Science Advances
AAAS

How Researchers Claim Novelty in Biomedical Science: A Taxonomy for Understanding Innovation

Xueqing Peng,[1] Huan He,[1] Rui Shi,[1] Vipina K. Keloth,[1] Lingfei Qian,[1] Yan Hu,[2] Jimin Huang,[1] Qianqian Xie,[1] Fares Alahdab,[3] Erin Leahey,[4] Brian Ondov,[1] Qiaozhu Mei,[5] Na Hong,[1,*] and Hua Xu[1,*]

[1]Department of Biomedical Informatics & Data Science, School of Medicine, Yale University, New Haven, CT.

[2]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX.

[3]Departments of Cardiology, Biomedical Informatics, Biostatistics, and Epidemiology, University of Missouri, Columbia, MO.

[4]School of Sociology, University of Arizona, Tucson, AZ.

[5]School of Information, University of Michigan, Ann Arbor, MI.

[*]Corresponding author. Email: hua.xu@yale.edu (H.X.); na.hong@yale.edu (N.H.)

Abstract

Scientific novelty is essential for progress, yet it is often hard to define and measure. In this study, we create *NOBL (NOvelty taxonomy for Biomedical Literature)*, a hierarchical taxonomy for classifying different types of novelty claims in biomedical research papers. Developed through an iterative, data-driven, expert-guided, and AI-assisted framework using a domain-stratified corpus of 12,701 PubMed articles, *NOBL* offers a fine-grained, interpretable schema of innovation types, validated for clarity, consistency, and coverage. We find that 6.47% of articles include at least one explicit novelty claim, with new findings being the most common type. We also identify three distinct innovation patterns across 130 biomedical fields. *NOBL* provides a systematic approach for researchers, reviewers, and funders to understand the innovation landscape of biomedical research and offers a new way to make scientific novelty more visible, measurable, and useful.

MAIN TEXT

Introduction

Scientific advances are driven by unexpected discoveries, periodic paradigm shifts, and continual methodological innovation. Central to this progress is the pursuit of novelty: ideas, methods, or perspectives that challenge established knowledge. Novelty fuels scientific understanding by uncovering previously unknown facts and catalyzing both technological advances and conceptual breakthroughs. (*1*) As Charles C. Gillispie noted, scientific revolutions are marked not only by the accumulation of new data and techniques, but also by "*a new way of seeing*" the foundational structure of science. (*2*) In the rapidly evolving field of biomedical research, novelty plays a particularly vital role, directly impacting human health through the reorientation of clinical practice, the reshaping of research priorities, and the development of new approaches to disease understanding, diagnosis, and treatment. (*3*)

Given its central role in scientific progress, systematically identifying and characterizing novelty in research publications is essential. Accurate novelty assessment can help guide readers along the trajectory of innovation, accelerate the learning process for early-career researchers, inform strategic decisions about future research directions, and play a critical role in peer review and funding evaluations. Yet, the question remains: *how can we better identify and understand the novelty of biomedical research?*

Existing definitions of novelty vary widely across the literature. In some contexts, novelty refers to entirely new phenomena, theories, or methodologies. (*4*, *5*) In others, it denotes novel combinations of existing elements, such as interdisciplinary integration or recontextualization of known findings, as demonstrated by AI-driven drug discovery. (*6–11*) These diverse forms of novelty highlight the need for a unified framework capable of capturing the heterogeneous and evolving nature of biomedical innovation. Prior attempts to summarize novelty categories using features such as keywords, co-authorship networks, or novelty components (*12–15*) are often inconsistent, subjective, and coarse-grained with limited flat (non-hierarchical) novelty types. Furthermore, they often fail to accommodate temporal evolution and are difficult to scale.

Despite its importance, a universally accepted definition of scientific novelty remains elusive. This ambiguity arises from three primary challenges. (a) **Domain-specific variation**. Different disciplines assess novelty through distinct conceptual lenses. For instance, in evolutionary biology, novelty might involve genetic or morphological divergence, (*16*) whereas in oncology it may pertain to methodological shifts such as liquid biopsy (*17*) or conceptual frameworks like the cancer stem cell hypothesis. (*18*) (b) **Temporal dynamics**. As biomedical knowledge advances, our understanding of what is novel must also evolve. The emergence of CRISPR (*19*) and AlphaFold (*20*) exemplifies how platform-level methods and AI-driven discovery tools, once rarely framed as standalone novelties, have become explicitly recognized as distinct novelty types in recent literature. (c) **Subjectivity**. Perceptions of novelty can differ among individuals, leading to inconsistencies in evaluation. This challenge is reflected in the need for multiple reviewers in peer review processes.

To overcome the aforementioned challenges, we propose a data-driven, hierarchical taxonomy to understand scientific novelty based on authors' claims in publications. First, to ensure comprehensive domain coverage, we conducted a large-scale analysis of biomedical literature (12,701 articles) to ensure coverage of all biomedical domains. Second, our framework employs a flexible structure and an AI-assisted approach that support efficient expansion and adaptation in future. Third, to minimize subjectivity of annotators in assessing which contributions are novel, we leverage a data-driven approach that focuses on direct statements of novelty extracted from conclusions of peer-reviewed articles. This comprehensive analysis across thousands of biomedical publications requires robust computational methods. Recent advances in artificial intelligence, particularly large language models (LLMs) (*21*) and natural language processing (NLP), provide

the technical foundation to systematically aggregate and categorize author-claimed novelty statements across a comprehensive corpus. These technologies enable consistent, reproducible analysis at the scale needed for building taxonomies with broad coverage and robustness, supporting fine-grained novelty analysis.

In this study, we present a biomedical novelty taxonomy, evaluate its structural validity, semantic coherence, and domain coverage, and apply it to characterize the distribution of novelty across diverse biomedical subfields. Our main contributions include:

(a) A biomedical scientific novelty taxonomy. To our knowledge, we introduce the first biomedical novelty taxonomy, *NOBL* (*NOvelty taxonomy for Biomedical Literature)*, for identifying novelty in biomedical literature, enabling structured and interpretable analysis.

(b) An LLM-assisted taxonomy development pipeline. We develop a scalable and reproducible pipeline that leverages LLMs to systematically identify, cluster, and organize authors' novelty claims.

(c) A publicly released annotated dataset. We release the *Biomedical Novelty Claim Corpus* of articles labeled with novelty types, supporting future work in innovation analysis and NLP-based literature mining.

(d) A biomedical domain novelty map. We analyze novelty type distribution across biomedical disciplines, offering insights into the structure and dynamics of biomedical innovation patterns.

Together, our framework and resources provide a foundation for identifying novel contributions at scale, tracking emerging research trends, and supporting innovation evidence-based decision-making by researchers, reviewers, and policymakers.

Results

1. *NOBL*

To systematically characterize innovation in biomedical research, we developed *NOBL*, which organizes diverse novelty claims into structured, interpretable categories (Fig. 1) based on our *Biomedical Novelty Claim Corpus* of 12,701 PubMed articles (data curation details in 4.1.1.). Complete definitions of *NOBL* are provided in Supplementary. *NOBL* defines three top-level categories: "Concept novelty," "Method and/or material novelty," and "Finding novelty," each divided further into second-level and third-level categories, forming a deep, multi-tiered hierarchy. In total, *NOBL* includes 31 subcategories that comprehensively capture the spectrum of innovation within the biomedical literature. This taxonomy developed from both expert-driven and language model-assisted annotation, offers a unified framework for reproducible and scalable novelty analysis. By standardizing the representation of novelty claims, *NOBL* advances the systematic study of scientific novelty across disciplines.

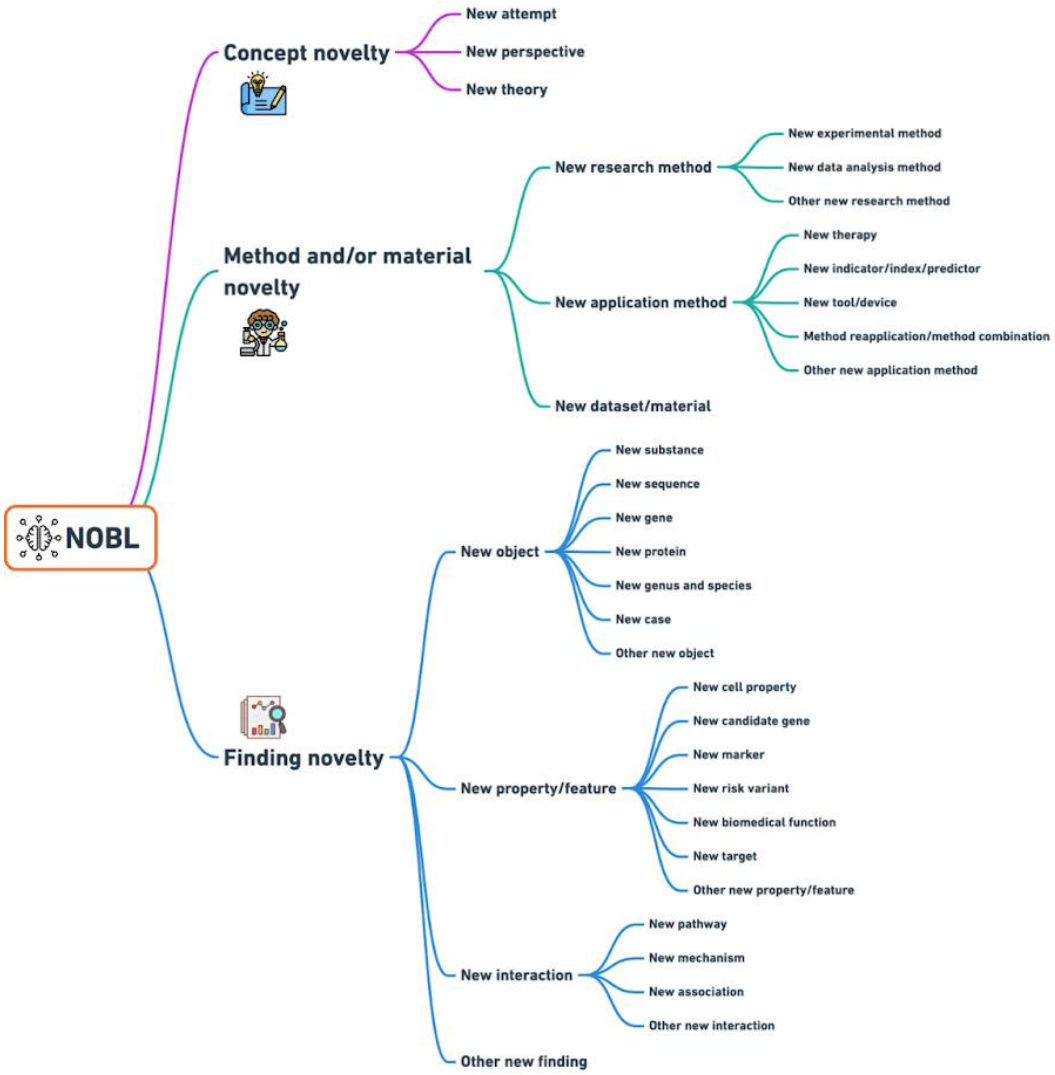


Fig. 1. The hierarchical structure of *NOBL*. Some icons were made by Vectors Tank, Uniconlabs, Eucalyp, sourced from www.flaticon.com.

2. *NOBL* Evaluation

To ensure the robustness and utility of *NOBL*, we performed a multi-faceted evaluation comprising structural, semantic, and coverage analyses.

2.1. Statistical and Topological Analysis

NOBL comprises 39 nodes distributed across three hierarchical levels (Table S4, Supplementary). It features 3 top-level nodes, 10 second-level nodes (including 5 leaf nodes and 5 internal nodes), and 26 leaf nodes at the third level, yielding a total of 31 leaf nodes and 8 internal nodes. The taxonomy reaches a maximum depth of three hierarchical levels, with an average depth of 2.84 and minimum of 2, reflecting a considered balance between structural granularity and cognitive tractability. The branching factor, the number of child nodes per internal node, ranges from 3 to 7, averaging 4.5, indicating a balanced yet diverse hierarchical structure. Leaf node distribution varies across top-level categories: "Concept novelty" contributes 3 leaf nodes (9.68%), "Method and/or material novelty" 9 (29.03%), and "Finding novelty" 19 (61.29%). This asymmetry reflects the relative complexity and emphasis of novelty types in the biomedical domain. Collectively, the structural profile of *NOBL* underscores its capacity to capture diverse biomedical innovations within a coherent and user-friendly hierarchical framework.

2.2. Semantic Evaluation

We assessed the semantic coherence of *NOBL* using an ontology validation framework that transforms hierarchical parent–child relationships into semantically equivalent natural language questions, for example, "*Is every new protein a new object?*" (Table S3, Supplementary). This approach enables evaluation of conceptual coherence by domain experts, independent of formal ontology syntax. Expert evaluation yielded a high rational agreement score of 0.81 (81% of questions were deemed rational by both evaluators) and a strong inter-evaluator agreement of 0.92 (for 92% of questions, both evaluators gave the same judgment either rational or irrational), suggesting robust alignment between the taxonomy's structure and expert understanding of biomedical novelty.

Discrepancies surfaced during evaluation were systematically resolved by refining the definitions and labels of novelty types. For instance, in response to the question "Is *every new candidate gene a new property/feature?*", evaluators identified ambiguity, prompting a revision of the definition for "New candidate gene." The updated description ("*Identifying a previously known gene as potentially associated with a particular trait, such as a disease or physical attribute, based on its genomic location or known function. Experimental or computational evidence suggests its role, making it a candidate for additional study.*") clarifies the functional nature of this novelty type and distinguishes it from the de novo identification encompassed by "New gene." These adjustments reflect the dynamic interplay between ontology structure and domain-specific semantic expectations, reinforcing the interpretability and practical applicability of *NOBL*.

2.3. Coverage Evaluation

To evaluate coverage, we applied *NOBL* to an independent set of 100 articles sampled from PubMed that were not included in the original corpus (details in 4.2.3.). Among sentences judged to express genuine novelty, the coverage score, which is defined as the proportion assigned to a well-defined novelty type excluding residual classes such as "Other new research method," "Other new application method," or "Other new finding," reached 0.88 (0.88 from both Annotator 1 and Annotator 2). The inter-annotator agreement was 0.86. Most sentences were mapped to a core novelty type rather than residual classes. These findings highlight the taxonomy's conceptual scope and its practical value in systematically capturing biomedical innovation.

3. *NOBL* Application

With the final validated version of *NOBL*, we conducted a systematic analysis of creativity patterns across the biomedical research domain and its subdomains, leveraging our *Biomedical Novelty Claim Corpus* of 12,701 PubMed articles (details in 4.1.1.). This analysis revealed several key insights into the prevalence and distribution of various types of novelty within biomedical science.

3.1. Distribution of Author-claimed Novelty Statements in Biomedical Articles

To investigate the nature of claimed novelties, we analyzed the distribution of *NOBL* categories on both high-level categories and their more granular subcategories to capture a comprehensive view of biomedical innovation patterns. Based on our *Biomedical Novelty Claim Corpus*, the *NOBL* distribution estimated across the entire PubMed is depicted in Fig. 2: "Concept novelty" (16.40%, red), "Method and/or material novelty" (32.86%, green), and "Finding novelty" (50.74%, blue). The dominance of "Finding novelty," particularly categories such as "New association" (5.75%) and "New mechanism" (6.95%), suggests a strong emphasis in current biomedical research on discovering novel interactions or relationships among existing entities. In contrast, "Concept novelty," encompassing foundational advances such as "New theory" (0.71%), was relatively rare. This aligns with expectations, as paradigm-shifting theories are inherently less frequent than applied or incremental innovations.

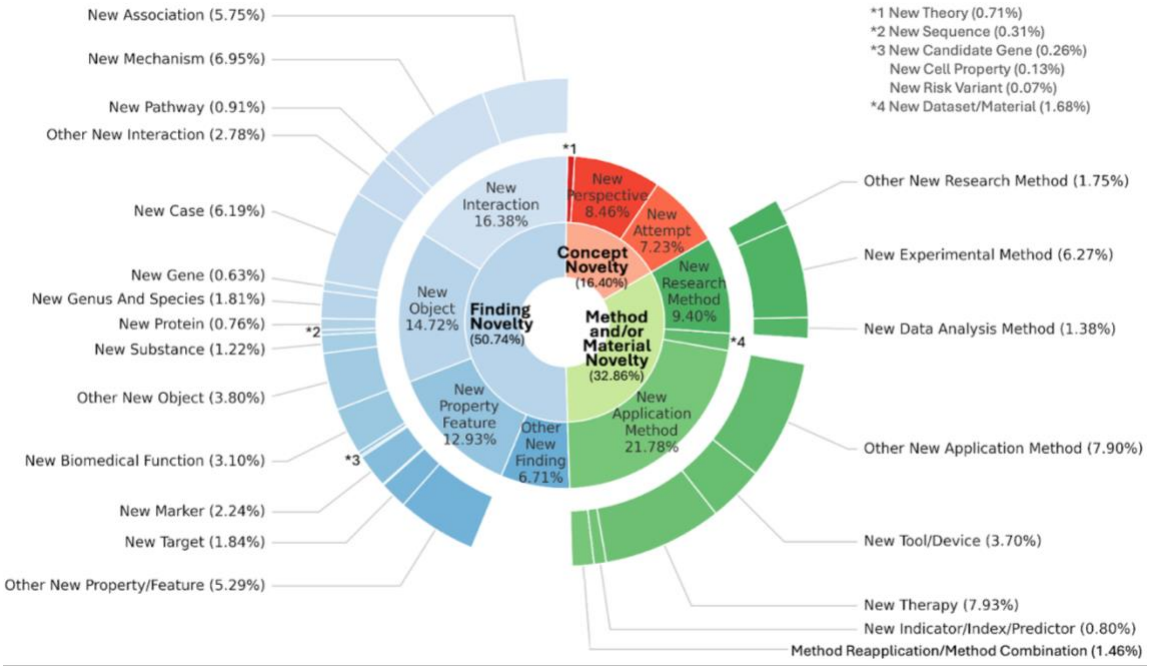


Fig. 2. The estimated global distribution of *NOBL* categories across the biomedical domain. Because the *Biomedical Novelty Claim Corpus* uses subdomain-balanced sampling, we adjusted its novelty distribution with subdomain prevalence statistics to estimate the true composition of novelty types across the entire PubMed database.

Overall, an estimated 6.47% of biomedical publications include explicit claims of novelty in their conclusion sections. However, this proportion varies substantially across subdomains. The five subdomains with the highest rates of author-claimed novelty are Microbiology (15.73%), Botany (14.29%), Parasitology (12.95%), Genetics, Medical (12.64%), and Chemistry (12.12%) (Table 1). In contrast, the lowest rates are observed in Primary Health Care (0.63%), Vital Statistics (1.02%), Library Science (1.09%), Family Planning Services (1.11%), and Hospitals (1.38%) (Table 1). A complete breakdown of novelty claim rates across all 130 biomedical subdomains is provided in Table S1 (Supplementary).

Subdomain	Articles Included † (*a*)	Articles with Innovation Keywords ‡ (*b*)	Sampled Articles § (*c*, from *b*)	Novel Articles After Review ‖ (*d*)	Estimated Novelty Rate ((*a* ÷ *b*) × (*c* ÷ *d*))
Microbiology	382,245	73,312	100	82	15.73%
Botany	155,789	26,494	100	84	14.29%
Parasitology	95,351	16,035	100	77	12.95%
Genetics, Medical	142,392	22,786	100	79	12.64%
Chemistry	682,805	114,964	100	72	12.12%
Nanotechnology	118,845	19,853	100	72	12.03%
Biotechnology	282,674	45,320	100	74	11.86%
Cell Biology	481,122	71,358	100	78	11.57%
Computational Biology	61,796	10,485	100	68	11.54%
Virology	143,450	23,182	100	70	11.31%
…	…	…	…	…	...
Audiology	42,383	2,255	100	36	1.92%
Obstetrics	191,676	17,350	100	21	1.90%
Orthodontics	175,66	1,132	100	26	1.68%
Photography	317	26	26	5	1.58%
Women's Health	16,159	1,068	100	22	1.45%
Hospitals	37,859	3,736	100	14	1.38%
Family Planning Services	4,746	440	100	12	1.11%
Library Science	1,933	234	100	9	1.09%
Vital Statistics	6,749	689	100	10	1.02%
Primary Health Care	50,198	3,980	100	8	0.63%
PubMed (All Subdomains) ¶	**19,606,567**	**2,204,267**	**12,701**	**6,086**	**6.47% #**

Table 1. Estimated Rate of Articles Containing Self-Claimed Innovation Across Biomedical Subdomains*

*For results across all 130 biomedical subdomains, see Table S1 in Supplementary.
†Number of articles included with extractable conclusion sections (alias: *a*).
‡Number of articles with conclusion sections containing innovation keywords (alias: *b*).
§Number of articles sampled for the *Biomedical Novelty Claim Corpus* from *b* (alias: *c*).
‖Number of articles annotated as genuinely novel (alias: *d*).
¶Refers to the entire PubMed database.
#Because the *Biomedical Novelty Claim Corpus* uses subdomain-balanced sampling, we adjusted the overall author-claimed novelty rate from subdomain-specific novelty rates with subdomain prevalence statistics to estimate the true novelty rate across PubMed.

3.2. Three Distinct Novelty Patterns Were Identified Across Biomedical Subdomains

To further explore domain-specific trends in biomedical innovation, we analyzed the *NOBL* profiles of 130 biomedical subdomains. For each subdomain, we first calculated its novelty distribution, represented as a three-dimensional vector of percentages corresponding to "Concept novelty," "Method and/or Material novelty," and "Finding novelty." These distributions, rather than the dominant novelty type alone, served as the input features for clustering. We then applied k-means unsupervised clustering with k = 3 on these vectors, which revealed three major clusters (Table 2 and Fig. 3). Each cluster was named according to its predominant novelty type: the Concept Novelty Cluster (33 subdomains; red dots), the Method and/or Material Novelty Cluster (41 subdomains;

green dots), and the Finding Novelty Cluster (56 subdomains; blue dots). For each subdomain, its *NOBL* profile is visualized as a horizontal stacked bar (Fig. 4), with purple, pink, and blue segments representing the proportions of "Concept novelty," "Method and/or material novelty," and "Finding novelty," respectively.

The Concept Novelty Cluster (red dots in Fig. 3; red cluster in Fig. 4) exhibited consistently elevated levels of "Concept novelty" relative to the global baseline of 16.40%, reaching a peak in Ethics (62.50%). These fields frequently prioritize theoretical and philosophical contributions over empirical discoveries. The Method and/or Material Novelty Cluster (green dots in Fig. 3; green cluster in Fig. 4) included subdomains such as Nuclear Medicine, Radiology, Biomedical Engineering, and Medical Informatics, which demonstrated markedly higher proportions of "Method and/or material novelty" (72.41%, 71.11%, 68.42%, and 68.00%, respectively) compared to the overall baseline of 32.86%. Advances in these areas typically involve the creation of datasets, development of research methods, or advancement of application tools. The Finding Novelty Cluster (blue dots in Fig. 3; blue cluster in Fig. 4) encompassed more traditional biomedical disciplines, such as Bacteriology, Internal Medicine, Embryology, Virology, and Communicable Diseases, where novelty stemmed primarily from empirical findings ("Finding novelty": 84.93%, 78.26%, 77.97%, 77.14% and 68.3%, respectively), exceeding the domain-wide baseline of 50.74%. These subfields often focus on elucidating disease mechanisms, identifying pathogenic processes, and discovering therapeutic targets.

Novelty Pattern Cluster	**#Subdomains**	**Concept novelty**	**Method and/or material novelty**	**Finding novelty**
Concept Novelty Cluster	33	**36.95%**	30.79%	32.26%
Method and/or Material Novelty Cluster	41	15.56%	**55.07%**	29.37%
Finding Novelty Cluster	56	13.38%	23.21%	**63.41%**
PubMed (All Subdomains)	**130**	**16.40%**	**32.86%**	**50.74%**

Table 2. Distribution of Novelty Types Across Three Distinct Novelty Pattern Clusters

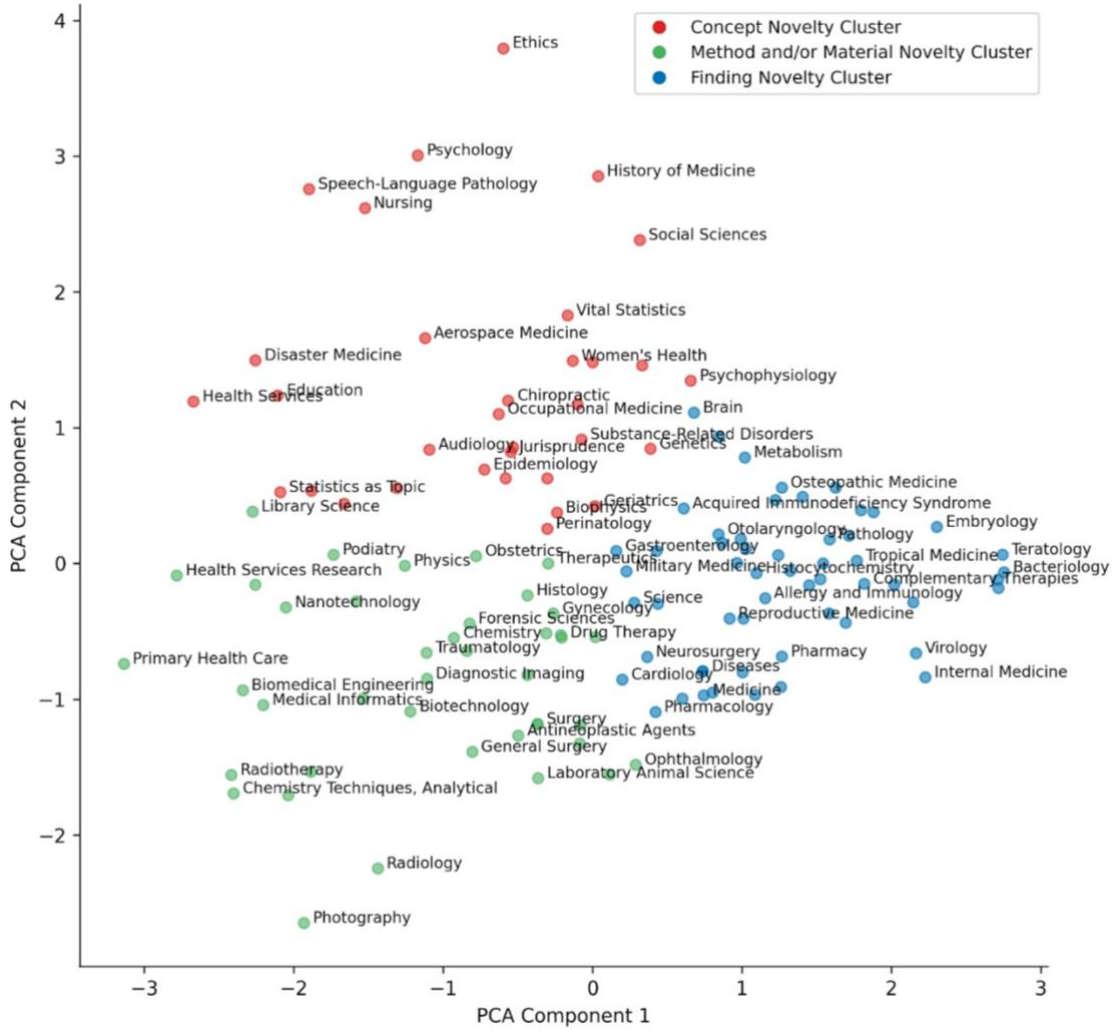
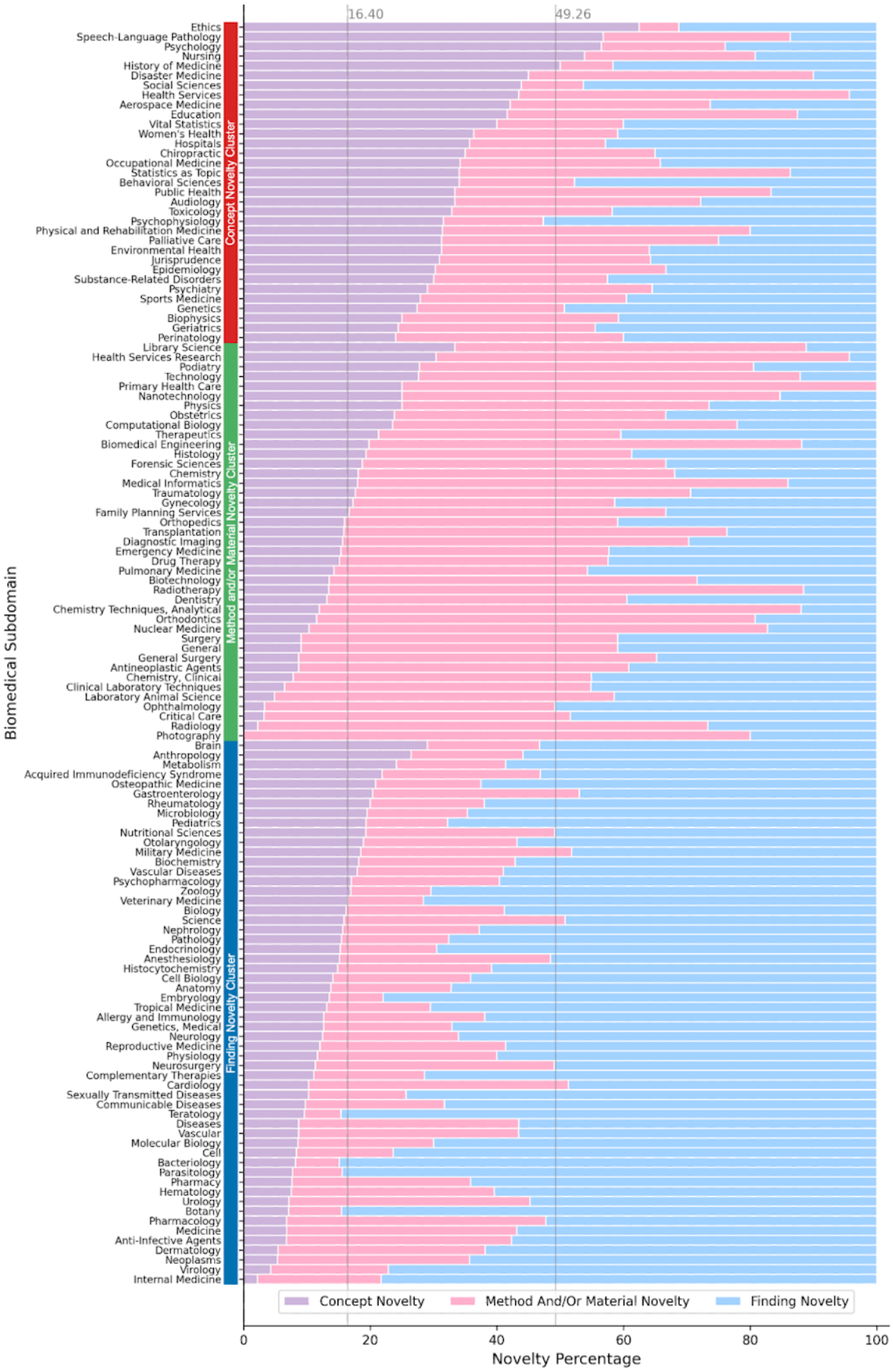


Fig. 3. Principal component analysis (PCA) of novelty patterns across biomedical subdomains. PCA was performed on the distribution of *NOBL* categories across 130 biomedical subdomains. The analysis reveals three distinct clusters, reflecting domain-specific innovation patterns. Each point represents a subdomain, color-coded by its assigned novelty pattern cluster: red for the Concept Novelty Cluster, green for the Method and/or Material Novelty Cluster, and blue for the Finding Novelty Cluster. To avoid excessive overlap, only a subset of labels is displayed.


16.40
49.26
Concept Novelty Cluster
Method and/or Material Novelty Cluster
Finding Novelty Cluster
Biomedical Subdomain
Ethics
Speech-Language Pathology
Psychology
Nursing
History of Medicine
Disaster Medicine
Social Sciences
Health Services
Aerospace Medicine
Education
Vital Statistics
Women's Health
Hospitals
Chiropractic
Occupational Medicine
Statistics as Topic
Behavioral Sciences
Public Health
Audiology
Toxicology
Psychophysiology
Physical and Rehabilitation Medicine
Palliative Care
Environmental Health
Jurisprudence
Epidemiology
Substance-Related Disorders
Psychiatry
Sports Medicine
Genetics
Biophysics
Geriatrics
Perinatology
Library Science
Health Services Research
Podiatry
Technology
Primary Health Care
Nanotechnology
Physics
Obstetrics
Computational Biology
Therapeutics
Biomedical Engineering
Histology
Forensic Sciences
Chemistry
Medical Informatics
Traumatology
Gynecology
Family Planning Services
Orthopedics
Transplantation
Diagnostic Imaging
Emergency Medicine
Drug Therapy
Pulmonary Medicine
Biotechnology
Radiotherapy
Dentistry
Chemistry Techniques, Analytical
Orthodontics
Nuclear Medicine
Surgery
General
General Surgery
Antineoplastic Agents
Chemistry, Clinical
Clinical Laboratory Techniques
Laboratory Animal Science
Ophthalmology
Critical Care
Radiology
Photography
Brain
Anthropology
Metabolism
Acquired Immunodeficiency Syndrome
Osteopathic Medicine
Gastroenterology
Rheumatology
Microbiology
Pediatrics
Nutritional Sciences
Otolaryngology
Military Medicine
Biochemistry
Vascular Diseases
Psychopharmacology
Zoology
Veterinary Medicine
Biology
Science
Nephrology
Pathology
Endocrinology
Anesthesiology
Histocytochemistry
Cell Biology
Anatomy
Embryology
Tropical Medicine
Allergy and Immunology
Genetics, Medical
Neurology
Reproductive Medicine
Physiology
Neurosurgery
Complementary Therapies
Cardiology
Sexually Transmitted Diseases
Communicable Diseases
Teratology
Diseases
Vascular
Molecular Biology
Cell
Bacteriology
Parasitology
Pharmacy
Hematology
Urology
Botany
Pharmacology
Medicine
Anti-Infective Agents
Dermatology
Neoplasms
Virology
Internal Medicine
Concept Novelty
Method And/Or Material Novelty
Finding Novelty
0
20
40
60
80
100
Novelty Percentage

246

Fig. 4. Distribution of *NOBL* categories across biomedical subdomains. A horizontal stacked bar plot illustrates the proportion of each *NOBL* type across 130 biomedical subdomains, revealing substantial variation in novelty composition and domain-specific innovation profiles. Each horizontal bar represents a subdomain, with segments color-coded by novelty type: purple for "Concept novelty," pink for "Method and/or material novelty," and blue for "Finding novelty." Vertical grey lines indicate the baseline proportions of each *NOBL* type from the full PubMed database. Vertical separators delineate three novelty pattern clusters: red for the Concept Novelty Cluster, green for the Method and/or Material Novelty Cluster, and blue for the Finding Novelty Cluster.

Discussion

In this study, we introduced *NOBL*, a new taxonomy for classifying novelty claims made by authors of biomedical literature. Developed through an iterative, data-driven, and LLM-assisted framework, *NOBL* is grounded in the *Biomedical Novelty Claim Corpus*, which is a domain-balanced dataset of 12,701 keyword-enriched conclusion sections from PubMed. We refined the taxonomy through expert review and large-scale LLM-assisted annotation with chain-of-thought prompting, supported by a web-based annotation platform. The final taxonomy underwent rigorous validation for structural clarity, semantic rationality, and coverage, and was applied across 130 biomedical subdomains to reveal domain-specific innovation patterns. We found that 6.47% of articles explicitly contained novelty claims, with "Finding novelty" being the most common. Clustering analysis further identified three major novelty profiles, each characterized by distinct innovation types, highlighting the taxonomy's effectiveness in uncovering diverse innovation dynamics across domains.

Our work addresses a fundamental gap in biomedical informatics and scientometrics: the lack of structured and scalable methods for identifying and analyzing types of scientific novelty. By offering a principled and interpretable taxonomy of innovation claims, *NOBL* bridges the disconnect between computational analysis and expert evaluation of novel scientific contributions. Our study offers five major advantages. First, *NOBL* provides the first systematic and fine-grained definition of biomedical novelty, addressing the longstanding lack of conceptual clarity and moving beyond simplistic binary judgments. Second, by leveraging a data-driven approach focusing on authors' novelty claims, our framework captures how researchers themselves define and present their contributions, offering a more direct and transparent perspective on innovation. This approach reduces interpretive ambiguity and supports scalable, reproducible analysis. Third, we release the *Biomedical Novelty Claim Corpus*, a large-scale, domain-stratified annotated dataset to support future research in novelty detection, literature curation, and trend analysis. Fourth, we present a scalable, LLM-assisted annotation framework embedded in our *web-based annotation platform*, accelerating taxonomy development and facilitating domain adaptation. Fifth, we demonstrate the practical value of *NOBL* by mapping innovation patterns across 130 biomedical subdomains, providing actionable insights into how novelty varies across disciplines.

For researchers, *NOBL* and its *web-based annotation platform* offer a practical framework to understand how their work fits into the broader novelty landscape and innovation patterns of their subdomain. By highlighting predominant novelty types in their research subfield, researchers can align their contributions with both established and emerging scientific directions, enhancing strategic positioning and potential impact. At the same time, identifying underrepresented novelty types reveals promising, underexplored areas well-suited for high-risk, high-reward investigations. These insights not only support more informed topic selection and project planning but also contribute to long-term career development. For reviewers and editors, *NOBL* introduces a unified vocabulary and structure for evaluating innovation claims. It promotes more transparent and consistent peer review by clarifying novelty types and enabling reproducible assessments. When integrated into editorial workflows, it helps streamline review criteria and supports fairer decision-

making. For policymakers and funding agencies, *NOBL* enables scalable monitoring of innovation across biomedical subdomains. By quantifying author-claimed novelty, it helps identify saturated areas and emerging trends, informing funding priorities and resource allocation. This supports a more balanced and evidence-based investment strategy across both foundational and frontier research.

Nonetheless, our study has limitations. First, author-claimed novelty, while reproducible, may be influenced by self-reporting bias, potentially omitting externally perceived innovation. Second, although our dataset is diverse and domain-balanced, it represents only a subset of PubMed. We plan to scale our analysis to the entire PubMed corpus. Third, our current annotation assigns one novelty label per sentence, which may oversimplify multi-faceted contributions. Future work will enable multi-label annotations to capture richer novelty structures.

We envision *NOBL* as a dynamic and extendable framework. To accommodate emerging innovation types, we include "other" categories as placeholders for future refinement. Looking forward, we aim to develop an automated, LLM-powered pipeline for continuous taxonomy updating and large-scale multi-label annotation. This would support real-time novelty mapping across the entire PubMed corpus, enhancing applications such as funding prioritization, novelty-aware peer review, and horizon scanning. Furthermore, we would explore integrating self-reported novelty with other novelty metrics, such as semantic novelty, to combine both qualitative and quantitative analysis into a more comprehensive framework for evaluating scientific innovation.

By operationalizing novelty as a structured, computable construct, *NOBL* lays the foundation for innovation-aware AI systems in science. It deepens our understanding of biomedical research dynamics and advances efforts to make novelty more transparent, measurable, and accountable. While developed in the biomedical domain, *NOBL* offers a generalizable approach that can be extended to other scientific disciplines such as materials science, computer science, and environmental studies. Its principles can guide the development of cross-domain discovery systems, intelligent literature recommenders, and data-driven grant evaluation tools.

Materials and Methods

In this section, we present a data-driven, iterative, and LLM-assisted approach to construct a biomedical author-claim novelty taxonomy, *NOBL*, aiming at enhancing our understanding of innovation within the biomedical research domain. The overall workflow of our proposed approach includes three phases: taxonomy development, evaluation, and application (Fig. 5).

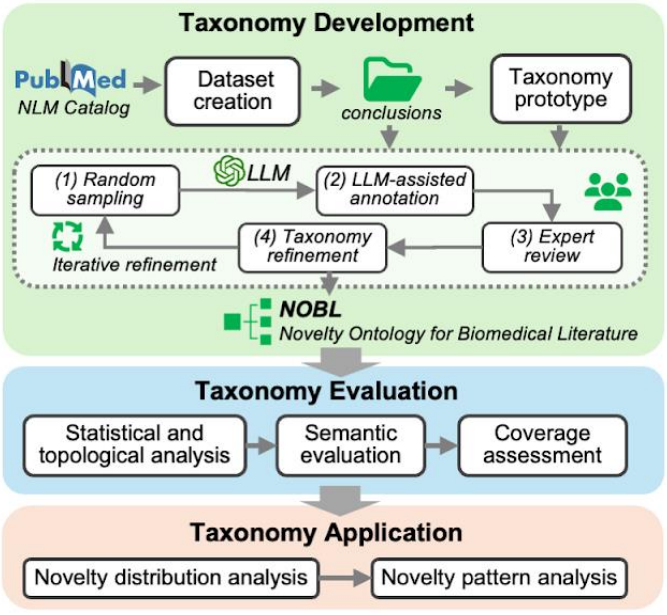


Fig. 5. The overall workflow.

1. Taxonomy Development

Following a data-driven approach, we curated a *Biomedical Novelty Claim Corpus* from PubMed conclusion section, serving as the foundation for our author-claim novelty taxonomy, *NOBL*. This corpus was assembled by collecting conclusion sections in which researchers explicitly claim novelty, guided by a set of novelty-related keywords. Author-claimed novelties were then extracted from these sections to populate and shape our *NOBL*. To streamline this otherwise labor-intensive and time-consuming process, we developed an iterative, LLM-assisted approach to identify and extract explicit novelty elements, enabling systematic and consistent *NOBL* refinement.

1.1. *Biomedical Novelty Claim Corpus* Curation

To ensure broad biomedical coverage, representativeness, and a high concentration of novelty claims, we employed a subdomain-balanced, keyword-guided sampling strategy. Drawing from PubMed articles, we extracted conclusion sections across 130 biomedical subdomains and filtered them using a refined set of novelty-related keywords. This yielded a corpus of 12,701 keyword-enriched conclusion sections, providing a representative and novelty-dense foundation for building *NOBL*.

PubMed served as the foundational data source for our study. We collected 36,502,590 articles published between 1781 and 2024 from PubMed,(*22*) downloaded from National Library of Medicine (NLM) via FTP on January 8, 2024. We utilized the pubmed_parser Python package to extract essential metadata elements, including PMID, title, abstract, publication year, and journal title.

Subdomains were delineated using 130 Broad Subject Terms (BSTs). Each journal in PubMed NLM Catalog is associated with a BSTs tag, which encompasses one or more Medical Subject Headings (MeSH) terms (*23*) and corresponds to a specific subdomain. We extracted the BSTs for all 36,431 journals in PubMed, resulting in the identification of 130 subdomains (Table S1, Supplementary). For example, the journal "Nature" is associated with the BST "Science," while the journal "The Lancet" is associated with the BST "Medicine." The BSTs were used to categorize the articles in the *Biomedical Novelty Claim Corpus* into their respective subdomains.

Conclusions were extracted as the key summarization of article contribution. The article metadata from PubMed contains two types of abstracts: structured and unstructured. Structured abstracts are organized into distinct sections, such as background, methods, results, and conclusions, which we can directly extract conclusions from. In contrast, the unstructured abstracts lack this organization, making it challenging to identify conclusions. To address this, we leveraged a BERT-based sentence classification tool (*24*) to identify conclusion sentences in unstructured abstracts. Sentences labeled as conclusion were then concatenated to obtain conclusion sections. Ultimately, we obtained 19,606,567 article conclusion sections, primarily due to the absence of abstracts in earlier papers. This total includes 4,665,642 conclusion sections from structured abstracts and 14,940,925 from unstructured abstracts.

Novelty-related keyword filtering strategy was applied to enhance the density of novelty claims in our corpus. We aimed to identify conclusion sections in which authors explicitly articulate novelty. For example, when a conclusion section contains keywords such as "novel" or "new," it is likely related to a specific novelty statement such as "novel method" or "new discovery," which are essential for taxonomy development. However, the presence of these keywords does not guarantee that the conclusion section articulates genuine novelty, such as "not novel" or "new patient." To address this, we adopted a heuristic refinement process to curate a high-precision list of novelty-related keywords, enabling more reliable filtering of conclusion sections for taxonomy development.

Novelty-related keywords were selected based on expert knowledge and empirical assessment of statement prevalence. We began with an initial list of 11 candidate keywords curated through expert discussion and literature review: "novel," "innovat," "breakthrough," "first," "new," "original," "uncommon," "unusual," "unexpected," "unprecedented," and "surpris." To evaluate their effectiveness, we manually reviewed a sample of keyword-flagged conclusion sections and assessed how often they conveyed explicit novelty claims. We finalized a set of four keywords including "novel," "innovat," "first," and "new" that are most indicative of genuine novelty claims (full details in Table S2, Supplementary).

Finally, we applied the refined set of novelty-related keywords to all extracted conclusion sections from PubMed and randomly sampled up to 100 keyword-flagged conclusion sections per subdomain. This resulted in a total of 12,701 conclusion sections comprising the *Biomedical Novelty Claim Corpus* used for *NOBL* construction. Some subdomains contained fewer than 100 eligible papers, and some papers were sampled across multiple subdomains.

1.2. Expert-driven LLM-Assisted *NOBL* Building

Building a comprehensive taxonomy of biomedical novelty from a large-scale dataset demands both expert insight and scalable annotation strategies beyond traditional manual approaches. To address this challenge, we developed a two-phase iterative framework that integrates domain expertise with LLM-assisted annotation for the creation and refinement of *NOBL*. The process begins with the creation of an expert-curated prototype and proceeds to large-scale refinement across the full corpus through a semi-automated, scalable pipeline.

Phase 1: Expert-Initiated Prototype Creation

We began by initiating a taxonomy prototype grounded in expert knowledge and prior literature on scientific novelty. This initial prototype was then refined through expert-led discussions and structured consensus meetings during a pre-annotation phase. Experts annotated representative samples and resolved disagreements through discussion, enabling clarification and alignment of novelty type definitions. The resulting taxonomy prototype served as the foundational framework for categorizing biomedical novelty and guided subsequent large-scale refinement efforts.

This taxonomy prototype creation process consists of two steps:

> **Step 1. Prototype initiation**: Leveraging domain expertise and insights, we initiated a hierarchical novelty taxonomy prototype and corresponding annotation guideline from existing works on novelty definitions. (*12–14*)
>
> **Step 2. Prototype refinement**: The initial prototype was applied to a representative subset (N = 300) from the *Biomedical Novelty Claim Corpus* by three experts (X.P., N.H., and R.S., with diverse backgrounds in biomedical science, information science, and computer science). The subset consisted of (1) 100 samples from the biology subdomain, (2) 100 from the medical subdomain, and (3) 100 randomly selected from other subdomains. Each sample was annotated with a novelty type from our taxonomy prototype by all three experts, with disagreements resolved through structured consensus meetings. Key revisions included clarification of ambiguous categories, introduction of new categories to improve coverage, merging of conceptually overlapping categories, elimination of redundant ones, and reorganization for enhanced structural clarity. These consensus-driven updates resulted in a refined and more robust taxonomy prototype.

Phase 2: Iterative Refinement with LLM Assistance

In this phase, we refined the taxonomy prototype across the entire *Biomedical Novelty Claim Corpus* (N = 12,701) using a consensus-driven, LLM-assisted framework designed for both scalability and precision.

To support high-throughput accurate annotation, we implemented two core innovations. First, we adopted a *chain-of-thought prompting* strategy (Fig. 6), which encouraged the language model to produce accurate and consistent annotations by incorporating the taxonomy structure, detailed novelty type definitions, and the target conclusion section into the prompt. Each output included both a predicted label and a structured justification in JSON format. Second, we developed a *web-based annotation platform* to streamline expert review (Fig. 7), significantly reducing manual effort while ensuring high-quality annotations. This tool accepts tab-separated (.TSV) files containing PMIDs and corresponding conclusion sections, and automatically generates novelty type suggestions from LLM using our tailored prompts and taxonomy definitions. Reviewers can accept these suggestions or make manual edits. More details and its source code are publicly available at GitHub repository (https://github.com/BIDS-Xu-Lab/novelty_taxonomy_for_biomedical_literature).

Input

System Prompt: You are an expert specializing in biomedical research analysis tasked with classifying sentences from PubMed abstracts based on their novelty.
User Prompt: Your task is to analyze the provided plaintext conclusion of a PubMed paper, focusing on segments where the authors explicitly claim novelty in their own research work using strong indicators such as "first," "novel," "innovation," or "new," where these terms directly refer to their own novel contributions in the whole content. The response should be formatted as JSON following the provided JSON below:

```
{
 "category": "Primary Label (only label) - Relevant Sub-label - Relevant Sub-sub-label",
 "reason": "Stepwise reasoning, first determine whether the text should be classified as "4 Not Novelty" based on the specified criteria. Then classify into one of the leaf node of the taxonomy tree."
}
Let's think step by step.
### Taxonomy:
{taxonomy}
### Input:
{conclusion}
### Output:
```

Output (Example PMID: 25234876)

```
{
 "category": "2 Method and/or Material Novelty - 2.2 New application method - 2.2.3 New tool/device",
 "reason": "The sentence explicitly mentions the development of an 'innovative tool,' which indicates the creation of a new tool or device. This aligns with the category of Method and/or Material Novelty, specifically under New application method as a New tool/device."
}
```

Fig. 6. The chain-of-thought prompting strategy with an example output.

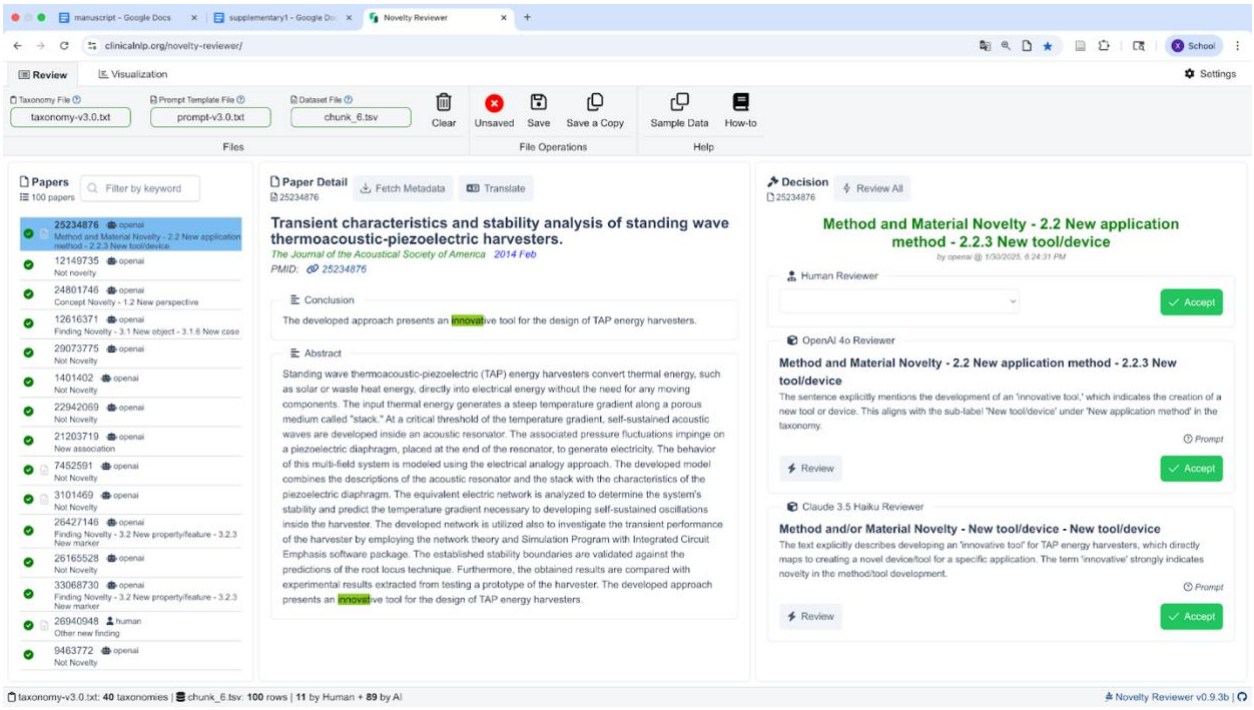


Fig. 7. The web-based annotation platform.

The taxonomy refinement process proceeded through the following steps:

Step 1. Sample Generation: For each round of annotation, we randomly selected a set of 500 unlabeled conclusion sections from the entire *Biomedical Novelty Claim Corpus*.

Step 2. Automated Labeling by LLM: We employed GPT-4o (*21*) in our *web-based annotation platform* to label the selected conclusion sections based on the latest version of the taxonomy and the *chain-of-thought prompt*. The LLM was instructed to assign a label using the taxonomy and provide a detailed explanation for each conclusion section, resulting in 500 annotated conclusion sections per iteration.

Step 3. Expert Review and Refinement: Experts reviewed the LLM-generated labels, corrected errors, and proposed further improvements to the taxonomy based on the types of errors observed. This step ensured the taxonomy remained accurate and adaptable across diverse biomedical subdomains.

Step 4. Consensus-Driven Updates: Weekly meetings were held where experts discussed and approved proposed changes. Only updates unanimously agreed upon were incorporated into the revised taxonomy and annotation guideline.

Step 5. Iterative Refinement: The process was repeated, starting from Step 1, to annotate the next chunk of conclusion sections from the corpus using the latest revised taxonomy. This iterative cycle continued until the entire corpus was annotated.

In the first few rounds of refinement, the taxonomy and the corresponding annotation guideline were updated weekly, incorporating changes such as adding new categories, merging similar ones, and refining novelty type definitions. With a higher consensus on both the taxonomy itself and the annotation guideline was achieved, updates transitioned to a monthly schedule, focusing on resolving edge cases and ensuring consistency. Throughout this process, we assessed the performance of the LLM by comparing its outputs against expert-reviewed annotations. The F1-score improved from 0.5367 to 0.8410, demonstrating the reliability of the LLM-assisted approach in refining the taxonomy. The entire annotation and refinement process spanned approximately six months. The final taxonomy comprises 3 top-level categories and 31 leaf-level subcategories,

effectively capturing a broad spectrum of novelty types in biomedical research. The final version of *NOBL* with its complete definitions and annotation guidelines are provided in Supplementary.

2. Taxonomy Evaluation

To assess the comprehensiveness and representational fidelity of the proposed *NOBL*, we conducted a multi-faceted evaluation comprising statistical and topological analysis, semantic evaluation, and coverage evaluation. These evaluations were performed by three domain experts (X.P., N.H., and R.S.) and two independent reviewers (V.K. and L.Q.) who were not involved in the taxonomy development.

2.1. Statistical and Topological Analysis

We first examined the structural properties of *NOBL* using standard Ontology Metrics (*25*, *26*). Statistical characteristics included the number and distribution of nodes, while key topological features encompassed depth (quantified as the minimum, maximum, and average number of hierarchical levels from the root to leaf nodes) and width (measured by the minimum, maximum, and average branching factors across all levels). Together, these metrics provided a quantitative summary of the taxonomy's structural complexity and semantic granularity.

2.2. Semantic Evaluation

To assess the semantic coherence of *NOBL*, we utilized the ontology evaluation tool Hootation (*27*), which converts formal ontology relationships into semantically equivalent natural language questions (Table S3, Supplementary). Each parent–child relationship in *NOBL* was translated into a corresponding question to evaluate subsumption rationality. Two independent biomedical informatics experts (V.K. and L.Q., not involved in the taxonomy development) assessed these questions and provided judgments on whether each relationship was semantically rational. We computed two agreement metrics: (1) inter-evaluator agreement, defined as the proportion of questions for which both evaluators gave the same judgment (rational or irrational), and (2) rational agreement, defined as the proportion of questions deemed rational by both evaluators. Revisions to *NOBL* were informed by evaluator comments and areas of disagreement. These refinements included clarifying novelty type definitions, refining label names to reduce ambiguity, and resolving disagreements through structured discussions to ensure improved semantic alignment.

2.3. Coverage Evaluation

To assess the concept coverage of *NOBL*, we assembled an out-of-sample evaluation dataset comprising 100 conclusion sections randomly sampled from PubMed articles not included in the original *Biomedical Novelty Claim Corpus*. Two domain experts (X.P. and N.H., both involved in the taxonomy development) independently annotated this dataset using *NOBL*. We computed two metrics: (1) coverage, defined as the proportion of sentences, among those identified as articulating genuine novelty, that could be classified into one of the predefined novelty labels, excluding residual classes such as "Other new research method", "Other new application method", or "Other new finding"; and (2) inter-annotator agreement, defined as the proportion of annotations in which both experts assigned the same label. This assessment provides a quantitative measure of the taxonomy's representational breadth within the biomedical literature, focusing on well-defined categories.

3. Taxonomy Application

We applied *NOBL* to the entire *Biomedical Novelty Claim Corpus* to analyze the author-claimed novelty rate and hierarchical distribution of novelty types across the biomedical research landscape and its constituent subdomains. We defined the author-claimed novelty rate as the proportion of conclusion sections in which novelty is explicitly claimed, relative to the total number of conclusion sections. Because of the subdomain-balanced sampling of our corpus, we adjusted the overall author-claimed novelty rate and the global *NOBL* distribution with subdomain-specific prevalence

statistics to estimate the original values for the whole PubMed database. While the subdomain-level author-claimed novelty rate and distribution were drawn directly from the corresponding subset of *Biomedical Novelty Claim Corpus*. To further investigate domain-specific innovation patterns, we performed k-means unsupervised clustering analysis with k = 3 across 130 biomedical subdomains based on each subdomain's *NOBL* profile. This analysis revealed three dominant clusters: (1) Concept Novelty Domains, (2) Method and/or Material Novelty Domains, and (3) Finding Novelty Domains. These clusters illuminate characteristic novelty signatures across biomedical subfields and highlight the value of *NOBL* for large-scale, structured novelty analysis.

References

1. T. S. Kuhn, The Structure of Scientific Revolutions (University of Chicago Press, Chicago, IL, 3rd ed., 1996).
2. C. C. Gillispie, The Nature of Science. Science 138, 1251–1253 (1962).
3. J. E. Hallsworth, Z. Udaondo, C. Pedrós-Alió, J. Höfer, K. C. Benison, K. G. Lloyd, R. J. B. Cordero, C. B. L. de Campos, M. M. Yakimov, R. Amils, Scientific novelty beyond the experiment. Microb. Biotechnol. 16, 1131–1173 (2023).
4. M. Packalen, J. Bhattacharya, Age and the Trying Out of New Ideas. J. Hum. Cap. 13, 341–373 (2019).
5. M. Packalen, J. Bhattacharya, "New Ideas in Invention" (National Bureau of Economic Research, Cambridge, MA, 2015); https://doi.org/10.3386/w20922.
6. M. H. N. Le, P. K. Nguyen, T. P. T. Nguyen, H. Q. Nguyen, D. N. H. Tam, H. H. Huynh, P. K. Huynh, N. Q. K. Le, An in-depth review of AI-powered advancements in cancer drug discovery. Biochim. Biophys. Acta Mol. Basis Dis. 1871, 167680 (2025).
7. A. Gangwal, A. Lavecchia, Unleashing the power of generative AI in drug discovery. Drug Discov. Today 29, 103992 (2024).
8. C. Chen, Y. Chen, M. Horowitz, H. Hou, Z. Liu, D. Pellegrino, Towards an explanatory and computational theory of scientific discovery. J. Informetr. 3, 191–209 (2009).
9. S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, A.-L. Barabási, Science of science. Science 359, eaao0185 (2018).
10. S. Kaplan, K. Vakili, The double-edged sword of recombination in breakthrough innovation. Strateg. Manag. J. 36, 1435–1457 (2015).
11. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical Combinations and Scientific Impact. Science 342, 468–472 (2013).
12. E. Leahey, J. Lee, R. J. Funk, What Types of Novelty Are Most Disruptive? Am. Sociol. Rev. 88, 562–597 (2023).
13. Y. Yan, S. Tian, J. Zhang, The impact of a paper's new combinations and new components on its citation. Scientometrics 122, 895–913 (2020).
14. F. Shi, J. Evans, Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. Nat. Commun. 14, 1641 (2023).
15. S. Mishra, V. I. Torvik, Quantifying Conceptual Novelty in the Biomedical Literature. - Lib Mag. Mag. Digit. Libr. Forum 22 (2016).
16. M. C. Gilbert, C. S. Lerose, A. J. Conith, R. C. Albertson, Breaking constraints: The development and evolution of extreme fin morphology in the Bramidae. Evol. Dev. 24, 109–124 (2022).
17. L. Ma, H. Guo, Y. Zhao, Z. Liu, C. Wang, J. Bu, T. Sun, J. Wei, Liquid biopsy in cancer current: status, challenges and future prospects. Signal Transduct. Target. Ther. 9, 336 (2024).
18. J. D. O'Flaherty, M. Barr, D. Fennell, D. Richard, J. Reynolds, J. O'Leary, K. O'Byrne, The cancer stem-cell hypothesis: its emerging role in lung cancer biology and its relevance

for future therapy. J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer 7, 1880–1890 (2012).
19. J. Y. Wang, J. A. Doudna, CRISPR technology: A decade of genome editing is only the beginning. Science 379, eadd8643 (2023).
20. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
21. R. Islam, O. M. Moushi, GPT-4o: The Cutting-Edge Advancement in Multimodal LLM. Institute of Electrical and Electronics Engineers (IEEE) [Preprint] (2024). https://doi.org/10.36227/techrxiv.171986596.65533294/v1.
22. K. Canese, S. Weis, PubMed: The Bibliographic Database. NCBI Handb. (2013).
23. C. E. Lipscomb, Medical Subject Headings (MeSH). Bull. Med. Libr. Assoc. 88, 265 (2000).
24. Y. Hu, Y. Chen, H. Xu, Towards More Generalizable and Accurate Sentence Classification in Medical Abstracts with Less Data. J. Healthc. Inform. Res. 7, 542–556 (2023).
25. A. Agárdi, L. Kovács, Property-Based Quality Measures in Ontology Modeling. Appl. Sci. 12, 12475 (2022).
26. M. Batet, D. Sánchez, "A Semantic Approach for Ontology Evaluation" in 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (2014; https://ieeexplore.ieee.org/document/6984466/), pp. 138–145.
27. M. Amith, F. J. Manion, M. R. Harris, Y. Zhang, H. Xu, C. Tao, Expressing Biomedical Ontologies in Natural Language for Expert Evaluation. Stud. Health Technol. Inform. 245, 838–842 (2017).

Acknowledgments

Funding: The authors acknowledge that they received no funding in support for this research.

Author contributions:

Conceptualization: HX, NH, XP

Methodology: XP, VK

Data: XP, YH

Annotation: XP, NH, RS, LQ, VK

Investigation: XP, NH

Visualization: XP, HH

Supervision: HX, NH, QM, BO, EL, JH, QX

Writing—original draft: XP, NH, HH

Writing—review & editing: HX, NH, QM, BO, EL, AF

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: All data are available in the main text or the supplementary materials. The *Biomedical Novelty Claim Corpus* is available on our Hugging Face repository:

https://huggingface.co/datasets/clinicalnlplab/biomedical_novelty_claim_corpus. The complete codebase, including the *web-based annotation platform*, is provided on our GitHub repository: https://github.com/BIDS-Xu-Lab/novelty_taxonomy_for_biomedical_literature.

Science Advances
AAAS

Supplementary Materials for

How Researchers Claim Novelty in Biomedical Science: A Taxonomy for Understanding Innovation

Xueqing Peng *et al.*

*Corresponding author. Email: hua.xu@yale.edu, na.hong@yale.edu

This PDF file includes:

Supplementary Text - Taxonomy Definition and Example

1 Concept Novelty

Introducing fundamentally new directions, interpretations, or theoretical frameworks in biomedical scientific research. It includes first-time explorations, fresh perspectives on existing knowledge, and novel theoretical frameworks, expanding scientific thought beyond new findings or methods.

1.1 New attempt

Exploring a new scientific research content which has not been explored before. (Includes phrases like 'first study', 'first investigation', 'first trial', etc.)

- PMID21983299: This is the first report to evaluate the reinstatement of food-seeking in mGlu5 receptor KO mice.
- PMID21726459: This study provides the first detailed phylogeographic investigation of a widespread species whose distribution spans virtually all of the major biogeographic barriers in eastern Australia.

1.2 New perspective

Introducing a fresh subjective perspective or unique way of interpreting the objective findings. (Includes phrases like 'new insights', 'new consideration', etc.)

- PMID36058144: This work defines the transcriptional programme associated with post-acute pulmonary sequelae and provides novel insights for targeted interventions and biomarker development.
- PMID32902946: This work may unfold a new area for highly efficient radioprotection by polymeric drugs instead of small-molecular agents.

1.3 New theory

Developing or proposing a novel theory that offers an original framework for understanding phenomena in a biomedical field.

- PMID16768790: We propose a new model of tissue-specific actions of octopamine, in which strengthening of peritoneal sheath contractions, coupled with relaxation of the oviduct, eases ovulation.
- PMID18559114: We, therefore, propose a new model involving the interdependent evolution of C2H2-ZNF gene subfamilies.

2 Method and/or Material Novelty

Developing new research methods, applications, or materials, including novel experimental or analytical techniques, new tools/devices, datasets, or the application of known methods in unique ways to address specific scientific challenges.

2.1 New research method

Creating a novel approach to conduct research, whether through experimental, analytical, or observational means.

2.1.1 New experimental method

Introducing a new way to conduct experiments, including techniques, procedures, or protocols.

- PMID20122672: The present injectable system using (15)O(2)-HbV was successfully utilized to measure CMRO(2) in rats, indicating that this new method could be useful for animal models to measure oxygen metabolism in the brain.
- PMID20346870: An improved two-step and an innovative one-step reaction for synthesizing (18)FECH in a high yield were reported.

2.1.2 New data analysis method

Developing a novel analytical technique for interpreting research data, including statistical methods, data science methods, algorithms, and models.

- PMID27107720: The newly derived method is suitable for calculating the exact amount of intra-family variation of the estimated breeding values and genetic values (comprising additive and dominance effects).
- PMID23001152: We present a new empirical Bayes shrinkage estimate of the dispersion parameters and demonstrate improved DE detection.

2.1.3 Other new research method

Other new research method that cannot be classified into the above categories.

- PMID24462077: This novel model of consensus conference allows the construction of consensual, evidence-based, explicit recommendations for therapies in a process that may also identify issues for further research, eventually fostering progress in the field.
- PMID21899420: Together, these two novel approaches enable us to attack heretofore intractable problems, such as phylogenetic inference for high-resolution vertebrate genomes, as we demonstrate on a set of six vertebrate genomes with 8,380 syntenic blocks.

2.2 New application method

Applying/translating research findings into practice, or combining existing knowledge or methods in innovative ways to form new interdisciplinary approaches.

2.2.1 New therapy

Developing a new therapeutic approach.

- PMID34113099: This study offers a novel strategy to provide safe and effective microRNAs-delivery nanoplatforms based on carbon dots for promising cardiac regeneration and disease therapy.
- PMID30213635: This study's results suggest that LEV associated with PPB could represent a new therapeutic strategy to reduce long-term consequences induced by SE that facilitate pharmacoresistant SRS.

2.2.2 New indicator/index/predictor

Defining a novel metric for evaluation or assessment, with specific application scenarios demonstrating its utility.

- PMID24890391: These results suggest that the TMS may be useful as a new reliable method for scoring locomotor function for SCI models.
- PMID28852426: This study validated a new 15-item FAAM-Sp ADL and FAAM-Sp Sport subscales, which can be used as a self-reported outcome measure in clinical practice and research for patients resident in Spain whose main language is Spanish.

2.2.3 New tool/device

Creating a new tool, device, or instrument for applications.

- PMID31825553: Our results could expand knowledge of the pheromones of plant bugs, and provide novel technologies to monitor and control this pest. ¬© 2019 Society of Chemical Industry.
- PMID25856858: The novel syringe facilitates cell transplantation across the subretina and extravascular spaces of the choroid using a minimally-invasive procedure. Human BM-MSC transplantation using this system ameliorates retinal degeneration in the animal model. This new transplantation system may increase the therapeutic effect of other cell-based therapies and therapeutic agents.

2.2.4 Method reapplication/method combination

Adapting an existing method to a new context.

- PMID21916149: This research is the first in Egypt to apply the re-engineering approach to public health systems.
- PMID27536102: We describe the novel application of cationic bolasomes to deliver ASOs into bacteria.

2.2.5 Other new application method

Other new application method that cannot be classified into the above categories.

- PMID24874812: Combining the unique structural and biochemical properties of native acellular scaffolds with subsequent recellularization techniques offers a novel platform for organ engineering and regeneration, for experimentation ex vivo and potential clinical application in vivo.

2.3 New dataset/material

Creating or curating a novel dataset or material essential for specific research.

- PMID32333361: We have created the first accurately annotated, non-synthetic, dataset of hip fluoroscopy.
- PMID25747341: In addition, this paper presents a new public online dataset which is made available to the research community with the aim of providing a common evaluation framework to overcome some of the current limitations identified in this survey.

3 Finding Novelty

Discovering new results and knowledge, including new objects, properties, and relationships within the biomedical domain.

3.1 New object

Discovering a new object or element that has not been previously identified.

3.1.1 New substance

Identification of a new chemical or biological substance.

- PMID23822618: The effect of this substance is not separate or additive but acts as a newly formed substance.

3.1.2 New sequence

Discovery of a novel nucleotide or amino acid sequence without requiring known biological function, which may be part of a gene or exist in non-coding regions.

- PMID15291097: The other 2 sequences are most probably new sequences not existing in GenBank DNA database, which have been registered in the GenBank DNA database as new sequences.

3.1.3 New gene

Discovery of a novel gene with a specific biological function, confirmed to participate in a genetic or cellular process. This includes both the sequence and its functional role in the organism.

3.1.4 New protein

Identification of a previously unknown protein, including its unique sequence, structure, and specific role in biological processes. This innovation focuses on the protein's functional and structural contributions within cells.

- PMID20447282: PaCRT is a new calcium-binding protein that interacts with potyviral HC-Pro.

3.1.5 New genus and species

Discovering a new genus or species.

- PMID30225172: The new genus and species Kwanzacetus khoisani shares a series of morphological features with Inia geoffrensis, including the combination of a frontal boss with nasals being lower on the anterior wall of the vertex, the laterally directed postorbital

process of the frontal, the anteroposterior thickening of the nuchal crest, and robust teeth with wrinkled enamel.

3.1.6 New case

The identification or documentation of a newly diagnosed patient with a disease that has been observed for the first time.

- PMID21862756: To our knowledge, the present report is the first description of a case of 3 primary cancers that includes a mesothelioma of the tunica vaginalis.

3.1.7 Other new object

Other new object that cannot be classified into the above categories.

- PMID21414775: Our study highlights the potential utilities of these two novel endophytic fungi as biodiesel feedstock.

3.2 New property/feature

Identification of a new attribute or characteristic of an existing object.

3.2.1 New cell property

Discovering a new functional or structural property of a cell.

- PMID19956550: The newly discovered immunological properties of NSPCs may have implications in assigning a new role of these cells as non-professional antigen presenting cells in the central nervous system.

3.2.2 New candidate gene

Identifying a previously known gene as potentially associated with a particular trait, such as a disease or physical attribute, based on its genomic location or known function. Experimental or computational evidence suggests its role, making it a candidate for additional study.

- PMID20157514: This study supports the candidacy of CD36 as a novel susceptibility gene for neovascular AMD. Replication of our results in other populations will provide further convincing evidence for the genetic association.

3.2.3 New marker

Discovering a novel marker with diagnostic or prognostic potential.

- PMID25993985: Our data suggest that expression of TGF-Œ≤1 is a novel prognostic marker for intrahepatic cholangiocarcinoma.

3.2.4 New risk variant

Identification of a new genetic variant linked to disease risk.

- PMID25003827: Through a systematic case-control study of microRNA binding site SNPs, we identified a new breast cancer risk variant rs8752 in HPGD in Chinese women. Further studies are warranted to investigate the underling mechanism for this association.

3.2.5 New biomedical function

Identifying a previously unknown role for an existing substance, for example, recognizing it as a new receptor ligand or regulator.

- PMID20137882: These results suggest that the -1C allele of the CD40 (-1C/T) gene polymorphism is a genetic factor that may determine an individual's susceptibility by ACS in Chinese. The CD40 -1C/T polymorphism is a novel regulator of CD40 expression.
- PMID32958054: These results demonstrated a novel REC8 function that suppressed tumor angiogenesis and progression by attenuation of VEGF in gastric cancer microenvironment.
- PMID2554665: In conclusion, P-8502 appears to be a new mu selective opioid receptor ligand, whereas P-8511 has no such selectivity.

3.2.6 New target

Identifying a new molecular or cellular target for therapeutic intervention.

- PMID36493879: The results reveal a novel molecular target that could be further explored for developing preventive, predictive, and individualized treatment strategies for UCB.

3.2.7 Other new property/feature

Other new property/feature that cannot be classified into the above categories.

- PMID9130710: These results provide the first biochemical evidence that a given ErbB receptor has distinct signaling properties depending on its dimerization.

3.3 New interaction

Identifying a new interaction or relationship among known objects, including both direct and indirect interactions.

3.3.1 New pathway

Discovery of a previously unknown pathway.

- PMID18388194: These results suggest that this new identified oxidative stress-FOXO3a-TR4 pathway is a fundamentally important mechanism regulating stress resistance and cell survival.

3.3.2 New mechanism

Identifying a novel mechanism underlying a process.

- PMID36057401: The present study demonstrates that a new mechanism for GSK-3Œ≤-mediated anti-fibrotic function in renal fibrosis through phosphorylation of SIRT6 to prevent its proteasomal degradation.

3.3.3 New association

Discovering a new correlation among variables.

- PMID19295951: This ecological study represents, to the authors' knowledge, the first statistically significant association between increasing rates of LNG EC distribution and decreasing abortion rates.

3.3.4 Other new interaction

Other new interaction that cannot be classified into the above categories.

- PMID19798981: For the first time, the authors confirmed that the direct interaction has occurred between heme-iron of myoglobin and additional metal ions, and saw about how the metal ions intension affects the direct interaction.

3.4 Other new finding

Other new finding that cannot be classified into the above categories. Uncovering new evidence that reveals or supports a scientific finding, or documenting rare events or conditions.

- PMID18783470: This clinical study showed for the first time the evidence of poor recovery times for the diabetic foot with neuropathy when assessing the foot under load. A temperature deficit (because of poor recovery to baseline temperature) suggests degeneration of thermoreceptors, leading to diminished hypothalamus-mediated activity in the diabetic neuropathic group.

Supplementary Text - Annotation Guideline

General rules:

1. Assign the novelty classification at the lowest layer based on our novelty taxonomy definition.
2. Focus on the self-mention novelty (keyword mention) rather than the entire conclusion sentence.
3. For papers with multiple novelties, annotate only the main novelty.
    a. In summary, we reveal, for the first time, a novel mechanism of circMTO1/miR-6893 in tumorigenesis and chemoresistance of cervical cancer. Our findings support the notion of developing the new therapies and biomarkers by targeting circMTO1 for cervical cancer. — New mechanism
4. Potential novelty directly related to one's own contribution should be included, even if it is not yet completed. E.g. may, might, should, could, possible, potential, etc.
    a. This may suggest an innovative approach to the treatment of pain in clinical practice. — New therapy
    b. BBOX1-AS1 facilitated the EC development and malignancy via miR-361-3p/COL1A1 axis, indicating BBOX1-AS1 could be a novel therapy target for the diagnostic of EC. — New target
    c. This provides novel clues as to the possible molecular herbicidal mechanism of berberine. — New mechanism
    d. Together, our results suggest for the first time that the genetic variants in NTRK2 may regulate eGFR. — New association
5. Reviews should be excluded.
    a. This review provides novel ideas and important aspects for the future research of ginsenosides for treating respiratory diseases. — Not Novelty
    b. RNAi is a very promising strategy that in principle will provide many new targets against HIV infection. The mechanism of sequence complementarity utilized by siRNAs against their targets provides a new approach to fight against HIV infection. However, this technology still needs many fine refinements before its potential for HIV treatment strategies can be utilized. This review discusses the possibilities of using siRNA as a therapeutic tool for HIV treatment. — Not Novelty
6. Mentions of keywords without novelty should be excluded.
    a. The development of DIVA-vaccines in the past 10 to 15 years has created, in principle, an excellent response instrument to counter intentional animal disease outbreaks. These developments have made our animal agriculture less vulnerable to agroterrorism. But we cannot relax; there are still many challenges, in particular with respect to integration of first line of defense, law enforcement, and early detection systems for animal diseases. — Not Novelty
    b. The 70% of screening rate in the group of pregnant women is still far from reaching the goal of 95% intended in the Plan de Salud Infantil in Navarra for 1993. A prevalence of 0.70 for positive HBsAg was observed. No geographical differences were observed in the distribution of HBV markers, but its prevalence was influenced by the type of residence, social class and risk exposure. A discussion follows whether a massive immunization program form newborns in Navarra is warranted. — Not Novelty

Specifical rules:

1. For a “first” study, if it is the first attempt at exploring this research content, assign the tag “New attempt”. Otherwise, assign tags based on the specific focus of the study, such as “Other new finding”.
    a. However, due to the complexity in defining a "traditional Sami" diet, and the limitations of our questionnaire for this purpose, the study should be considered exploratory, a first attempt to relate a "traditional Sami" dietary pattern to health endpoints. — New attempt
    b. This is the first study to demonstrate an association between NIRS-based StO2 measurements and peristaltic activity visualized by ultrasound imaging. — New association
    c. To our knowledge this is the first report of a metazoan metaphase-anaphase transition being delayed in response to DNA damage. — Other new finding
    d. This is the first study to show that KPs can help with fracture healing by promoting osteogenic differentiation, and it also suggests that KPs can be used as a nutritional supplement to accelerate fracture healing. — Other new finding
2. For all tags, we should prioritize specific tags first, followed by others such as “Other new research method”, “Other new application method”, “Other new object”, “Other new property/feature”, “Other new interaction”, and “Other new finding”.
    a. This novel model of consensus conference allows the construction of consensual, evidence-based, explicit recommendations for therapies in a process that may also identify issues for further research, eventually fostering progress in the field. — Other new research method
    b. The direct 3D printing method is tested and validated for various microfluidic components that can be combined on a chip depending on the specific requirements of the experiment. The ease of use and production opens up the potential of microfluidics to a wide range of users, especially in biomedical research. Our demonstration of its use as a cytotoxicity screening system and as an assay for osteoblastic differentiation shows the methods potential in the development of novel biomedical applications. With the presented method, we aim to disseminate microfluidics as a standard method in biomedical research, thus improving the reproducibility and transferability of results to clinical applications. — Other new application method
    c. Our study highlights the potential utilities of these two novel endophytic fungi as biodiesel feedstock. — Other new object

Subdomain	Articles Included † (*a*)	Articles with Innovation Keywords ‡ (*b*)	Sampled Articles § (*c*, from *b*)	Novel Articles After Review ‖ (*d*)	Estimated Novelty Rate ((*a* ÷ *b*) × (*c* ÷ *d*))
Microbiology	382245	73312	100	82	15.73%
Botany	155789	26494	100	84	14.29%
Parasitology	95351	16035	100	77	12.95%
Genetics, Medical	142392	22786	100	79	12.64%
Chemistry	682805	114964	100	72	12.12%
Nanotechnology	118845	19853	100	72	12.03%
Biotechnology	282674	45320	100	74	11.86%
Cell Biology	481122	71358	100	78	11.57%
Computational Biology	61796	10485	100	68	11.54%
Virology	143450	23182	100	70	11.31%
Biomedical Engineering	294199	43473	100	76	11.23%
Molecular Biology	759160	120690	100	70	11.13%
Biochemistry	1111216	157250	100	77	10.90%
Genetics	193885	28508	100	73	10.73%
Zoology	112378	16356	100	71	10.33%
Chemistry Techniques, Analytical	159981	22619	100	67	9.47%
Biophysics	381140	46785	100	76	9.33%
Biology	603016	80589	100	68	9.09%
Anti-Infective Agents	65398	9611	100	59	8.67%
Bacteriology	40330	4724	100	73	8.55%
Science	714115	95652	100	63	8.44%
Allergy and Immunology	515731	64121	100	63	7.83%
Embryology	76403	9642	100	59	7.45%
Neoplasms	828368	107985	100	56	7.30%
Hematology	226829	30476	100	53	7.12%
Pathology	252068	25204	100	71	7.10%
Physics	141601	14285	100	68	6.86%
Complementary Therapies	82782	8957	100	63	6.82%
Histology	33751	3701	100	62	6.80%
Tropical Medicine	105035	11660	100	61	6.77%
Cell	14097	1299	100	72	6.63%
Veterinary Medicine	283654	27986	100	67	6.61%
Toxicology	308630	30384	100	67	6.60%
Clinical Laboratory Techniques	62467	6470	100	62	6.42%
Technology	112123	12374	100	58	6.40%
Histocytochemistry	23396	2021	100	74	6.39%
Metabolism	201525	22131	100	58	6.37%
Antineoplastic Agents	45278	6230	100	46	6.33%
Therapeutics	175306	23524	100	47	6.31%
Endocrinology	318185	33007	100	59	6.12%
Environmental Health	595480	56424	100	64	6.06%
Brain	225941	21987	100	62	6.03%
Diagnostic Imaging	208539	19251	100	64	5.91%
Neurology	947688	99664	100	56	5.89%
Ophthalmology	337471	31815	100	61	5.75%
Physiology	516362	49320	100	60	5.73%
Reproductive Medicine	174132	16972	100	58	5.65%
Dermatology	159407	16207	100	55	5.59%
Medical Informatics	102200	11361	100	50	5.56%
Pharmacology	615396	76259	100	44	5.45%
Rheumatology	110392	11695	100	50	5.30%
Teratology	6816	677	100	52	5.16%
Chemistry, Clinical	29371	2932	100	51	5.09%
Anatomy	59120	5170	100	58	5.07%
Internal Medicine	86426	9357	100	46	4.98%
Nephrology	128825	12549	100	51	4.97%

Nutritional Sciences	325670	27982	100	57	4.90%
Medicine	1324712	144724	100	44	4.81%
Drug Therapy	176474	25418	100	33	4.75%
Psychophysiology	144059	11806	100	57	4.67%
Pharmacy	61759	7361	100	39	4.65%
Communicable Diseases	208821	23322	100	41	4.58%
Laboratory Animal Science	8730	972	100	41	4.56%
Nuclear Medicine	86547	6738	100	58	4.52%
Gastroenterology	316469	29026	100	49	4.49%
Neurosurgery	120990	10180	100	53	4.46%
Psychopharmacology	64716	5956	100	47	4.33%
Radiotherapy	68930	5698	100	52	4.30%
Diseases	1563	141	100	46	4.15%
Vascular	1563	141	100	46	4.15%
Forensic Sciences	26960	2113	100	48	3.76%
Sexually Transmitted Diseases	29703	2851	100	39	3.74%
Vascular Diseases	382780	36130	100	39	3.68%
Jurisprudence	48978	4258	100	42	3.65%
Cardiology	485225	45282	100	39	3.64%
Urology	183501	15868	100	42	3.63%
Transplantation	119159	11323	100	38	3.61%
Ethics	26195	2875	100	32	3.51%
Social Sciences	80715	6769	100	41	3.44%
General Surgery	707878	51587	100	46	3.35%
History of Medicine	24559	3414	100	24	3.34%
Geriatrics	133671	9888	100	45	3.33%
Podiatry	5144	468	100	36	3.28%
Orthopedics	262331	19233	100	44	3.23%
Statistics as Topic	52358	3830	100	44	3.22%
Behavioral Sciences	183589	13408	100	44	3.21%
Psychology	303438	20691	100	46	3.14%
Anthropology	15466	1421	100	34	3.12%
Radiology	249848	17298	100	45	3.12%
Perinatology	52890	6548	100	25	3.10%
Acquired Immunodeficiency Syndrome	50930	4896	100	32	3.08%
Pediatrics	436550	42700	100	31	3.03%
General	10476	710	100	44	2.98%
Surgery	10476	710	100	44	2.98%
Pulmonary Medicine	265321	21754	100	35	2.87%
Otolaryngology	148552	11367	100	37	2.83%
Substance-Related Disorders	69101	4814	100	40	2.79%
Critical Care	74202	6284	100	31	2.63%
Dentistry	281047	19408	100	38	2.62%
Public Health	350935	29475	100	30	2.52%
Psychiatry	353249	27859	100	31	2.44%
Epidemiology	114675	8480	100	33	2.44%
Gynecology	217450	18065	100	29	2.41%
Palliative Care	34775	2617	100	32	2.41%
Sports Medicine	90004	5016	100	43	2.40%
Education	75944	7539	100	24	2.38%
Health Services Research	104966	10682	100	23	2.34%
Traumatology	124668	8556	100	34	2.33%
Aerospace Medicine	12047	721	100	38	2.27%
Physical and Rehabilitation Medicine	93517	6063	100	35	2.27%
Speech-Language Pathology	32557	1616	100	44	2.18%
Chiropractic	2281	124	100	40	2.17%
Health Services	144445	13055	100	23	2.08%
Nursing	248145	19658	100	26	2.06%
Occupational Medicine	59290	3213	100	38	2.06%
Osteopathic Medicine	1984	166	100	24	2.01%

Disaster Medicine	6695	668	100	20	2.00%
Emergency Medicine	70029	5335	100	26	1.98%
Military Medicine	19966	1448	100	27	1.96%
Anesthesiology	113472	6666	100	33	1.94%
Audiology	42383	2255	100	36	1.92%
Obstetrics	191676	17350	100	21	1.90%
Orthodontics	17566	1132	100	26	1.68%
Photography	317	26	26	5	1.58%
Women's Health	16159	1068	100	22	1.45%
Hospitals	37859	3736	100	14	1.38%
Family Planning Services	4746	440	100	12	1.11%
Library Science	1933	234	100	9	1.09%
Vital Statistics	6749	689	100	10	1.02%
Primary Health Care	50198	3980	100	8	0.63%
PubMed (All Subdomains) ¶	**19606567**	**2204267**	**12701**	**6086**	**6.47%** #

Table S1. Estimated Rate of Articles Containing Self-Claimed Innovation Across Biomedical Subdomains

†Number of articles included with extractable conclusion sections (alias: *a*).
‡Number of articles with conclusion sections containing innovation keywords (alias: *b*).
§Number of articles sampled for the *Biomedical Novelty Claim Corpus* from *b* (alias: *c*).
||Number of articles annotated as genuinely novel (alias: *d*).
¶Refers to the entire PubMed database.
#Because the *Biomedical Novelty Claim Corpus* uses subdomain-balanced sampling, we adjusted the overall author-claimed novelty rate from subdomain-specific novelty rates with subdomain prevalence statistics to estimate the true novelty rate across PubMed.

Keyword	Keyword Detected Rate †	Novelty Detected Rate ‡	Included
novel	**1,847/100,000**	**78/100**	**yes**
innovat	**271/100,000**	**52/100**	**yes**
breakthrough	43/100,000	6/100	no
first	**3,971/100,000**	**67/100**	**yes**
new	**4,076/100,000**	**63/100**	**yes**
original	254/100,000	26/100	no
uncommon	306/100,000	2/100	no
unusual	170/100,000	42/100	no
unexpected	139/100,000	5/100	no
unprecedented	23/100,000	17/100	no
surpris	80/100,000	3/100	no

Table S2. Novelty-Related Keywords and Their Association With Genuine Novelty

†Number of papers containing the keyword out of 100,000 randomly sampled conclusion sentences from the PubMed database.

‡Number of papers identified as genuinely novel (based on expert annotation) among 100 randomly sampled papers that contained the keyword.

Semantically equivalent natural language question	Evaluator 1	Evaluator 2
Every new interaction is a finding novelty	yes	yes
Every new substance is a new object	yes	yes
Every new gene is a new object	yes	yes
Every new tool/device is a new application method	yes	yes
Every new sequence is a new object	yes	yes
Every other new object is a new object	yes	yes
Every new experimental method is a new research method	yes	yes
Every new association is a new interaction	no	no
Every new data analysis method is a new research method	yes	yes
Every new dataset/material is a method and/or material novelty	yes	yes
Every new perspective is a concept novelty	yes	no
Every new property/feature is a finding novelty	yes	yes
Every new pathway is a new interaction	no	yes
Every other new research method is a new research method	yes	yes
Every new target is a new property/feature	yes	yes
Every new candidate gene is a new property/feature	no	no
Every new biomedical function is a new property/feature	yes	yes
Every new indicator/index/predictor is a new application method	no	no
Every new protein is a new object	yes	yes
Every method reapplication/method combination is a new application method	yes	yes
Every new case is a new object	yes	no
Every new object is a finding novelty	yes	yes
Every other new interaction is a new interaction	yes	yes
Every other new property/feature is a new property/feature	yes	yes
Every new research method is a method and/or material novelty	yes	yes
Every other new finding is a finding novelty	yes	yes
Every new mechanism is a new interaction	yes	yes
Every new genus and species is a new object	yes	yes
Every new risk variant is a new property/feature	yes	yes
Every new cell property is a new property/feature	yes	yes
Every new attempt is a concept novelty	no	no
Every other new application method is a new application method	yes	yes
Every new therapy is a new application method	yes	yes
Every new theory is a concept novelty	yes	yes
Every new application method is a method and/or material novelty	yes	yes
Every new marker is a new property/feature	yes	yes

Table S3. Semantic Evaluation

Ontology Metrics	Value
Number of Nodes	
First-level	3 (Leaf nodes †: 0, Internal nodes ‡: 3)
Second-level	10 (Leaf nodes: 5, Internal nodes: 5)
Third-level	26 (Leaf nodes: 26, Internal nodes: 0)
Total Number	39 (Leaf nodes: 31, Internal nodes: 8)
Depth of Leaf Nodes	
Minimum	2
Maximum	3
Mean §	2.84
Width of Internal Nodes (Branching Factors) ‖	
Minimum	3
Maximum	7
Mean ¶	4.5
Distribution of Leaf Nodes in Main Categories	
Concept Novelty	3 (9.68%)
Method and Material Novelty	9 (29.03%)
Finding Novelty	19 (61.29%)
Total Number	31 (100%)

Table S4. Statistical and Topological Characteristics of *NOBL*

†Leaf nodes: Nodes without children.

‡Internal nodes: Nodes with children.

§Mean depth of leaf nodes: The mean depth of all leaf nodes. (Total Depth ÷ Number of Leaf Nodes = 88 ÷ 31 ≈ 2.84; Total Depth = (Number of Leaf Nodes at Level 2 × 2) + (Number of Leaf Nodes at Level 3 × 3) = (5 × 2) + (26 × 3) = 88)

‖Width of internal nodes (branching factors): The number of direct child nodes for each internal node.

¶Mean width of internal nodes: The mean width of all internal nodes. (Total Width ÷ Number of Internal Nodes = 36 ÷ 8 = 4.5).