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

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

1. 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.

2. Results

2.1. *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 2. *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.

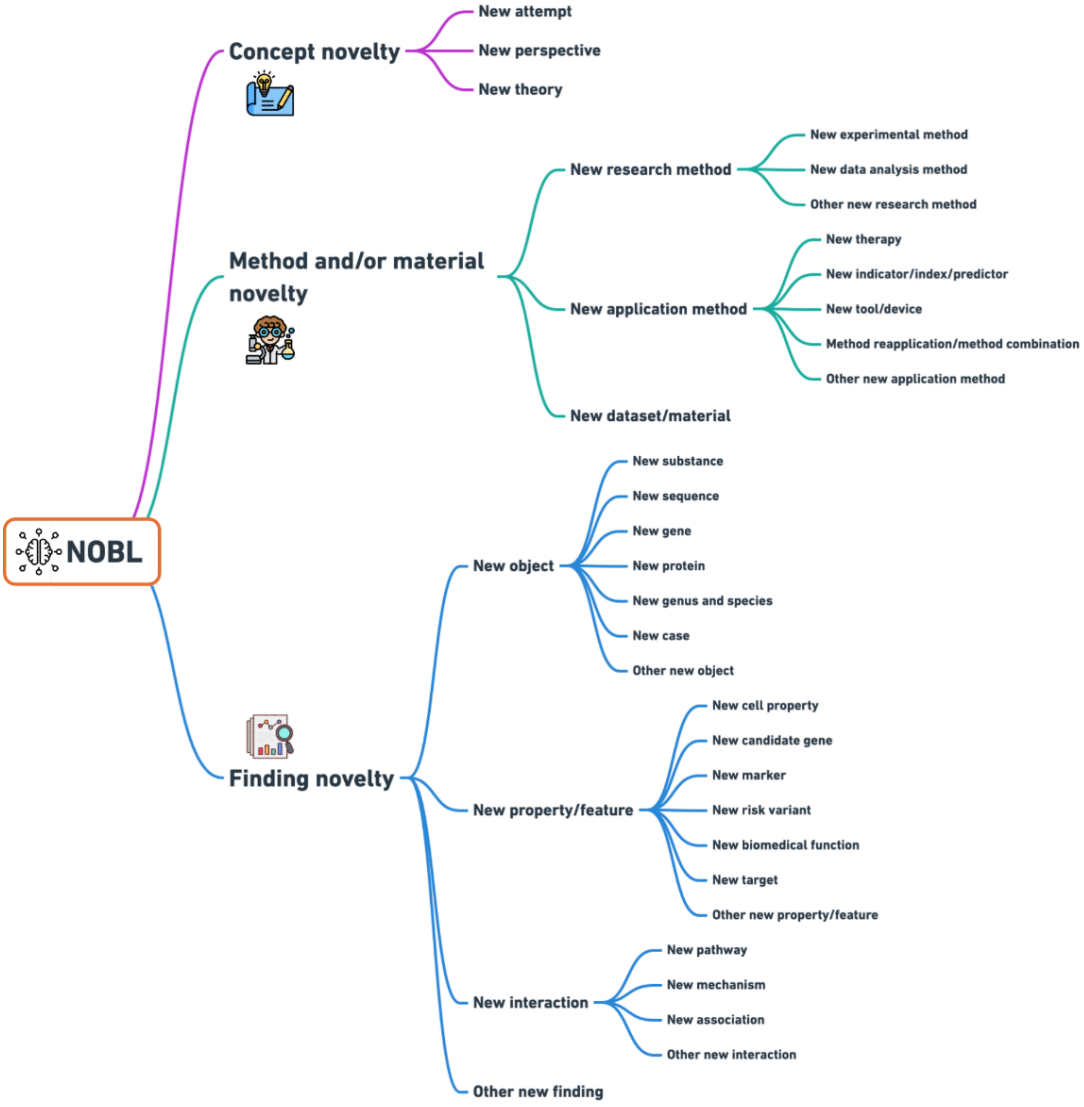

Fig. 1. The hierarchical structure of *NOBL*. Some icons were made by Vectors Tank, Uniconlabs, Eucalyp, sourced from www.flaticon.com.

2.2. *NOBL* Evaluation

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

2.2.1. Statistical and Topological Analysis

NOBL comprises 39 nodes distributed across three hierarchical levels (Table S4, Supplementary 1). 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.

2.2.2. 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 1). 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*.

2.2.3. 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.

2.3. *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.

2.3.1. 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.

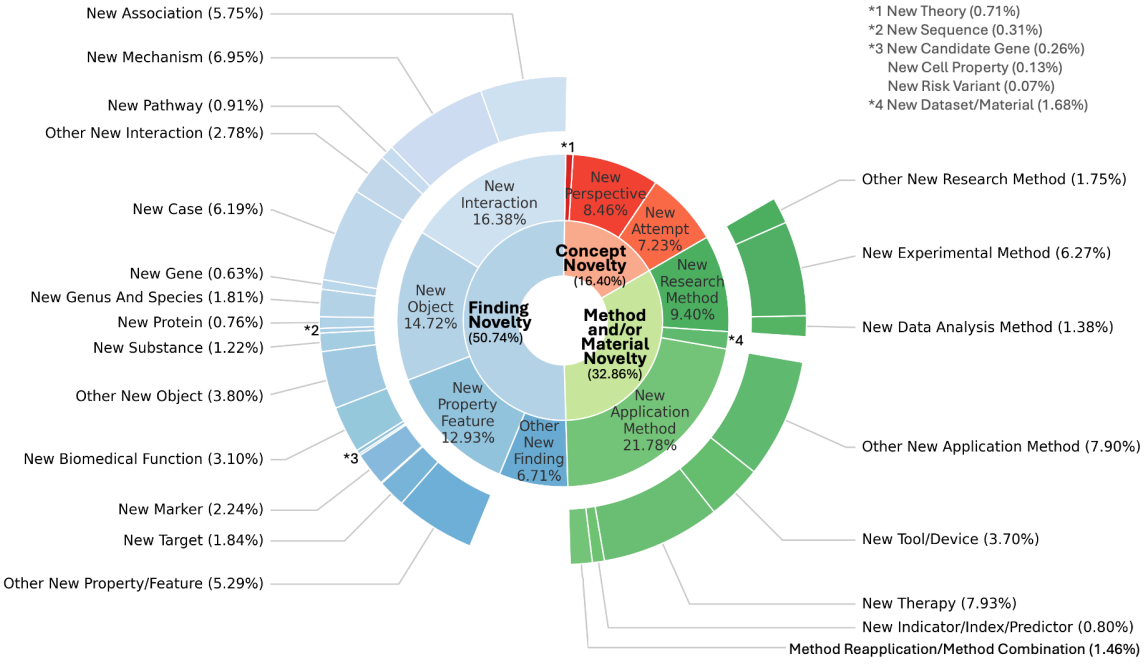

Fig. 2. The estimated global distribution of *NOBL* categories across the biomedical domain. Because the *Biomedical Novelty Claim Corpus* uses subdomain-balanced sampling, we adjusted its novelty distribution with subdomain prevalence statistics to estimate the true composition of novelty types across 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 1).

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 1.

†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.

2.3.2. 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

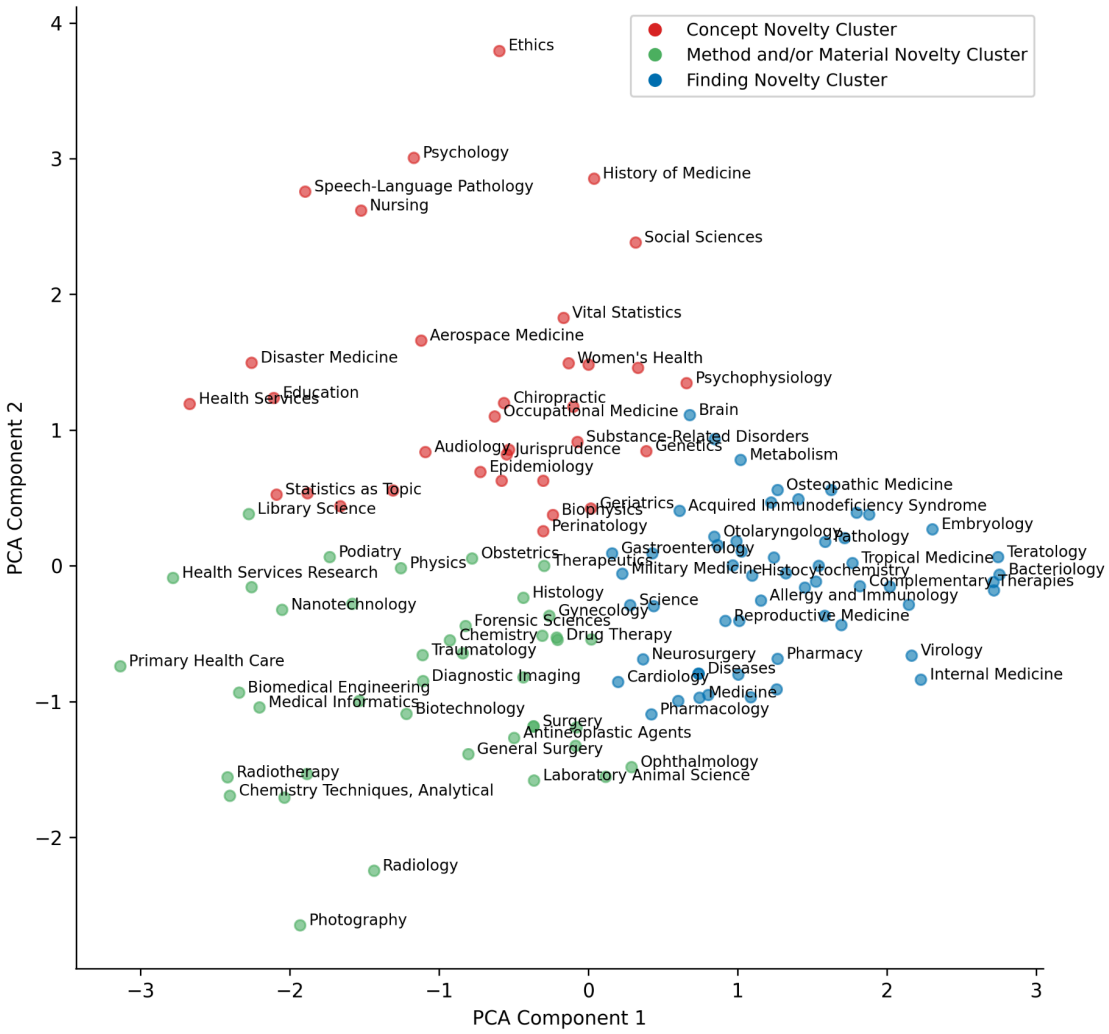

Fig. 3. Principal component analysis (PCA) of novelty patterns across biomedical subdomains. PCA was performed on the distribution of *NOBL* categories across 130 biomedical subdomains. The analysis reveals three distinct clusters, reflecting domain-specific innovation patterns. Each point represents 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.

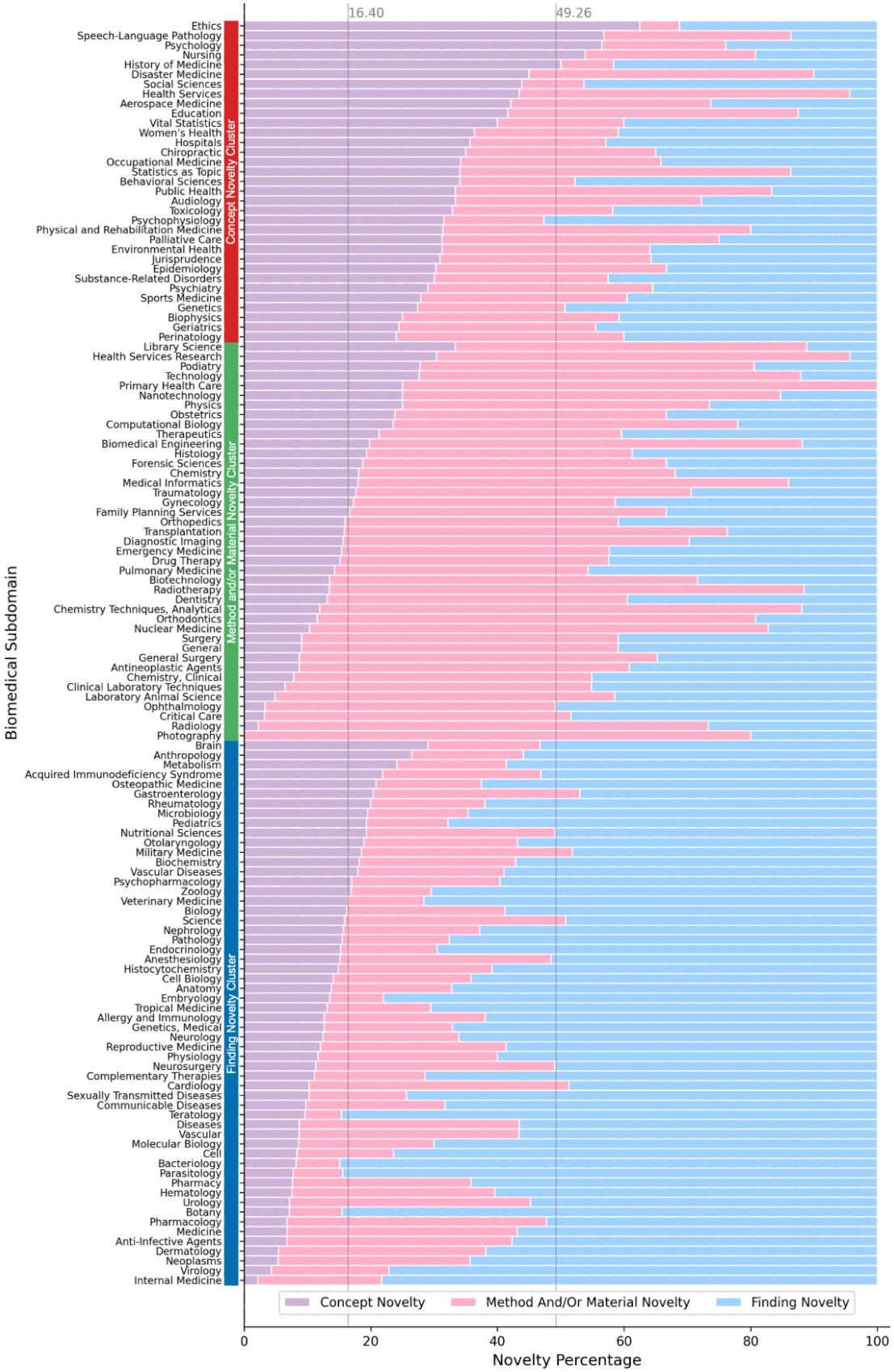
16.40
49.26
Concept Novelty Cluster
Method and/or Material Novelty Cluster
Finding Novelty Cluster
Biomedical Subdomain
Ethics
Speech-Language Pathology
Psychology
Nursing
History of Medicine
Disaster Medicine
Social Sciences
Health Services
Aerospace Medicine
Education
Vital Statistics
Women's Health
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

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.

3. 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.

4. Materials & 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).

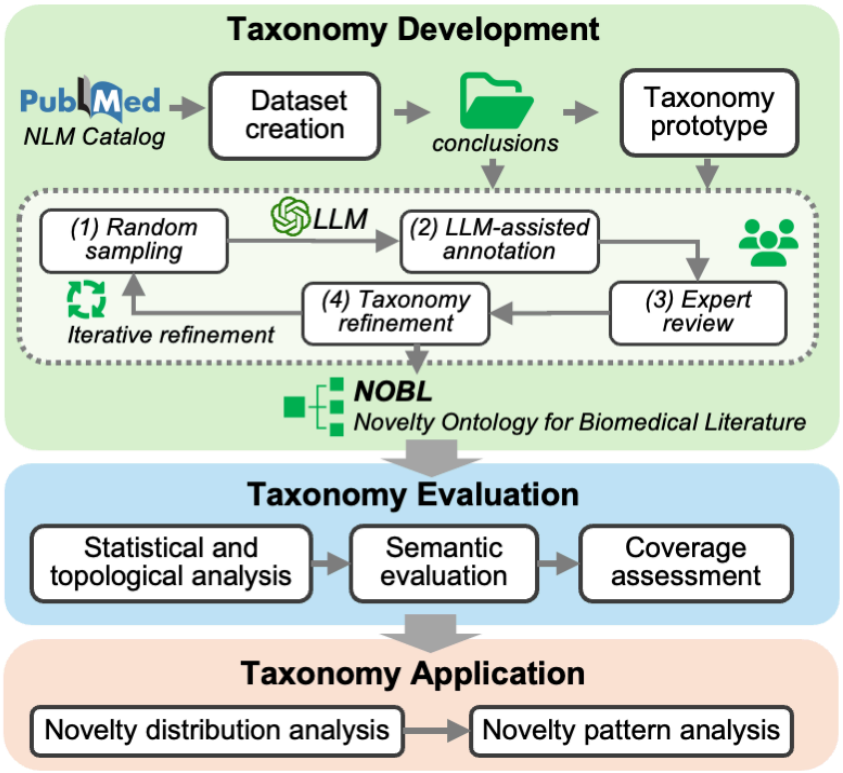

Fig. 5. The overall workflow.

4.1. 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.

4.1.1. *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 1). 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 1).

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.

4.1.2. 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.

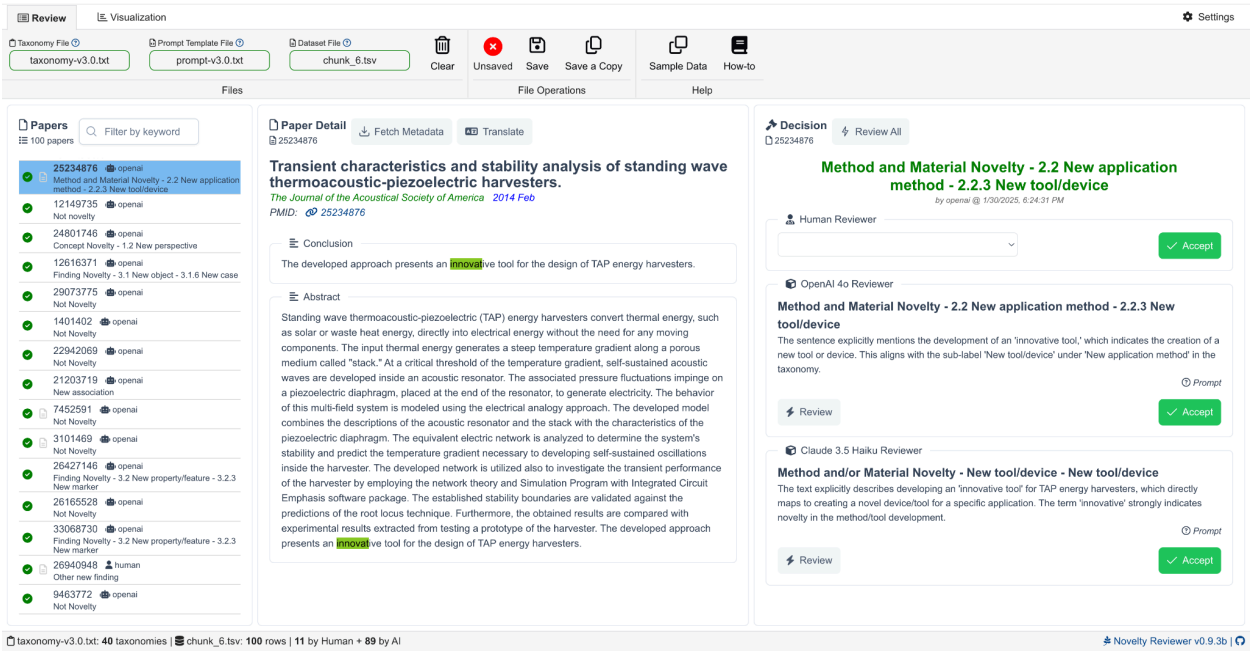

Fig. 7. The web-based annotation platform.

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 and 3.

4.2. 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.

4.2.1. 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.

4.2.2. 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 1). 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.

4.2.3. 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.

4.3. 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.

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.