Author-Mobi: A Mobility Dataset of PubMed Authors

Dongin Nam[1], Tzu-Kun Hsiao[1,2], Deepanshu Malhotra[1], Malik Oyewale Salami[1], Liri Fang[1,3], Xiaoliang Jiang[1], Vetle I. Torvik[1]

[1] School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA
[2] Present affiliation: Scholar Support and Data Services, University of Mississippi, University, MS, USA
[3] Present affiliation: Amazon, Seattle, WA, USA

ABSTRACT

We present the Author-Mobi dataset, which integrates MapAffil 2018 with Author-ity 2018 to build a timestamped, name-disambiguated author-affiliation dataset covering 13.7 million authors across 120 million author name instances from PubMed, linked NSF/NIH grants, and ORCID employment/education records. Using an exponential smoothing technique, Author-Mobi reconstructs researchers' primary institution, city, country, and discipline over time, with particular value for historical periods in which PubMed affiliation coverage is incomplete. Newly inferred affiliation information accounted for 77-92% and 58-70% of author name instances, during 1946-1987 and 1988-2013, respectively. We evaluate the reconstructed histories against a Large Language Model (LLM)-assisted "silver standard" derived from publicly available career histories of 192 Nobel laureates in Physiology or Medicine. Author-Mobi and the LLM-based silver standard agreed on institution and city 84% and 90% during 1946-1987 and 1988-2013, respectively. This shows high accuracy in the presence of missing data, although Nobel laureates are not representative of the average author in terms of productivity. These rates represent a lower bound accuracy of Author-Mobi because the LLM can make mistakes. Some of these disagreements are due to publication lags where the institution transition agrees but Author-Mobi lags behind a year or two. The resulting longitudinal affiliation trajectories support studies of scientific careers and mobility. Author-Mobi is freely available in the Illinois Data Bank: https://databank.illinois.edu/datasets/IDB-4273770?code=FNvLExGBwqyGyCxnFdr9lSbAEIXEBjN5I0Zc9QhZ_f4

Background & Summary

The mobility of highly skilled workers plays an important role in the movement of human capital, knowledge, and expertise across organizations and geographic regions. Mobility across organizational boundaries has been studied in relation to knowledge spillovers and innovation, with studies examining the movement of inventors and R&D workers between firms [1-3]. Geographic mobility can also facilitate knowledge transfer and recombination, expand social and professional networks, and create value for both individuals and organizations [4]. At a broader regional scale, the movement of highly skilled workers can have important consequences for innovation and the distribution of human capital across both origin and destination regions [5].

Researchers represent a particularly important population in this context because their movements across institutions, cities, and countries can shape the circulation of scientific knowledge and human capital [6, 7]. Scientific research contributes to innovation through the diffusion of knowledge, development of human capital, and commercialization of new technologies [8-11]. Recent evidence further highlights the

importance of university-trained scientists in supporting firms' research and inventive activity [12]. Because scientific knowledge and capabilities are partially embodied in researchers, their movement across institutions and geographic locations can therefore redistribute scientific human capital [6, 13]. Researcher mobility has been associated with changes in collaboration networks, scientific productivity, and research impact, although these effects vary across mobility types and contexts [14, 15]. Prior studies have examined mobility patterns at multiple scales. In the United States, academic mobility has been associated with movements toward more research-intensive institutions and urban locations [16], while large-scale analyses of life scientists have identified structured patterns of intercity and international mobility, including strong national border effects and differences in how countries participate in international scientific exchange [17].

Studying researcher mobility, however, requires reliable longitudinal information about where individual researchers are affiliated over time. Bibliometric data offer the scale needed for large-scale mobility research, but reconstructing career trajectories from publication records introduces challenges related to author disambiguation, incomplete affiliation coverage, and the timing of observed affiliations [15, 18]. Major bibliographic systems such as Dimensions, OpenAlex, Lens, and Web of Science (WoS) provide structured affiliation information and increasingly incorporate standardized institution identifiers such as Global Research Identifier Database (GRID) and Research Organization Registry (RoR). However, the temporal information is often not provided as a ready-made, per-year author affiliation timeline in these repositories. For example, OpenAlex's author-level affiliation history covers roughly the most recent decade [19]. Researchers have therefore developed different methods and datasets to reconstruct mobility from bibliographic records. Hajikhani and Suominen [20] released a WoS-based dataset covering Finland-based researchers from 2008 to 2018, representing city- and disciplinary-level mobility through origin-destination matrices and networks. Dobre et al. [21] developed a customizable framework for reconstructing annual researcher affiliation histories, carrying the most recently observed affiliation through years without publications, and released the accompanying PySpark workflow through the Elsevier Data Repository [22]. Wei et al. [18] integrate Microsoft Academic Graph (MAG) and SciSciNet to construct standardized annual career timelines for approximately 793,000 scholars, from which they identify nearly 2.9 million mobility events and classify mobility by geographic scope, institutional prestige, and career stage. Scholarly Migration Database[1] reconstructs researcher migration histories from Scopus and OpenAlex and provides openly accessible bilateral migration flows and rates at national and subnational levels, with further disaggregation by field and gender [23, 24]. These resources demonstrate the growing feasibility of reconstructing researcher mobility at scale, while also reflecting differences in temporal reconstruction methods, geographic and institutional resolution, and the level at which reconstructed mobility data are made available.

Author-Mobi complements these existing efforts by providing a publicly available, author-level dataset of yearly affiliation trajectories across multiple dimensions, including institution, city, country, and discipline. Such characteristics enable researchers to construct and examine mobility measures and researcher careers based on different dimensions from individual longitudinal trajectories. Author-Mobi is constructed by linking MapAffil 2018, which standardizes affiliation information, with Author-ity 2018, which disambiguates PubMed author identities. Author-Mobi also incorporates supplementary information from linked National Science Foundation (NSF) and National Institutes of Health (NIH) grants and Open Researcher and Contributor ID (ORCID) records. Also, because Author-Mobi is based on freely downloadable PubMed bibliographic data, public dissemination is possible for these reconstructed author-level records.

Beyond mobility events, longitudinal information about individual researchers provides a basis for examining broader patterns in scientific careers. Longitudinal researcher-level data have enabled studies

[1] https://www.scholarlymigration.org/

of academic career progression and changes in cohort composition [25], as well as investigations linking scientists' career and research histories to the evolution of scientific fields [26]. Research careers can also be examined as sequences of institutional and occupational states, rather than solely through individual mobility events or publication outcomes [27, 28]. Large-scale analyses of biomedical publication histories have further examined how scientific productivity, publication career longevity, interdisciplinarity, and the adoption of new ideas vary across researchers' careers [29-32]. Related research has also combined survey information with longitudinal publication records to examine how the timing and sequencing of life-course events relate to academic outcomes [33]. Building on these approaches, Author-Mobi provides an additional longitudinal affiliation dimension, enabling researchers to examine institutional persistence, observed career duration, and the timing and sequence of institutional and geographic affiliations alongside mobility.

A central objective of Author-Mobi is to reconstruct historical affiliation trajectories during the periods when author-level affiliation information in PubMed was incomplete. The National Library of Medicine (NLM) did not consistently record first-author affiliations until 1988 and began recording affiliations for all authors when available in 2014, leaving substantial gaps in earlier bibliographic records [34, 35]. To address these gaps, Author-Mobi combines affiliation evidence from surrounding years and linked external records to infer researchers' primary institutional, geographic, and disciplinary affiliations. Based primarily on the December 2018 PubMed snapshot, the resulting dataset covers approximately 13.7 million disambiguated authors and nearly 120 million author name instances from PubMed and linked NSF, NIH, and ORCID records. We further assess the reconstructed histories through an external LLM-assisted silver-standard benchmark constructed from publicly documented career histories of Nobel laureates in Physiology or Medicine. Figure 1 illustrates how Author-Mobi reconstructs an author's affiliation trajectory over time.

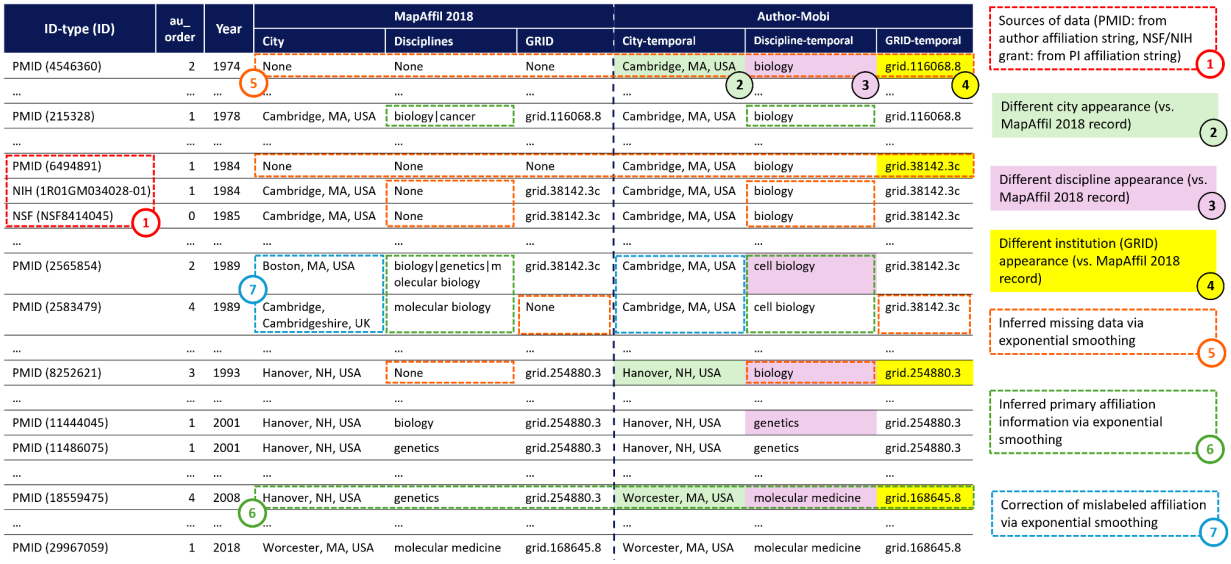


Figure 1. An illustrative snippet of Author-Mobi records for Victor Ambros (Nobel laureate in Physiology or Medicine, AUID = 4546360_2) showing his longitudinal affiliation trajectory and the reconstruction of missing affiliation information ("None"). Continuous records have been truncated with an ellipsis (...), and instances are ordered chronologically. Red boxed cells indicate source data, while highlighted cells indicate changes from the previous record (green: city, pink: discipline, yellow: institution). The trajectory shows institutional and geographic changes over time, including MIT in Cambridge, MA in 1974; Harvard University in Cambridge, MA in 1984 and 1989; Dartmouth College in Hanover, NH in 1993 and 2001; and the University of Massachusetts in Worcester, MA in 2008. Author-Mobi also reconstructs missing or potentially erroneous affiliation information using surrounding records. For example, for the 1989 record (PMID 2583479),

Author-Mobi inferred Cambridge, MA, USA, whereas MapAffil 2018 mapped the record to Cambridge, Cambridgeshire, UK, reflecting an indexing error in the underlying source.

Methods

In this section, we first describe the two core datasets, MapAffil 2018 and Author-ity 2018, that serve as the foundation of Author-Mobi. We then describe supplementary linked records from NSF and NIH grants and ORCID that provide additional affiliation information beyond PubMed. Finally, we describe the temporal affiliation reconstruction procedure used to construct Author-Mobi.

MapAffil: Mapping affiliation strings to institutions, cities, and disciplines

The MapAffil 2018 dataset[2] was curated based on the information of the PubMed author affiliations mapped to geographical locations. The MapAffil bibliographic tool [36] was used to build this dataset, which takes advantage of a rule-based algorithm that is based on a dictionary consisting of 24,000 city/state/country names and 1.1 million word n-grams indicating a unique country/US state for disambiguation. The algorithm extracts the cities, GRID identifiers, and disciplines from the raw affiliation string. After applying this algorithm to the December 2018 PubMed snapshot, which is an annual baseline file of all PubMed citation records provided by NLM, the MapAffil 2018 dataset was built. As a result, this dataset consists of 52,931,957 author-paper rows with cleaned affiliation strings and 18 feature columns such as *PMID, au_order* (i.e., order of the author in a given paper)*, lastname, firstname, affiliation, grid, city, state, country, lat* (i.e., latitude)*,* and *lon* (i.e., longitude). For more information, refer to MapAffil 2018 dataset.

Author-ity: Disambiguating author names in PubMed

The Author-ity 2018 dataset[3] is a PubMed author name disambiguated dataset. This dataset was derived from the Author-ity software [37], which utilizes a pairwise probabilistic model across different attributes in MEDLINE. It disambiguated author identities collected on the December 2018 PubMed snapshot (same as the MapAffil 2018), consisting of a total of 29.1 million articles and 114.2 million author name instances. Each instance row is uniquely indexed by PMID and author position information (e.g., 10786286_3 is the third author name on PMID 10786286). Records were grouped based on the last name and first name initial (including some close variants) and then inferred to a cluster of predicted individuals. Each cluster has a unique Author ID (AUID; the PMID of the earliest paper in the cluster and the author name position) and includes 19 more features such as *name variants, email addresses, ORCIDs, top 20 most frequent Medical Subject Headings (MeSH), co-author names, grant IDs,* and *citation counts*. For more details, refer to Author-ity 2018 dataset.

Linked datasets: NSF, NIH, and ORCID

While most of our data is based on PMIDs, some author instances lacked relevant affiliation information in PubMed. In this case, the missing affiliation data had to be harvested from external resources such as NSF/NIH grants and ORCIDs. To do this, the Author-linked data for Author-ity 2009 dataset[4] was used, which links PubMed author IDs from Author-ity 2009 dataset with principal investigators (i.e., PI) on NSF/NIH grants and inventors on USPTO patents to connect researchers beyond PubMed articles. Records were each matched based on a probability model, and only the ones that had more than a 0.5

[2] https://doi.org/10.13012/B2IDB-2556310_V1
[3] https://doi.org/10.13012/B2IDB-2273402_V1
[4] https://doi.org/10.13012/B2IDB-4370459_V1

probability score were retained. For our new dataset, the PIs on NSF/NIH grants were only considered. In addition, ORCIDs Mapped to PubMed Authors dataset[5] was used, which is also a linked dataset that bridges PubMed author IDs with ORCID's employment and education records based on a rule-based matching system.

These additional linked datasets were essential in the process of linking external sources (in this case, NSF, NIH, and ORCID) beyond PubMed information. This enabled our dataset to be more robust by preventing valuable information from being lost.

Author-Mobi

Each row in Author-Mobi represents an author instance from PubMed or a linked external source. For all records sharing the same AUID and year, the same temporally inferred primary city, institution, and discipline are assigned. Also, it includes the linked data as records alongside PubMed records. However, when there was missing affiliation information in the PubMed data, relevant information from external sources were collected based on the aforementioned linked datasets.

The joining process of the two datasets was based on a shared key (i.e., PMID and author position). While Author-ity 2018 provided the disambiguated author name cluster, MapAffil 2018 provided the geocoded affiliation information for an author. Considering this, for a given disambiguated author, all cities, disciplines, and institutions were collected, respectively. To infer the value for a specific author *A* in a target year *T*, we designed an exponential smoothing technique that assigns exponentially decaying weights to observations from surrounding years *Y*. Thus, observations in the target year *T* receive the maximum weight (i.e., 1), and the weighted halves for each yearly span (0.5 for *T±1,* 0.25 for *T±2,* etc.). For example, assuming that an author instance has *Boston, MA, USA* for its city attribute in the year *2010* (*T*), *Boston, MA, USA* will receive 1 as the weight; if the same city is observed in 2011, the weight is 0.5; in 2012, 0.25; and so on. These weights are summed per candidate, and the highest score is selected. The selected candidate is treated as the author's primary affiliation value for that year. For missing data (e.g., non-first authors in PubMed prior to 2014), the first author's affiliation was considered as a candidate, however, with less weight. Weights also vary depending on the quality of the information and were adjusted through qualitative assessment. For instance, PubMed Central identifiers (PMCIDs) receive less weight as they tend to be less reliably assigned to the correct author. Excessively long affiliation strings were also downweighted because they typically are a result of multiple institutions conflated as one chunk.

The final dataset contains 14 attributes for each author instance. Records are grouped by Author-ity Author ID and ordered by year to form longitudinal affiliation trajectories. Original affiliation information is retained alongside the temporally inferred fields, including *grid-temporal*, *discipline-temporal*, *City-temporal*, and *Country-temporal*, as well as consistency and mobility indicators such as *CitySame*, *CountrySame*, and *Moved*.

Data Records

Author-Mobi contains city, institution, and discipline for 119,966,689 name instances across 30,638,055 unique records from PubMed, ORCID, NSF grants, and NIH grants. For each AUID-year, all corresponding records are assigned the same temporally inferred primary city, institution, and discipline. The main dataset contains 14 attributes, including record and author identifiers, source information, original affiliation information, temporally inferred affiliation fields (*grid-temporal, discipline-temporal,*

[5] https://doi.org/10.13012/B2IDB-9246015_V1

City-temporal, and *Country-temporal*), and consistency and mobility indicators (*CitySame, CountrySame,* and *Moved*). The dataset will be made publicly available through a public research data repository.

The majority of name instances are from PubMed (114,445,384 instances, 95.4%), followed by ORCID (2,922,037 instances, 2.4%). Note that an author may occur multiple times in the dataset if the author appears on more than one PubMed, ORCID, or NSF/NIH record. Table 1 presents a breakdown of the number of name instances and the number of unique authors by source.

Table 1. Distribution of Author-Mobi records across harvested sources

Source	Number of Name Instances	(%)	Number of Unique Authors	(%)
PubMed	114,445,384*	(95.4)	13,702,526**	(100)
ORCID	2,922,037	(2.4)	518,648	(3.8)
NIH grant	1,825,538	(1.5)	138,925	(1.0)
NSF grant	773,730	(0.6)	49,420	(0.4)
Total	119,966,689	(100)	Not Applicable	

*The number of instances is slightly larger than Author-ity 2018 (114.2 million) because Author-Mobi has some duplicate authors.
**The total number of unique authors is slightly smaller than Author-ity 2018 (13,722,485) because some authors have no affiliation information whatsoever.

The records range from the years 1786 to 2030. The dataset is based on a PubMed snapshot collected in December 2018, and very few records go beyond 2018 due to errors (e.g., ORCID typos or anticipated dates, prepublications, grant extensions, etc.). For example, grants' amendment dates may go beyond 2018 (e.g., NSF1333564), and ORCID records may list future dates such as expected graduation years. *ORCID ID: 0000-0002-1816-1721* shows that the person started their medical scientist training program in 2018 and anticipates completing the program in 2026.

As addressed in the Methods section, the Author-Mobi dataset extends MapAffil 2018. In MapAffil 2018, data is available only for first authors of articles published before 2014, except for data harvested beyond PubMed, including PubMed Central (PMC) and Microsoft Academic Graph (MAG). This limitation is addressed in Author-Mobi by inferring cities, institutions, and disciplines for all authors based on their linked-records. Of the 119,966,689 name instances in Author-Mobi, 44% (52,811,275 instances) originated from MapAffil 2018, while 56% (67,155,414 instances) were newly inferred. As shown in Figure 2, Author-Mobi substantially supplements data for instances at years prior to 2014: Between 1867 and 2013, the number of newly inferred name instances increased from 311 to 3,178,477 (left panel in Figure 2). However, the proportion inferred varied systematically by time periods: 1946-1987: 77-92% inference; 1988-2013: 58-70%; after 2013, it goes gradually down to 7% in 2018 (right panel in Figure 2).

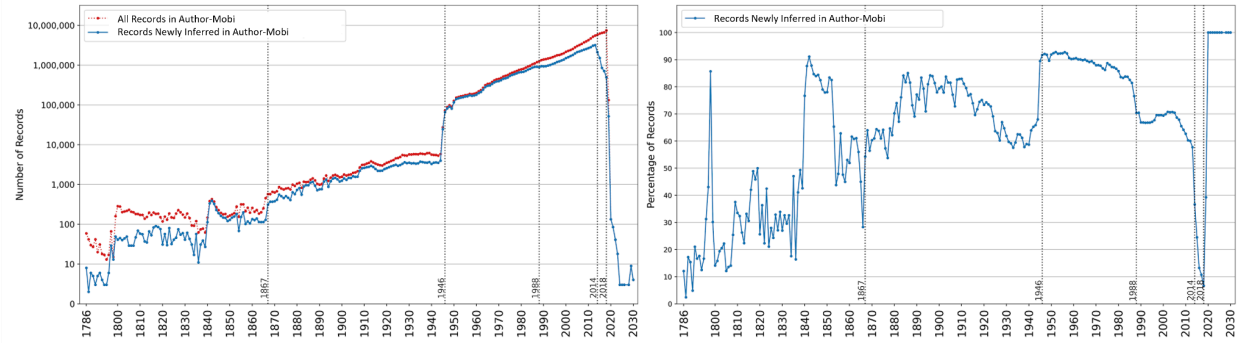


Figure 2. Number of Name Instances Newly Inferred in Author-Mobi (left) and Percentages of Name Instances Newly Inferred in Author-Mobi (right) over Time. PubMed provides access to MEDLINE and OLDMEDLINE [38], dating back to 1966 and 1949, respectively [39, 40]. 1988 is the year that the US National Library of Medicine (NLM) started consistently recording first-author affiliations. 2014 is the year that NLM started recording affiliations for all authors when available. 2018 is the data collection year. Records with years beyond 2018 are due to preprints and future dates listed in grants and ORCID records.

Technical Validation

Nobel laureate-based silver standard construction

For external validation, we constructed an LLM-assisted silver-standard dataset from publicly available career histories of Nobel laureates in Physiology or Medicine. We selected this Nobel category because Author-Mobi is primarily derived from PubMed records. Nobel laureates constitute a selective population and are not representative of the broader PubMed author population, particularly with respect to scientific prominence, productivity, and career length. Nevertheless, their education and employment histories are comparatively well documented in public sources, making it possible to construct a relatively large and systematically assembled longitudinal benchmark rather than relying on a small number of manually curated case studies. We identified 232 laureates through 2025, of whom 192 (82.8%) were ultimately linked to at least one usable Author-ity 2018 author clusters (AUIDs) and included in the silver standard benchmark.

The initial laureate list and corresponding Wikipedia article URLs were retrieved from Wikidata using the Wikidata Query Service (query.wikidata.org) with a SPARQL query for recipients of the Nobel Prize in Physiology or Medicine. Laureates were then linked to candidate Author-ity 2018 AUIDs. Candidate generation primarily used PMIDs extracted from each laureate's Wikipedia article, which were matched to PMID-author records in Author-ity 2018. When PMID overlap did not yield a usable candidate, additional candidates were identified from Author-ity name variants using surname, given-name or first-initial, middle-initial, and publication year range consistency. Candidate AUIDs with no usable Author-Mobi institution, city, country, or discipline information were excluded. This filtering stage was based solely on whether Author-Mobi contained usable affiliation information and not on agreement with the subsequently constructed silver standard. The remaining candidates were manually reviewed for identity consistency. Multiple AUIDs were retained when more than one AUID could be confidently attributed to the same laureate. The final linkage contained 195 AUIDs for 192 laureates. Most laureates (190/192; 99.0%) were represented by a single AUID, whereas one laureate (0.5%) were linked to two

AUIDs and another (0.5%) to three AUIDs, reflecting cases in which a laureate's publications were split across multiple valid Author-ity clusters.

Public biographical text was collected independently of Author-Mobi predictions from each laureate's Wikipedia article. Each Wikipedia URL was converted to a page title, and the rendered article content was retrieved through the MediaWiki API. The introductory paragraphs were retained together with sections whose headings included *biography, career, life, research,* or *education.* Citation markers and editing elements were removed before the selected text was combined into a single biographical input for each laureate. Relevant biographical text was successfully extracted for all 192 linked laureates.

The biographical text was transformed into structured affiliation histories using GPT-5.4. For each laureate, the model produced a chronological JSON timeline containing institution, location, position or role, employment or education category, primary or secondary affiliation status, discipline, start and end years, and supporting evidence. The supplied Wikipedia text served as the baseline source. GPT-5.4's web search feature was used to verify or complete affiliation details and was restricted to nobelprize.org, britannica.com, encyclopedia.com, and pubmed.ncbi.nlm.nih.gov. Web search was required for each generation request, with at most three search calls, and generation used a temperature of 0 to minimize stochastic variation and discourage creative generation, as the task required conservative extraction of factual affiliation information. Returned source URLs were retained with the extracted timeline records to preserve provenance. The extraction prompt restricted employment records to clearly supported research-based appointments, including explicitly identified postdoctoral positions. Doctoral and professional medical education was retained, whereas undergraduate and master's-level education was excluded. The model was instructed to assign at most one primary research employment affiliation at a given time; concurrent substantive research appointments could be represented as secondary affiliations. Years were populated only when they were explicitly supported by the supplied text or an allowed web source, and unresolved or unsupported values were left missing rather than inferred. Locations and disciplines were similarly recorded only when supported by the source evidence. Table 2 provides an example of the structured silver-standard timeline for Victor Ambros, who is also used as the illustrative author in Figure 1. The final LLM-assisted extraction produced 890 primary research employment records across 193 laureates. Among these records, institution information was available for 889 (99.9%), geographic location (city and country) for 676 (76.0%), and discipline for 579 (65.1%). Both start and end years were available for 625 records (70.2%), while 230 (25.8%) contained only a start year, 6 (0.7%) contained only an end year, and 29 (3.3%) contained neither.

Table 2. Example of an LLM-assisted silver standard career history. Selected fields from the structured timeline generated for Victor Ambros. Missing values ("-") reflect information not explicitly supported by the source material at the extraction stage. These affiliation records were subsequently used to construct the silver standard histories for comparison with Author-Mobi; institution and location information was normalized using MapAffil.

Start	End	Institution	Location	Discipline	Category	Source
-	1979	Massachusetts Institute of Technology	-	biology	Education	Wikipedia
1979	1985	Massachusetts Institute of Technology	Cambridge, MA, USA	-	Employment	Mixed
1985	1992	Harvard University	Cambridge, MA, USA	-	Employment	Web
1992	2007	Dartmouth Medical School	-	-	Employment	Mixed
2008	-	University of Massachusetts Medical School	Worcester, MA, USA	Molecular medicine	Employment	Mixed

For comparison with Author-Mobi, the extracted affiliation intervals were converted to AUID-year histories. Explicit start-end intervals were treated as inclusive. When only a primary-employment start year was available, the affiliation was propagated through the year preceding the next known primary employment start; if no subsequent start was available, it was propagated through the last Author-Mobi year observed for that AUID. Start-only secondary employment and education records were assigned only to their reported start year, while end-only records were assigned only to their reported end year. Entries containing neither a start nor an end year were not assigned to any year. Primary-employment, secondary-employment, and education histories were retained separately. The extracted institution and location information was then normalized using MapAffil and mapped back to the corresponding AUID-year histories for comparison with Author-Mobi. Discipline was evaluated separately using the discipline field extracted directly from the LLM-assisted silver-standard records, because the affiliation strings used for MapAffil normalization frequently lacked department- or field-level information, which is needed for discipline parsing. The resulting evaluation dataset comprised 195 AUIDs representing 192 laureates and 7,492 unique AUID-years. Of these, 6,820 AUID-years (91.0%) contained a primary silver-standard affiliation.

Dimension-specific affiliation coverage in Author-Mobi and the silver standard across the 7,492 benchmark AUID-years was also examined (Table 3). Author-Mobi shows higher coverage across all four dimensions, while the silver standard showed lower coverage, particularly for discipline (65.76% vs. 98.99%) and city (83.84% vs. 97.90%). These estimates describe the availability of affiliation information in each source rather than its accuracy or agreement with the other source.

Table 3. Affiliation information coverage within the Nobel laureate-based silver standard

Dimension	Author-Mobi Coverage	Silver Standard Coverage
Institution	86.25%	75.31%
City	97.90%	83.84%
Country	99.75%	89.90%
Discipline	98.99%	65.76%

Note that percentages use 7,492 benchmark AUID-years as the denominator.

Affiliation agreement with the silver standard

We evaluated Author-Mobi against year-matched primary affiliations in the silver standard at the AUID-year level. For each affiliation dimension, an AUID-year was classified as a match, mismatch, or non-comparable. Agreement was calculated only among comparable AUID-years:

$$Agreement = \frac{matches}{matches + mismatches}$$

AUID-years with missing or insufficiently resolved information were generally treated as non-comparable rather than as errors. We also report the number of evaluable AUID-years and their proportion of the full benchmark (i.e., evaluable overlap) to indicate the sample underlying each agreement estimate.

Institution agreement required an exact match between normalized institution identifiers (e.g., GRID). Country agreement required exact normalized country equality. City agreement additionally allowed equivalent geographic hierarchies, such as cases where one source contained a more specific district-level representation of the same city (e.g., Manhattan, New York, NY, USA). Insufficient geographic resolution was treated as non-comparable unless the available state- or country-level information directly contradicted the other source, in which case the comparison was scored as a mismatch. Discipline was compared directly with the discipline field extracted by the LLM-assisted silver standard using exact normalized text matching. Because discipline labels can differ substantially in terminology and granularity, discipline was treated as a diagnostic comparison rather than as a primary basis for mobility validation. Overall AUID-year agreement and comparability coverage across the four affiliation dimensions are summarized in Table 4.

Table 4. AUID-year-level agreement between Author-Mobi and the silver standard.

Dimension	Comparable AUID-years	Evaluable Overlap	Matches	Mismatches	Agreement (95% CI)
Institution	5,311	70.89%	4,544	767	85.56% [81.69-89.12%]
City	6,182	82.51%	5,280	902	85.41% [81.95-88.70%]
Country	6,719	89.68%	6,398	321	95.22% [93.08-97.03%]
Discipline	4,865	64.94%	1,901	2,964	39.08% [32.73-44.88%]

Total AUID-year is 7,492; Agreement is calculated among comparable AUID-years only; evaluable overlap is the proportion of all AUID-years that were comparable. Related but nonidentical disciplines (e.g., “neurology” and “neurobiology”) were counted as mismatches. Confidence intervals are 95% laureate-clustered bootstrap intervals for agreement.

Across 7,492 AUID-years, institution and city information were comparable for 70.89% and 82.51% of AUID-years, respectively, with agreement rates of 85.56% and 85.41% (Table 4). Country information exhibited higher comparability coverage (89.68%) and agreement (95.22%), indicating substantial consistency between Author-Mobi and the silver standard at the country level. These results provide evidence that Author-Mobi captures institutional and geographic information consistent with independently built biographical records for the evaluable benchmark cases. In contrast, discipline agreement was 39.08%. This lower agreement should be interpreted cautiously because discipline was evaluated through strict normalized-label matching, which treats related but nonidentical terms (e.g., “neurology” and “neurobiology”) as mismatches. The observed discrepancies may therefore reflect both

differences in disciplinary assignments and differences in terminology or granularity between the two sources.

To examine agreement across individual longitudinal affiliation histories, we additionally calculated agreement separately for each AUID using the same year-level matching criteria. For each AUID, agreement was calculated as the proportion of its comparable years classified as matches. AUIDs without comparable years for a given affiliation dimension were excluded. We then calculated macro mean agreement across evaluable AUIDs:

$$MacroAgreement_d = \frac{1}{N_d} \sum_{i=1}^{N_d} Agreement_{i,d}$$

where N_d denotes the number of evaluable AUIDs for dimension *d*. AUIDs without comparable years are excluded. This approach assigns equal weight to each AUID, regardless of its publication history length. We additionally report the median and interquartile range (IQR) to characterize variation across individual (i.e., AUID) histories. Confidence intervals for the macro mean were estimated using 1,000 laureate-clustered bootstrap resamples.

Table 5. AUID-level agreement between Author-Mobi and the silver standard.

Dimension	Comparable AUID	Median Agreement (IQR)	Macro Mean Agreement 95% CI
Institution	167	93.10% (83.10-97.89%)	83.05% [79.15-86.88%]
City	180	93.33% (82.30-98.30%)	83.64% [80.11-87.29%]
Country	182	100.00% (95.65-100%)	95.23% [93.28-96.91%]
Discipline	168	30.99% (0.00-78.87%)	39.85% [33.84-45.89%]

At the AUID-level, macro mean agreement was 83.05% for institution, 83.64% for city, and 95.23% for country (Table 5), broadly consistent with the pooled AUID-year results. Median agreement was higher at 93.10%, 93.33%, and 100.00%, respectively, indicating that many individual affiliation histories exhibited high agreement with the silver standard, while lower-agreement histories reduced the macro means. In contrast, discipline exhibited substantially lower macro mean agreement (39.85%) and median agreement (30.99%), with a wide interquartile range (0.00-78.87%), indicating considerable variation across individual histories. This suggests that the discipline dimension requires further refinement and validation compared to the other dimensions, although disagreement may partly reflect differences in terminology or granularity under strict label matching. Restricting the analysis to laureates represented by a single AUID yielded similar macro mean estimates for institution (84.14%), city (84.55%), country (95.61%), and discipline (39.87%) (data not shown).

We further examined the temporal consistency of affiliation agreement across five-year calendar intervals (Figure 3). Agreement was calculated separately for institution, city, and country within each interval. Country agreement remained consistently high, from 90.97% to 100% across the displayed intervals. Institution and city, on the other hand, showed greater variation, particularly in earlier periods, followed by generally higher agreement in later decades. For example, in 1930-1934, institution and city agreement

were 76.47% and 67.57%, respectively. By 2000-2004, these values had increased to 87.83% and 91.92%. In the most recent interval, 2015-2018, institution agreement remained high at 90.00%, while city agreement was 85.12%. To further assess the reconstructed histories during periods of substantial affiliation coverage expansion, we calculated agreement separately for 1946-1987 and 1988-2013. During 1946-1987, agreement was 83.70% for institution, 83.67% for city, and 93.54% for country. The corresponding values for 1988-2013 were 89.15% for institution, 90.15% for city, and 97.37% for country. These results show how closely the inferred affiliations matched the silver standard during the periods when the temporal model added the most affiliation information to Author-Mobi.

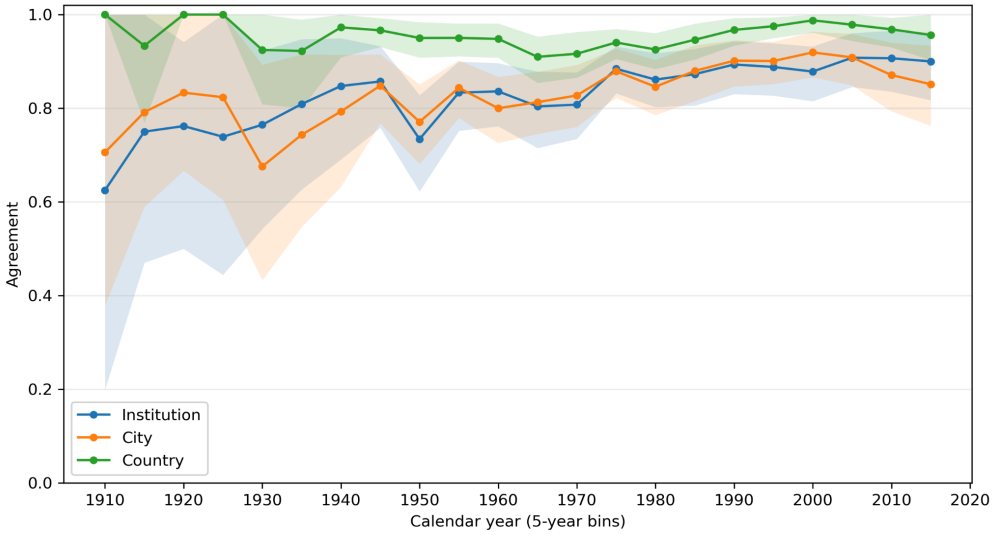


Figure 3. Temporal agreement between Author-Mobi and the silver standard. Agreement for institution, city, and country across five-year calendar intervals. Agreement represents the proportion of evaluable AUID-years classified as matches. Shaded bands represent 95% bootstrap confidence intervals. Only intervals containing at least 20 evaluable AUID-years for each of the three dimensions are shown.

Discussion

We developed Author-Mobi to provide longitudinal affiliation histories for approximately 13.7 million disambiguated researchers by integrating MapAffil 2018, Author-ity 2018, and linked NSF/NIH grants and ORCID records. The temporal model provides newly inferred affiliation information for 56% of author-name instances, substantially expanding the availability of historical affiliation information. External validation against an LLM-assisted silver standard showed high agreement for institutional and geographic affiliations, including the periods when the temporal model contributed the most significantly to our dataset. These findings support the value of Author-Mobi for studying scientific career and researcher mobility.

By providing temporally ordered affiliation histories, Author-Mobi supports broader investigations into scientific careers and researcher mobility. Previous research has highlighted the importance of temporal dependencies in researchers' mobility trajectories and institutional hierarchies in shaping scientific careers [41, 42]. Linking Author-Mobi with complementary demographic information could enable investigations into differences in mobility, career progression, and collaboration across ethnicity, gender, and age [31,

43-48]. These longitudinal trajectories could also support investigations into how institutional mobility and changing collaboration patterns relate to scientific novelty, disruption, and impact [49-54]. For example, Bhattacharya et al. [55] used disambiguated MEDLINE publication histories of Nobel laureates in Physiology or Medicine, which is the same Nobel category used in our external validation, to study changes in productivity, impact, and novelty across career stages. Author-Mobi could extend this type of analysis by incorporating longitudinal institutional and geographic affiliation information. Integration with citation links, citation contexts, and automated IMRaD section recognition could further support studies of how researchers' citation practices and use of prior knowledge evolve across career stages and following institutional or geographic mobility, providing insights into knowledge diffusion and scientific impact [56-58]. Linking longitudinal affiliation histories with CRediT author-contribution statements could facilitate investigations into the division of scientific labor across research teams, the allocation of scientific credit, and how individual contribution roles evolve over scientific careers [59-62]. Furthermore, longitudinal coauthorship and affiliation histories could support studies of mentor-mentee relationships and academic genealogy [63, 64]. Together, these applications extend Author-Mobi's potential beyond individual mobility events to broader investigations of scientific careers and research communities.

There are also places that require improvements for the proposed dataset. Potential errors in author disambiguation, affiliation normalization, and linked source records may propagate into the inferred affiliation histories. Also, our external evaluation relies on an LLM-assisted silver standard rather than a manually verified gold standard, and Nobel laureates represent a selective population that may not reflect the broader PubMed author population. Furthermore, agreement estimates are based on evaluable AUID-years and measure reconstructed affiliation states rather than directly validating the identification or timing of mobility transitions. Regarding the low discipline agreement, this may partly reflect differences in terminology and granularity under strict label matching, highlighting the need for further refinement and validation.

For the next step, we are currently harvesting additional affiliation records from OpenAlex to expand the available evidence for adding and refining authors' affiliation histories. We are also developing a more direct evaluation of researcher mobility, which requires assessing affiliation transitions and their timing beyond year-level affiliation agreements. Future work will include manual validation of the LLM-assisted silver standard and further refinement of the discipline dimensions. These efforts aim to improve the completeness and reliability of Author-Mobi and strengthen its utility for longitudinal studies of scientific careers and researcher mobility. Author-Mobi is freely available in the Illinois Data Bank[6].

Acknowledgments

Research reported in this publication was supported by the DATA-CORE of the National Bureau of Economic Research (NBER) project "Innovation in an Aging Society" funded by the National Institutes of Health (Award Number P01AG039347). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

[6] https://databank.illinois.edu/datasets/IDB-4273770?code=FNvLExGBwqyGyCxnFdr9lSbAEIXEBjN5I0Zc9QhZ_f4

References

[1] Marx, M., Strumsky, D. & Fleming, L. Mobility, skills, and the Michigan non-compete experiment. *Management Science* 55, 875–889 (2009). https://doi.org/10.1287/mnsc.1080.0985

[2] Kaiser, U., Kongsted, H. C. & Rønde, T. Does the mobility of R&D labor increase innovation? *Journal of Economic Behavior & Organization* 110, 91–105 (2015). https://doi.org/10.1016/j.jebo.2014.12.012

[3] Braunerhjelm, P., Ding, D. & Thulin, P. Labour market mobility, knowledge diffusion and innovation. *European Economic Review* 123, 103386 (2020). https://doi.org/10.1016/j.euroecorev.2020.103386

[4] Choudhury, P. Geographic mobility, immobility, and geographic flexibility: a review and agenda for research on the changing geography of work. *Academy of Management Annals* 16, 258–296 (2022). https://doi.org/10.5465/annals.2020.0242

[5] Faggian, A., Rajbhandari, I. & Dotzel, K. R. The interregional migration of human capital and its regional consequences: a review. *Regional Studies* 51, 128–143 (2017). https://doi.org/10.1080/00343404.2016.1263388

[6] Bozeman, B., Dietz, J. S. & Gaughan, M. Scientific and technical human capital: an alternative model for research evaluation. *International Journal of Technology Management* 22, 716–740 (2001). https://doi.org/10.1504/IJTM.2001.002988

[7] Edler, J., Fier, H. & Grimpe, C. International scientist mobility and the locus of knowledge and technology transfer. *Research Policy* 40, 791–805 (2011). https://doi.org/10.1016/j.respol.2011.03.003

[8] Jaffe, A. B. Real effects of academic research. *American Economic Review* 79, 957–970 (1989). http://www.jstor.org/stable/1831431

[9] Narin, F., Hamilton, K. S. & Olivastro, D. The increasing linkage between U.S. technology and public science. *Research Policy* 26, 317–330 (1997). https://doi.org/10.1016/S0048-7333(97)00013-9

[10] Cohen, W. M., Nelson, R. R. & Walsh, J. P. Links and impacts: the influence of public research on industrial R&D. *Management Science* 48, 1–23 (2002). https://doi.org/10.1287/mnsc.48.1.1.14273

[11] Jaffe, A. B., Shupp, L. B. & Tartari, V. Real effects of academic research revisited. *NBER Working Paper* No. 35017 https://doi.org/10.3386/w35017 (2026).

[12] Arora, A., Belenzon, S., Cioaca, L. C., Sheer, L. & Zhang, H. The effect of public science on corporate R&D. *NBER Working Paper* No. 31899 https://doi.org/10.3386/w31899 (2023).

[13] Zellner, C. The economic effects of basic research: evidence for embodied knowledge transfer via scientists' migration. *Research Policy* 32, 1881–1895 (2003). https://doi.org/10.1016/S0048-7333(03)00080-5

[14] Petersen, A. M. Multiscale impact of researcher mobility. *Journal of the Royal Society Interface* 15, 20180580 (2018). https://doi.org/10.1098/rsif.2018.0580

[15] Netz, N., Hampel, S. & Aman, V. What effects does international mobility have on scientists' careers? A systematic review. *Research Evaluation* 29, 327–351 (2020). https://doi.org/10.1093/reseval/rvaa007

[16] Yan, E., Zhu, Y. & He, J. Analyzing academic mobility of U.S. professors based on ORCID data and the Carnegie Classification. *Quantitative Science Studies* 1, 1451–1467 (2020). https://doi.org/10.1162/qss_a_00088

[17] Verginer, L. & Riccaboni, M. Cities and countries in the global scientist mobility network. *Applied Network Science* 5, 38 (2020). https://doi.org/10.1007/s41109-020-00276-0

[18] Wei, Y., Zhang, X., Ho, A. T.-K. & Wang, X. Academics' mobility revisited: validating dimensions and unpacking inequalities. *Scientometrics* 131, 3485–3513 (2026). https://doi.org/10.1007/s11192-026-05641-x

[19] OpenAlex. Authors overview. *OpenAlex Help Center* https://help.openalex.org/data/authors/ (2026).

[20] Hajikhani, A. & Suominen, A. Dataset on researcher's mobility by utilising bibliometric data: case of Finland-based scholars for the duration of 2008–2018. *Data in Brief* 34, 106764 (2021). https://doi.org/10.1016/j.dib.2021.106764

[21] Dobre, S., Herbert, R., Ding, A. S. & Pohl, H. A bibliometric analysis using a newly developed model and a customizable research tool: a case study of researcher mobility in Sweden. *PLOS ONE* 19, 1–25 (2024). https://doi.org/10.1371/journal.pone.0308147

[22] Ding, A. & Dobre, S. Researcher mobility—customizable research tool. Mendeley Data V1 https://doi.org/10.17632/jmy6pt5zhj.1 (2024).

[23] Akbaritabar, A., Theile, T. & Zagheni, E. Bilateral flows and rates of international migration of scholars for 210 countries for the period 1998–2020. *Scientific Data* 11, 816 (2024). https://doi.org/10.1038/s41597-024-03655-9

[24] Akbaritabar, A., Theile, T. & Zagheni, E. Bilateral flows and rates of subnational and international migration of scholars worldwide by gender and field. *MPIDR Working Paper* WP 2026-029 https://doi.org/10.4054/MPIDR-WP-2026-029 (2026).

[25] Aksnes, D. W., Kahn, S., Reiling, R. B. & Ulvestad, M. E. S. Examining career trajectories of Norwegian PhD recipients: slower progression for women academics but not a leaky pipeline. *Studies in Higher Education* 50, 1797–1820 (2025). https://doi.org/10.1080/03075079.2024.2400545

[26] Azoulay, P., Fons-Rosen, C. & Graff Zivin, J. S. Does science advance one funeral at a time? *American Economic Review* 109, 2889–2920 (2019). https://doi.org/10.1257/aer.20161574

[27] Vinkenburg, C. J., Connolly, S., Fuchs, S., Herschberg, C. & Schels, B. Mapping career patterns in research: a sequence analysis of career histories of ERC applicants. *PLOS ONE* 15, 1–19 (2020). https://doi.org/10.1371/journal.pone.0236252

[28] Edwards, K. A., Acheson-Field, H., Rennane, S. & Zaber, M. A. Mapping scientists' career trajectories in the survey of doctorate recipients using three statistical methods. *Scientific Reports* 13, 8119 (2023). https://doi.org/10.1038/s41598-023-34809-1

[29] Yu, H., Marschke, G., Ross, M. B., Staudt, J. & Weinberg, B. A. Publish or perish. *Journal of Human Resources* 58, 1307–1346 (2023). https://doi.org/10.3368/jhr.59.2.1219-10630R1

[30] Berkes, E., Marion, M., Milojević, S. & Weinberg, B. A. Slow convergence: career impediments to interdisciplinary biomedical research. *Proceedings of the National Academy of Sciences* 121, e2402646121 (2024). https://doi.org/10.1073/pnas.2402646121

[31] Packalen, M. & Bhattacharya, J. Age and the trying out of new ideas. *Journal of Human Capital* 13, 341–373 (2019). https://doi.org/10.1086/703160

[32] Cheng, W. & Weinberg, B. A. Old and connected versus young and creative: networks and the diffusion of new scientific ideas. *NBER Working Paper* No. 33030 https://doi.org/10.3386/w33030 (2024).

[33] Zhao, X., Derrick, G. E., Sugimoto, C. R. & Larivière, V. Starting a PhD and becoming a parent: gendered life-course trajectories in academic careers. *Higher Education* (2026). https://doi.org/10.1007/s10734-026-01668-4

[34] National Library of Medicine. pubmed_250101.dtd: Affiliation element. *NLM DTD*. https://dtd.nlm.nih.gov/ncbi/pubmed/doc/out/250101/el-Affiliation.html

[35] National Library of Medicine. Author, corporate author, and collaborator affiliation display changes. *NLM Technical Bulletin* 395, e9 https://www.nlm.nih.gov/pubs/techbull/nd13/nd13_author_affiliation_display.html (2013).

[36] Torvik, V. I. MapAffil: a bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. *D-Lib Mag*. 21, 11/12 (2015). https://doi.org/10.1045/november2015-torvik

[37] Torvik, V. I. & Smalheiser, N. R. Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data* 3, 11 (2009). https://doi.org/10.1145/1552303.1552304

[38] National Center for Biotechnology Information. About PubMed. *PubMed* https://pubmed.ncbi.nlm.nih.gov/about/ (2025).

[39] National Library of Medicine. About MEDLINE. *MEDLINE* https://www.nlm.nih.gov/medline/medline_overview.html (2025).

[40] National Library of Medicine. OLDMEDLINE data. *NLM Databases* https://www.nlm.nih.gov/databases/databases_oldmedline.html (2024).

[41] Vaccario, G., Verginer, L. & Schweitzer, F. The mobility network of scientists: analyzing temporal correlations in scientific careers. *Applied Network Science* 5, 36 (2020). https://doi.org/10.1007/s41109-020-00279-x

[42] Clauset, A., Arbesman, S. & Larremore, D. B. Systematic inequality and hierarchy in faculty hiring networks. Science Advances 1, e1400005 (2015). https://doi.org/10.1126/sciadv.1400005

[43] Torvik, V. I. & Agarwal, S. Ethnea—an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. Presented at the *International Symposium on Science of Science* (Washington, DC, 2016). https://hdl.handle.net/2142/88927

[44] Smith, B. N., Singh, M. & Torvik, V. I. A search engine approach to estimating temporal changes in gender orientation of first names. In Proc. 13th *ACM/IEEE-CS Joint Conf. on Digital Libraries* (JCDL '13) 199–208 (ACM, 2013). https://doi.org/10.1145/2467696.2467720

[45] Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. Bibliometrics: global gender disparities in science. *Nature* 504, 211–213 (2013). https://doi.org/10.1038/504211a

[46] Way, S. F., Larremore, D. B. & Clauset, A. Gender, productivity, and prestige in computer science faculty hiring networks. In Proc. 25th *International Conference on World Wide Web* (WWW '16) 1169–1179 (International World Wide Web Conferences Steering Committee, 2016). https://doi.org/10.1145/2872427.2883073

[47] Mishra, A., Lee, H., Jeoung, S., Torvik, V. I. & Diesner, J. Patterns of diversity in biomedical coauthorships: an analysis across authors' ethnicity, gender, age, and expertise. PLOS ONE 20, 1–22 (2025). https://doi.org/10.1371/journal.pone.0316890

[48] Weinberg, B. A. & Galenson, D. W. Creative careers: the life cycles of Nobel laureates in economics. *NBER Working Paper* No. 11799 https://doi.org/10.3386/w11799 (2005).

[49] Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. Atypical combinations and scientific impact. *Science* 342, 468–472 (2013). https://doi.org/10.1126/science.1240474

[50] Mishra, S. & Torvik, V. I. Quantifying conceptual novelty in the biomedical literature. *D-Lib Magazine* 22, 9/10 (2016). https://doi.org/10.1045/september2016-mishra

[51] Funk, R. J. & Owen-Smith, J. A dynamic network measure of technological change. *Management Science* 63, 791–817 (2017). https://doi.org/10.1287/mnsc.2015.2366

[52] Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* 566, 378–382 (2019). https://doi.org/10.1038/s41586-019-0941-9

[53] Lin, Y., Evans, J. A. & Wu, L. New directions in science emerge from disconnection and discord. *Journal of Informetrics* 16, 101234 (2022). https://doi.org/10.1016/j.joi.2021.101234

[54] Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive over time. *Nature* 613, 138–144 (2023). https://doi.org/10.1038/s41586-022-05543-x

[55] Bhattacharya, J., Bollyky, P., Goldhaber-Fiebert, J. D., Holom, G. H., Packalen, M. & Studdert, D. M. Resting on their laureates? Research productivity among winners of the Nobel Prize in Physiology or Medicine. *NBER Working Paper* No. 31352 https://doi.org/10.3386/w31352 (2023).

[56] Fang, L., Salami, M. O., Weber, G. M. & Torvik, V. I. uCite: the union of nine large-scale public PubMed citation datasets with reliability filtering. *Data in Brief* 60, 111535 (2025). https://doi.org/10.1016/j.dib.2025.111535

[57] Hsiao, T.-K. & Torvik, V. I. OpCitance: citation contexts identified from the PubMed Central open access articles. *Scientific Data* 10, 243 (2023). https://doi.org/10.1038/s41597-023-02134-x

[58] Nam, D., Hsiao, T.-K. & Torvik, V. I. Combining document structure and language model for IMRaD classification. Poster presented at the *International Conference of Science of Science and Innovation* (ICSSI) 2026 (29 June 2026).

[59] Larivière, V., Pontille, D. & Sugimoto, C. R. Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT). *Quantitative Science Studies* 2, 111–128 (2021). https://doi.org/10.1162/qss_a_00097

[60] Lu, C., Zhang, Y., Ahn, Y.-Y., Ding, Y., Zhang, C. & Ma, D. Co-contributorship network and division of labor in individual scientific collaborations. *Journal of the Association for Information Science and Technology* 71, 1162–1178 (2020). https://doi.org/10.1002/asi.24321

[61] Shen, H.-W. & Barabási, A.-L. Collective credit allocation in science. *Proceedings of the National Academy of Sciences* 111, 12325–12330 (2014). https://doi.org/10.1073/pnas.1401992111

[62] Malhotra, D., Jacobs, K. A. Jr & Torvik, V. I. Classifying author contribution statements using sentence embeddings. Poster presented at the *International Conference of Science of Science and Innovation (ICSSI) 2026* (30 June 2026).

[63] Salami, M. O., Fang, L. & Torvik, V. I. Can LLMs identify advisor-advisee relationships in coauthorship network based on publication metadata? In Proc. 2026 *ACM/IEEE Joint Conf. on Digital Libraries* (JCDL '26) (ACM, 2026, To appear Oct. 13-16). https://doi.org/10.1145/3805696.3846065

[64] Cugmas, M., Mali, F. & Kronegger, L. Longitudinal patterns of scientific collaboration in doctoral studies. *Scientometrics* 129, 1055–1077 (2024). https://doi.org/10.1007/s11192-023-04900-5