computation

MDPI

Article

Mining Scholarly Impact: A Career-Adaptive Higher-Order Aggregation Framework for Multi-Domain Researcher Classification

Hafiza Zarafshan Mukhtiar [1], Muhammad Tanvir Afzal [2,3,4], Ahmad Sami Al-Shamayleh [5], Ghulam Mustafa [2,*], Isabel De la Torre Díez [6], Eduardo Silva Alvarado [7,8,9,10,11] and Helena Garay [7,9,12]

1 Department of Computing, Air University Kamra, Attock 43570, Pakistan; zarafshan.cs@aack.au.edu.pk
2 Faculty of Computing, Shifa Tameer-e-Millat University, Islamabad 44000, Pakistan; tanvir.afzal@gisma.com
3 Department of Computing and Data Science, GISMA University of Applied Sciences, 14469 Potsdam, Germany
4 Department of Computer Science, Gulf University for Science and Technology (GUST), P.O. Box 7207, Hawally 32093, Kuwait
5 Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19111, Jordan; a.alshamayleh@ammanu.edu.jo
6 eHealth and Telemedicine Group, University of Valladolid, 47011 Valladolid, Spain; isator@uva.es
7 Departamento de Ciências da Computação, Universidade Internacional do Cuanza, Estrada Nacional 250, Bairro Kaluapanda, Cuito, Bié, Angola; eduardo.silva@uneatlantico.es (E.S.A.); helena.garay@uneatlantico.es (H.G.)
8 Fundación Universitaria Internacional de Colombia, Bogotá 111321, Colombia
9 Departamento de Soluciones Tecnológicas y Sistemas, Escuela Politécnica Superior, Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain
10 Departamento de Ciencias de la Computación, Universidad Internacional Iberoamericana, Campeche 24560, Mexico
11 Department of Computer Science, Universidad Internacional Iberoamericana, Arecibo, PR 00613, USA
12 Departamento de Ciencias de la Computación, Universidad de La Romana, La Romana 22000, Dominican Republic
* Correspondence: ghulam.mustafa.ssc@stmu.edu.pk

Abstract

Assessing scholarly contribution across disciplines is challenging because bibliometric indicators capture different aspects of research activity and citation practices vary by field. This study evaluates a weighted cubic-mean composite (Hr) and a career-adaptive extension (DHr) across civil engineering, computer science, mathematics, and neuroscience. The analysis considers 4479 researchers and uses recognition by professional societies as an observable benchmark, while recognizing that awards are imperfect proxies for research quality. To reduce selection optimism, we performed nested cross-validation with researcher-level grouping, selecting candidate indicators and metric pairs within inner folds and evaluating the fixed DHr reference formulation on held-out outer folds. On out-of-fold predictions, Hr AUCs were 0.556, 0.692, 0.768, and 0.713, while DHr AUCs were 0.559, 0.690, 0.767, and 0.712 across the four domains, respectively. Paired bootstrap confidence intervals for the Hr–DHr AUC differences included zero in every domain; therefore, these results do not establish that career-adaptive weighting improves discrimination over static aggregation. Because screening and metric-pair selection were conducted separately in each discipline, these AUCs reflect within-domain validation across four datasets and do not demonstrate transfer to an unseen discipline. The findings support further investigation of composite bibliometric indicators, but independent validation, transparent benchmarking, and fairness assessment are needed before consequential use.

Keywords: bibliometric indices; researcher impact assessment; *h*-index variants; career-adaptive weighting; higher-order mean aggregation; cubic mean integration; Hr-index; DHr-index; multi-domain evaluation; within-domain validation; temporal weighting

Check for updates

Academic Editor: Shengkun Xie

Received: 8 October 2026
Revised: 4 November 2026
Accepted: 6 November 2026
Published: 11 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid growth of scientific publications has intensified the need for reliable and systematic methods to evaluate researcher impact. With an increasing volume of scholarly output across disciplines, distinguishing between mere productivity and meaningful scientific influence has become a complex challenge [1]. Effective researcher evaluation mechanisms are essential for multiple stakeholders, including funding agencies, hiring committees, journal editors, and academic institutions. Such evaluations influence critical decisions related to promotions, tenure, research funding, editorial appointments, and prestigious awards [2,3]. Consequently, the development of fair, transparent, and robust author-ranking methodologies remains a central issue in bibliometrics [2,4].

Over the past decades, numerous metrics have been proposed to quantify scholarly impact. Basic author-level indicators include publication and citation counts, although both depend on database coverage, disciplinary conventions, and career length [5,6]. More advanced methods incorporated structural information such as co-authorship networks [7,8] and hybrid frameworks integrating multiple indicators [5]. Among these contributions, the h-index introduced by Hirsch [5] gained widespread adoption due to its simplicity and ability to jointly reflect productivity and citation impact. However, the h-index has well-documented limitations, including insensitivity to highly cited publications, inability to differentiate closely ranked researchers, and, critically, a complete absence of temporal sensitivity; it treats a paper published last year identically to one published three decades ago [9].

To address these shortcomings, a wide range of h-index variants have been proposed [9]. These include the a-index and ar-index [10], the contemporary h-index [11], raw h-rate [12], m-quotient [5], f-index and t-index [13], Maxprod index [14], q^2-index [15], and the w-index [16], among others. Despite this proliferation, three fundamental problems persist. First, no variant has demonstrated consistent superiority across multiple scientific disciplines; performance reported in one domain may not transfer to another. Second, existing variants address temporal limitations only partially, typically through simple age normalization rather than career-stage sensitivity. Third, virtually all studies evaluate indices in isolation, overlooking the complementary information that structured aggregation of multiple indices could provide [6].

A relevant limitation in this literature is that cross-domain evaluations of career-sensitive aggregation frameworks remain limited. Most proposed metrics are validated on synthetic case studies [6] or small, domain-homogeneous datasets [9,15], which restricts generalizability and prevents objective cross-disciplinary comparison. Crucially, citation distributions vary substantially across scientific fields: a highly cited civil engineer operates in a fundamentally different citation environment than a prolific mathematician or neuroscientist. Any proposed aggregation framework should therefore be evaluated across heterogeneous citation environments; performance in one field should not be assumed to transfer to another.

Career stage is another relevant source of heterogeneity. Cumulative indicators may disadvantage some early-career researchers because they have had less time to accumulate publications and citations, while long publication histories can increase cumulative scores for later-career researchers. These are potential limitations of cumulative metrics rather than universal effects for every researcher. A weighting mechanism that adapts to career length, rather than treating all researchers as occupying the same career stage, is a plausible design hypothesis. Whether this approach improves discrimination or fairness requires empirical testing against static and established indicators.

To address these interrelated gaps, this study proposes a career-adaptive weighted cubic mean framework for researcher impact assessment, evaluated across four scientific

domains. The framework makes two distinct and independently verifiable contributions. First, it introduces a consistent evaluation protocol applied separately to 4479 researchers spanning civil engineering, computer science, mathematics, and neuroscience. The four-domain coverage evaluates the framework in multiple disciplinary settings, but does not test transfer of a model selected in one discipline to an unseen discipline. Award recognition from professional societies is used as an observable recognition outcome, not as a definitive or bias-free measure of research quality. Second, it introduces the DHr-index, a career-stage adaptive extension of the weighted cubic mean that employs a sigmoid weighting function to dynamically modulate parameter contributions based on researcher career length, formally defined as

$$\mathrm{DHr} = \left(w_a(t)\, a^3 + w_b(t)\, b^3\right)^{1/3} \tag{1}$$

where the adaptive weights $w_A(t)$ and $w_B(t)$ are governed by a career-stage sigmoid:

$$\sigma(t) = \frac{1}{1 + e^{-0.1(t-20)}} \tag{2}$$

with t denoting career length in years and the shift parameter $\alpha = 0.4$ controlling the magnitude of career-stage reweighting. This formulation is designed to introduce career-stage sensitivity into this particular composite; it does not imply that all other composite approaches lack temporal components.

The original exploratory analysis evaluated 64 bibliometric parameters, then examined all 45 pairwise combinations of the top-10 parameters under seven statistical aggregation models. In the revised evaluation, all screening and pair selection is nested within the training folds; the original full-data rankings are reported only as exploratory context. The revised nested cross-validation analysis evaluates the selected Hr and DHr formulations on held-out researchers. It finds no statistically significant paired AUC difference between Hr and DHr in any domain; consequently, the results do not establish a predictive advantage for career-adaptive weighting.

The original full-data screening and aggregation summaries are retained only as exploratory descriptions; all confirmatory performance claims are based on the revised nested-CV comparison reported in the Results section. The remainder of this paper is organized as follows. Section 2 reviews related literature and identifies critical methodological gaps. Section 3 presents the proposed career-adaptive framework and its mathematical foundations. Section 4 reports results and comparative analysis across all four domains. Section 5 discusses interpretation, limitations, and responsible use. Section 6 concludes with directions for future research.

2. Related Work

The evaluation of scientific contribution is not a purely measurement problem: the appropriate evidence depends on the decision being made and the contributions being valued. Research assessment can combine expert peer review and narrative evidence with quantitative indicators, including publication and citation counts, field- and time-normalized citation measures, collaboration-network measures, usage and attention indicators, and data-driven prediction models. These approaches answer different questions. Peer review can consider originality and context but may be costly and affected by reviewer judgment; citation-based indicators are reproducible at scale but reflect field conventions, time since publication, database coverage, and patterns of recognition; network and alternative-attention measures capture other forms of visibility but do not directly establish scholarly quality. Composite indicators can summarize several dimensions, but their weights and validation targets introduce additional choices that must be justified. Responsible-assessment guidance consequently recommends contextual interpretation and warns against replacing expert judgment

with a single number [2,17,18]. This section critically compares established bibliometric indices, domain-specific evaluation studies, and composite or data-driven approaches before identifying the narrower contribution examined here.

2.1. Foundational Bibliometric Indicators

Basic evaluation approaches often use publication count [5], which measures productivity but fails to account for research quality or influence. Citation count was subsequently introduced to reflect academic visibility and impact; however, citation-based metrics are sensitive to self-citations, disciplinary citation norms, and time-dependent accumulation effects [18,19]. Moreover, raw citation counts conflate positive scholarly recognition with critical or corrective referencing, reducing their reliability as standalone indicators of genuine impact.

The introduction of the *h*-index by Hirsch [5] marked a significant milestone by combining productivity and citation impact into a single interpretable measure. Despite its widespread adoption, the *h*-index carries several well-documented limitations: insensitivity to highly cited publications beyond the Hirsch core, limited discriminative power among closely ranked researchers, systematic bias against early-career scholars, and, critically for the present work, complete insensitivity to career stage and temporal dynamics [6,20]. A researcher who produced seminal work a decade ago receives the same *h*-index credit as one producing equivalent output today, despite the fundamentally different career contexts these represent.

2.2. h-Index Variants and Their Limitations

To address these shortcomings, numerous variants of the *h*-index have been proposed [9]. The *a*-index and *ar*-index [10] incorporate average citation counts within the Hirsch core, providing sensitivity to highly cited outliers. The contemporary *h*-index [11] introduces time weighting to reward recent output, while the *m*-quotient [5] normalizes by career length. The *f*-index and *t*-index [13] attempt to account for citation distribution shape, and the Maxprod index [14] and q^2-index [15] incorporate both publication count and citation volume. The *w*-index [16] focuses on widely cited papers to reduce sensitivity to marginal contributions.

Despite this proliferation, no single variant is accepted as a universally sufficient measure across disciplines. Each variant addresses particular limitations while retaining other blind spots. Many temporal variants use fixed age adjustments rather than explicitly varying the relative contribution of indicators across career stages. The sigmoid approach examined here is one possible design choice; its added value must be tested empirically rather than assumed from the form of the weighting function.

2.3. Domain-Specific Evaluation Studies

Recent work continues to emphasize that author-level indicators capture different dimensions of scholarly performance and may respond differently to publication volume, career stage, and disciplinary context. Laimon [21] discussed a volume-normalized efficiency measure alongside an impact-concentration indicator, illustrating the value of distinguishing cumulative impact from the structure of a publication portfolio. In a complementary responsible-assessment perspective, Şener et al. [22] documented how metric-based evaluation rules can impose unequal burdens on early-career researchers. These studies reinforce the need to interpret any composite against its explicit target and to avoid treating a single benchmark as a comprehensive measure of merit.

Several studies have attempted systematic comparison of bibliometric indices within specific scientific domains. Ayaz [23] evaluated the *h*-complete index using award-based

civil engineering datasets, establishing the award-validation methodology that subsequent studies in this line of work have adopted. Raheel et al. [24] examined citation intensity and publication age factors within the same domain, reporting improved discriminative performance for certain *h*-index variants under age-adjusted evaluation.

Beyond civil engineering, Ain et al. [25] evaluated bibliometric variants in mathematics, while Ameer et al. [26] examined variants in neuroscience. These studies provide historical, discipline-specific benchmarks rather than evidence of general transferability. Independent analyses of the limitations of author-level indicators and field normalization further show why rankings must be interpreted in their disciplinary and career-stage contexts [3,27]. Differences in the strongest-performing indicators across datasets motivate transparent domain-specific comparisons, but do not establish that any composite index is universally preferable.

2.4. Aggregation and Machine Learning Approaches

Researcher assessment has increasingly considered composite indicators, field-normalized measures, and data-driven analysis. Ioannidis et al. [18] provide a transparent example of combining several citation dimensions, while Waltman and van Eck [27] discuss normalization and its methodological implications. The Leiden Manifesto [2] emphasizes the need for transparent, contextualized interpretation, especially when quantitative measures inform consequential decisions. These independent studies provide relevant context for testing whether the proposed weighted-cubic combination offers an empirically useful addition to existing indicators. They do not establish that a new composite will improve out-of-sample prediction. The present study therefore examines Hr and its career-adaptive extension using nested cross-validation and treats observed differences cautiously.

The literature review motivates examining a common aggregation framework across several domains and testing whether career-stage weighting adds value. However, claims that no prior study has addressed these questions require a complete and current literature search and should not be stated categorically without one. The present study evaluates one such framework across four domains, using award recognition as the outcome and nested cross-validation for the primary performance comparison.

Recent research has further highlighted the importance of considering career stage, disciplinary differences, model stability, and the broader consequences of metric-based research evaluation when developing quantitative indicators of scholarly impact. Laimon (2026) proposes recent measures of research impact that account for factors such as research volume, career stage, and discipline, providing a relevant basis for comparing different approaches to researcher assessment [21]. This perspective is particularly relevant to the DHr-index, which incorporates career-adaptive weighting, although the effectiveness of such weighting must be established through systematic empirical evaluation. In addition, Meysami, Lotfi, and Saleem (2026) examine stability, generalization, regularization, and robustness in complex learning problems, highlighting the importance of evaluating predictive models beyond their apparent classification performance [28]. These considerations are relevant when comparing interpretable bibliometric indicators with more complex learning-based approaches, particularly where reliable generalization to unseen researchers is essential. Furthermore, Şener et al. (2025) investigate metric-based research evaluation and its implications for early-career researchers, drawing attention to the potential relationship between evaluation systems and researchers' career stages [22]. Taken together, these studies motivate a broader assessment framework in which researcher-impact indicators are evaluated not only in terms of predictive discrimination but also in relation to disciplinary comparability, career-stage sensitivity, stability, and responsible use. Within this framework, the Hr-index and DHr-index should be assessed through transparent valida-

tion procedures, with particular attention to whether career-adaptive weighting improves out-of-sample performance without introducing systematic disadvantages for particular groups of researchers.

Recent articles in *Informatics* provide complementary methodological context for the present researcher-level assessment problem. Ivanova et al. [29] and Omarsaib et al. [30] used bibliometric approaches to characterize research on artificial intelligence in education, while Favara et al. [31] examined research activity involving language-processing tools in public health. Akpan et al. [32] combined bibliometric performance analysis with geographic and research-investment context, making disciplinary and regional differences particularly salient. Haque et al. [33] integrated bibliometric reflection with content analysis, Nastasa et al. [34] mapped research patterns in medical data governance, and Wang and Li [35] analyzed co-authorship, co-citation, and thematic evolution in AI marketing. These seven studies evaluate research landscapes or publication patterns; none should be interpreted as a direct validation of individual-researcher ranking by Hr or DHr.

A related methodological strand in the journal addresses the representation, combination, and interpretation of multidimensional evidence. Kurasova et al. [36] investigated embeddings and visualization for decision support, Wu et al. [37] proposed an adaptive weighted ensemble for an educational classification task, and Xu et al. [38] reviewed human–AI decision support and the need to calibrate reliance on automated recommendations. Their relevance to the present study is conceptual rather than benchmark-level: they motivate transparent multivariate modeling, testing adaptive weighting rather than assuming its benefit, and maintaining expert oversight when quantitative scores inform consequential decisions. Unlike these topic-mapping and decision-support studies, the present work evaluates whether a specified pairwise cubic composite and its career-dependent extension discriminate award-recognition outcomes among individual researchers. Its nested cross-validation findings do not establish general superiority or deployment readiness.

Table 1 integrates a critical comparison of researcher-assessment approaches with representative studies, their evaluation strategies, and the limitations relevant to Hr and DHr. It distinguishes the validity of a scoring method from the validity of award recognition as a benchmark: predicting awards does not, by itself, establish an unbiased measure of scholarly merit.

Independent Evidence on Normalization, Indicator Construction, and Responsible Interpretation

Several independent studies provide important qualifications for composite researcher-level assessment. Waltman et al. [39] compared field- and year-normalization procedures, while Waltman and van Eck [27] systematically examined alternatives for normalizing citation impact across fields. Waltman, Yan, and van Eck [40] showed that some recursively weighted field-normalized measures can remain sensitive to the field-classification scheme. Percentile-based bibliometric scores also require explicit handling of ties and discrete citation distributions [41]; Thelwall [42] proposed practical field-normalized measures with uncertainty estimates. These studies support careful specification of reference populations, transformations, and uncertainty when interpreting Hr and DHr, rather than assuming that combining indicators automatically resolves differences between disciplines.

Other independent work highlights threats to validity that are particularly relevant when bibliometric indices are used for consequential researcher assessment. Bartneck and Kokkelmans [43] investigated potential manipulation of the *h*-index through self-citation, and Ferrara and Romero [44] proposed a self-citation-discounted impact measure. King et al. [45] documented variation in self-citation patterns across researchers and fields. Aksnes et al. [3] distinguished citation impact from the broader construct of research quality,

while Quaderi and Duine [46] examined risks of misusing bibliometric measures in research assessment. Collectively, these findings reinforce the need to treat award-status AUC as an association with recognition, disclose data and self-citation limitations, and avoid using Hr or DHr as a standalone proxy for scholarly merit. None of these studies constitutes direct external validation of the proposed indices.

2.5. Critical Analysis and Research Gap Identification

Table 1 compares the principal assessment approaches, representative studies, validation strategies, and remaining limitations. The comparison makes clear that aggregation, cross-domain testing, and career-stage adjustment are distinct design choices whose incremental value must be evaluated rather than presumed.

Four critical gaps emerge from this analysis. First, many prior studies evaluate individual bibliometric indices, while established composite approaches demonstrate that combining indicators is not itself novel [18,47]. While comparative ranking studies identify high-performing metrics, they do not explore how complementary indices interact or how structured aggregation can capture multidimensional aspects of scholarly influence that no single index encodes.

Second, although cross-disciplinary composite-indicator studies exist [18], the present award-recognition setting merits evaluation across the four selected domains. Studies conducted in civil engineering, mathematics, or neuroscience report results that cannot be assumed to generalize, yet most authors present findings as broadly applicable. The absence of cross-domain validation with consistent methodology represents the most significant gap in the current bibliometric aggregation literature.

Third, temporal sensitivity in existing indices is handled through simple normalization rather than career-stage modeling. Age-adjusted variants divide scores by career length, which fails to distinguish between an early-career researcher producing breakthrough work and a late-career researcher whose accumulated output dominates cumulative metrics. A continuous, bounded, and theoretically interpretable sigmoid-based career-stage weighting function has not previously been applied to bibliometric aggregation.

Fourth, statistical validation must account for model selection, multiple comparisons, and uncertainty; earlier exploratory results in the present study did not meet that standard. Performance is typically reported as raw ranking counts without significance testing, confidence intervals, or AUC analysis, making it impossible to assess whether reported improvements reflect genuine discriminative capacity or sampling variation.

These four gaps collectively underscore the need for a principled framework that

- Systematically identifies high-performing bibliometric indicators through ranking-based screening across multiple domains;
- Evaluates parameter interactions through structured pairwise aggregation under multiple statistical models;
- Integrates complementary metrics using a career-adaptive higher-order weighting mechanism grounded in a theoretically motivated functional form;
- Validates all results using both discriminative ranking scores and AUC-ROC analysis with statistical significance testing across four scientifically distinct domains.

The proposed framework investigates these gaps by extending weighted cubic-mean aggregation with a career-adaptive formulation and evaluating it across civil engineering, computer science, mathematics, and neuroscience. The nested cross-validation analysis provides an internal assessment of discrimination and uncertainty, but does not establish generalizability, superiority over established indicators, or suitability for consequential researcher assessment without independent external validation.

Table 1. Integrated critical comparison of bibliometric assessment approaches, representative studies, validation strategies, and implications for Hr and DHr.

Approach/Representative Studies	Focus and Intended Strength	Evaluation/Evidence	Limitation or Trade-Off	Implication for Hr and DHr
Individual indicators: Hirsch (2005) [5]; Bornmann (2010) [9]; Alonso et al. [6]	General-purpose, interpretable citation and *h*-index summaries	Citation ranking and comparison of index variants	Single measures omit some impact dimensions and can depend on career length and field norms	Use strong individual-metric baselines; combining metrics does not guarantee improvement
Award-based single-index comparisons: Ayaz (2016) [23]; Raheel (2018) [24]; Ain (2019) [25]; Ameer (2019) [26]	Tests bibliometric variants in civil engineering, mathematics, and neuroscience	Recognition-based comparisons within individual disciplines	Limited cross-domain transferability; awards depend on nomination and eligibility processes	Treat award status as an observed recognition outcome, not unbiased scholarly merit
Career-stage and authorship-aware assessment: Hirsch (2005) [5]; Schreiber (2008–2009) [48,49]; Laimon (2026) [21]	Accounts for coauthorship, publication/citation characteristics, or career-related effects	Indicator comparisons, largely within civil engineering	Context-specific estimates and limited evidence of transferable gains	Evaluate career-sensitive weighting and domain-specific performance empirically
Composite citation indicators: Ioannidis et al. (2016) [18]	Combines multiple citation-profile dimensions	Composite scoring and researcher ranking	Component selection, scaling, correlation, and weighting influence scores	Higher-order aggregation is a specific hypothesis, not a novel concept of combining indicators
Field-normalized measures [19]	Adjusts for field- and time-dependent citation practices	Reference-set and field-based normalization	Depends on field definitions, publication coverage, and reference groups	Four-domain evaluation alone does not establish field-invariant validity
Data-driven and composite methods: Ioannidis et al. (2016) [18]; Bornmann and colleagues [9]; stability considerations [28]	Multivariate statistical and nonlinear classification approaches	Logistic-regression ranking and classification accuracy	Selection bias, interpretability, and fairness require scrutiny	Use leakage-resistant nested validation and comparable out-of-fold baselines
Recent journal research (2024–2026) [29,30,35,37]	Bibliometric mapping, collaboration analysis, and adaptive data modeling	Research-landscape studies and task-specific models	Typically not direct researcher-level award-recognition benchmarks	Provides methodological context rather than direct evidence of DHr superiority
Present study: Hr/DHr	**Weighted cubic aggregation and career-dependent weights across four domains**	**4479 researchers; award recognition; nested cross-validation and paired AUC comparisons**	**Internal validation only; award bias; small and statistically inconclusive incremental differences**	**Requires independent validation, fair comparisons, and subgroup analysis before consequential use**

3. Methodology

Building upon the identified research gaps, this study proposes a career-adaptive higher-order aggregation framework for researcher impact evaluation across multiple scientific domains. The methodology is designed to systematically identify high-performing bibliometric indicators across heterogeneous disciplinary environments, analyze their pairwise interactions, and integrate complementary metrics using a theoretically motivated career-stage adaptive weighting mechanism.

Figure 1 illustrates the overall architectural workflow of the proposed framework. The process consists of eight sequential stages:

1. Multi-Domain Selection and Dataset Collection;
2. Parameter Computation;
3. Ranking-Based Screening;
4. Individual Parameter Analysis;
5. Pairwise Statistical Aggregation;
6. Hr-Index Formulation (Static);
7. DHr-Index Formulation (Career-Adaptive);
8. Multi-Domain Statistical Validation.

Each stage is described in detail in the following subsections.

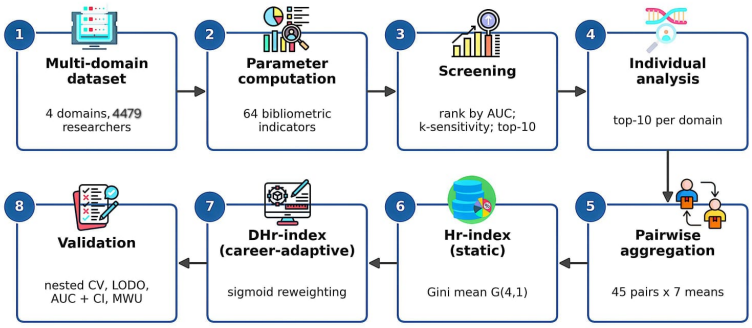

Figure 1. Architectural workflow of the proposed career-adaptive higher-order aggregation framework for multi-domain researcher impact evaluation. The framework extends prior single-domain approaches by incorporating four scientific domains and a career-stage adaptive weighting branch (Stage 7).

3.1. Multi-Domain Selection and Dataset Description

A key methodological departure from prior work is the simultaneous evaluation of the proposed framework across four scientifically distinct domains: civil engineering, computer science, mathematics, and neuroscience. This multi-domain design allows the proposed framework to be examined under a consistent procedure in four fields, although it does not by itself establish generalizability to other disciplines or settings.

Each domain was selected on the basis of three criteria: (i) availability of a dataset containing substantial numbers of award recipients and non-recipients, (ii) recognition by internationally established professional societies providing a publicly verifiable external benchmark, and (iii) sufficient bibliometric diversity to test whether aggregation behavior generalizes across citation environments. Civil engineering and neuroscience represent citation-intensive experimental disciplines, while mathematics represents a low-citation-volume theoretical field providing variation in citation environments for comparative evaluation.

The complete multi-domain dataset comprises 4479 researchers. Summary statistics for each domain are presented in Table 2.

Table 2. Summary statistics of the multi-domain researcher dataset used for cross-domain framework evaluation.

Domain	Total	Awardees	Non-Awardees	Awarding Societies
Civil Engineering	1180	590	590	ASCE, CSCE, ACI, IC
Computer Science	1184	609	575	ACM, IEEE
Mathematics	1049	525	524	AMS, IMU
Neuroscience	1066	529	537	SfN, IBRO
Total	4479	2253	2226	

Award status in each domain is determined by recognition from the respective professional societies listed in Table 2. We chose award status because it is observable and can be independently checked against public society records, enabling a reproducible classification target across the four domains. It is not a neutral gold standard: awards differ in eligibility, nomination procedures, geographic reach, career-stage requirements, and the kinds of contribution they recognize. Awardee/non-awardee classification therefore evaluates alignment with an existing recognition process, not research quality in its entirety. This benchmark also cannot establish whether an index would be fair or appropriate for hiring, promotion, or funding. Bibliometric metadata were extracted using Publish or

Perish software (version 8), which retrieves data from Google Scholar; using one retrieval workflow improves procedural consistency but does not eliminate database coverage or author-disambiguation limitations [3,50].

Non-awardee comparison-group construction. The retained dataset records verify that non-awardees were assembled as discipline-specific comparison groups corresponding to the same four broad fields as the awardees, with the counts reported in Table 2. However, the retained documentation does not record the historical sampling frame or a reproducible rule showing whether these researchers were selected randomly, consecutively, by convenience, or by matching to individual awardees. We therefore do not describe them as matched controls and cannot establish comparability beyond broad disciplinary membership and the observed group sizes. This limitation is substantive: differences in career length, institution, geography, subfield, visibility, or award eligibility may contribute to the observed award-status discrimination. The analysis is consequently interpreted as a comparison within the available assembled groups, not as an estimate from a matched or representative case–control sample.

Comparison-group construction and temporal alignment. The awardee and non-awardee groups were compared within the four stated disciplines, but the available records do not establish a documented individual-level matching protocol for career stage, institution, geographic region, or other potential confounders. Likewise, the precise bibliometric extraction date and its temporal relationship to each recipient's award date are not verified. It therefore cannot be assumed that all publication and citation measures were observed before award recognition; some values may incorporate post-award scholarly activity or citation accumulation. This possibility introduces temporal ambiguity and prevents interpretation of the reported discrimination as prospective award prediction. Accordingly, the AUC estimates are treated as retrospective associations with observed award status, and any future predictive evaluation would require award-date-aligned, pre-award feature snapshots and explicitly defined comparison-group selection criteria.

Bibliometric data provenance and reproducibility. The researcher-level bibliometric datasets for civil engineering, computer science, mathematics, and neuroscience are publicly accessible at https://github.com/Dr-GMustafa/Scientometrics_Related_Dataset (accessed on 5 November 2024). The public dataset provides a basis for examining the reported researcher-level measures and disciplinary comparisons. Nevertheless, public access to derived datasets should not be conflated with complete reproducibility of the original Google Scholar/Publish or Perish extraction and computation of all 64 measures. The present documentation does not independently establish the precise retrieval date or date range, the duplicate-detection and missing-record rules, the author-disambiguation procedure, or whether self-citations were included or excluded, nor is availability of a complete executable pipeline for deriving every indicator established here. These details should be verified against the original collection records and source code before making stronger reproducibility claims. Because Google Scholar profiles and citation counts may change over time, independent re-extraction may also yield different results.

Use of AI-assisted language-editing tools. During manuscript preparation, the authors used ChatGPT 5.6 (OpenAI) and Claude Sonnet 5.5 (Anthropic) to assist with English-language clarity, sentence structure, readability, and academic phrasing, and Grammarly 1.203.0.0 for grammar, spelling, punctuation, and proofreading. The authors reviewed and edited the resulting suggestions and retain full responsibility for the accuracy, originality, and integrity of the final manuscript. This statement describes the disclosed language-editing uses of these tools; it does not attribute data collection, statistical analysis, or study design to them.

3.2. Parameter Computation

A total of 64 bibliometric indicators were computed for each researcher across all four domains. These indicators are categorised into four primary groups:

1. Primitive Parameters: basic counts including total publications, total citations, career length, and derived ratios;
2. Publication- and Citation-Based Parameters: *h*-index and its variants incorporating citation distribution structure;
3. Author Count-Based Parameters: fractional and normalized variants accounting for multi-authorship;
4. Publication Age-Based Parameters: temporally adjusted indices incorporating career length or publication recency.

The complete set of evaluated parameters is illustrated in Figure 2, which provides a structured overview of the metric classification framework. The same 64 parameters were computed for all 4479 researchers across all four domains, providing a common set of candidate indicators for separate within-domain evaluations, rather than a held-out cross-domain transfer test.

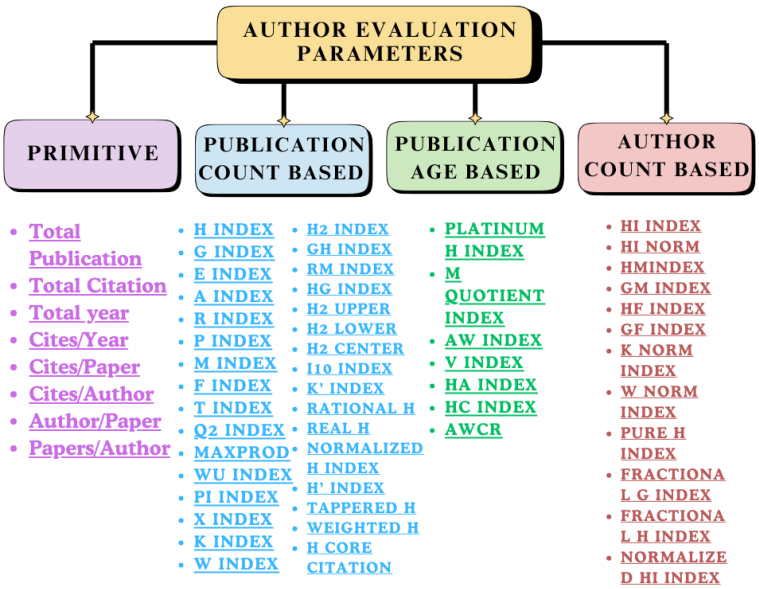

Figure 2. Conceptual categorisation of the 64 bibliometric parameters evaluated across all four domains. Parameters are grouped into four categories: primitive, publication- and citation-based, author count-based, and publication age-based.

3.3. Ranking-Based Parameter Screening

Domain-Specific Selection and Validation Boundary

For each discipline, candidate indicator screening and pair selection were performed using data from that discipline, and the revised nested-CV protocol restricted these choices to its training folds. The DHr reference formulation used the prespecified settings $\alpha = 0.4$, sigmoid slope 0.1, career midpoint 20 years, and clipping limits $[0.05, 0.95]$; these settings were not re-estimated from the outer evaluation folds. No indicator set, pair, or parameter configuration was learned from three domains and evaluated unchanged in the fourth. Thus, the four-domain design assesses within-domain performance in multiple fields rather than out-of-domain transfer.

To assess the discriminative capability of each bibliometric indicator independently within each domain, a ranking-based screening procedure was applied to all 64 computed

parameters. Let $P = \{p_1, p_2, \ldots, p_{64}\}$ denote the set of evaluated parameters and let $R^{(d)} = \{r_1, r_2, \ldots, r_{N_d}\}$ represent the set of N_d researchers in domain d, where $d \in \{\text{CE, CS, MA, NS}\}$.

For each parameter $p_j \in P$ and domain d, researchers were sorted in descending order based on their parameter values, producing a domain-specific ranked list:

$$L_j^{(d)} = \text{sort}_{\downarrow}\left(p_j(r_i^{(d)})\right), \quad i = 1, 2, \ldots, N_d \tag{3}$$

The discriminative score for parameter p_j in domain d is defined as

$$S^{(d)}(p_j) = \sum_{i=1}^{k} \mathbb{I}\left(r_i^{(d)} \in A^{(d)}\right) \tag{4}$$

where $A^{(d)}$ represents the set of award-winning researchers in domain d, $\mathbb{I}(\cdot)$ is the indicator function, and $k = 100$ is the evaluation threshold. The value $S^{(d)}(p_j)$ measures the number of awardees appearing within the top 100 ranked positions for parameter p_j in domain d. Parameters with higher $S^{(d)}(p_j)$ values identify more awardees at this particular cut-off in the observed dataset. The threshold $k = 100$ was retained to maintain comparability with the original top-100 screening and to provide an interpretable fixed-size shortlist; it is not a statistically optimized, uniquely justified, or field-invariant cut-off. Because $S^{(d)}(p_j)$ depends on k, the top-ten indicators and their subsequent pairs may change at another cut-off. The revised analysis reports pooled out-of-fold awardee counts at $k = 50$, 100, and 200 as a descriptive check of the final ranking outputs, but it does *not* rerun the original 64-indicator screening or the 45-pair search at each k. Thus, these counts cannot be interpreted as evidence of feature-selection stability. The original full-data top-ten and pair rankings remain exploratory.

Based on this criterion, the ten parameters with the highest discriminative scores within each domain were selected for subsequent pairwise interaction analysis. This domain-specific selection acknowledges that citation environments differ substantially across fields; the optimal discriminative parameter in civil engineering need not be optimal in mathematics or neuroscience. The complete screening procedure is summarised in Algorithm 1.

Algorithm 1 Multi-Domain Ranking-Based Parameter Screening

Input: Domains D, researchers $R^{(d)}$, parameter set P, awardee sets $A^{(d)}$
Output: Top-10 discriminative parameters per domain
for each domain $d \in D$ **do**
    for each parameter $p_j \in P$ **do**
        Rank $R^{(d)}$ in descending order by p_j
        Select top $k = 100$ researchers
        Compute $S^{(d)}(p_j)$ = awardees in top k
    end for
    Sort parameters by $S^{(d)}(p_j)$ in descending order
    Select top-10 parameters $T^{(d)}$
end for

3.4. Pairwise Statistical Aggregation Framework

Following the domain-specific screening procedure, the ten highest-performing parameters per domain were selected for pairwise interaction analysis. Let $T^{(d)} = \{p_1^{(d)}, p_2^{(d)}, \ldots, p_{10}^{(d)}\}$ denote this reduced domain-specific parameter set.

3.4.1. Pairwise Combination

All possible unordered parameter pairs were generated for each domain using the binomial coefficient:

$$\binom{10}{2} = 45 \tag{5}$$

This produced 45 distinct two-parameter combinations per domain, yielding 180 pairwise combinations in total across the four domains. For each pair $(p_i^{(d)}, p_j^{(d)})$, a composite metric was constructed using each of the statistical aggregation functions defined below.

3.4.2. Statistical Aggregation Models

Seven aggregation models were applied to each parameter pair, as summarised in Table 3. These span a range of mean-type functions from sub-linear (harmonic) to super-linear (contra-harmonic), enabling systematic comparison of how aggregation order affects discriminative capacity under skewed citation distributions.

Table 3. Statistical aggregation models applied to each pairwise parameter combination.

Model	Abbr.	Definition
Arithmetic Mean	AM	$\frac{a_1 + a_2}{2}$
Harmonic Mean	HM	$\frac{2}{\frac{1}{a_1} + \frac{1}{a_2}}$
Geometric Mean	GM	$\sqrt{a_1 a_2}$
Quadratic Mean	QM	$\sqrt{\frac{a_1^2 + a_2^2}{2}}$
Cubic Mean	CM	$\sqrt[3]{\frac{a_1^3 + a_2^3}{2}}$
Contra-Harmonic	CHM	$\frac{a_1^2 + a_2^2}{a_1 + a_2}$
Logarithmic Mean	LM	$\frac{a_2 - a_1}{\ln a_2 - \ln a_1}$

Each composite metric was evaluated using the ranking-based discriminative score $S^{(d)}$ defined in Section 3.3, applied consistently across all parameter pairs, aggregation models, and domains.

3.5. Hr-Index: Static Weighted Cubic Mean Formulation

Based on the aggregation analysis, the weighted cubic mean is adopted as the foundation of the proposed composite index. For two parameters a and b from the identified best-performing pair within a given domain, magnitude-proportional weights are defined as

$$w_a = \frac{a}{a+b}, \qquad w_b = \frac{b}{a+b} \tag{6}$$

where $w_a + w_b = 1$. This normalization ensures proportional contribution of each index based on its relative magnitude, avoiding the equal-weight assumption of the standard cubic mean.

The Hr-index is then defined as

$$\mathrm{Hr} = \left(w_a\, a^3 + w_b\, b^3\right)^{1/3} \tag{7}$$

The cubic power is a design choice intended to increase the influence of larger input values relative to smaller ones; it is not derived from a general theorem that cubic aggregation is optimal for bibliometric data. Because the inputs are cubed, their scale and normalization materially affect the composite, and the resulting value is not invariant to arbitrary rescaling of either metric. The weights are magnitude-dependent and sum to one, but this does not make the score a validated measure of research quality. The choice of cubic aggregation should therefore be treated as a transparent modeling hypothesis whose value must be established by held-out comparisons against alternatives.

Mathematical Interpretation and Scale Sensitivity

Substituting the magnitude-proportional weights from Equation (6) into Equation (7) gives the equivalent expression

$$\mathrm{Hr} = \left(\frac{a^4 + b^4}{a + b} \right)^{1/3}, \qquad a, b \geq 0, \quad a + b > 0. \tag{8}$$

This expression shows that the larger component receives greater influence through both its magnitude-dependent weight and its cubic term. The amplification is an intentional feature of the proposed aggregation design, intended to examine whether emphasizing relatively strong indicator values improves ranking discrimination. It is not, by itself, evidence of superior validity or predictive performance. It may also increase sensitivity to unusually large values and to differences in component scales. When both inputs are zero, the index is defined by continuity as $\mathrm{Hr} = 0$.

Input scaling therefore requires explicit consideration. In the revised nested-cross-validation evaluation, the selected bibliometric measures are transformed into fold-specific percentile scores using mappings estimated only from the relevant training data; the transformed values are used before computing the magnitude-dependent weights and the cubic terms. The resulting operational score is denoted $\mathrm{Hr}_{\mathrm{pct}}$. This transformation reduces the direct influence of the original measurement units and prevents the held-out evaluation data from determining the normalization mapping. Nevertheless, percentile transformation does not eliminate every distributional or field-comparability concern. In particular, the original unnormalized Hr and operational $\mathrm{Hr}_{\mathrm{pct}}$ are distinct quantities and should not be assumed numerically equivalent or interchangeable.

3.6. DHr-Index: Career-Adaptive Temporal Extension

The Hr-index defined in Equation (7) treats all researchers as occupying the same implicit career stage, assigning weights purely on the basis of parameter magnitude. This assumption is theoretically problematic: early-career researchers with concentrated high-impact output are structurally different from late-career researchers with sustained but distributed productivity, and a bibliometric composite index should reflect this distinction.

To address this limitation, we introduce the DHr-index, a career-stage adaptive extension of the Hr-index that dynamically modulates parameter weights based on researcher career length t (measured in years of active publication).

3.6.1. Career-Stage Sigmoid Function

The career-stage factor is modelled using a logistic sigmoid function:

$$\sigma(t) = \frac{1}{1 + e^{-0.1(t-20)}} \tag{9}$$

where t denotes career length in years. This function maps career length to a continuous value in $(0, 1)$: researchers with $t < 20$ years (early career) yield $\sigma(t) < 0.5$, while those with $t > 20$ years (late career) yield $\sigma(t) > 0.5$. The inflection point at $t = 20$ years is a prespecified modeling choice used in the original formulation; it should not be interpreted as a universally valid career midpoint or as empirically established by the present evaluation.

The sigmoid was selected as a bounded, smooth parameterization, not because the present data establish that it is the uniquely appropriate career function. The midpoint of 20 years and slope of 0.1 specify the location and steepness of the transition; these are modeling choices that require sensitivity analysis and external validation. The nested procedure selects candidate indicators and metric pairs inside the inner training folds. The DHr career-weighting settings themselves are treated here as a prespecified reference formulation rather than as a data-driven hyperparameter search.

The DHr formulation evaluated in the revised analysis uses the prespecified reference settings $\alpha = 0.4$, sigmoid slope 0.1, career midpoint 20 years, and clipping interval $[0.05, 0.95]$. These values were carried forward from the original formulation and were not selected by optimizing performance on the outer folds. Accordingly, all DHr results reported in the revised manuscript—including DHr scores, AUCs, top-k awardee counts, and Hr–DHr statistical comparisons—were calculated using this same fixed reference configuration; the DHr career-weighting parameters were not re-tuned within or across the nested-CV folds. In the nested-CV analysis, the data-dependent selection steps are indicator screening and ordered metric-pair selection within the inner training folds; the resulting Hr and fixed-reference DHr scores are then evaluated on the corresponding outer fold. This clarification replaces the previous wording that could be read as implying a documented multi-parameter DHr grid search. No such grid search is claimed in the present manuscript.

Rationale and Scope of Career-Weighting Parameters

The four reference settings have distinct mathematical roles rather than independently established optimal values. The 20-year midpoint places the sigmoid's neutral transition at $\sigma(t) = 0.5$, so the additive career shift is zero at that career length; it is a modeling reference point, not an empirically verified boundary between career stages. The slope of 0.1 determines how gradually the adjustment changes with career length, avoiding an abrupt step at the midpoint while leaving its exact steepness open to validation. The shift magnitude $\alpha = 0.4$ limits the absolute additive adjustment to less than 0.2 because $0 < \sigma(t) < 1$; it was adopted as a reference magnitude, not demonstrated to be optimal. The clipping interval $[0.05, 0.95]$ keeps both complementary component weights away from zero and one, preventing either indicator from being entirely suppressed. These choices influence the resulting scores and may interact with the underlying component distributions and the order of the indicators.

The original values specify an illustrative reference formulation and should not be confused with independently established optimal settings. The present revision does not claim that midpoint, slope, shift magnitude, or clipping limits were tuned by a systematic grid search. Instead, the nested-CV procedure isolates the data-dependent screening and metric-pair selection from outer-fold performance estimation while holding the DHr reference settings fixed. Consequently, the current evidence does not establish sensitivity or robustness to alternative career-weighting settings.

3.6.2. Adaptive Weight Computation

Given the career-stage factor $\sigma(t)$, the base magnitude weights from Equation (6) are adjusted as follows:

$$w_a^{(t)} = \text{clip}(w_a + \alpha\,(0.5 - \sigma(t)),\ 0.05,\ 0.95) \tag{10}$$

$$w_b^{(t)} = 1 - w_a^{(t)} \tag{11}$$

where $\alpha = 0.4$ is the career-stage shift parameter controlling the magnitude of temporal reweighting, and the clip operation enforces minimum and maximum weight bounds to prevent degeneracy. The adaptive shift term $\alpha(0.5 - \sigma(t))$ has a direct interpretation:

- For **early-career** researchers $(\sigma(t) < 0.5)$, the shift is positive, boosting the weight of parameter a favoring structural impact metrics such as h^2-center that capture concentrated citation output;
- For **late-career** researchers $(\sigma(t) > 0.5)$, the shift is negative, boosting the weight of parameter b favoring sustained metrics such as the normalized h-index that reflect long-term consistent productivity;

- At the **career midpoint** ($t = 20$, $\sigma(t) = 0.5$), the shift vanishes and the DHr-index reduces exactly to the Hr-index.

3.6.3. DHr-Index Definition

The DHr-index is formally defined as

$$\mathrm{DHr}(a, b, t) = \left(w_a^{(t)}\, a^3 + w_b^{(t)}\, b^3 \right)^{1/3} \tag{12}$$

where $w_a^{(t)}$ and $w_b^{(t)}$ are the career-adaptive weights from Equations (10) and (11), a and b are the values of the two selected bibliometric parameters, and t is the researcher's career length in years. The pair is ordered: the first metric listed by the candidate-pair procedure is assigned to a, and the second to b. Consequently, swapping a and b can change DHr because the career shift is applied to the weight of the first-listed metric. This ordering convention must be preserved for reproducibility; DHr is not invariant to exchanging the inputs. At $t = 20$, the shift term is zero and DHr equals Hr for the same ordered pair.

Component Ordering and Asymmetry

The DHr construction is deliberately asymmetric with respect to its component positions: for a fixed career length t, $\mathrm{DHr}(a, b, t)$ generally differs from $\mathrm{DHr}(b, a, t)$. The career-dependent adjustment is applied to the first component's weight, with the second weight defined by complementarity; thus, the first and second positions have different mathematical roles. In the candidate-pair representation described above, a denotes the first-listed indicator and b the second-listed indicator. For reproducibility, the order used for each evaluated pair must be retained when calculating and reporting DHr; an unordered set of two indicator names alone is insufficient to reconstruct the score. This positional convention specifies how a recorded ordered pair is evaluated, but does not itself establish that the original candidate-pair enumeration followed a substantive, prespecified criterion for choosing which indicator should occupy each position. The current results therefore do not establish that reversing a pair would preserve rankings, discrimination, or selected configurations. An explicit prespecified indicator-ordering rule and an evaluation of both orientations are important requirements for future replication and robustness assessment. We do not interpret the observed DHr results as evidence of order invariance or of an optimal orientation.

The temporal weighting framework is illustrated in Figure 3, which shows the sigmoid career-stage function, the resulting adaptive weight curves, and the empirical career length distributions across all four domains that motivated the choice of inflection point.

3.7. Validation Strategy

The evaluation of bibliometric indices requires transparent benchmarks and out-of-sample testing to limit selection optimism. The revised primary validation uses nested cross-validation and paired bootstrap uncertainty estimates.

3.7.1. Layer 1 Award-Based External Benchmark

Researcher impact is evaluated by examining the ability of each bibliometric index to distinguish award recipients from non-recipients within each domain. Awards conferred by the professional societies listed in Table 2 serve as observable domain-specific recognition outcomes. They are publicly verifiable, but may reflect nomination practices, professional visibility, networks, eligibility rules, and other sources of selection bias; they are therefore not treated as neutral ground truth for research quality.

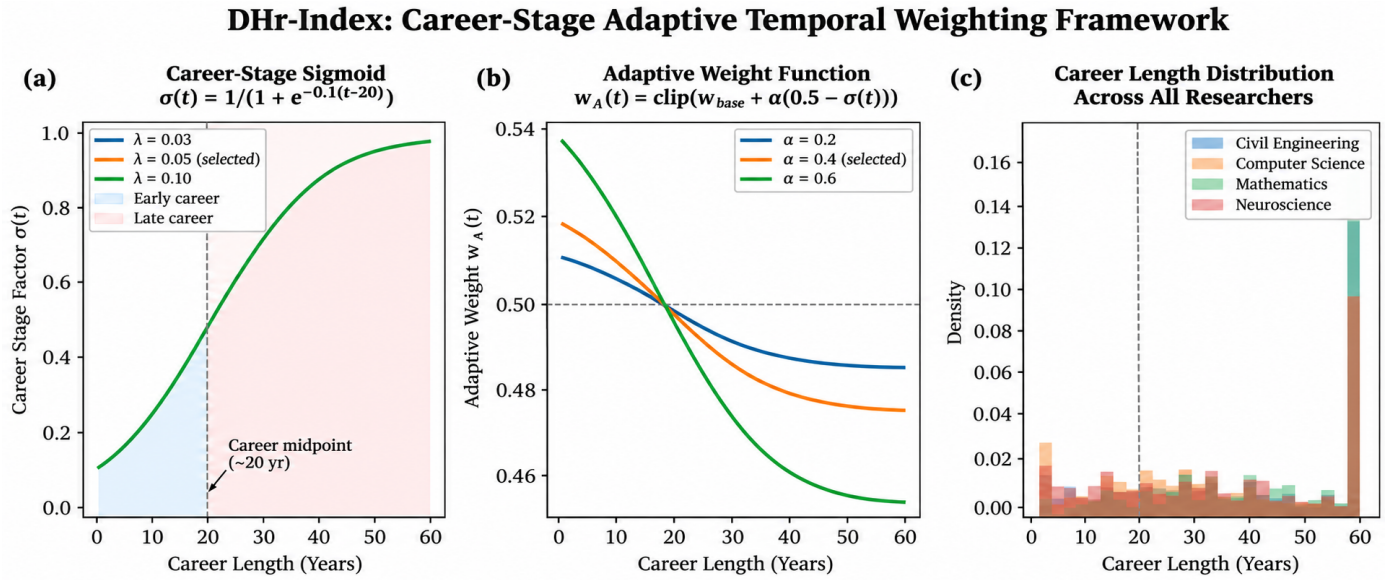

Figure 3. Illustration of the DHr career-adaptive weighting function. (**a**) Career-stage sigmoid function $\sigma(t)$ for the illustrated decay constants, with the reference setting identified in the legend as selected. (**b**) Adaptive weight function $w_A(t)$ for different values of the shift parameter α; $\alpha = 0.4$ is identified as the selected reference setting. (**c**) Distribution of career lengths across the four datasets; the colors represent Civil Engineering, Computer Science, Mathematics, and Neuroscience, respectively. The revised nested-CV analysis uses the prespecified DHr reference settings stated in Section 3.6; the alternative curves are illustrative only and are not candidate settings selected separately in each outer fold.

The top-100 awardee count $S^{(d)}$, defined in Section 3.3, is retained as a historical screening criterion and a secondary, threshold-specific descriptive ranking summary, facilitating comparison with prior bibliometric aggregation work [24,42]. The primary revised performance comparison instead uses AUC on nested-cross-validation out-of-fold predictions, which does not require choosing k. This distinction does not retroactively make the original top-100 screening independent of the award labels.

3.7.2. Layer 2 AUC-ROC Analysis

To complement ranking-based evaluation with a threshold-independent discriminative measure, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is computed for each composite index across all domains. Unlike ranking-based scores that depend on the choice of k, AUC-ROC evaluates discriminative capacity across all possible classification thresholds, ranging from 0.5 (random classification) to 1.0 (perfect discrimination). This provides a complete and threshold-independent characterisation of index performance.

3.7.3. Statistical Uncertainty and Direct Comparison

The revised primary evaluation uses nested cross-validation and paired bootstrap confidence intervals for the difference in AUC between Hr and DHr. The Mann–Whitney statistics from the original full-data analysis were not retained because they could not be reconciled with the reported AUCs and group sizes using the verified materials available for this revision. No U statistic or p-value is inferred from a rounded AUC. Statistical significance against a random-ranking baseline, if reported in a future analysis, would not by itself demonstrate superiority over existing indicators or validate award status as a measure of research quality.

4. Results

This section presents exploratory full-data screening summaries followed by the primary out-of-sample comparison. The full-data screening and aggregation figures are retained for transparency but are not used to support confirmatory performance claims. The primary comparison uses nested cross-validation and paired bootstrap uncertainty estimates.

4.1. Cross-Domain Parameter Screening Results

The ranking-based screening procedure was applied to all 64 bibliometric parameters within each of the four domains. For each parameter, researchers were ranked in descending order and the number of award recipients appearing within the top 100 positions was recorded as the discriminative score $S^{(d)}(p_j)$.

Figure 4 presents the comparative performance of all parameters across all four domains. The results reveal both within-domain stratification and important cross-domain variation in parameter performance.

Within civil engineering, the h^2-center index achieved the highest discriminative score, identifying 53 awardees within the top 100 ranked positions. The k'-index (49) and h^2-lower index followed closely, reinforcing the discriminative relevance of second-order h-index variants in this domain. Parameters such as the P-index (8), A-index (8), and m-index (6) displayed minimal discriminative effectiveness, suggesting poor alignment with externally validated recognition in civil engineering.

In computer science, the h^2-upper index achieved the highest discriminative score (73 awardees), followed by the fractional g-index and k-index. In mathematics, the normalized h-index led with 93 awardees, the highest single-parameter score observed across all four domains, while in neuroscience the h^2-upper index again dominated with 76 awardees.

Three cross-domain observations are of particular note. First, the top-performing parameter differs across all four domains: h^2-center in civil engineering, h^2-upper in computer science and neuroscience, and normalized h-index in mathematics. This descriptive full-data pattern motivates domain-specific screening, but does not itself establish out-of-sample performance or generalizability. Second, higher-order h-index variants performed well in the original screening, but this full-data comparison is exploratory and does not establish universal superiority over raw counts. Third, the magnitude of the original full-data screening scores varies substantially from 93 in mathematics to 53 in civil engineering. These differences may reflect citation practices, dataset composition, award processes, and other domain-specific factors; the current analysis does not isolate their causes.

4.2. Individual Analysis of Top-Performing Parameters

To further examine the behavior of the most discriminative metrics per domain, the top ten parameters identified in the screening stage were analyzed individually. Figure 5 presents their relative discriminative scores across all four domains.

Within each domain, the top-10 parameters exhibit a characteristic performance drop between the first and tenth ranked parameters. In civil engineering this drop spans from 53 to approximately 31 awardees; in computer science from 73 to approximately 55; in mathematics from 93 to approximately 74; and in neuroscience from 76 to approximately 55. This stratification describes the observed full-data rankings; it does not independently validate the top-10 threshold, which remains a design choice evaluated within the revised nested-CV procedure.

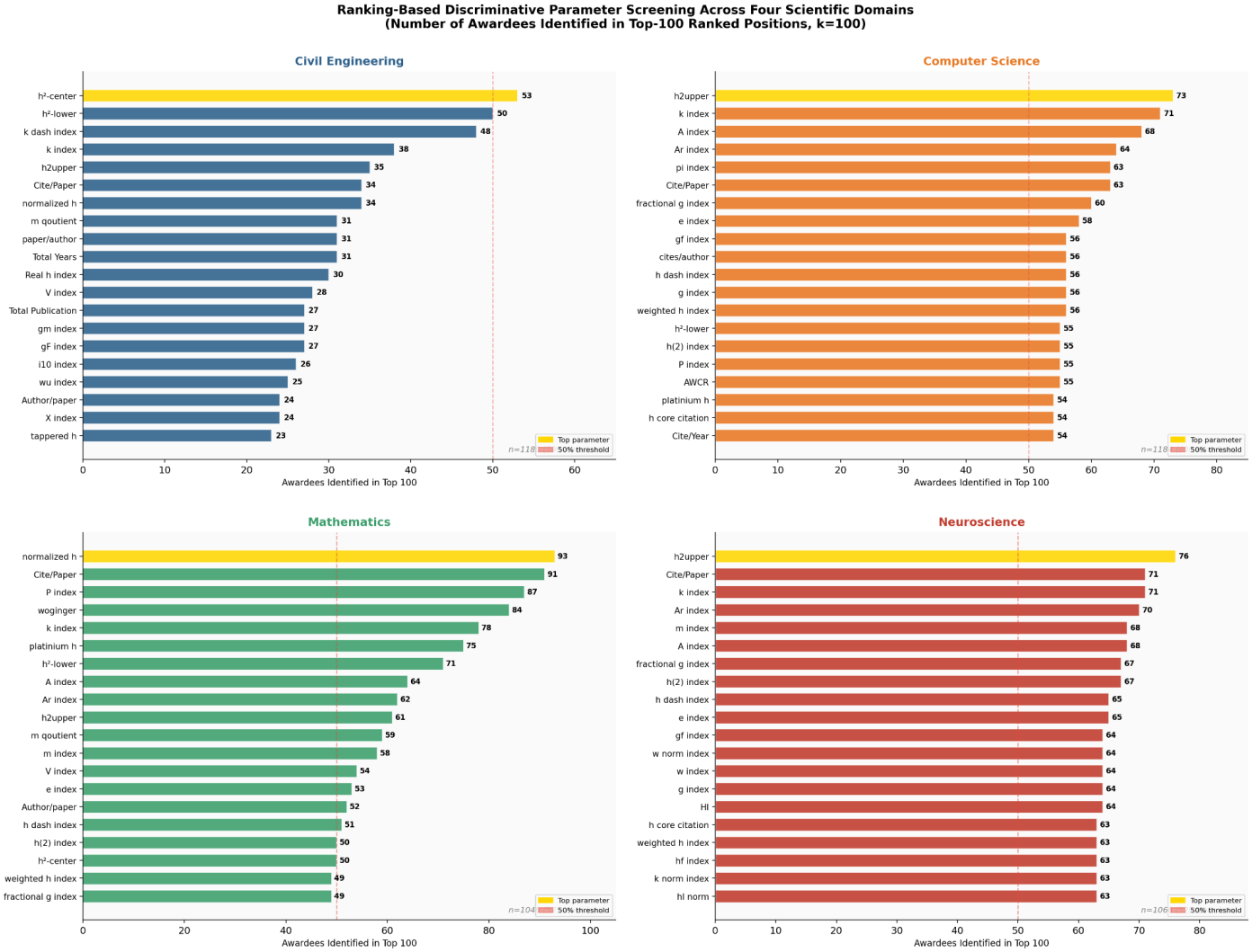

Figure 4. Ranking-based discriminative parameter screening across four scientific domains. Parameters are ranked by the number of awardees identified within the top 100 positions ($k = 100$). Blue, orange, green, and red bars represent Civil Engineering, Computer Science, Mathematics, and Neuroscience, respectively; the gold bar indicates the top-ranked parameter in each domain. The dashed vertical line marks the 50% threshold (50 awardees identified within the top 100 positions). Results reveal substantial cross-domain variation in parameter performance, confirming that no single metric dominates uniformly across disciplines.

The diversity of top-ranked metrics in the original full-data screening is consistent with the possibility that different indicators capture different aspects of scholarly activity. Because these rankings are exploratory, they motivate—but do not validate—the pairwise aggregation strategy described in Section 3.4.

4.3. Statistical Aggregation Model Comparison

To examine the effect of aggregation strategy on discriminative performance, seven statistical models were applied to the h^2-center × Real-*h* parameter pair across all four domains. This pair was selected for a descriptive cross-domain comparison based on its previously reported performance in the civil engineering analysis and its appearance among the top ten indicators in at least three domains. This provides a consistent descriptive comparison across domains.

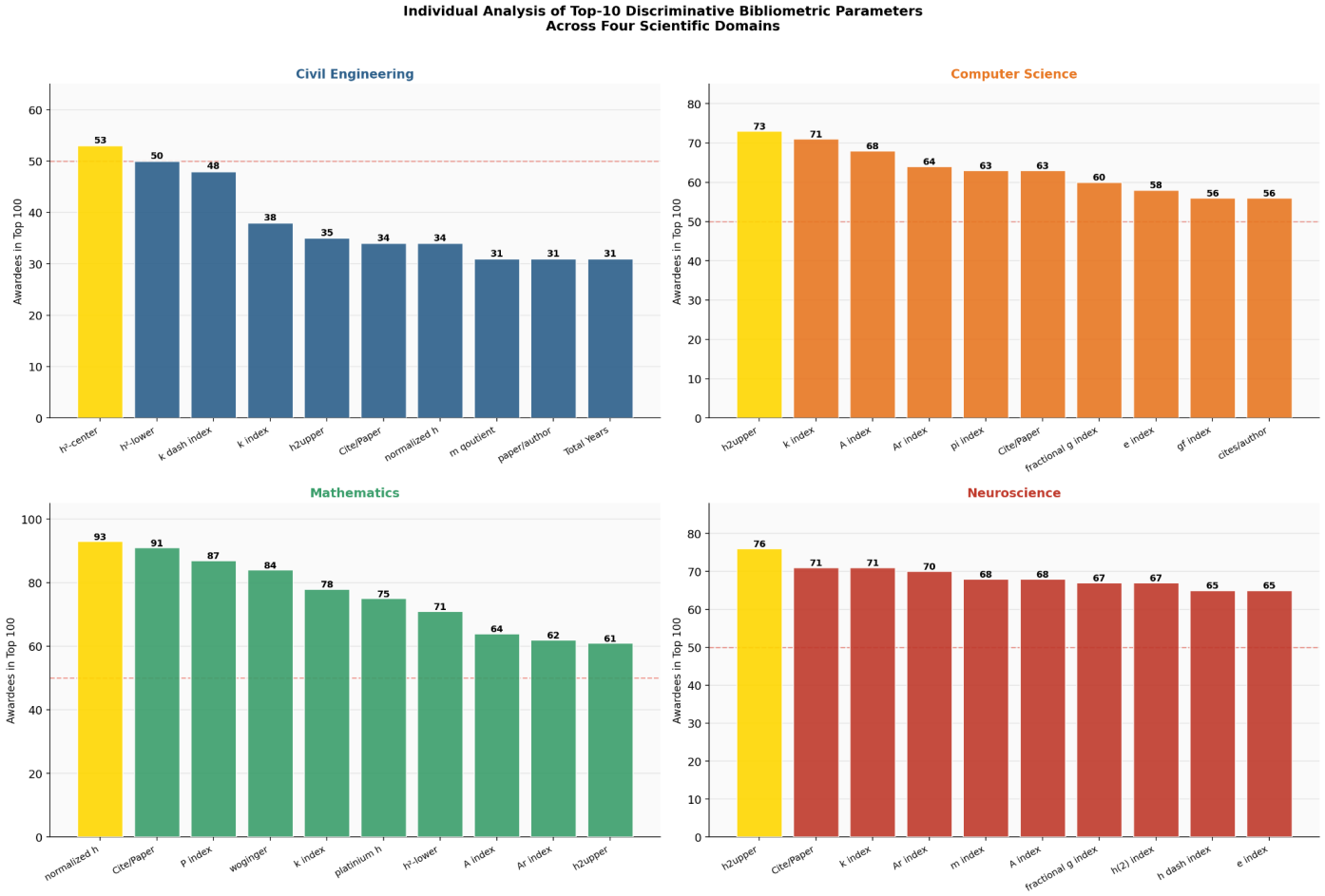

Figure 5. Individual analysis of the top-10 discriminative bibliometric parameters per domain. Blue, orange, green, and red bars represent Civil Engineering, Computer Science, Mathematics, and Neuroscience, respectively. Gold bars indicate the highest-scoring parameter in each domain, and the horizontal dashed line marks the 50-awardee reference threshold. The observed variation is descriptive and motivates examining domain-specific selection; it does not establish that such selection is necessary or that selected indicators transfer to unseen disciplines.

Figure 6 presents the discriminative scores of all seven aggregation models plus the Hr and DHr indices for this parameter pair across all four domains.

The original full-data comparison suggested differences among aggregation functions, but because pair selection and evaluation were not fully separated in that analysis, these results are exploratory. The revised nested-CV comparison does not evaluate every aggregation function under a single held-out protocol; therefore, no domain-general superiority claim for cubic or contra-harmonic means is made here.

The original comparison of 45 parameter pairs under seven aggregation functions was exploratory and susceptible to selection optimism from multiple candidate comparisons. Therefore, the previously observed maximum improvement should not be interpreted as an independently validated performance gain. In the revised analysis, parameter screening and pair selection occur within the training portion of nested cross-validation, and primary performance conclusions rely on held-out predictions. The revised confirmatory comparison concerns the Hr and DHr formulations; it does not establish that cubic aggregation is superior to every alternative mean. This distinction separates exploratory model development from out-of-sample performance assessment and avoids overstating improvements arising from multiple comparisons.

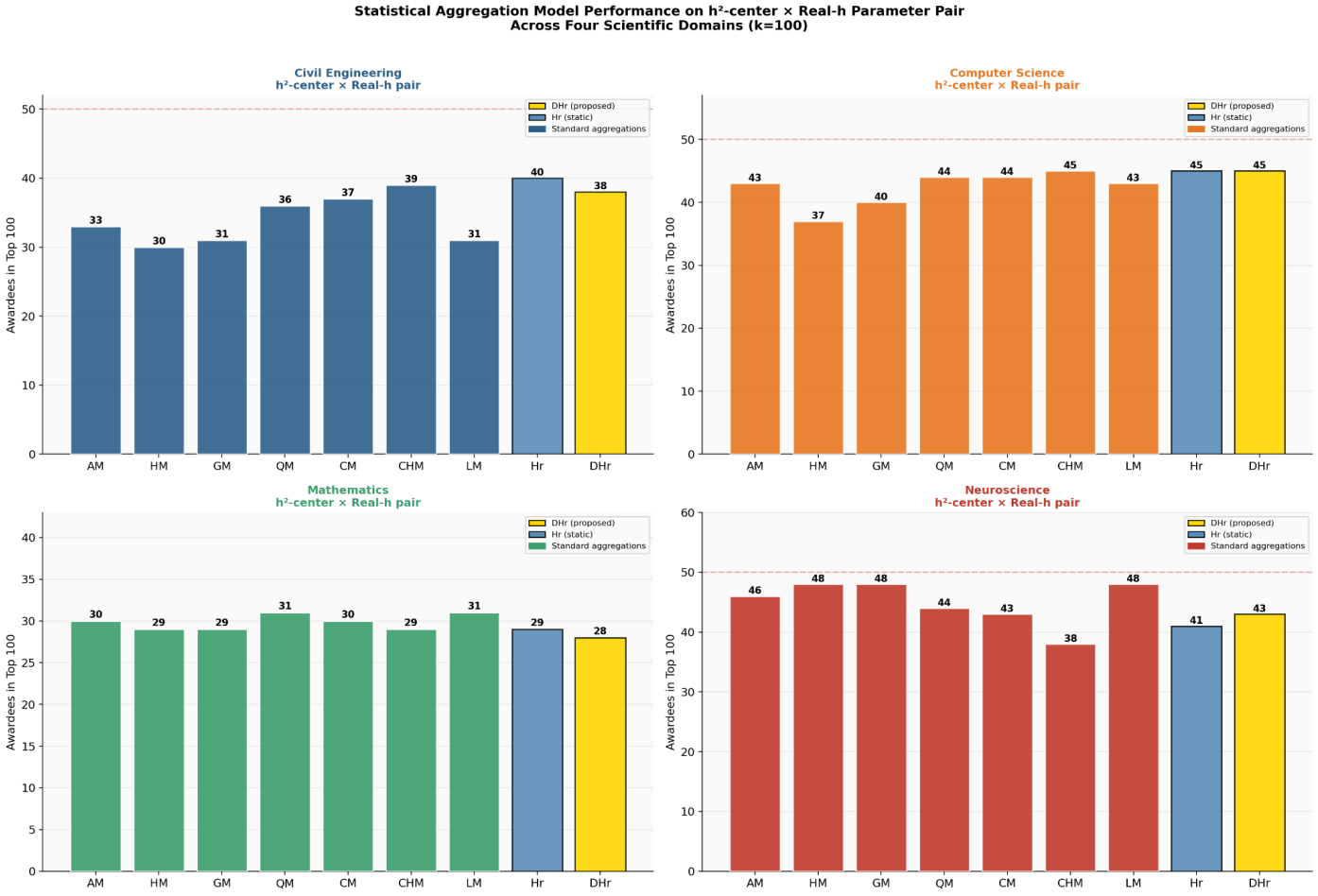

Figure 6. Comparative performance of seven statistical aggregation models and the proposed Hr and DHr indices applied to the h^2-center × Real-*h* parameter pair across four domains. Domain-colored bars (blue, orange, green, and red) denote the seven standard aggregation models for Civil Engineering, Computer Science, Mathematics, and Neuroscience, respectively; the light-blue bar indicates Hr, and the gold bar indicates DHr. The horizontal dashed line marks the 50-awardee reference threshold. Original full-data comparison of aggregation functions; these results are exploratory and do not establish held-out superiority.

The logarithmic mean (LM) exhibits inconsistent behavior across domains—competitive in some and weak in others—suggesting sensitivity to the specific citation distribution shape of each field and limiting its utility as a general-purpose aggregation model.

In the original full-data comparison, Hr and DHr sometimes ranked competitively relative to the standard CHM. Because this comparison was not fully held out, it is exploratory and does not establish generalizable superiority. Whether magnitude-proportional weighting or career-stage adaptation adds useful information remains an empirical question for independent evaluation.

4.4. Cross-Domain Performance Heatmap

To provide a comprehensive view of aggregation model performance across all domains and parameter pairs simultaneously, Figure 7 presents a cross-domain performance heatmap for all nine aggregation strategies evaluated on the domain-specific best parameter pair.

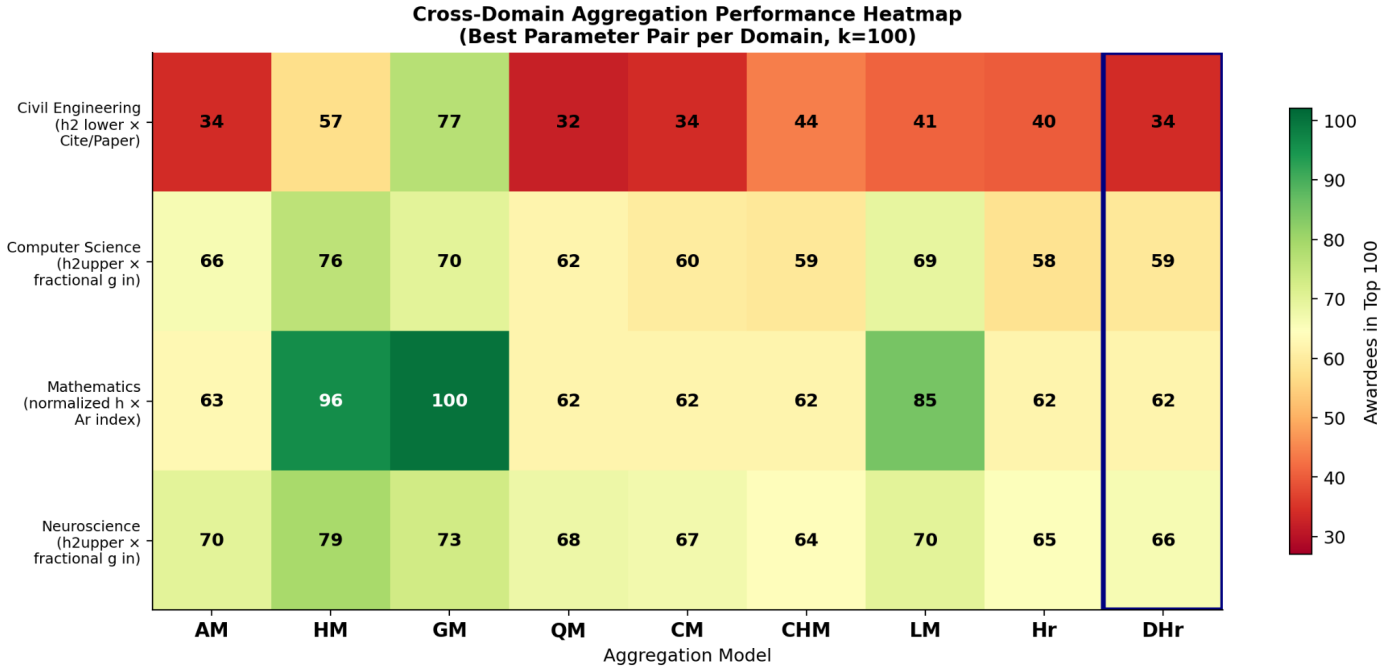

Figure 7. Cross-domain aggregation performance heatmap. Each cell reports the number of awardees identified in the top 100 positions ($k = 100$) for a given aggregation model (columns) applied to the best parameter pair per domain (rows). The numbers inside the cells are the corresponding counts; white and black text is used only for readability against the background color. The blue outline highlights the DHr column. The original full-data heatmap is exploratory and is not used to support comparative out-of-sample performance claims. Color scale from red (low) to green (high).

The heatmap is a descriptive summary of the original full-data analysis. Because candidate selection and evaluation were not fully separated, it should not be used to infer out-of-sample superiority among aggregation models. The nested-CV comparison does not show a consistent DHr advantage over Hr.

4.5. Hr-Index and DHr-Index Performance Results

Table 4 reports the primary nested-CV comparison of the best single metric, Hr, and DHr using pooled out-of-fold predictions.

Table 5 reports paired DeLong tests for the three pairwise AUC comparisons. Unadjusted p-values are shown first and Holm-adjusted p-values across the 12 comparisons are in parentheses. None of the pairwise comparisons remain statistically significant after this family-wise adjustment. Because the tests use pooled out-of-fold predictions, they compare the observed ranking scores but do not fully capture uncertainty from retraining the entire model-selection procedure; they are therefore interpreted alongside, rather than instead of, the paired bootstrap analysis.

Table 4. Nested cross-validation comparison of the best single metric, Hr, and DHr using pooled out-of-fold predictions. Parentheses show stratified-bootstrap 95% confidence intervals for each AUC; the final column shows the paired-bootstrap Hr–DHr difference and its 95% confidence interval. AUCs describe prediction of award-recognition status, not research quality.

Domain	Best Single AUC (95% CI)	Hr AUC (95% CI)	DHr AUC (95% CI)	Hr–DHr Difference (95% CI)
Civil Engineering	0.574 (0.540, 0.605)	0.556 (0.520, 0.589)	0.559 (0.524, 0.591)	−0.0034 (−0.0165, 0.0096)
Computer Science	0.691 (0.660, 0.723)	0.692 (0.661, 0.723)	0.690 (0.660, 0.722)	0.0014 (−0.0092, 0.0120)
Mathematics	0.782 (0.754, 0.811)	0.768 (0.742, 0.798)	0.767 (0.740, 0.798)	0.0011 (−0.0035, 0.0056)
Neuroscience	0.727 (0.698, 0.758)	0.713 (0.683, 0.743)	0.712 (0.681, 0.743)	0.0008 (−0.0029, 0.0045)

Table 5. Paired DeLong tests comparing pooled out-of-fold AUCs. Each cell reports the unadjusted p-value followed by the Holm-adjusted p-value in parentheses across 12 pairwise tests.

Domain	Best Single vs. Hr	Best Single vs. DHr	Hr vs. DHr
Civil Engineering	0.214 (1.000)	0.210 (1.000)	0.606 (1.000)
Computer Science	0.927 (1.000)	0.972 (1.000)	0.795 (1.000)
Mathematics	0.035 (0.347)	0.020 (0.242)	0.641 (1.000)
Neuroscience	0.037 (0.347)	0.025 (0.278)	0.656 (1.000)

Figure 8 visualizes the nested-CV AUC values in Table 4.

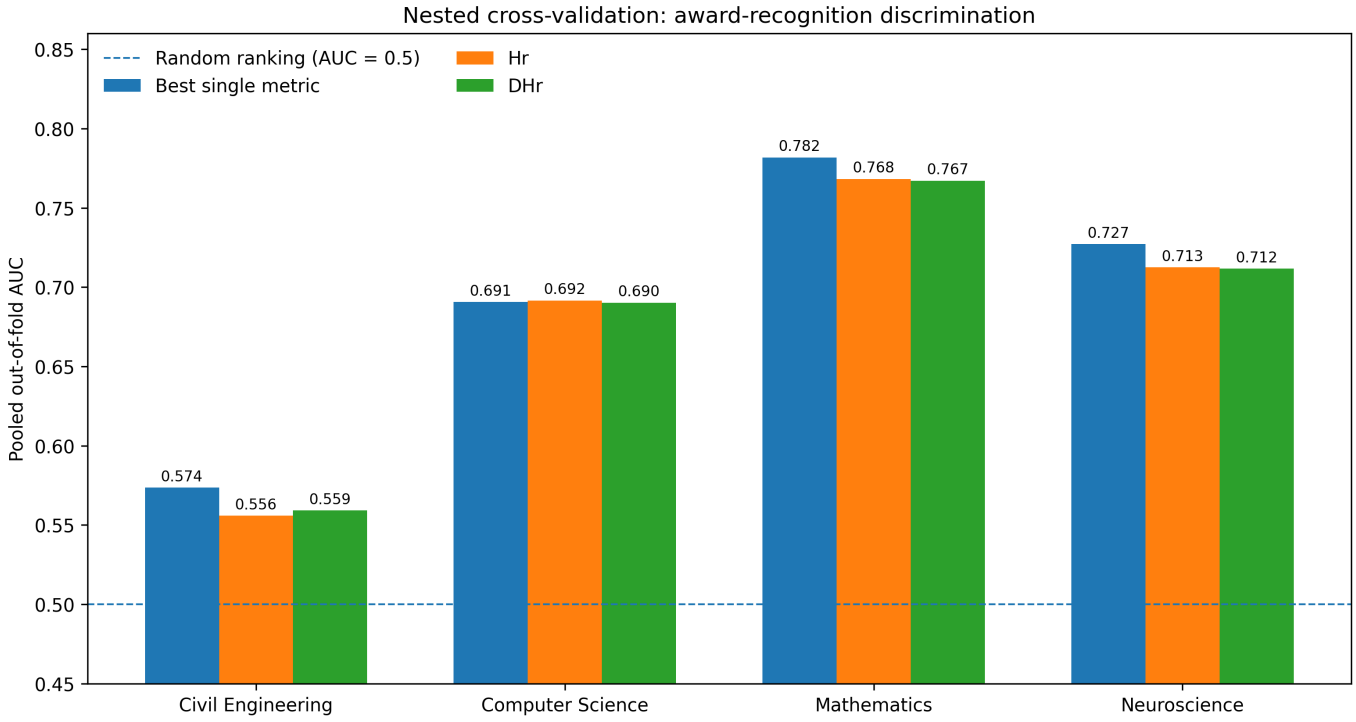

Figure 8. AUC comparison for the best single metric, Hr, and DHr using pooled nested-CV out-of-fold predictions. The horizontal dashed line indicates AUC = 0.5.

The nested-CV results show no consistent advantage for DHr. DHr has a slightly higher point-estimate AUC than Hr in civil engineering, while Hr has slightly higher point estimates in computer science, mathematics, and neuroscience. The paired bootstrap confidence intervals for all four AUC differences include zero. The best single metric has the highest point estimate in civil engineering, mathematics, and neuroscience; in computer science, Hr is only marginally higher than the best single metric. These point-estimate differences should not be interpreted as proof of superiority: the confidence intervals overlap substantially, and the paired-bootstrap Hr–DHr intervals include zero in every domain. The results do not support a claim that DHr is generally superior to Hr or to established single indicators.

Figure 9 presents the number of awardees among the top 100 researchers under each pooled nested-CV out-of-fold ranking.



Figure 9. Top-100 awardee counts for the best single metric, Hr, and DHr calculated from pooled nested-CV out-of-fold predictions. These are secondary threshold-specific summaries and should be interpreted alongside AUCs and their uncertainty.

4.5.1. Sensitivity to the Top-k Cut-off

Because the top-100 count depends on a single ranking threshold, we examined the pooled out-of-fold rankings at $k = 50$, 100, and 200. Table 6 reports the number of awardees in each top-k set for the best single metric, Hr, and DHr. The relative ordering varies with both domain and cut-off; no index consistently identifies more awardees at every threshold. These counts describe sensitivity of the final ranking outcomes, not sensitivity of the feature-screening or pair-selection decisions; they are not independent confirmatory tests. The original screening considered multiple indicators, pairs, and aggregation functions; selection within the inner folds reduces reuse of the outer test folds for selection, but it does not eliminate the need for independent replication.

Table 6. Top-k awardee counts from pooled out-of-fold scores for three ranking cut-offs. Counts are threshold-dependent descriptive summaries.

Domain	k	Best Single Metric	Hr	DHr
Civil Engineering	50	18	12	7
	100	30	35	40
	200	92	108	104
Computer Science	50	30	31	35
	100	68	68	73
	200	137	135	138
Mathematics	50	43	42	41
	100	90	88	88
	200	169	171	172
Neuroscience	50	24	27	27
	100	61	66	63
	200	142	140	142

4.5.2. Screening Threshold, Search-Space Sensitivity, and Multiplicity

The choice of $k = 100$ in the original screening is a fixed reporting convention rather than an empirically established optimum. The top-k analysis in Table 6 examines the final out-of-fold rankings at three cut-offs, not the sensitivity of the *screening and selection pipeline* to k. Consequently, the present evidence does not establish that the top-ten indicators, the selected metric pairs, or the preferred aggregation configuration would remain unchanged if screening were repeated at $k = 50$ or $k = 200$. We do not report such selection-stability estimates because the corresponding reruns have not been performed. A fully responsive future sensitivity study would repeat the complete training-fold-only screening and pair-selection procedure at prespecified values of k, report top-ten overlap and pair-selection frequencies, and evaluate the resulting pipelines on untouched outer folds.

The original search comprised 64 individual indicators per domain, followed by $\binom{10}{2} = 45$ pairs and seven aggregation functions (up to 315 pair–function configurations per domain, before considering any additional tuning). This count describes the candidate search space, not 315 independent hypothesis tests: the candidates are correlated and the stages of selection are dependent. A larger search space can increase selection optimism when the same award labels are used for both choosing and evaluating configurations. The original full-data best-of-search values are therefore descriptive and must not be interpreted as multiplicity-adjusted confirmatory evidence. The revised nested-cross-validation protocol limits this particular optimism by restricting selection to training data and reserving outer folds for evaluation; however, we have not conducted a separate empirical search-space-size sensitivity experiment, so robustness to the number of screened indicators or candidate combinations remains unknown.

Multiplicity correction must be matched to the family of *inferential tests*, rather than mechanically applied to the number of candidates inspected during descriptive ranking. Table 5 reports Holm-adjusted p-values across its 12 prespecified paired AUC comparisons. That correction does not cover the original 64-indicator screening, the 45 pairs, the seven aggregation functions, or any unreported threshold-dependent selections, and it should not be described as doing so. We make no formal significance claim for the exploratory screening or aggregation search. If future work performs hypothesis testing across those candidates, it should define the relevant test families in advance and use an appropriate family-wise or false-discovery-rate procedure; independent external validation would still be needed to assess transportability.

4.6. Statistical Validation

4.6.1. Score Distribution Analysis

Figure 10 presents score distributions from the original full-data analysis. These descriptive plots are distinct from the revised nested-CV evaluation and should not be used to infer out-of-sample performance.

The original full-data score distributions are retained only as descriptive context because parameter selection and evaluation were not fully separated in that analysis. They are not used as confirmatory evidence or as support for the revised out-of-sample performance claims.

4.6.2. Nested Cross-Validation Performance Comparison

The revised evaluation used five outer folds and three inner folds, with author-level grouping. Feature screening and metric-pair selection were conducted within the inner training folds, while the DHr career-weighting settings were held at the prespecified reference configuration; all reported AUCs below are calculated from pooled out-of-fold predictions. For avoidance of doubt, the reported DHr results use the fixed settings $\alpha = 0.4$,

sigmoid slope 0.1, career midpoint 20 years, and clipping interval [0.05, 0.95] throughout the nested-CV evaluation; these settings were not tuned using outer-fold performance. This evaluation is separate from the original full-data analysis.

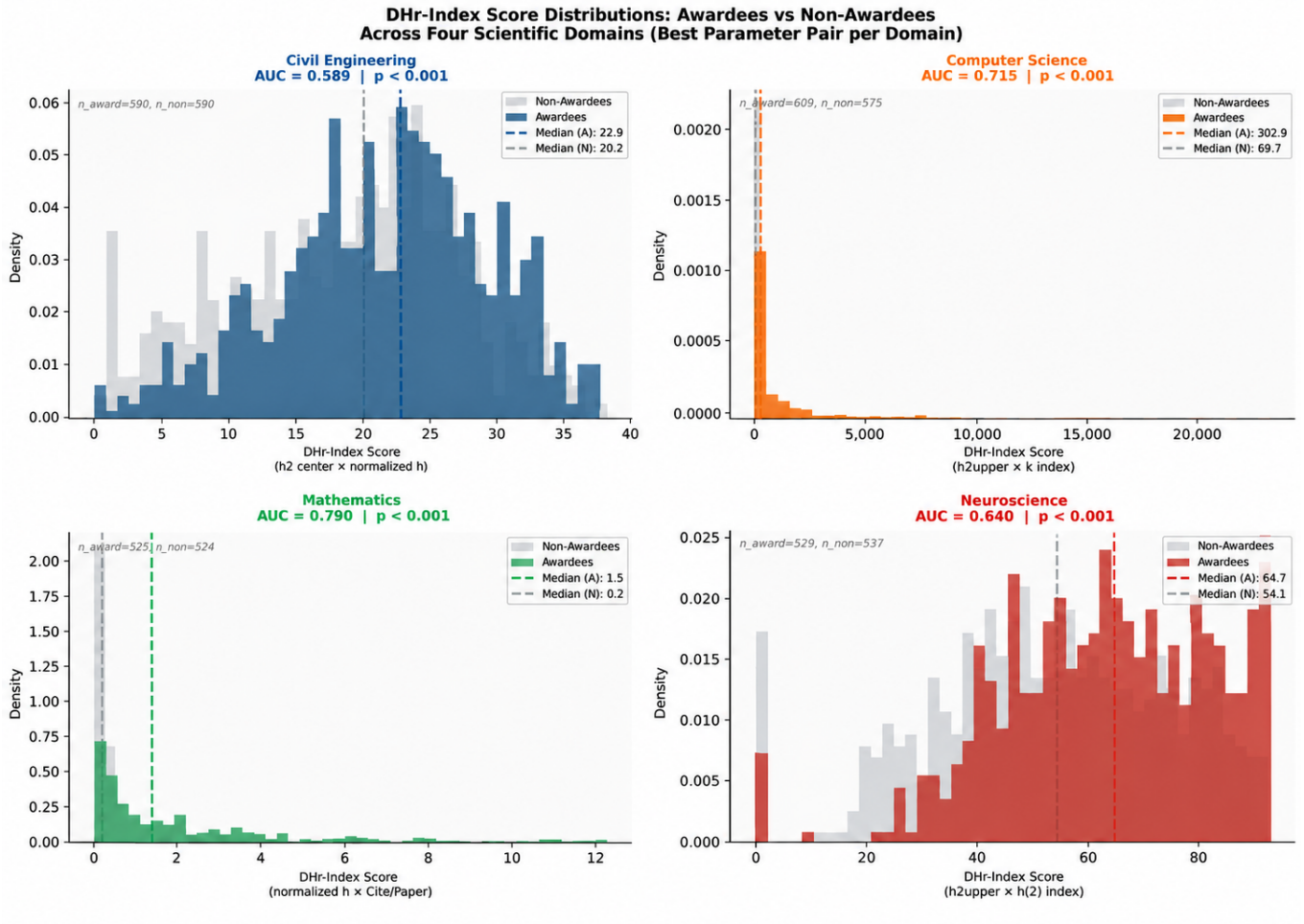

Figure 10. DHr-index score distributions for awardees (colored) and non-awardees (grey) across four domains. Vertical dashed lines indicate group medians. These descriptive original-analysis distributions are not the basis for the revised out-of-sample performance claims.

The nested cross-validation design separates model selection from performance estimation. The five outer folds provide held-out evaluation subsets, while the three inner folds are used for parameter screening and metric-pair selection. The DHr career-weighting settings are fixed at the prespecified reference configuration rather than optimized on the outer evaluation data. Selection operations are restricted to the corresponding training data, and the resulting configurations are evaluated on researchers excluded from selection. Pooled out-of-fold predictions are then used to calculate AUC estimates and statistical comparisons. This procedure addresses the in-sample evaluation concern arising when the best-performing combinations are selected and assessed on the same observations. Importantly, an outer-test researcher's award label is not used to screen the 64 candidate indicators, select an indicator pair, choose the aggregation configuration, or otherwise use that researcher's outcome during model selection for that researcher's evaluation fold. The outer-fold predictions are generated only after these decisions have been made using the corresponding training partition. Thus, the outer folds serve as internal held-out test samples, not as a separately recruited external cohort. This distinction limits claims to internally validated, within-domain discrimination; it does not establish performance on independently collected researchers or unseen disciplines.

Because the original researcher-level DHr scores used to produce the earlier Mann–Whitney table could not be verified against the reported statistics, the inconsistent U statistics and associated p-values have been removed rather than reconstructed.

The paired bootstrap confidence intervals for the Hr–DHr AUC difference include zero in all four domains. The point estimates are small and change direction across domains. Thus, the nested-CV results do not provide statistically significant evidence that DHr improves AUC over Hr. Nor should significance against a random-ranking baseline be interpreted as evidence that either composite is superior to established indicators. The original Mann–Whitney statistics are not reported because they could not be reconciled with the stated group sizes and AUCs using the available verified score files. The retained materials do not provide a complete score-level audit trail sufficient to reconstruct whether the earlier discrepancy arose from a transcription, versioning, or analysis-output error; we therefore cannot identify its historical source with confidence. Rather than speculate, we removed the unverifiable statistics and based the revised inferential comparison on reproducible out-of-fold AUCs, paired bootstrap intervals, and paired DeLong tests. No U statistic or p-value has been inferred from rounded AUC values.

The nested-CV comparison is the primary performance evidence in this revision. Original full-data results and figures are retained only where explicitly identified as exploratory and are not used to support the revised out-of-sample performance claims.

5. Discussion

The findings of this study yield several important insights that advance both the theoretical foundations and practical applicability of bibliometric composite index design. This section interprets the results in light of the four principal findings established above, situates them within the broader bibliometric literature, and addresses the methodological implications of the proposed framework.

5.1. Cross-Domain Parameter Behavior

Scope of Multi-Domain Evaluation

The same index formulations and evaluation protocol were examined in four disciplines, but indicator screening and candidate-pair selection were carried out separately for each discipline within its training data, while the DHr career-weighting settings were held at the prespecified reference configuration. Consequently, the reported nested cross-validation AUCs estimate within-domain discrimination for previously held-out researchers under domain-specific indicator and pair selection. They do not estimate performance when indicators or other data-dependent choices learned from one or more source disciplines are transferred unchanged to a new target discipline. A leave-one-domain-out design would instead hold out an entire discipline, select indicators and ordered pairs using only the other three disciplines, lock the prespecified scoring procedure, and then evaluate it in the held-out discipline. No such transfer experiment was conducted here; claims of cross-domain generalization are therefore not supported.

The original full-data screening identified different top-ranked indicators across domains. These patterns are descriptive and may motivate hypotheses about field differences, but they do not by themselves establish why the indicators differ or demonstrate that the same features will be selected in independent samples. Field-normalized comparisons and external validation are needed before drawing substantive conclusions about cross-field citation distributions.

The domain-specific variation cautions against assuming that a parameter pair selected in one field will transfer unchanged to another. The observed variation raises questions about whether fixed parameter pairs transfer between fields, but does not test transfer

directly. Domain-specific screening was used in this study; its necessity or superiority over a transferable, prespecified pair has not been established. Studies that validate composite indices in a single domain and claim general applicability should therefore be interpreted with caution [9,27].

5.2. Why Higher-Order Aggregation Works

The attraction of cubic aggregation is that it offers a transparent way to combine two indicators while allowing larger values to contribute more strongly. However, the fact that bibliometric distributions are often skewed does not itself establish that a cubic mean is superior: that claim requires a held-out comparison under a prespecified evaluation procedure. The revised nested-CV results presented here focus on the selected Hr and DHr formulations and do not independently establish a domain-general ranking of all aggregation functions.

Earlier full-data comparisons motivated the use of cubic aggregation, but these comparisons should be treated as exploratory unless the full selection process is repeated inside cross-validation. The present revised results do not justify recommending higher-order means as a default across disciplines.

5.3. Interpreting the Ceiling Effect in High-Performing Domains

The top-100 comparison does not show a uniform advantage for either composite: Hr and DHr identify more awardees than the best single metric in civil engineering, DHr does so in computer science, the best single metric leads in mathematics, and Hr leads in neuroscience. These rankings change with the top-k threshold (Table 6), so they should not be summarized as a general composite advantage or a general ceiling effect.

Top-k ranking counts and AUC capture different aspects of performance. AUC assesses ranking discrimination over all possible thresholds, whereas top-k counts focus on a single cut-off. Neither metric, on its own, establishes that an index measures research quality or improves on alternatives. In the nested-CV results, the AUC differences between Hr and DHr are small and statistically inconclusive in all four domains.

These differences illustrate that top-k counts and AUC summarize different aspects of ranking performance. The present analysis does not establish that the composites provide distributional robustness or other benefits beyond the reported metrics; such benefits would need to be defined in advance and evaluated directly.

Interpretation of Statistical and Practical Significance

Statistical significance should not be equated with practically meaningful improvement. Although the evaluated indices show different levels of award-status discrimination across the four disciplines, some AUC estimates are relatively close to 0.5 and therefore indicate limited discrimination. The 95% AUC confidence intervals in Table 4 and the paired DeLong comparisons with Holm-adjusted p-values in Table 5 help distinguish observed differences from sampling uncertainty. The AUC differences among DHr, Hr, and the best single indicators are generally modest, and the adjusted comparisons do not establish statistically significant superiority. These results should therefore be understood as within-domain associations with award recognition, not evidence that either composite offers a practically superior measure of research quality.

5.4. Career-Adaptive Weighting: Theoretical and Practical Implications

The DHr-index explores whether the relative contribution of two bibliometric indicators can vary with career length. This is a plausible design hypothesis, but the presence of a sigmoid weighting function does not establish that the resulting score is more valid or

predictive. In the nested cross-validation analysis, the AUC differences between DHr and the static Hr index were small, their direction varied by domain, and all paired bootstrap confidence intervals included zero. The current results therefore do not demonstrate a consistent advantage for career-adaptive weighting.

Interpretation of the Incremental Value of Career Adaptation

The original exploratory Civil Engineering top-100 comparison reported a decrease in identified awardees from 65 under the static formulation to 56 after career adjustment. This specific ranking outcome does not support an advantage for the adaptive index. Moreover, the revised nested-CV AUCs for Hr versus DHr are 0.556 versus 0.559 in Civil Engineering, 0.692 versus 0.690 in Computer Science, 0.768 versus 0.767 in Mathematics, and 0.713 versus 0.712 in Neuroscience. These small, mixed differences, together with the absence of statistically significant paired improvements, do not demonstrate incremental predictive value from career adjustment. The DHr-index is therefore presented as an exploratory mechanism for incorporating career length, not as an empirically superior replacement for static Hr. Establishing whether the adaptive weighting has practical value would require prespecified career-stage-stratified comparisons and external validation; neither is established by the present analysis.

The parameters and metric pairs were selected within the inner folds to reduce information leakage. The revised scores use fold-specific percentile transformations of bibliometric indicators, so they should be described as operational percentile-normalized versions (Hr_pct and DHr_pct) rather than assumed equivalent to unnormalized scores from the original full-data analysis. Future work should test whether any apparent benefit persists on independent external datasets and examine subgroup performance before using such indices in consequential decisions.

Recent work provides useful context for interpreting these results without implying direct empirical comparability. Laimon's research-impact measures [21] highlight that publication volume and the structure of a researcher's output can influence conclusions drawn from cumulative citation measures. The DHr formulation addresses a different design question—whether a career-dependent weighting of two selected indicators adds discrimination—and was not benchmarked directly against Laimon's measures. Therefore, the present findings should not be interpreted as evidence of comparative superiority or career-stage fairness.

Şener et al. [22] examine how metric-based assessment systems may create barriers for researchers at different career stages. This underscores why a career-adaptive formula should not be assumed to correct career-related inequities merely because it contains a time-weighting function. The current award-recognition analysis does not test differential effects on early- and late-career subgroups; such effects require dedicated subgroup evaluation before use in consequential assessment.

The discussion of interpretable indices versus more complex predictive methods also benefits from the stability and generalization perspective considered by Meysami, Lotfi, and Saleem [28]. Although their study is not a bibliometric benchmark, it motivates distinguishing in-sample discrimination from robustness to model selection, regularization choices, and shifts in data. Nested cross-validation addresses one source of selection optimism here, but it does not demonstrate stability across institutions, time periods, indicator definitions, or alternative DHr parameter settings. Interpretability alone is not a substitute for these checks, just as model complexity alone is not evidence of stronger generalization.

5.5. Limitations and Generalizability

No Unseen-Domain Transfer Validation

Although four domains were studied, each was screened and tuned separately. The nested-CV results therefore quantify internal within-domain performance, not generalization to a discipline excluded from indicator selection, pair selection, or tuning. Leave-one-domain-out testing with a prespecified transfer protocol is needed to evaluate such generalization. The current results should be described as a multi-domain evaluation of a common framework, not as evidence of successful cross-domain transfer.

Several limitations constrain interpretation. First, the four datasets were assembled around professional-award recognition and therefore represent a selected population and a particular recognition process, not a representative sample of all researchers. Second, the primary outcome is award status, which may be influenced by nomination practices, eligibility rules, professional networks, visibility, geography, career stage, and historical inequities. Third, bibliometric values retrieved from Google Scholar may be affected by coverage differences, author disambiguation, document-type classification, and citation practices. Fourth, nested cross-validation reduces selection optimism within the available data, but it is internal validation and does not substitute for independent external validation across institutions, countries, databases, or time periods. Fifth, the nested-CV scores are operational percentile-normalized versions of the indices (Hr_pct and DHr_pct); their results should not be treated as direct validation of the original unnormalized full-data scores. Finally, the current primary comparison focuses on Hr and DHr and does not provide a uniform held-out comparison of every aggregation function considered in the original exploratory analysis. An additional limitation is that the top-ten screening and candidate-pair search were not repeated at alternative k values, and no search-space-size sensitivity analysis was conducted; thus, selection stability and the influence of search breadth are unquantified. Holm correction applies only to the 12 reported paired AUC tests, not the exploratory candidate search. The DHr formulation is also position-dependent: the available evaluation does not quantify the effect of reversing a and b, and the first-listed/second-listed convention alone does not verify a substantive prespecified orientation rule in the original implementation. Replication should record ordered indicator pairs explicitly and assess orientation sensitivity. The reference DHr midpoint (20 years), sigmoid slope (0.1), shift magnitude ($\alpha = 0.4$), and clipping limits ($[0.05, 0.95]$) have not been accompanied by a reported factorial or one-at-a-time sensitivity analysis; consequently, parameter stability and the influence of each setting remain unquantified. The nested training-fold selection protocol should not be interpreted as proof of robustness across these four parameter dimensions. These limitations mean that the results support further methodological investigation, not general deployment or claims of universal applicability.

Data-processing documentation and replicability. Although the four-domain datasets are publicly linked in the Data Availability Statement, the precise collection timestamp, record-level deduplication and missingness rules, author-identity verification, treatment of self-citations, and availability of the full 64-indicator computation scripts are not sufficiently documented to guarantee exact end-to-end replication. This limits the extent to which the published scores can be independently reconstructed from raw search results. We therefore distinguish reproducibility of analyses using the released dataset from reproducibility of the original data acquisition and feature-construction pipeline; future releases should include versioned extraction dates, processing rules, indicator definitions, and executable scripts.

5.6. Composite Indicators and Responsible Research Assessment

The present study should be situated within the broader literature on composite bibliometric indicators and responsible research assessment. Ioannidis, Klavans, and Boyack

(2016) combined multiple citation indicators to represent researchers' citation profiles and showed that composite rankings can identify different researchers from those identified by individual indicators [18]. The contribution here is therefore not the general idea of combining indicators, but the evaluation of a particular weighted-cubic framework and its career-dependent extension across four research domains.

Composite construction alone does not guarantee valid or fair comparison. In particular, Ioannidis et al. provide a standardized author-level citation database with field and subfield information, whereas field- and source-normalized citation indicators address systematic differences in citation practices [19,51]. Such methods provide important comparators, but this study does not claim to have benchmarked DHr directly against them. Citation practices vary across disciplines, publication years, document types, and database coverage; field normalization and transparent reporting of definitions and limitations are important when comparing researchers [19,51]. The Leiden Manifesto and the Declaration on Research Assessment (DORA) emphasize that quantitative indicators should be used in context and should support, rather than replace, informed qualitative judgment [2,17].

Because award dates have not been linked to verified pre-award bibliometric snapshots, temporal leakage cannot be excluded. The comparisons should not be interpreted as forecasts of future awards, and the absence of documented matching beyond discipline limits causal or merit-based interpretations.

A key limitation is the use of award status as the outcome for validation. Awards are observable recognition outcomes, but they are not neutral or exhaustive measures of research quality. Award decisions may reflect scholarship and influence, but can also be shaped by visibility, institutional and professional networks, nomination practices, gender-related inequities, country or regional representation, career stage, and society-specific eligibility criteria. These are plausible sources of bias rather than subgroup effects quantified in the present dataset. Discrimination between awardees and non-awardees should therefore be interpreted as prediction of an existing recognition outcome, not proof that an index measures scientific merit independently of the processes that produced that outcome. Moreover, the available documentation does not establish individual-level matching of non-awardee controls on career stage, institution, geographic region, or subfield. Although comparisons are organized by discipline, residual differences in these characteristics could confound the observed associations and limit the interpretation of AUC.

If an index trained or evaluated against award recognition is used in hiring, promotion, or funding decisions, historical differences in recognition may be reproduced or amplified, potentially disadvantaging groups underrepresented among award recipients. This study does not establish that Hr or DHr is free from such effects. Future work should examine performance across relevant subgroups, including gender and country or region where reliable and ethically appropriate data are available, and should report subgroup sample sizes, uncertainty, and limitations. Until independent external validation, comparisons with prespecified baselines, incremental-value analysis, and systematic fairness evaluation are available, DHr should be regarded as an exploratory composite indicator rather than a standalone decision rule.

6. Conclusions

This study examined a weighted cubic-mean composite (Hr) and a career-adaptive extension (DHr) across civil engineering, computer science, mathematics, and neuroscience. The revised evaluation used separate within-domain nested cross-validation with researcher-level grouping, with candidate indicators and metric pairs selected within the inner folds of each respective domain, while the DHr career-weighting settings were held at the prespecified reference configuration. The resulting out-of-fold AUCs were 0.556, 0.692,

0.768, and 0.713 for Hr and 0.559, 0.690, 0.767, and 0.712 for DHr across the four domains, respectively. Paired bootstrap confidence intervals for the Hr–DHr AUC difference included zero in every domain. The analysis therefore does not establish that career-adaptive weighting consistently improves discrimination over static aggregation, nor does it establish superiority over the best individual metric.

The results support treating weighted-cubic aggregation and career-dependent weighting as hypotheses for further evaluation, not as validated replacements for established bibliometric indicators. Award status was used as an observable recognition outcome; it should not be interpreted as a definitive measure of research quality or as a bias-free ground truth. Any consequential application would require independent external validation, transparent and reproducible computation, direct comparisons with prespecified baselines, subgroup/fairness analysis, and qualitative assessment of researchers' contributions and circumstances. Quantitative indicators may provide useful supporting evidence, but should not replace expert judgment.

Future work should focus on three priorities: independent external validation on newly assembled researcher samples; explicit tests of cross-domain transfer using locked source-domain selections; and robustness analyses of the prespecified DHr career-weighting settings, including subgroup and fairness assessment where appropriate. These extensions are more important than expanding model complexity before the basic validity and transportability of the proposed composite have been established.

Author Contributions: Conceptualization, H.Z.M. and G.M.; methodology, H.Z.M.; software, H.Z.M.; validation, H.Z.M., G.M. and M.T.A.; formal analysis, H.Z.M.; investigation, H.Z.M.; resources, G.M.; data curation, H.Z.M.; writing—original draft preparation, H.Z.M.; writing—review and editing, G.M., M.T.A., A.S.A.-S., I.D.l.T.D., E.S.A. and H.G.; visualization, H.Z.M.; supervision, M.T.A.; project administration, M.T.A.; funding acquisition, A.S.A.-S., I.D.l.T.D., E.S.A. and H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FUNDACIÓN UNIVERSIDAD EUROPEA DEL ATLÁNTICO (CIF: ESG39764972), C/ISABEL TORRES 21, 39011 Santander, Spain.

Data Availability Statement: The researcher-level bibliometric datasets for the four disciplines examined in this study are publicly available at https://github.com/Dr-GMustafa/Scientometrics_Related_Dataset (accessed on 5 November 2024). Availability of the complete source-data extraction history and executable code used to compute all 64 indicators has not been independently established in this manuscript.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT 5.6 (OpenAI) and Claude Sonnet 5.5 (Anthropic) to improve English-language clarity, sentence structure, and readability, and Grammarly 1.203.0.0 for grammar, spelling, punctuation, and proofreading. The authors reviewed and edited the AI-assisted output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bihari, A.; Tripathi, S.; Deepak, A. A review on h-index and its alternative indices. *J. Inf. Sci.* **2023**, *49*, 624–665. [CrossRef]
2. Hicks, D.; Wouters, P.; Waltman, L.; De Rijcke, S.; Rafols, I. Bibliometrics: The Leiden Manifesto for research metrics. *Nature* **2015**, *520*, 429–431. [CrossRef]
3. Aksnes, D.W.; Langfeldt, L.; Wouters, P. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open* **2019**, *9*, 2158244019829575. [CrossRef]
4. Yudianto, M.R.A.; Jati, H.; Arifin, F. Bibliometric Analysis of Aspect-based Sentiment Analysis Research: Trends, Challenges, and Future Directions. *Int. J. Mod. Educ. Comput. Sci.* **2026**, *18*, 68–86. [CrossRef]
5. Hirsch, J.E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572. [CrossRef]

6. Alonso, S.; Cabrerizo, F.J.; Herrera-Viedma, E.; Herrera, F. H-Index: A review focused in its variants, computation and standardization for different scientific fields. *J. Inf.* **2009**, *3*, 273–289. [CrossRef]
7. Nuermaimaiti, R. Statistical Models for Frequency Distributions of Count Data with Applications to Scientometrics. Ph.D. Thesis, University of Leeds, Leeds, UK, 2023.
8. Kuchanskyi, O.; Amangeldi, M. From citation counts to computational science: Methods, data, and applications in scientometrics . *Sci. Anal. J. High. Sch. Kazakhstan* **2026**, *3*, 93–107. [CrossRef]
9. Bornmann, L.; Mutz, R.; Daniel, H.D. The h index research output measurement: Two approaches to enhance its accuracy. *J. Inf.* **2010**, *4*, 407–414. [CrossRef]
10. Jin, B.; Liang, L.; Rousseau, R.; Egghe, L. The R-and AR-indices: Complementing the h-index. *Chin. Sci. Bull.* **2007**, *52*, 855–863. [CrossRef]
11. Sidiropoulos, A.; Katsaros, D.; Manolopoulos, Y. Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics* **2007**, *72*, 253–280. [CrossRef]
12. Burrell, Q.L. On the h-index, the size of the Hirsch core and Jin's A-index. *J. Inf.* **2007**, *1*, 170–177. [CrossRef]
13. Tol, R. The h-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics* **2009**, *80*, 317–324. [CrossRef]
14. Kosmulski, M. MAXPROD-a new index for assessment of the scientific output of an individual, and a comparison with the h-index. *Cybermetr. Int. J. Sci. Inf. Bibliometr.* **2007**, *35*, 5.
15. Cabrerizo, F.J.; Alonso, S.; Herrera-Viedma, E.; Herrera, F. Q2-Index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. *J. Inf.* **2010**, *4*, 23–28. [CrossRef]
16. Wu, Q. The w-index: A measure to assess scientific impact by focusing on widely cited papers. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 609–614. [CrossRef]
17. San Francisco Declaration on Research Assessment. San Francisco Declaration on Research Assessment (DORA), 2012. Declaration Developed in 2012 and Publicly Released in 2013. Available online: https://sfdora.org/read/ (accessed on 5 November 2024).
18. Ioannidis, J.P.A.; Klavans, R.; Boyack, K.W. Multiple Citation Indicators and Their Composite across Scientific Disciplines. *PLoS Biol.* **2016**, *14*, e1002501. Correction in *PLoS Biol.* **2016**, *14*, e1002548. [CrossRef]
19. Waltman, L.; van Eck, N.J. Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics* **2013**, *96*, 699–716. [CrossRef]
20. Panaretos, J.; Malesios, C. Assessing scientific research performance and impact with single indices. *Scientometrics* **2009**, *81*, 635–670. [CrossRef]
21. Laimon, M. Beyond cumulative impact: The he-index and a volume-normalized efficiency metric for research evaluation. *Front. Res. Metr. Anal.* **2026**, *11*, 1793664. [CrossRef]
22. Şener, H.S.; Kaya, İ.S.; Köksal, M.S.; Taşkın, Z. Closing the door behind: Metric-based research evaluation systems and gatekeeping towards young researchers. *Scientometrics* **2025**, *130*, 2291–2310. [CrossRef]
23. Ayaz, S.; Afzal, M.T. Identification of conversion factor for completing-h index for the field of mathematics. *Scientometrics* **2016**, *109*, 1511–1524. [CrossRef]
24. Raheel, M.; Ayaz, S.; Afzal, M.T. Evaluation of h-index, its variants and extensions based on publication age & citation intensity in civil engineering. *Scientometrics* **2018**, *114*, 1107–1127. [CrossRef]
25. Ain, Q.U.; Riaz, H.; Afzal, M.T. Evaluation of h-index and its citation intensity based variants in the field of mathematics. *Scientometrics* **2019**, *119*, 187–211. [CrossRef]
26. Ameer, M.; Afzal, M.T. Evaluation of h-index and its qualitative and quantitative variants in Neuroscience. *Scientometrics* **2019**, *121*, 653–673. [CrossRef]
27. Waltman, L.; van Eck, N.J. A systematic empirical comparison of different approaches for normalizing citation impact indicators. *J. Inf.* **2013**, *7*, 833–849. [CrossRef]
28. Meysami, M.; Lotfi, A.; Saleem, S. Deep learning and the geometry of compactness in stability and generalization. *Mach. Learn. Appl.* **2025**, 23, 100820. [CrossRef]
29. Ivanova, M.; Grosseck, G.; Holotescu, C. Unveiling Insights: A Bibliometric Analysis of Artificial Intelligence in Teaching. *Informatics* **2024**, *11*, 10. [CrossRef]
30. Omarsaib, M.; Mitha, S.B.; Vahed, A.; Mohamed, G.M. Mapping the AI Surge in Higher Education: A Bibliometric Study Spanning a Decade (2015–2025). *Informatics* **2025**, *12*, 137. [CrossRef]
31. Favara, G.; Barchitta, M.; Maugeri, A.; Magnano San Lio, R.; Agodi, A. The Research Interest in ChatGPT and Other Natural Language Processing Tools from a Public Health Perspective: A Bibliometric Analysis. *Informatics* **2024**, *11*, 13. [CrossRef]
32. Akpan, I.J.; Offodile, O.F.; Akpanobong, A.C.; Kobara, Y.M. A Comparative Analysis of Virtual Education Technology, E-Learning Systems Research Advances, and Digital Divide in the Global South. *Informatics* **2024**, *11*, 53. [CrossRef]
33. Haque, A.; Akther, N.; Khan, I.; Agarwal, K.; Uddin, N. Artificial Intelligence in Retail Marketing: Research Agenda Based on Bibliometric Reflection and Content Analysis (2000–2023). *Informatics* **2024**, *11*, 74. [CrossRef]

34. Nastasa, I.V.; Artamonov, A.R.; Busnatu, S.S.; Minca, D.G.; Andronic, O. From Innovation to Regulation: Insights from a Bibliometric Analysis of Research Patterns in Medical Data Governance. *Informatics* **2025**, *12*, 66. [CrossRef]
35. Wang, W.; Li, Z. The Evolution of Artificial Intelligence in Marketing: A Bibliometric Analysis of Three Decades (1992–2025). *Informatics* **2026**, *13*, 67. [CrossRef]
36. Kurasova, O.; Budzys, A.; Medvedev, V. Exploring Multidimensional Embeddings for Decision Support Using Advanced Visualization Techniques. *Informatics* **2024**, *11*, 11. [CrossRef]
37. Wu, Y.; Zhang, L.; Li, B.; Zhang, W. MCD-Temporal: Constructing a New Time-Entropy Enhanced Dynamic Weighted Heterogeneous Ensemble for Cognitive Level Classification. *Informatics* **2025**, *12*, 134. [CrossRef]
38. Xu, G.; Murthy, S.V.; Jia, B. Enhancing Intuitive Decision-Making and Reliance Through Human–AI Collaboration: A Review. *Informatics* **2025**, *12*, 135. [CrossRef]
39. Waltman, L.; van Eck, N.J.; van Leeuwen, T.N.; Visser, M.S.; van Raan, A.F.J. Towards a new crown indicator: An empirical analysis. *Scientometrics* **2011**, *87*, 467–481. [CrossRef]
40. Waltman, L.; Yan, E.; van Eck, N.J. A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics* **2011**, *89*, 301–314. [CrossRef]
41. Waltman, L. On the calculation of percentile-based bibliometric indicators. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 372–379. [CrossRef]
42. Thelwall, M. Three practical field normalised alternative indicator formulae for research evaluation. *J. Inf.* **2017**, *11*, 128–151. [CrossRef]
43. Bartneck, C.; Kokkelmans, S. Detecting h-index manipulation through self-citation analysis. *Scientometrics* **2011**, *87*, 85–98. [CrossRef]
44. Ferrara, E.; Romero, A.E. Scientific impact evaluation and the effect of self-citations: Mitigating the bias by discounting the h-index. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 2332–2339. [CrossRef]
45. King, M.M.; Bergstrom, C.T.; Correll, S.J.; Jacquet, J.; West, J.D. Men set their own cites high: Gender and self-citation across fields and over time. *Socius* **2017**, *3*, 2378023117738903. [CrossRef]
46. Quaderi, N.; Duine, M. The misuse of bibliometrics in research assessment infrastructures. *Inf. Serv. Use* **2024**, *44*, 23–26. [CrossRef]
47. Forthmann, B.; Doebler, P.; Mutz, R. Why summing up bibliometric indicators does not justify a composite indicator. *Scientometrics* **2024**, *129*, 7475–7499. [CrossRef]
48. Schreiber, M. A modification of the h-index: The hm-index accounts for multi-authored manuscripts. *J. Inf.* **2008**, *2*, 211–216. [CrossRef]
49. Schreiber, M. Fractionalized counting of publications for the g-index. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 2145–2150. [CrossRef]
50. Harzing, A.W.K.; Van der Wal, R. Google Scholar as a new source for citation analysis. *Ethics Sci. Environ. Politics* **2008**, *8*, 61–73. [CrossRef]
51. Ioannidis, J.P.A.; Baas, J.; Klavans, R.; Boyack, K.W. A standardized citation metrics author database annotated for scientific field. *PLoS Biol.* **2019**, *17*, e3000384. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.