scientific data

Article in Press

Check for updates

OPEN

DATA DESCRIPTOR

A U.S. inventor–patent linkage dataset for National Inventors Hall of Fame inductees

Jianming Guo[1,2], Mengting Gao[1,2], Yue Yang[1,2], Haotian Hu[2,3], Jia Kong[2,4], Alex Jie Yang[1,2] ✉ & Sanhong Deng[1,2] ✉

Elite inventors are important subjects for studying technological change, inventor careers, and innovation networks. However, honorary institutions usually list only representative inventions or seed patents, and general-purpose patent databases do not directly provide audited inventor-level portfolios for externally defined recognition-based inventor groups. Here, we present a U.S. inventor–patent linkage dataset for inductees of the National Inventors Hall of Fame (NIHF). Starting from official NIHF records and U.S. induction-related seed patents, we construct an NIHF–PatentsView public baseline and combine it with broad retrieval of U.S. patent candidates from InCoPat, record-level evidence variables, PatentSBERTa-based semantic clustering, and manual adjudication. The final curated release contains 28,330 inventor–patent links representing 27,155 unique U.S. patent identifiers and covers 638 NIHF inductees. Compared with a public baseline based only on NIHF seed patents and PatentsView inventor identifiers, the dataset retains most baseline links while substantially expanding inventor and patent coverage. The dataset supports studies of NIHF inventors' U.S. patent portfolios, patenting careers, technology trajectories, collaboration networks, and innovation activities, and can be linked to external patent metadata through standardized patent identifiers.

Background & Summary

Patent data are a central empirical foundation for studying technological change, knowledge diffusion, inventor careers, collaboration networks, and innovation impact[1,2]. With the development of large-scale patent databases and inventor-disambiguation resources, including the USPTO, PatentsView, and Lens, researchers can now systematically access patent-level metadata, inventor–patent links, technology classifications, citation relations, and applicant or assignee information[3–5]. These resources have substantially improved the reusability of patent-based data for innovation studies. However, for inventor groups defined by external recognition systems, general-purpose patent databases do not usually provide complete and audited group-level patent portfolios. Researchers still need to connect individuals, representative inventions, and inventor entities across honorary records and patent databases, while handling name ambiguity, historical coverage limits, database boundaries, and mismatched links[6–9].

This study focuses on inductees of the National Inventors Hall of Fame (NIHF). We treat NIHF induction as an operational definition of a recognition-based, institutionally defined elite inventor population[10,11]. This definition differs from identifying elite inventors by patent counts, forward citations, highly cited patents, corporate positions, or other innovation awards[12–15]. NIHF induction reflects technological contribution and public recognition as assessed by an honorary institution. We do not treat NIHF inductees as an unbiased sample of all important inventors, nor do we claim that NIHF recognition replaces other definitions of elite inventors. Rather, the value of NIHF lies in its stable inductee list, induction years, technology descriptions, and representative seed patents, which make it a traceable population for studying recognition-based elite inventors.

Official NIHF records are an important starting point for constructing patent data for this population. NIHF pages typically list inductee names, induction years, descriptions of technological contributions, and seed patents associated with representative inventions[11]. These seed patents are valuable for identity confirmation

[1]School of Information Management, Nanjing University, Nanjing, China. [2]Jiangsu International Joint Informatics Laboratory, Nanjing University, Nanjing, China. [3]Jiangsu Academy of Agricultural Sciences, Nanjing, China. [4]School of Information, Guizhou University of Finance and Economics, Guiyang, China. ✉e-mail: alexjieyang@outlook.com; sanhong@nju.edu.cn

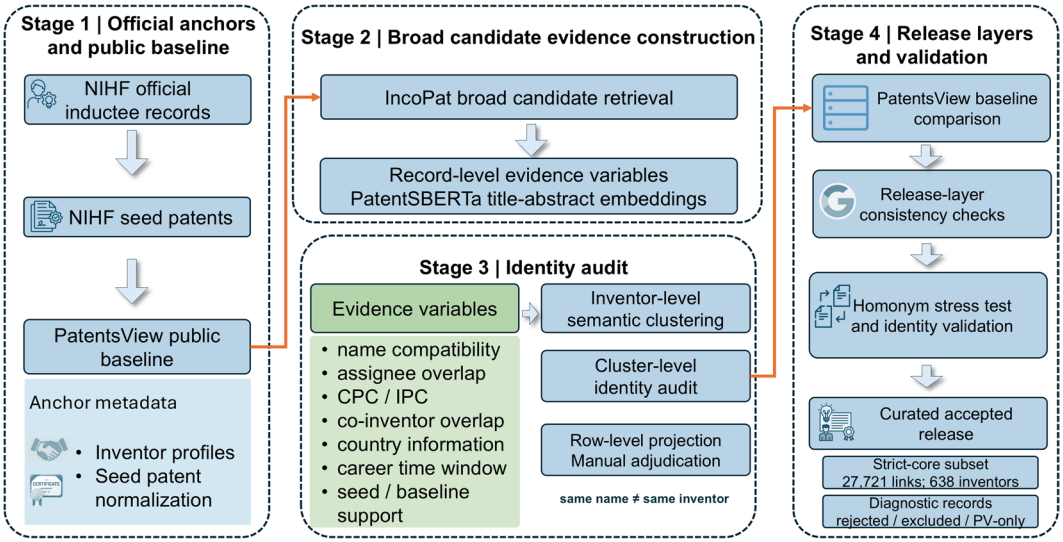


Fig. 1 Overview of the NIHF inventor–patent linkage construction workflow. NIHF records and seed patents define the target population and official identity anchors. The NIHF–PatentsView public baseline provides an external public benchmark. Broad InCoPat retrieval, record-level evidence variables, and PatentSBERTa patent-text embeddings support inventor-level semantic clustering, cluster-level identity audit, row-level adjudication, and final release assembly. The final outputs are the curated accepted release, the strict-core subset, and diagnostic records.

because they provide person–patent anchors explicitly associated with inductees by NIHF. However, NIHF records are not designed to provide a complete patent portfolio for each inductee. For many inductees, seed patents represent selected inventions rather than all identifiable patenting activity. Therefore, if the research goal is to reconstruct NIHF inventors' patent portfolios, career spans, technology trajectories, or collaboration networks, official NIHF records alone are insufficient. Award-linked patents, including patents associated with Nobel Prizes and the National Inventors Hall of Fame, have been used in prior patent-text research; however, such work focuses on representative award-linked patents rather than audited inventor-level patent portfolios[16].

A transparent and reproducible alternative is to map NIHF seed patents to PatentsView, identify the corresponding disambiguated inventor identifiers, and then retrieve all patents linked to those inventor identifiers. We explicitly construct this NIHF–PatentsView public baseline to simulate what researchers could obtain using only NIHF seed patents and PatentsView inventor identifiers. This baseline is valuable because it is based entirely on public data resources, is reproducible, and provides a benchmark for evaluating the added coverage of the present dataset. The resulting NIHF–PatentsView baseline covers 167 NIHF inventors and 8,665 inventor–patent links.

The NIHF–PatentsView baseline also has clear boundaries. First, it depends on whether seed patents can be accurately mapped to PatentsView records; if seed patents are missing, formatted inconsistently, or outside the mappable scope, the subsequent inventor-level expansion is limited. Second, PatentsView inventor identifiers are highly valuable for research on modern U.S. granted patents, but their coverage scope, historical boundaries, and data organization affect portfolio identification for some early or cross-period NIHF inductees. Third, inferring the target inventor's PatentsView identity from seed patents depends on the accuracy of the seed-to-inventor mapping. For these reasons, we treat the PatentsView baseline as a public, reproducible benchmark and recall-oriented diagnostic source, not as a gold standard for the complete NIHF patent portfolio.

Comparable honor-based career-output datasets have been constructed for Nobel laureates in the natural sciences, illustrating the research value of standardized curated datasets for officially recognized elite populations[17]. To address the limitations described above, we construct a curated inventor–patent linkage dataset for NIHF inductees. The goal is to provide an auditable, reusable, and clearly bounded linkage resource for an externally defined honorary inventor population. Specifically, we use official NIHF records and seed patents as identity anchors, construct the NIHF–PatentsView public baseline as a public reference, and internally use InCoPat for broad candidate retrieval to improve the recall of potentially relevant patent records. Candidate records are then organized and audited using name compatibility, assignee/applicant bridges, co-inventor bridges, technology support, career-window plausibility, seed/baseline support, and patent-text semantic evidence.

The construction process pays particular attention to homonyms and boundary cases. Candidate records are not admitted into the final data layer solely because they match an inventor name or exceed a composite-score threshold. Instead, we use record-level evidence variables and PatentSBERTa patent-text embeddings to organize inventor-level semantic clusters, and we evaluate whether candidate clusters correspond to the target NIHF inventor by examining applicants or assignees, co-inventors, technology topics, timelines, and official seed anchors. Records with conflicting, insufficient, or benchmark-discordant evidence are retained in a review-required queue and manually adjudicated. The overall construction workflow is shown in Fig. 1. The final released dataset is therefore organized around accepted inventor–patent links.

The final curated accepted dataset contains 28,330 inventor–patent links and 27,155 unique U.S. patents, covering 638 NIHF inductees. A strict-core subset containing 27,721 links is provided for analyses requiring more conservative identity precision. Rejected, excluded, and benchmark-discordant records are retained separately for methodological review, error analysis, and baseline comparison.

The dataset can be used to study NIHF inventors' patent portfolios, patenting-career spans, technology trajectories, seed-to-expanded portfolio differences, and collaboration network structures. Because the public release provides standardized person_key, inventor_id, and patent_id_clean fields, users can link the dataset to external patent metadata to add citations, applicants or assignees, technology classifications, full texts, legal-status information, patent-value measures, or other patent-level variables. Because of third-party licensing restrictions, the repository provides redistributable linkage files, baseline-comparison files, diagnostic records, derived evidence indicators, data dictionaries, and protocol documentation, but not the restricted InCoPat source fields. This release strategy is intended to balance data reuse, methodological transparency, and data-use boundaries.

Users should also consider the interpretive boundaries of the target population. NIHF induction reflects historical, institutional, geographic, organizational, and technology-field selection mechanisms. This dataset should therefore be understood as a specialized resource for a recognition-based elite inventor population, not as a representative sample of the general inventor population. Studies of average inventor behavior, general innovation ecosystems, or unbiased inventor distributions should combine this dataset with larger general-purpose patent databases. For research on recognized technological contributors, honorary recognition, inventor careers, and elite inventor innovation trajectories, the dataset provides a curated and audited linkage layer.

Methods

The data-construction workflow consists of six main stages: defining the NIHF target inventor population and curating official seed-patent anchors; constructing the NIHF–PatentsView public baseline; expanding the candidate patent set through broad InCoPat retrieval; generating record-level evidence variables and patent-text semantic representations; conducting inventor-level identity audit and row-level adjudication; and assembling the curated accepted release, the strict-core subset, and diagnostic records. The workflow was designed to produce an auditable and reusable inventor–patent linkage resource for an externally defined honorary inventor population.

NIHF target population and official seed anchors. We used official records from the National Inventors Hall of Fame (NIHF) as the source for the target population. The NIHF inductee list was accessed in March 2026. From the official records, we curated inductee names, induction years, NIHF technology domains, birth years, death years when available, and seed patents associated with induction-related inventions. The final public target table contains 653 NIHF inductees. Each inductee was assigned a stable person_key, and the public release also provides an inventor_id as the release-level inventor identifier.

NIHF seed patents are patents explicitly associated with inductees' representative inventions in the official NIHF records. We treated these seed patents as official identity anchors. When multiple seed patents were associated with the same inductee, all identifiable patent numbers were retained and the number of seed patents was recorded at the inventor level. A small number of inductees lacked usable seed patents or had seed patents that could not be reliably mapped; these inductees remain in the target inventor table but do not necessarily appear in the final curated accepted release.

Of the 653 official NIHF inductees, 638 are represented in the final curated accepted dataset. The remaining 15 uncovered inductees comprise 2 for whom neither usable seed patents nor candidate records were available and 13 whose retrieved candidates did not meet the final acceptance criteria.

Patent identifiers were standardized before cross-source linkage. Patent numbers may contain country prefixes, patent-type prefixes, kind codes, spaces, punctuation, leading zeros, or version suffixes. The public field patent_id_clean is therefore treated as a normalized string rather than a numeric variable. For special U.S. patent types, patent-type information retained in patent_id_original_public_if_safe is used to preserve prefix-aware patent identity. These identifiers and prefix-aware identity checks were used to connect NIHF seed anchors, PatentsView baseline records, candidate records, accepted links, and diagnostic records.

Public baseline from PatentsView. To clarify the added value of the dataset relative to existing public patent resources, we constructed an NIHF–PatentsView public baseline. This baseline simulates a transparent and reproducible alternative workflow: first mapping NIHF seed patents to PatentsView patent records; then identifying the corresponding PatentsView disambiguated inventor identifiers on those seed patents; and finally retrieving all patents associated with those inventor identifiers to obtain a PatentsView-based inventor–patent linkage set[2,9,18]. The PatentsView source used in this study was the metadata snapshot released in 2025 with data coverage through 31 December 2024.

The PatentsView baseline was constructed using patent_id_clean as the main matching field, with prefix-aware checks for special U.S. patent types. For seed patents that could be found in PatentsView, we extracted the associated disambiguated inventor identifiers and linked them to NIHF inductees using the seed-patent identity and inventor-name information. If multiple PatentsView inventor identifiers were mapped to the same NIHF inductee, patent records associated with these identifiers were pooled and deduplicated using person_key together with prefix-aware patent identity. The resulting baseline links were compared with the curated accepted release at the same person_key + patent_id_clean level and classified as intersection, our_only, or pv_only. intersection indicates links present in both the PatentsView baseline and the curated accepted release; our_only indicates links present only in the curated accepted release; and pv_only indicates links present in the PatentsView baseline but not retained in the curated accepted release.

We used the PatentsView baseline as a public reproducible benchmark and recall-oriented diagnostic source, not as a gold standard. This treatment is appropriate for two reasons. First, PatentsView inventor identifiers are valuable for research on modern U.S. granted patents, but their coverage scope, historical boundaries, and data organization affect portfolio identification for some early or cross-period NIHF inductees. Second, expanding from NIHF seed patents to PatentsView inventor identifiers depends on the mappability of seed patents and the accuracy of seed-to-inventor mapping.

Broad candidate retrieval from InCoPat. Official NIHF records and the PatentsView baseline are not sufficient to reconstruct NIHF inductees' patent portfolios on their own, mainly because of structural limits in coverage and construction logic. PatentsView publicly covers U.S. patent records from 1976 onward, while the core inventive activity of many NIHF inductees occurred earlier, leaving parts of their patent portfolios outside the scope of this baseline[19]. The NIHF–PatentsView baseline constructed in this study covers 167 NIHF inventors and 8,665 inventor–patent links, substantially fewer than the full target population and the broader patent activity represented in NIHF records. In addition, the seed-patent-to-PatentsView-identifier expansion depends on whether seed patents can be matched and whether the target inventor identity can be mapped completely. For these reasons, we internally used the licensed InCoPat database for broad candidate patent retrieval[20].

The purpose of the InCoPat retrieval stage was to improve recall of potentially relevant candidate records, not to determine whether inventor–patent links were valid. Records obtained through InCoPat were required to pass subsequent evidence construction, semantic auditing, or human adjudication before entering the final release.

InCoPat queries focused on U.S. patent records. The internal query protocol recorded U.S. patent restrictions, including conditions such as pt = (4) and PNC = US, and used IN-field searches based on inventor-name variants. Name variants included standardized full names, inverted names, initial-based forms, middle-name and middle-initial forms, hyphen- and punctuation-normalized forms, and approximate matching patterns designed to capture spelling variation in patent records. Direct publication-number searches were used to retrieve official seed patents when available.

Date windows for retrieval were constructed from available biographical and patent-activity information. When birth years, death years, seed-patent years, or other activity-period information were available, we used these fields to form plausible application/publication date windows. The goal was to reduce obviously impossible name noise without imposing windows so narrow that true patents would be omitted. For inductees with limited biographical or activity-period information, candidate retrieval remained relatively broad.

The InCoPat raw exports include publication and application identifiers, application and publication dates, titles, abstracts, first claims, inventor strings, first-inventor fields, inventor country or region, applicants/assignees, normalized applicant fields, IPC/CPC classifications, grant dates, family records, priority information, and related metadata. These fields were used only for internal evidence construction and identity auditing. Under the database-use license, we do not redistribute the InCoPat raw export tables. The public release contains only redistributable linkage files, release-layer labels, derived evidence flags, reason-code summaries, PatentsView comparison indicators, validation summaries, minimal derived citation metadata, data dictionaries, and protocol documentation. Researchers with licensed InCoPat access can use the documented search and derivation protocols to reproduce the restricted-input portions of the workflow.

Evidence variables and semantic patent representations. For each candidate inventor–patent pair, we constructed multidimensional record-level evidence variables and used them jointly in rule-based filtering and cluster-level identity audit[6–9,21]. These variables supported candidate construction, identity auditing, and accept/reject/review decisions; they were not used as a single additive-score classifier. The rule layer included broad-recall candidate generation, name compatibility gates, official seed-anchor support, PatentsView baseline comparison, PatentSBERTa-based semantic clustering, and consistency checks based on applicants/assignees, co-inventors, technology classifications, and career windows. Some rules, such as name gates and obvious conflict indicators, could trigger low-risk rejection. Other signals, such as seed-anchor support and semantic-cluster coherence, supported acceptance or manual review when combined with other evidence. Where present, the fields match_score and match_level are retained only as descriptive summaries of the underlying evidence; they do not determine final inclusion. Final labels are based on the combined rule-based, cluster-level, and manually adjudicated evidence. Table 1 summarizes the main evidence dimensions used for candidate construction, identity-cluster audit, and row-level adjudication. For descriptive technology counts, each accepted patent was assigned to a single CPC section, operationally defined as the first section code in the exported CPC string; patents were not multiply counted across sections. When CPC information was unavailable, IPC Main Class and then the NIHF technology label were used as fallbacks.

In addition to the structured evidence dimensions listed in Table 1, we encoded English patent titles and first claims using AI-Growth-Lab/PatentSBERTa[22]; documented fallback text was used when necessary. The model produced 768-dimensional vectors that were L2-normalized before cosine similarity was calculated within each inventor's candidate set. Candidate sets containing at most 2,500 embedded records were represented by an exact pairwise similarity graph, whereas larger sets used an exact brute-force cosine k-nearest-neighbour graph (k = 30). Edges with cosine similarity of at least 0.65 were retained, and connected components were treated as semantic clusters. No minimum cluster size was imposed. Singletons were retained, and records lacking embeddings were evaluated using the remaining non-semantic evidence. Sensitivity checks over thresholds from 0.45 to 0.70 are documented in the repository. Strong target anchors comprised official NIHF seed patents, direct InCoPat matches to those seed patents, and records assigned to the highest-support rule category during initial

Evidence dimension	Main input	Construction and role in audit
Name compatibility	NIHF inductee names; patent inventor strings	Inventor names were normalized for case, punctuation, spacing, abbreviations, middle names, and inverted order. This evidence identifies whether a candidate record has basic name compatibility. It is used for candidate screening and weak identity support but cannot determine acceptance on its own.
Official seed anchor	NIHF seed patent identifiers	This signal checks whether the candidate patent matches an official NIHF seed patent using normalized, prefix-aware patent identity. Official seed anchors provide strong positive identity evidence.
PatentsView baseline support	NIHF–PatentsView public baseline	This signal checks whether a person_key + patent_id_clean pair appears in the public baseline. It provides public benchmark support but is not treated as a gold-standard label.
Assignee/applicant bridge	Applicants/assignees in candidate patents, seed patents, baseline records, and accepted clusters	This signal checks whether a candidate record shares or continues organizations, firms, universities, research institutions, or technology-transfer entities associated with the target inventor's seed, baseline, or accepted cluster. It supports stable identity-cluster identification and helps exclude unrelated same-name clusters.
Co-inventor bridge	Co-inventors in candidate patents and known identity anchors	This signal checks whether a candidate record shares co-inventors with seed patents, PatentsView baseline records, or accepted clusters. Stable co-inventor networks provide important identity support.
Technology support	IPC/CPC classifications; NIHF technology fields; patent-text semantic representations	This signal evaluates whether candidate records fall within the target inventor's technology trajectory through technology-class overlap, domain consistency, and text-semantic similarity. It is especially useful for separating homonym clusters with distinct technology topics.
Career-window plausibility	Birth/death years; seed-patent years; application/publication years	This signal checks whether patent years fall within an interpretable activity period for the target inventor. It helps exclude obviously implausible records and identify boundary records requiring manual review.
Country/geographic consistency	Inventor country/region or address fields, where available	When available, country or geographic information was used to check whether the candidate record is consistent with the target inventor's known activity locations. This evidence was used only as auxiliary support.
Diagnostic conflict indicators	Baseline disagreement; name conflict; sparse evidence; homonym risk	These indicators mark records that may require manual review, including PatentsView-only disagreement, same-name records separated by technology, organization, or co-inventor structure, and records with insufficient evidence for stable classification.

Table 1. Evidence dimensions used in inventor–patent identity auditing.

screening. PatentsView support was retained as a separate evidence dimension. Full parameter settings and code locations are documented in the repository.

PatentSBERTa outputs were used for inventor-level semantic clustering and identity-audit diagnostics. Their role was to help determine whether a candidate set contained a technology-coherent target identity cluster, or whether it contained one or more same-name alternative identity clusters unrelated to the target inventor. For example, when a group of candidate patents formed a stable combination of applicants, co-inventors, and technology topics but was clearly separated from NIHF seed anchors and the target inventor's technology trajectory, those records were treated as a potential homonym cluster rather than being admitted into the accepted release solely because of name matching. A one-evidence-dimension-at-a-time ablation analysis is provided in Supplementary Table S1 as a process-level sensitivity diagnostic.

Cluster-level identity audit and row-level adjudication. Identity auditing was conducted at the person_key level. For each NIHF inductee, we integrated candidate patents, official seed anchors, PatentsView baseline records, structured evidence variables, and patent-text semantic representations into a common audit space. The audit goal was to determine whether candidate records belonged to the stable identity network of the target NIHF inventor, rather than merely whether a similar name appeared in a patent record.

Cluster-level identity audit followed four principles. First, seed-anchor consistency: records that directly matched official seed patents, or that were coherent with seed or baseline clusters in technology, applicants/assignees, and co-inventors, were treated as strong positive candidates. Second, metadata continuity: candidate patents that were continuously connected to the target inventor's known organizations, firms, co-inventors, or technology-transfer entities provided identity support. Third, semantic coherence: candidate patents that were semantically and technologically consistent with the target inventor's core technology trajectory strengthened acceptance evidence. Fourth, homonym separation: records were marked as homonym-risk or rejected candidates when they had the same or a similar inventor name but formed a separate cluster in applicant/assignee networks, co-inventor networks, technology topics, and timelines, and lacked seed or baseline support for the target identity.

Candidate records were provisionally accepted when the evidence dimensions in Table 1 consistently supported the target identity. They were rejected when name, organizational, co-inventor, technological, or temporal evidence indicated a distinct identity. Records with conflicting or incomplete evidence were referred for manual review.

Records requiring review were adjudicated at the row level as accepted, rejected, or uncertain/excluded. Accepted records had sufficient evidence for inclusion in the curated release. Rejected records had evidence supporting exclusion as invalid or non-target links. Uncertain/excluded records lacked sufficient evidence for reliable inclusion and were therefore excluded for quality-control purposes. Each reviewed record received a final label and could be assigned one or more reason codes documenting issues such as PatentsView disagreement, name conflict, technology-topic deviation, or insufficient evidence. The public release provides both a reason-code occurrence summary and a mutually exclusive human-adjudication outcome summary.

Two reviewers independently assessed the sampled inventor–patent records using a common adjudication protocol. Each reviewer assigned an accept/reject label without access to the other reviewer's label. Disagreements were resolved through joint review by the two reviewers. Inter-reviewer agreement was evaluated using raw agreement and unweighted Cohen's κ. Sampling was conducted without replacement within each validation block. The sampling code, random seed, and adjudication protocol are available in the repository.

Human adjudication should be understood as part of the data-construction and curation process. It was used to resolve boundary records that could not be stably classified by automatic evidence rules and cluster-level audit alone. The Technical Validation section further reports release-layer consistency, manual validation, PatentsView baseline comparison, and the homonym stress test.

Public release assembly and release-layer definitions. The final public release was assembled using person_key together with prefix-aware canonical patent identity. For the same inventor–patent pair appearing in multiple evidence sources or intermediate paths, records were deduplicated and source indicators were merged. When a record had multiple forms of support, such as official seed-anchor status, PatentsView baseline support, rule evidence, or human-reviewed acceptance, this information was retained through public-safe source flags or final decision codes. Both application_year and publication_year are provided in the public linkage tables. When a single year was needed for descriptive statistics, application_year was used first; if it was missing, publication_year was used as a fallback, and the selected source was recorded in year_field_used.

The curated accepted dataset is recommended for most analyses. It combines automatically accepted links with links accepted after manual review and excludes rejected, uncertain, and diagnostic-only records. The strict-core subset retains links with stronger identity evidence for analyses that are especially sensitive to false positives. Diagnostic records comprise rejected, uncertain, PatentsView-discordant, and other records retained for methodological review. They are not validated inventor–patent associations and should not be added to the main analysis dataset without case-level verification.

The final release underwent several consistency checks. Person-key integrity checks ensured that each person_key in accepted links, strict-core links, PatentsView comparison records, and diagnostic records corresponded to a target inductee record. Schema validation examined row counts, key fields, missing values, and duplicate keys across public files. For the curated accepted release and the strict-core subset, we checked for duplicate inventor–patent identities using person_key together with prefix-aware patent identity; for diagnostic records, we checked the uniqueness of diagnostic_record_id. Restricted-field stripping checks confirmed that public CSV files do not contain raw InCoPat titles, abstracts, claims, inventor strings, applicant/assignee names, family records, priority information, legal-status fields, addresses, or raw export metadata. Consistency checks verified that the public files matched the reported totals for accepted links, unique patents, covered inventors, strict-core links, adjudication outcomes, seed-anchor retention, and PatentsView comparison metrics.

The final public package contains the target inductee table, curated accepted inventor–patent links, the strict-core subset, the PatentsView baseline comparison files, review and diagnostic records, reason-code summaries, the data dictionary, the release summary, and protocol documentation. This release design explicitly separates valid inventor–patent links, a conservative accepted subset, and diagnostic records, allowing users to choose an appropriate data layer for their research goals while avoiding the misuse of broad candidate records or excluded records as valid patent portfolios.

Data Overview

All descriptive statistics in this section are based on the final curated accepted dataset, which contains 28,330 inventor–patent links and 27,155 unique U.S. patents for 638 of the 653 NIHF inductees. The remaining 15 inductees are outside the accepted release because usable candidate evidence was unavailable or no retrieved link met the final acceptance criteria.

Figure 2 summarizes the descriptive structure of the curated accepted release. At the inventor level, accepted patent portfolios are highly heterogeneous and right-skewed, as shown in Fig. 2a. Across covered inventors, the mean number of unique accepted patents is 44.4, the median is 17, the interquartile range is 5 to 54, and the maximum is 1,004. This distribution shows that NIHF inductees differ substantially in the size of their identifiable patent portfolios: most inventors have relatively limited but traceable portfolios, while a small number have very large patent portfolios.

We define an inventor's patenting-career span as the number of years from the earliest to the latest observed patent year, calculated as latest patent year − earliest patent year +1. Patent year is based first on application_year; when application_year is missing, publication_year is used as a fallback. As shown in Fig. 2b, the median patenting-career span among covered inventors is 27 years, the interquartile range is 12 to 38 years, and the mean is 26.8 years. This indicates that many NIHF inductees have identifiable patenting activity over long periods, making the dataset suitable for studying career length, stage-specific inventive activity, and portfolio evolution among recognition-based elite inventors.

Figure 2c shows that patent years in the curated accepted release range from 1800 to 2025. After applying the year-field fallback procedure, only one accepted link lacks a usable patent year. This broad temporal range reflects the historical span of NIHF inductees across early invention history and the modern patent system. The substantial presence of pre-1976 patents also illustrates why a simple linkage between NIHF and PatentsView alone would be insufficient for reconstructing NIHF inventors' patent portfolios.

Technology-field statistics use one assigned CPC section per unique accepted patent. The assigned section is operationally defined as the first valid CPC section code in the exported CPC string, so each patent contributes at most once to the section distribution. As shown in Fig. 2d, the largest assigned sections are Electricity (H; 6,746 patents), Physics (G; 4,874), Chemistry and metallurgy (C; 4,842), and Performing operations and

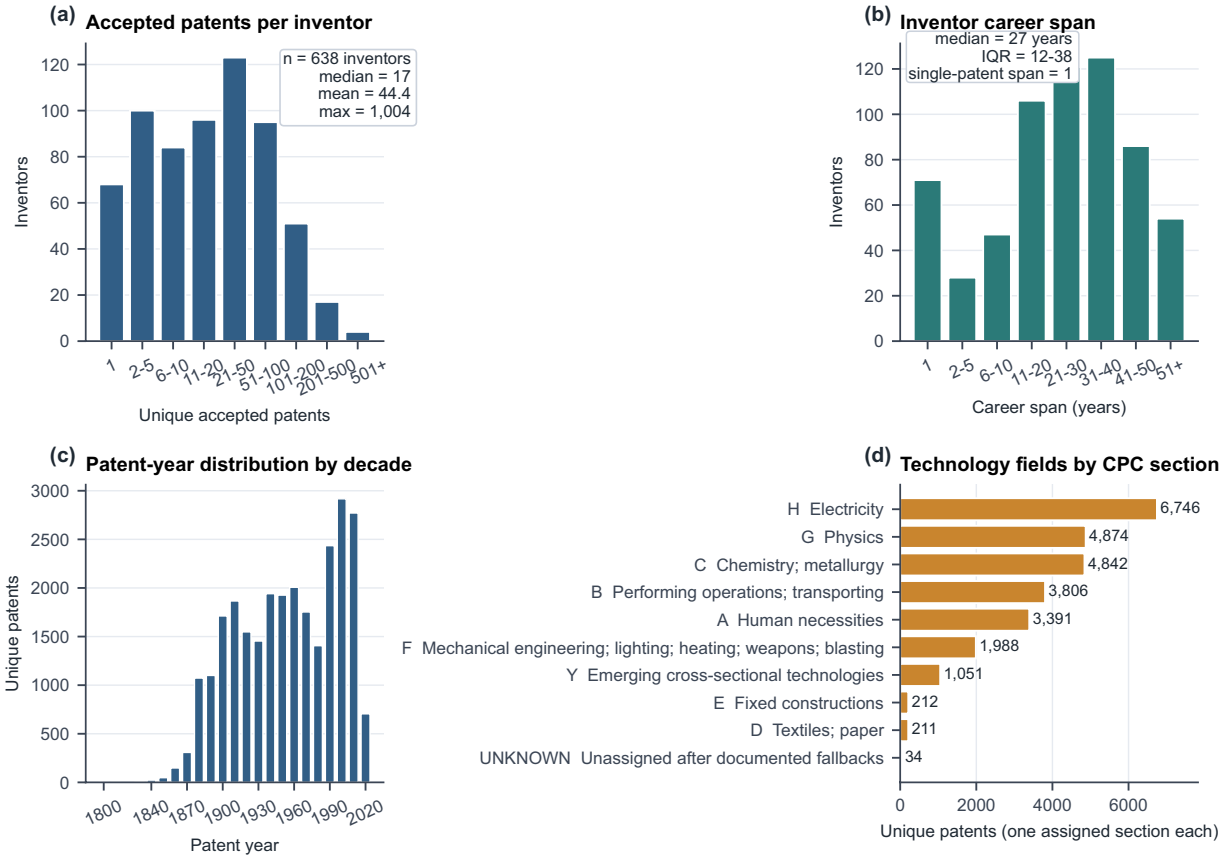


Fig. 2 Descriptive overview of the curated accepted release. (**a**) Distribution of accepted patents per inventor. (**b**) Inventor patenting-career span. (**c**) Patent-year distribution by decade. (**d**) CPC-section distribution using one assigned section per patent. All panels are based on the curated accepted release and exclude candidate and diagnostic records.

transporting (B; 3,806). The single-section counts sum to the 27,155 unique accepted patents. Supplementary Table S2 reports lifetime forward-citation distributions using the same section-assignment rule.

Data Records

The dataset is available at Zenodo (https://doi.org/10.5281/zenodo.22157727)[23]. The core analysis files described below correspond to repository release v1.2.0. The repository is organized into data/, validation/, process/, derived_metadata/, docs/, and reports/. The data/ folder contains the core inventor and inventor–patent linkage files; validation/ contains public-safe historical validation samples and final case-level adjudication summaries; process/ contains the historical evidence-dimension sensitivity diagnostic supporting Supplementary Table S1; derived_metadata/ contains citation-source coverage, citation edges, and single-CPC-section assignments; and docs/ and reports/ contain data dictionaries, provenance notes, protocols, release manifests, checksums, and quality-control reports.

The default analysis file is data/02_nihf_inventor_patent_links_curated_accepted_v1_2_0.csv, which contains the final curated accepted inventor–patent links. The file data/03_nihf_inventor_patent_links_strict_core_v1_2_0.csv provides a more conservative accepted subset for studies requiring higher identity precision or robustness checks. The file data/05_nihf_review_and_diagnostic_records_v1_2_0.csv contains rejected, uncertain/excluded, benchmark-discordant, and other diagnostic records. These records are not confirmed links in the released accepted layer. The PatentsView comparison files are provided for public benchmark comparison and should not be interpreted as gold-standard labels.

The public release uses three main linkage keys: person_key, inventor_id, and patent_id_clean. person_key is the stable inventor-level key used across release files. inventor_id is the public NIHF inventor identifier. patent_id_clean is a normalized string field used for cross-file linkage; users should preserve leading zeros and consult patent_id_original_public_if_safe when patent-type information is required. For accepted inventor–patent links, person_key together with prefix-aware patent identity uniquely identifies a link.

All patent identifiers should be read as strings rather than numeric variables. This includes patent_id_clean, patent_id, focal_patent_id_clean, citing_patent_id_clean, and seed patent identifiers. Some identifiers may contain leading zeros; converting them to integers may remove leading zeros and break cross-file joins.

File	Description	Rows
01_nihf_inductees_v1_2_0.csv	Target table of NIHF inductees, including stable person_key, public inventor_id, names, induction years, NIHF technology fields, and seed patent identifiers.	653
02_nihf_inventor_patent_links_curated_accepted_v1_2_0.csv	Main curated inventor–patent linkage file. Each row represents a link accepted by the construction and adjudication workflow.	28,330
03_nihf_inventor_patent_links_strict_core_v1_2_0.csv	More conservative strict-core accepted subset for high-precision analyses or robustness checks.	27,721
04_nihf_patentsview_baseline_comparison_v1_2_0.csv	Link-level comparison between the curated release and the NIHF–PatentsView public baseline.	28,838
04a_nihf_patentsview_baseline_comparison_summary_v1_2_0.csv	Summary metrics for the PatentsView comparison used in the manuscript.	10
05_nihf_review_and_diagnostic_records_v1_2_0.csv	Row-level diagnostic records, including rejected, uncertain/excluded, and benchmark-discordant records. These records are not confirmed links in the released accepted layer.	194,430
06_nihf_adjudication_reason_code_summary_v1_2_0.csv	Non-mutually exclusive multilabel counts for human-adjudication reason codes.	32
06a_nihf_human_adjudication_outcome_summary_v1_2_0.csv	Mutually exclusive summary of human-adjudication outcomes.	7
07_nihf_data_dictionary_v1_2_0.csv	Variable-level data dictionary for public release files.	109
08_nihf_release_summary_v1_2_0.csv	Automated release-consistency report comparing computed values with archived release values.	25

Table 2. Core data files.

Data layer	Analysis status	Rows	Recommended use
Curated accepted release	Accepted links; default analysis dataset	28,330	Default layer for analyses of patent portfolios, career spans, technology fields, time series, and collaboration networks.
Strict-core subset	Accepted links; conservative subset	27,721	More conservative layer for studies requiring higher identity precision.
Review/diagnostic records	Not accepted	194,430	Validation, error analysis, and benchmark-discordance review only; should not be merged with accepted links as valid samples.
PatentsView baseline comparison	Benchmark records only	28,838	Added-value, overlap, and recall-oriented diagnostic comparison.

Table 3. Data layer boundaries.

Table 2 lists the core data files in the data/ folder. Full variable definitions are provided in data/07_nihf_data_dictionary_v1_2_0.csv; additional variables used in validation, process, and derived metadata files are documented in docs/DATA_DICTIONARY_ADDENDUM_v1_2_0_validation_extension.csv.

Table 3 summarizes the boundaries and recommended uses of the main data layers. Full variable definitions, data types, and example values are provided in 07_nihf_data_dictionary_v1_2_0.csv.

In addition to the data/ folder, the release provides several public-safe auxiliary files. The validation/ folder contains stratified audit samples and results stripped of restricted third-party fields. The process/historical_diagnostics/ folder contains the evidence-dimension ablation diagnostic supporting Supplementary Table S1. These auxiliary files support validation, process transparency, and downstream extension.

Technical Validation

This section evaluates the internal consistency of the released data layers, manual adjudication outcomes, stratified validation results, comparison with the NIHF–PatentsView public baseline, homonym separation, and remaining validation and coverage limitations. The PatentsView baseline is used as a public and reproducible benchmark rather than as a gold standard.

Release-layer consistency checks. The final curated accepted dataset contains 28,330 inventor–patent links and 27,155 unique prefix-aware U.S. patents, covering 638 NIHF inductees; the strict-core subset contains 27,721 links and 26,549 unique patents. Both accepted layers were deduplicated using person_key together with prefix-aware patent identity. The diagnostic layer contains 194,430 records and remains separate from the valid inventor–patent linkage layers.

Consistency checks confirmed that the public files matched the reported totals for accepted links, unique patents, covered inventors, strict-core links, adjudication outcomes, seed-anchor retention, and PatentsView comparison metrics. No duplicate person_key + patent_id_clean records were found in the accepted or strict-core files, and no duplicate diagnostic_record_id values were found in the diagnostic records. These checks confirm clear boundaries among the default accepted layer, the conservative subset, and diagnostic records.

Human adjudication of review-required records. All review-required records were manually adjudicated. Records entered the review-required queue when they involved name ambiguity, PatentsView baseline

Validation component	N	Confirmed/agreement	Rejected/disagreement	Estimate
Accepted-link validation sample	300	287	13	95.67% (Wilson 95% CI 92.7–97.5%)
Strict-core subset of accepted-link validation sample	239	236	3	98.74% (Wilson 95% CI 96.4–99.6%)
Excluded/diagnostic case-level adjudication	300	31	269	10.33% confirmed valid (Wilson 95% CI 7.4–14.3%)
Inter-reviewer agreement overall	600	587	13	97.83%; $\kappa = 0.956$ (95% CI 0.931–0.979)

Table 4. Manual validation, case-level adjudication, and inter-reviewer agreement. Notes: Reviewers labelled records independently of each other but were not fully blinded to the pre-existing pipeline/release classifications. The accepted-link validation sample was drawn before the audit findings were incorporated into the final release; no additional validation sample was drawn after these findings were incorporated.

disagreement, unclear semantic-cluster boundaries, sparse structured evidence, or conflicting evidence. Within the main review-required queue, 15,155 records received human labels. At this stage, 5,418 were labelled accepted, 8,995 were labelled rejected, and 742 were marked as uncertain/excluded because the available evidence was insufficient for reliable inclusion. These queue-level counts describe the manual-review stage; final release status also incorporates the validation and case-level adjudication described below. The public release provides both a mutually exclusive human-adjudication outcome summary and a non-mutually exclusive reason-code summary.

Stratified manual validation of accepted and excluded records. The two reviewers worked independently with respect to each other's labels, but the review materials were not fully blinded to the pre-existing pipeline/release classification. Manual validation comprised two stratified blocks of 300 records each: one sampled from the curated accepted dataset and the other from excluded or diagnostic records. The sampling design deliberately included boundary cases, common-name records, homonym-risk cases, and other records expected to be difficult to classify. Block-level and stratum-specific results are therefore reported as assessments of the sampled records rather than as estimates of the dataset-wide false-positive rate, false-negative rate, or recall.

Across the 600 sampled records, the two reviewers assigned identical labels to 587 records, corresponding to a raw agreement of 0.978 and an unweighted Cohen's κ of 0.956 (bootstrap 95% CI, 0.931–0.979). The accepted-link audit confirmed 287 of 300 sampled accepted links. Within the strict-core subset represented in the same validation sample, 236 of 239 records were confirmed. The accepted-link audit was completed before the audit-identified discrepancies were incorporated into the final release; no new post-correction validation sample was drawn.

The excluded/diagnostic block was subsequently resolved through case-level adjudication. Of the 300 sampled excluded or diagnostic records, 31 were confirmed as valid links and 269 were rejected. Because the sampling design intentionally enriched difficult boundary, common-name, homonym-risk, and diagnostic records, this proportion should not be interpreted as a dataset-wide false-negative rate or recall estimate.

Table 4 summarizes the accepted-link audit, the strict-core subset result, the case-level adjudication of the excluded/diagnostic block, and inter-reviewer agreement.

NIHF–PatentsView public baseline comparison. We further compared the final curated accepted release with the NIHF–PatentsView public baseline at the inventor–patent-link level. The NIHF–PatentsView baseline covers 167 NIHF inventors and 8,665 inventor–patent links, whereas the final curated accepted dataset covers 638 NIHF inventors and 28,330 inventor–patent links. The intersection contains 8,157 links. The curated accepted release contains 20,173 additional curated-only links and covers 471 NIHF inventors who are not represented in the PatentsView baseline.

The final curated accepted release retains 94.1% of PatentsView baseline links (baseline retention proxy = 0.941), while 71.2% of final curated links lie outside the public baseline (added-value share = 0.712). The Jaccard similarity between the two link sets is 0.283. Figure 3a compares NIHF inventor coverage in the PatentsView baseline and the curated accepted release; Fig. 3b compares inventor–patent link coverage; and Fig. 3c reports the intersection, our-only links, PV-only links, and added-value metrics. These results show that the NIHF seed patents plus PatentsView inventor identifiers provide a valuable, transparent, and reproducible public baseline, but that relying only on this workflow would substantially limit the coverage of NIHF inductees and inventor–patent links.

PatentsView baseline records are not treated as gold-standard labels9,18. The 508 PatentsView-only links are retained as diagnostic records rather than being automatically classified as true omissions from the curated release. Full review of these PV-only records showed that they appeared in the candidate pool but were manually adjudicated as rejected diagnostic records. PV-only records are therefore more appropriately interpreted as benchmark-discordance diagnostics than as direct false negatives of the curated release.

Homonym stress test: Adi Shamir. Homonyms are a core risk in inventor–patent linkage construction. We use Adi Shamir as a homonym stress test to examine whether the revised workflow can distinguish the target NIHF inventor identity cluster from same-name alternative identity clusters. Adi Shamir is a 2002 Turing Award laureate, co-inventor of RSA, and cryptographer. His target identity network is centered on MIT, the Weizmann Institute, Yeda Research and Development, RSA, public-key cryptography, authentication, digital signatures, RFID security, and side-channel or fault attacks.

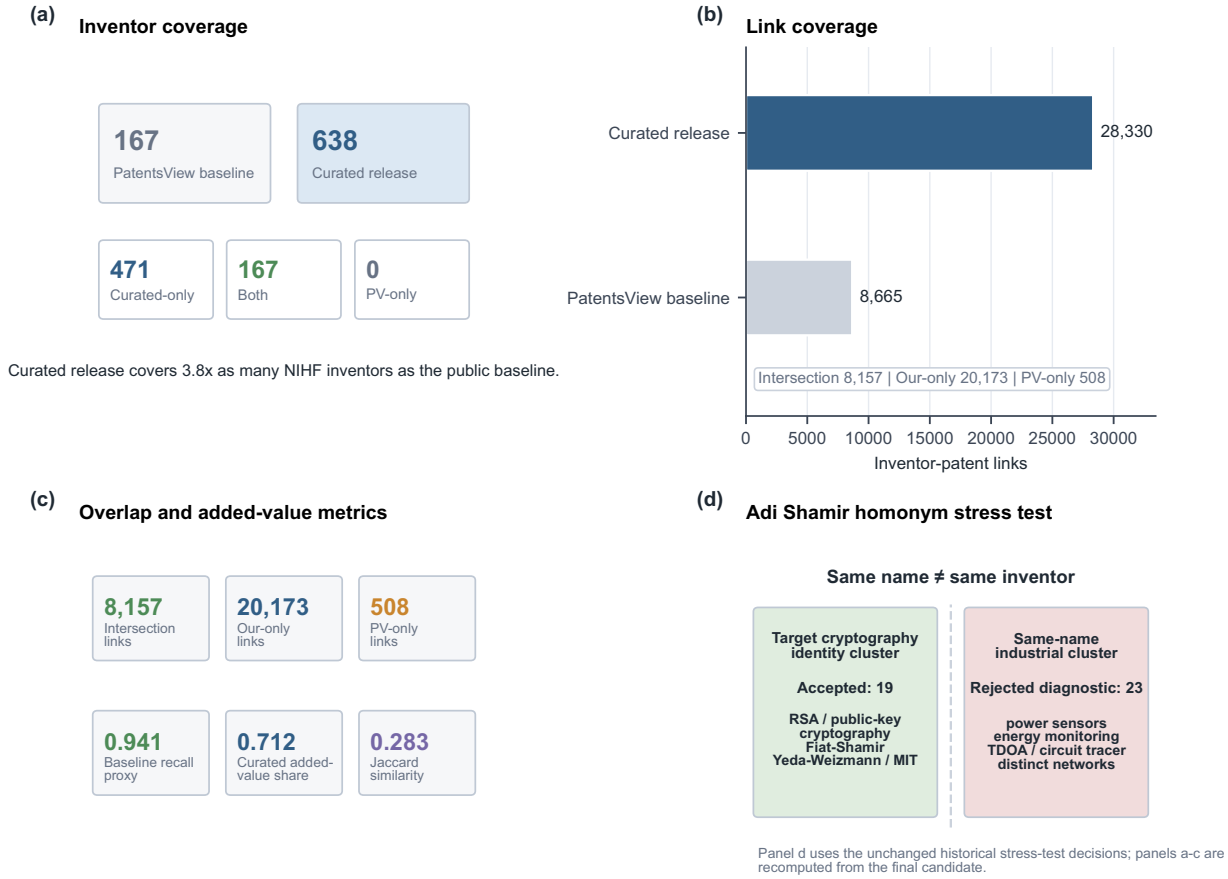


Fig. 3 PatentsView baseline comparison and homonym diagnostics. (**a**) NIHF inventor coverage in the PatentsView baseline and the curated accepted release. (**b**) Inventor–patent link coverage in the two datasets. (**c**) Intersection, our-only, PV-only, baseline retention proxy, added-value share, and Jaccard similarity. (**d**) Adi Shamir homonym stress test, distinguishing the target cryptography identity records from same-name industrial-technology records.

In the 65-record Adi Shamir stress-test set, 19 links were accepted as belonging to the target cryptography identity, whereas 46 were classified as rejected and retained only in the diagnostic records. Of the 46 rejected records, 23 formed the distinct industrial-technology identity network highlighted in Fig. 3d. These records match "Adi Shamir" at the name level, and some also appear in the NIHF–PatentsView public baseline, but they form a separate cluster in applicant/assignee networks, co-inventor combinations, and technology topics. One group is associated with Panoramic Power, AeroScout, power sensors, wireless localization, circuit tracing, or industrial monitoring, and repeatedly involves co-inventors and organizations unrelated to the target cryptography identity. This group lacks consistent support from NIHF seed anchors, the MIT/Yeda/Weizmann institutional chain, or the cryptography-related technology trajectory.

The Adi Shamir stress test shows that name matching, simple score thresholds, and even external PatentsView inventor identifiers are not sufficient by themselves to determine the target NIHF inventor identity in high-risk homonym settings. The final workflow uses technology topics, co-inventor networks, applicant/assignee networks, official seed anchors, PatentsView baseline status, and human audit jointly to classify identity clusters. Figure 3d shows the separation between the target cryptography identity records and same-name industrial-technology records. This case illustrates why retaining diagnostic records is necessary for reviewing homonym risks.

Limitations of validation. The validation results do not imply that the dataset is free of error. Because the audit intentionally enriched boundary, common-name, homonym-risk, and diagnostic strata, the unweighted block-level proportions do not estimate the overall false-positive rate, false-negative rate, or recall of the complete release. Instead, the audit identifies the release layers and evidence conditions in which residual uncertainty is concentrated. Curated non-strict links and common-name or homonym-risk cases require particular caution. Excluded and diagnostic records may also contain omitted valid links, especially among uncertain/excluded and previously human-rejected records. Users conducting high-precision or high-recall case studies should inspect accepted, strict-core, audit, and diagnostic records together.

Measure	Official seed anchors	Added non-seed links	N/denominator	Interpretation
Inventor–patent links	711	27,619	28,330 total links	Link-level counts.
Unique prefix-aware U.S. patents	610	26,556	Categories may overlap across inventors	Patent-level auxiliary counts.
Patent year, median (IQR)	1954 (1913.5–1976.5)	1963 (1925–1999)	711 seed links; 27,619 added links	Application year with publication-year fallback.
First added minus first seed year, median (IQR)	Not applicable	−3 (−11–1)	569 inventors with both seed and added links	Negative values indicate added links predating the earliest retained seed link.
Added-link share matching any seed assigned CPC section, median (IQR)	Not applicable	0.709 (0.375–0.902)	567 inventors with usable seed and added CPC assignments	Technology-overlap proxy using one assigned CPC section per patent.

Table 5. Official seed anchors versus added non-seed links in the final curated release. Notes: Seed anchors are official NIHF-linked U.S. patents retained in the final release. Technology overlap is descriptive and is not a direct semantic distance measure.

Usage Notes

Users should select the data file that best matches their research objective. For most analyses, we recommend 02_nihf_inventor_patent_links_curated_accepted_v1_2_0.csv. Studies requiring greater protection against false positives may instead use 03_nihf_inventor_patent_links_strict_core_v1_2_0.csv as the primary sample or as a robustness check. Review and diagnostic files are intended for error analysis and case-level inspection and should not be treated as confirmed patent portfolios. The v1.2.0 release incorporates the final adjudicated decisions used in this manuscript, so users do not need to apply separate correction files to the default analysis layer.

The dataset provides an inventor–patent linkage structure for NIHF inductees rather than a redistribution of complete patent metadata. Users can connect inventor-level information through person_key and link accepted patents to external patent databases through patent_id_clean to add citations, applicants/assignees, CPC/IPC classifications, patent full texts, claims, legal-status information, patent-value measures, or other patent-level variables. When linking to external data sources, users should account for differences in patent-number formats, kind codes, publication and application numbers, historical coverage, and database licensing conditions. All patent identifiers should be read as strings rather than numeric variables, including patent_id_clean, patent_id, focal_patent_id_clean, citing_patent_id_clean, and seed patent identifiers. Some patent identifiers may contain leading zeros; automatic conversion to integers may remove these zeros and break cross-file joins.

The file derived_metadata/15_nihf_accepted_patent_citation_edges_public_safe_v1_2_0.csv can be used to calculate lifetime forward-citation counts for accepted patents. The accompanying citation coverage summary file can be used to examine citation-source coverage by application year, application decade, and CPC section[24]. However, citation_year should not be treated as a complete citing-year variable. Because this field is missing for most citation edges, the citation edge file should not be used for complete annual citation trajectories, citation-window analyses, annual accumulation measures, or citation life-cycle analyses. Studies requiring complete citation timing should reconstruct citing-year information from external patent databases.

Supplementary Table S2 summarizes lifetime forward-citation counts using source-covered patents only. Of the 27,155 unique accepted U.S. patents, 26,580 have citation-source coverage and 575 are source-unavailable. Among source-covered patents, 4,974 have zero observed forward citations (18.7%), and the median is 6 citations (IQR 1–23). Section-specific summaries are descriptive and are not normalized for patent age, cohort, or field.

Table 5 provides a descriptive comparison between official seed-anchor links and the additional non-seed links recovered in the final curated release. The added links extend inventor portfolios both before and after the representative NIHF seed patents, while retaining substantial overlap with the seed patents in assigned technology fields.

The dataset can also be combined with co-inventor information to construct collaboration networks and inventor career trajectories. Based on accepted inventor–patent links, users can identify an inventor's co-inventors across career stages, repeated collaboration relationships, team-size changes, and collaboration structures within and across technology fields[25]. Figure 4 illustrates this downstream reuse with Robert S. Langer, Jr., showing how accepted links can be combined with co-inventor information to reconstruct repeated collaboration trajectories, NIHF induction timing, induction-related patents, and cumulative patenting activity.

Coverage is not uniform across the NIHF target population. Fifteen of the 653 inductees remain without accepted portfolios: 2 lack usable seed or candidate records, and 13 have retrieved candidates but no link meeting the final acceptance criteria. Historical OCR, name variation, and cross-source record availability may still limit recall for some early inventors.

A separate selection boundary arises from the definition of the target population itself. NIHF inductees constitute a recognition-based, institutionally defined elite inventor population, rather than an unbiased sample of all important inventors. NIHF induction may reflect historical period, technology field, geography, gender, organizational type, institutional recognition mechanisms, and source visibility. The dataset is therefore suitable for studying recognized technological contributors, honorary recognition, patent portfolios, elite inventor careers, and collaboration trajectories, but should not be used on its own to infer average behavior or unbiased distributions in the broader inventor population. Studies addressing general inventor behavior should combine this dataset with larger general-purpose patent databases.

Although the curated accepted release has undergone multistage evidence auditing, human adjudication, and stratified re-audit, a small number of mismatches, omissions, or insufficient-evidence boundary cases may

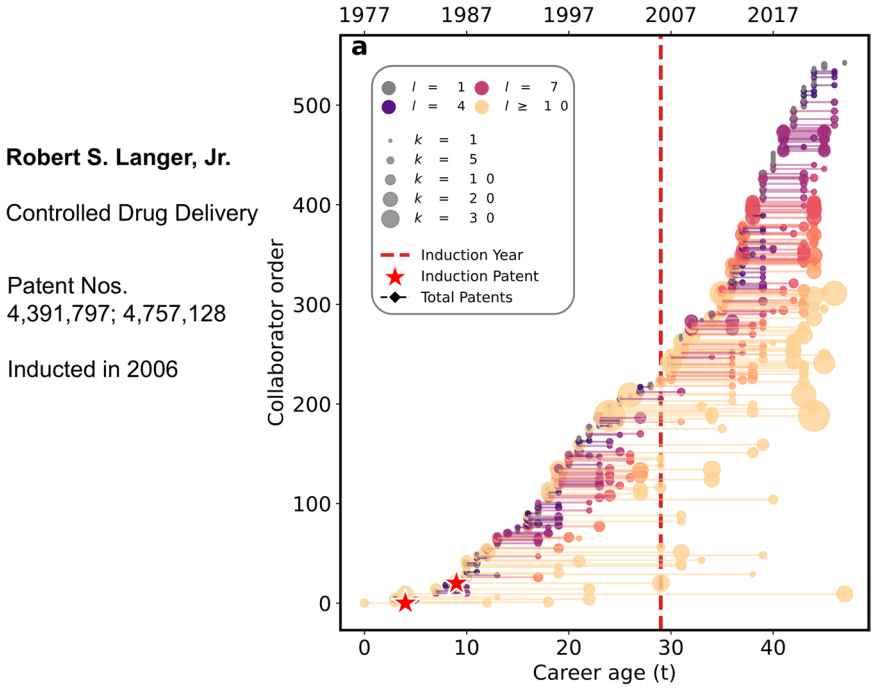


Fig. 4 Downstream reuse example: repeat-collaboration trajectory of Robert S. Langer, Jr. Horizontal lines represent co-inventors ranked by total collaboration strength; line length represents collaboration duration; point size represents collaboration strength; the red dashed line marks the NIHF induction year; red stars mark induction-related patents; and the black dashed line and diamonds show cumulative patent counts. The figure illustrates how accepted inventor–patent links can be combined with co-inventor information to construct inventor-level career collaboration profiles.

remain. For high-precision case studies, users should inspect the relevant inventor's accepted links, strict-core subset, validation files, and diagnostic records together, especially for inventors with common names, frequent institutional mobility, or elevated homonym risk. Diagnostic records are retained to support review and methodological transparency; they are not validated or accepted inventor–patent links and should not be merged into the default analysis sample without additional case-level review. Due to third-party database licensing and data-use boundaries, raw InCoPat export fields are not redistributed.

Data availability

The dataset can be accessed at Zenodo (https://doi.org/10.5281/zenodo.22157727)[23]. The repository contains the core inventor and inventor–patent linkage tables, stratified validation samples and summaries, historical process diagnostics, minimal citation metadata, data dictionaries, release manifests, third-party data notes, search protocols, patent-identifier documentation, and quality-control reports.

Code availability

The code used to construct and validate the released dataset is available on GitHub (https://github.com/Jianming-Guo/NIHF/tree/Jianming-Guo-patch-1).

Received: 26 March 2026; Accepted: 10 September 2026;
Published: xx xx xxxx

References

1. Jaffe, A. B. & de Rassenfosse, G. Patent citation data in social science research: overview and best practices. *J. Assoc. Inf. Sci. Technol.* **68**, 1360–1374 https://doi.org/10.1002/asi.23731 (2017).
2. Toole, A. A., Jones, C. & Madhavan, S. PatentsView: an open data platform to advance science and technology policy. *USPTO Economic Working Paper* No. 2021-1. https://doi.org/10.2139/ssrn.3874213 (2021).
3. United States Patent and Trademark Office. USPTO Open Data Portal. https://data.uspto.gov/ (accessed 7 July 2026).
4. United States Patent and Trademark Office. PatentsView. https://www.uspto.gov/ip-policy/economic-research/patentsview (accessed 7 July 2026).
5. The Lens. Lens.org. https://www.lens.org/ (accessed 7 July 2026).
6. Raffo, J. & Lhuillery, S. How to play the "Names Game": patent retrieval comparing different heuristics. *Res. Policy* **38**, 1617–1627 https://doi.org/10.1016/j.respol.2009.08.001 (2009).
7. Li, G.-C. *et al*. Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Res. Policy* **43**, 941–955 https://doi.org/10.1016/j.respol.2014.01.012 (2014).
8. Morrison, G., Riccaboni, M. & Pammolli, F. Disambiguation of patent inventors and assignees using high-resolution geolocation data. *Sci. Data* **4**, 170064 https://doi.org/10.1038/sdata.2017.64 (2017).

9. Binette, O. *et al*. Estimating the performance of entity resolution algorithms: lessons learned through PatentsView.org. *Am. Stat.* **77**, 370–380 https://doi.org/10.1080/00031305.2023.2191664 (2023).
10. United States Patent and Trademark Office. USPTO and National Inventors Hall of Fame programs. https://www.uspto.gov/learning-and-resources/ip-programs-and-awards/uspto-and-national-inventors-hall-fame-programs (accessed 7 July 2026).
11. National Inventors Hall of Fame. National Inventors Hall of Fame Inductee List. https://www.invent.org/inductees/list (accessed 5 March 2026).
12. Bell, A., Chetty, R., Jaravel, X., Petkova, N. & Van Reenen, J. Who becomes an inventor in America? The importance of exposure to innovation. *Q. J. Econ.* **134**, 647–713 https://doi.org/10.1093/qje/qjy028 (2019).
13. Moretti, E. The effect of high-tech clusters on the productivity of top inventors. *Am. Econ. Rev.* **111**, 3328–3375 https://doi.org/10.1257/aer.20191277 (2021).
14. Blomkvist, K., Kappen, P. & Zander, I. Superstar inventors—towards a people-centric perspective on the geography of technological renewal in the multinational corporation. *Res. Policy* **43**, 669–682 https://doi.org/10.1016/j.respol.2013.12.003 (2014).
15. Caviggioli, F. & Forthmann, B. Reach for the stars: disentangling quantity and quality of inventors' productivity in a multifaceted latent variable model. *Scientometrics* **127**, 7015–7040 https://doi.org/10.1007/s11192-022-04328-3 (2022).
16. Arts, S., Hou, J. & Gomez, J. C. Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures. *Res. Policy* **50**, 104144 https://doi.org/10.1016/j.respol.2020.104144 (2021).
17. Li, J., Yin, Y., Fortunato, S. & Wang, D. A dataset of publication records for Nobel laureates. *Sci. Data* **6**, 33 https://doi.org/10.1038/s41597-019-0033-6 (2019).
18. United States Patent and Trademark Office. Final release of PatentsView metadata, pre-grant and granted (12/31/2024). *Zenodo* https://doi.org/10.5281/zenodo.15783125 (2025).
19. United States Patent and Trademark Office. PatentsView Granted Patent Long Text Data. https://data.uspto.gov/bulkdata/datasets/pvgpattxt (accessed 7 July 2026).
20. InCoPat. InCoPat Global Patent Database. https://www.incopat.com/ (accessed 28 August 2026).
21. Yin, D., Motohashi, K. & Dang, J. Large-scale name disambiguation of Chinese patent inventors (1985–2016). *Scientometrics* **122**, 765–790 https://doi.org/10.1007/s11192-019-03310-w (2020).
22. Bekamiri, H., Hain, D. S. & Jurowetzki, R. PatentSBERTa: a deep NLP based hybrid model for patent distance and classification using augmented SBERT. *Technol. Forecast. Soc. Change* **206**, 123536 https://doi.org/10.1016/j.techfore.2024.123536 (2024).
23. Guo, J. & Yang, A. J. A. U. S. inventor–patent linkage dataset for National Inventors Hall of Fame inductees (Version v1_2_0) [Dataset]. *Zenodo* https://doi.org/10.5281/zenodo.22157727 (2026).
24. Higham, K., Kotula, H., Scharfmann, E., Gong, S. & de Rassenfosse, G. A dataset of scientific citations in U.S. patent Office Actions. *Sci. Data* **13**, 325 https://doi.org/10.1038/s41597-026-06720-7 (2026).
25. Yang, A. J., Guo, J., Shi, Y., Deng, S. & Zhao, S. X. Repeat collaboration and scientific innovation: evidence from dynamic ego networks of Nobel laureates. *Humanit. Soc. Sci. Commun.* **12**, 1620 https://doi.org/10.1057/s41599-025-05887-5 (2025).

Acknowledgements

The authors would like to thank the Nanjing University International Collaboration Initiative.

Author contributions

Conceptualization (J.G, A.J.Y), Methodology (J.G), Software (H.H, J.K), Validation (M.G, Y.Y), Resources (M.G, Y.Y), Data Curation (H.H, J.K), Writing - Original Draft (J.G), Writing - Review & Editing (A.J.Y), Visualization (J.G, A.J.Y), Supervision (S.D, A.J.Y), Project administration(S.D), Funding acquisition (S.D).

Funding

This work was supported by the Key Project of the National Social Science Fund of China (Grant No. 25ATQ008), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. 26CXJH0151), the National Natural Science Foundation of China for Young Scholars (Ph.D. research project, Grant No. 724B2015), the Social Science Foundation of Jiangsu Province (Grant No. 24TQC005), and the National Natural Science Foundation of China (Grant No. 72564009).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-08324-7.

Correspondence and requests for materials should be addressed to A.J.Y. or S.D.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.