymaking. *Research Policy, 49*(1), Article 103877.

Buechele, S., Buenstorf, G., Huegel, M., Koenig, J., & Theissen, M. (2026). Is the dominance of graduates from top-tier universities among tenured faculty driven by prestige or output? Evidence from 50 years of university appointments in Germany. *Research Policy, 55*(3), 105411.

Buenstorf, G., & Geissler, M. (2014). Like Doktorvater, like son? Tracing role model learning in the evolution of German laser research. *Jahrbücher für Nationalökonomie und Statistik, 234*(2–3), 158–184.

Buenstorf, G., & Heinisch, D. P. (2020). When do firms get ideas from hiring PhDs? *Research Policy, 49*(3), 103913.

Buenstorf, G., Koenig, J., & Otto, A. (2023). Expansion of doctoral training and doctorate recipients' labour market outcomes: Evidence from German register data. *Studies in Higher Education, 48*(8), 1216–1242. https://doi.org/10.1080/03075079.2023.2188397

Buenstorf, G., Koenig, J., & Otto, A. (2025). Keeping up with the Max Plancks? Germany's quest for university excellence and the role of public research institutes in doctoral education. *Scientometrics, 130*(1), 67–108.

Cabanes, B., Plantec, Q., Le Masson, P., & Weil, B. (2024). Basic or applied research in university–industry collaborations? An analysis of research orientations in knowledge creation partnerships. *European Management Review*. https://doi.org/10.1111/emre.12696

Cantner, U., Kalthaus, M., & Yarullina, I. (2024). Outcomes of science-industry collaboration: Factors and interdependencies. *The Journal of Technology Transfer, 49*(2), 542–580.

Caravale, A., Moscati, P., Duran-Silva, N., Grimau, B., & Rondelli, B. (2023). Developing a digital archaeology classification system using natural language processing and machine learning techniques. *Archeologia e Calcolatori, 34*(2), 9–32.

Celone, A., Cammarano, A., Caputo, M., & Michelino, F. (2022). Features of sustainability-oriented innovations: a content analysis of patent abstracts. *Sustainability, 14*(23), 15492.

Chams, N., Guesmi, B., & Folch, M., Cubel, R., & Gil Roig, J. M. (2024). Transfer versus co-production: Knowledge as 'MEANS' to sustainability as an 'END'. *Research Evaluation, 33*, rvae032.

Clark, B. R. (1998). The entrepreneurial university: Demand and response. *Tertiary Education and Management, 4*(1), 5–16.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2270–2282). Association for Computational Linguistics.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and impacts: The influence of public research on industrial R&D. *Management Science, 48*(1), 1–23.
Culp, M., Johnson, K., & Michailides, G. (2007). ada: An R package for stochastic boosting. *Journal of Statistical Software, 17*, 1–27.
De Marchi, V. (2012). Environmental innovation and R&D cooperation: Empirical evidence from Spanish manufacturing firms. *Research Policy, 41*(3), 614–623.
de Rassenfosse, G., Dernis, H., Guellec, D., Picci, L., & van Pottelsberghe de la Potterie, B. (2013). The worldwide count of priority patents: A new indicator of inventive activity. *Research Policy, 42*(3), 720–737.
Di Maria, E., De Marchi, V., & Spraul, K. (2019). Who benefits from university-industry collaboration for environmental sustainability? *International Journal of Sustainability in Higher Education, 20*(6), 1022–1041.
Duede, E., Dolan, W., Bauer, A., Foster, I., & Lakhani, K. (2024). Oil & water? Diffusion of AI within and across scientific fields. arXiv:2405.15828.
Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: From national systems and "mode 2'' to a triple helix of university-industry-government relations. *Research Policy, 29*(2), 109–123.
European Patent Office. (2023). Data catalog patstat global—2023 spring edition.
European Patent Office. (2024). Patent Index 2023.
Eurostat. (2024). R&D personnel. Retrieved 27 June 204, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=R%26D_personnel.
Favot, M., Vesnic, L., Priore, R., Bincoletto, A., & Morea, F. (2023). Green patents and green codes: How different methodologies lead to different results. *Resources, Conservation & Recycling Advances, 18*, 200132.
Gaughan, M., & Robin, S. (2004). National science training policy and early scientific careers in France and the United States. *Research Policy, 33*(4), 569–581.
Gómez, J., & Vázquez, P.-P. (2022). An empirical evaluation of document embeddings and similarity metrics for scientific articles. *Applied Sciences, 12*(11), 5664.
Haščič, I. & Migotto, M. (2015). Measuring environmental innovation using patent data. Technical report, OECD. OECD Environment Working Paper No. 89.
Heinisch, D. P., Koenig, J., & Otto, A. (2020). A supervised machine learning approach to trace doctorate recipients' employment trajectories. *Quantitative Science Studies, 1*(1), 94–116.
Huberty, M., Serwaah, A., & Zachmann, G. (2014). A flexible, scaleable approach to the international patent 'name game. Working Papers, 850, Bruegel.
Jaffe, A. B. (1989). Real effects of academic research. *The American Economic Review,* 957–970.
Ji, Z., Guo, S., Qiao, Y., & McDougal, R. A. (2024). Automating literature screening and curation with applications to computational neuroscience. *Journal of the American Medical Informatics Association, 31*(7), 1463–1470.
Kaiser, U., Kongsted, H. C., Laursen, K., & Ejsing, A.-K. (2018). Experience matters: The role of academic scientist mobility for industrial innovation. *Strategic Management Journal, 39*(7), 1935–1958.
König, J. (2022). Postdoctoral employment and future non-academic career prospects. *PLOS ONE, 17*(12), 1–17.
Liaw, A. (2002). Classification and regression by randomforest R news.
Lissoni, F., Sanditov, B., & Tarasconi, G. (2006). The Keins database on academic inventors: Methodology and contents. KITeS Working Papers 181, KITeS, Centre for Knowledge, Internationalization and Technology Studies. Milano.
Mansfield, E. (1995). Academic research underlying industrial innovations: sources, characteristics, and financing. *The Review of Economics and Statistics,* 55–65.
Maraut, S., & Martínez, C. (2014). Identifying author–inventors from Spain: Methods and a first insight into results. *Scientometrics, 101*(1), 445–476.
Markard, J., Raven, R., & Truffer, B. (2012). Sustainability transitions: An emerging field of research and its prospects. *Research Policy, 41*(6), 955–967.
Marx, M., & Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal, 41*(9), 1572–1594.
Marx, M., & Fuegi, A. (2022). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy, 31*(2), 369–392.
Matthies, A.-L., Koen, H., & Leskošek, V. (2022). Applying transdisciplinary sustainability transitions research in international social work doctoral training. *Social Work Education, 41*(7), 1541–1559.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (pp. 105–142). Academic Press.

Meyer, M. (2006). Knowledge integrators or weak links? An exploratory comparison of patenting researchers with their non-inventing peers in nano-science and technology. *Scientometrics, 68*, 545–560.
Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between us technology and public science. *Research Policy, 26*(3), 317–330.
Nast, C., Llopis, O., Yankova, D., & D'Este, P. (2025). Sourcing insights elsewhere: The positive influence of academic engagement on scientific impact. *Technovation, 139*, 103112.
Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy, 36*(7), 1016–1034.
OECD (2018). Education at a glance 2018: OECD Indicators. Technical report, OECD Publishing.
Perkmann, M., Salandra, R., Tartari, V., McKelvey, M., & Hughes, A. (2021). Academic engagement: A review of the literature 2011–2019. *Research Policy, 50*(1), 104114.
Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., Krabel, S., Kitson, M., Llerena, P., Lissoni, F., Salter, A., & Sobrero, M. (2013). Academic engagement and commercialisation: A review of the literature on university–industry relations. *Research Policy, 42*(2), 423–442.
Pezzoni, M., Lissoni, F., & Tarasconi, G. (2014). How to kill inventors: testing the massacrator algorithm for inventor disambiguation. *Scientometrics, 101*(1), 477–504.
Plantec, Q., Cabanes, B., Le Masson, P., & Weil, B. (2023). Early-career academic engagement in university-industry collaborative PhDs: Research orientation and project performance. *Research Policy, 52*(9), 104856.
Popp, D. (2012). The role of technological change in green growth. National Bureau of Economic Research. Working Paper 18506.
Porter, M. E., & van der Linde, C. (1995). Toward a new conception of the environment-competitiveness relationship. *Journal of Economic Perspectives, 9*(4), 97–118.
Rafols, I. (2014). Knowledge integration and diffusion: Measures and mapping of diversity and coherence. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 169–190). Springer International Publishing.
Roach, M., & Cohen, W. M. (2013). Lens or prism? Patent citations as a measure of knowledge flows from public research. *Management Science, 59*(2), 504–525.
Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., Taubert, N., Bausenwein, T., & Stahlschmidt, S. (2024). The data infrastructure of the german kompetenznetzwerk bibliometrie: An enabling intermediary between raw data and analysis. Preprint.
Schmoch, U., Dornbusch, F., Mallig, N., Michels, C., Schulze, N., & Bethke, N. (2012). Vollständige erfassung von patentanmeldungen aus universitäten. bericht an das bundesministerium für bildung und forschung (bmbf). Fraunhofer Institut System und Innovationsforschung. Technical report.
Schoen, A., Heinisch, D., & Buenstorf, G. (2014). Playing the 'name game' to identify academic patents in Germany. *Scientometrics, 101*, 527–545.
Schot, J., & Steinmueller, W. E. (2018). Three frames for innovation policy: R&D, systems of innovation and transformative change. *Research Policy, 47*(9), 1554–1567.
Shibayama, S. (2019). Sustainable development of science and scientists: Academic training in life science labs. *Research Policy, 48*(3), 676–692.
Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2022). Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on empirical methods in natural language processing*.
Spender, J.-C., & Grant, R. M. (1996). Knowledge and the firm: Overview. *Strategic Management Journal, 17*(S2), 5–9.
Stephens, J. C., Hernandez, M. E., Román, M., Graham, A. C., & Scholz, R. W. (2008). Higher education as a change agent for sustainability in different cultures and contexts. *International Journal of Sustainability in Higher Education, 9*(3), 317–338.
Stern, N., & Valero, A. (2021). Innovation, growth and the transition to net-zero emissions. *Research Policy, 50*(9), 104293.
Toole, A. A., & Czarnitzki, D. (2010). Commercializing science: Is there a university "Brain Drain'' from academic entrepreneurship? *Management Science, 56*(9), 1599–1614.
Trencher, G., Yarime, M., McCormick, K. B., Doll, C. N. H., & Kraines, S. B. (2013). Beyond the third mission: Exploring the emerging university function of co-creation for sustainability. *Science and Public Policy, 41*(2), 151–179.
Yang, Y. (2024). Impact of institutional support and green knowledge transfer on university students' absorptive capacity and green entrepreneurial behavior: The moderating role of environmental responsibility. *PLOS ONE, 19*, 1–19.

Yström, A., Plantec, Q., Ferrary, M., Miller, K., Rennings, M., & Sarpong, D. (2025). From science management to innovation management: New forms of science-industry relations and knowledge transfer. *Technovation, 145*, 103257.

Zucker, L. G., Darby, M. R., & Torero, M. (2002). Labor mobility from academe to commerce. *Journal of Labor Economics, 20*(3), 629–660.

Zumbach, D. & Bauer, P. C. (2021). Deeplr: tto the "deepl" translation API. Retrieved from https://github.com/zumbov2/deeplr.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.