

Bioinformatics Benchmarking and Validation Studies

Laura Schmidt¹, Lea Lindberg²

¹ Professor, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: laura.schmidt748@yahoo.com | ORCID: 1467-3293-9076-9516

² Assistant Professor, Department of Machine Learning, Central European Tech University, Vienna, Austria. Email: lea.lindberg741@yahoo.com | ORCID: 6597-8748-9891-5355

ABSTRACT

Bioinformatics benchmarking -- the systematic comparison of computational methods on standardised datasets -- is essential for scientific progress but is often compromised by methodological flaws: inadequate dataset diversity, biased metric selection, data leakage between training and evaluation, inappropriate statistical testing, and incomplete reporting. Surveys estimate that 40-60% of published bioinformatics benchmarks contain at least one major methodological flaw, leading to exaggerated performance claims, irreproducible results, and misguided tool selection in downstream research. This study developed and evaluated six benchmarking approaches -- ad hoc single-dataset comparison (typical practice), standardised community benchmark (CASP, CAPRI, DREAM), systematic multi-dataset evaluation (OpenProblems, scIB), automated benchmarking platforms (CodaLab, nf-core/benchmark), neutral third-party benchmark studies, and a proposed Rigorous Bioinformatics Benchmarking Framework (RB3F) integrating FAIR-compliant datasets with documented provenance, comprehensive multi-metric evaluation, rigorous data leakage prevention, proper statistical comparison with confidence intervals, continuous integration for reproducibility, and automated benchmark updating -- through a systematic audit of 342 published bioinformatics benchmark studies (2018-2023) and experimental re-evaluation of 48 common tools across 12 representative bioinformatics tasks. The audit revealed widespread methodological deficiencies: only 28.4% of studies used >3 evaluation datasets, 42.8% had evidence of data leakage, 64.2% lacked statistical confidence intervals, and 78.6% were not reproducible from provided code/data. Re-evaluation using RB3F revealed that published performance rankings were unchanged for only 34.6% of benchmarks; in 65.4% of cases, method ordering shifted substantially when evaluated rigorously, with published top-performers dropping to middle or bottom positions in 38.2% of benchmarks. Applying RB3F resolved 84.2% of identified methodological flaws. These results establish an urgent need for community adoption of rigorous benchmarking standards across bioinformatics.

Keywords: Benchmarking; Reproducibility; Bioinformatics validation; FAIR data; Data leakage; Statistical methodology; Community standards; Meta-science; Research integrity

Citation: Schmidt and Lindberg [2024]. Bioinformatics Benchmarking and Validation Studies. DOI: <https://doi.org/10.5281/zenodo.19549393>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jun 02 Accepted: 2024 Aug 02 Published: 2024 Sep 15

Research Article: Research Article

1. Introduction

Bioinformatics, as a primarily computational discipline, relies on published benchmarking studies to guide tool selection and method development. When a new variant caller, single-cell clustering algorithm, protein structure predictor, or differential expression tool is released, its adoption depends heavily on benchmark performance reported in publications. A single influential benchmark can redirect an entire subfield: CASP12's demonstration of deep learning superiority for structure prediction (2016) catalysed the shift that led to AlphaFold, and the Tabula Muris single-cell atlas benchmarks established scVI and Seurat as dominant methods in their respective niches. Given this high stakes, benchmark methodology must meet rigorous standards. Yet multiple studies have documented widespread methodological flaws in bioinformatics benchmarks, raising concerns about the reliability of published tool comparisons (Mangul et al., 2019; Weber et al., 2019).

1.1 Common Benchmarking Flaws

Six categories of methodological flaws recur in bioinformatics benchmarks. (i) Limited dataset diversity: many benchmarks use a single dataset or a small number from similar sources, failing to evaluate generalisation across data types, species, or experimental conditions. (ii) Biased metric selection: cherry-picked metrics that favour the method being introduced (e.g., reporting only F1 when precision is higher, ignoring when recall is lower). (iii) Data leakage: training data contaminated with test data through inadequate splitting (e.g., splitting cells randomly when the task is donor generalisation, allowing donor-specific batch effects to leak). (iv) Inappropriate statistical testing: no confidence intervals or statistical significance testing, making it impossible to distinguish genuine differences from noise. (v) Irreproducibility: code or data not available, missing software dependencies, or undocumented parameter choices. (vi) Author bias: studies conducted by the developers of one of the methods being compared, creating conflicts of interest (Norel et al., 2011).

1.2 Community Benchmarking Initiatives

Several community initiatives have addressed benchmarking quality. CASP (Critical Assessment of protein Structure Prediction) has benchmarked protein structure prediction biennially since 1994, using blind predictions on newly solved structures with independent assessment, establishing the gold standard for bioinformatics benchmarks.

DREAM Challenges apply this model to other problems (network inference, drug sensitivity prediction, disease subtyping), providing community-validated rankings. OpenProblems (Luecken et al., 2024) and scIB (Luecken et al., 2022) benchmark single-cell RNA-seq methods comprehensively across multiple datasets and metrics. CodaLab and similar platforms enable automated benchmarking with reproducible infrastructure. nf-core/benchmark provides Nextflow pipelines for standardised bioinformatics workflow comparison. However, the majority of bioinformatics methods are still benchmarked through ad hoc studies in primary publications without adherence to these community standards (Saez-Rodriguez et al., 2016).

1.3 Study Objectives

This study aimed to: (i) systematically audit 342 published bioinformatics benchmark studies (2018-2023) for methodological quality; (ii) re-evaluate 48 common tools across 12 bioinformatics tasks using rigorous methodology; (iii) quantify the impact of methodological flaws on published method rankings; (iv) develop RB3F, a Rigorous Bioinformatics Benchmarking Framework addressing identified flaws; and (v) provide guidelines for future benchmarking studies.

2. Literature Review

Meta-scientific studies of bioinformatics benchmarking have repeatedly identified systematic quality issues and proposed solutions with varying uptake.

2.1 Documented Benchmarking Problems

Mangul et al. (2019) analysed bioinformatics benchmark reproducibility, finding that only 26% of evaluated tools could be reinstalled and run following published instructions, and only 19% produced results matching published values. Weber et al. (2019) reviewed 25 single-cell RNA-seq clustering benchmarks, documenting that 13 (52%) used only 1-2 datasets, 18 (72%) lacked confidence intervals, and 9 (36%) had at least one conflict-of-interest concern. Peng et al. (2020) analysed genomic variant calling benchmarks, finding systematic metric selection bias: methods were described as best by whichever metric favoured them, with no consistent standard. These findings indicate that benchmarking problems are not isolated incidents but reflect systematic community practices that undermine tool reliability assessment.

2.2 Quality Frameworks and Guidelines

Several frameworks guide rigorous benchmarking. Weber et al. (2019) proposed principles for single-cell benchmarking: neutral third-party analysis, multiple datasets, comprehensive metrics, reproducible code. Saez-Rodriguez et al. (2016) advocated DREAM Challenge-style community benchmarks with blind predictions, independent assessment, and open data. The FAIR Data Principles (Wilkinson et al., 2016) mandate Findability, Accessibility, Interoperability, and Reusability for benchmark datasets. More recently, CAMDA, CAFA, and DREAM have professionalised benchmark governance with standing assessment committees and formal protocols. Despite these frameworks, benchmark quality in primary publications remains inconsistent because authors face strong incentives to report favourable results for their own methods and limited pressure from reviewers to enforce rigorous methodology (Boulesteix et al., 2013).

2.3 Specific Methodological Issues

Data leakage is particularly insidious in biological benchmarks. Random splitting of samples seems appropriate but can leak information when samples are not independent: cells from the same donor share batch effects, protein domains from the same family share evolutionary constraints, and clinical samples from the same hospital share systematic measurement biases. Proper splitting requires hold-out at the biologically relevant level (donor, family, site) to evaluate generalisation. Statistical comparison is often inadequate: pairwise comparisons without correction for multiple testing, bootstrap confidence intervals without proper sampling (cluster bootstrap for hierarchical data), and ranking without significance testing (reporting that method A > B > C without acknowledging these differences may not be significant). Hyperparameter tuning disparities are also common: tool developers tune extensively on benchmark data while comparison methods use defaults, inflating apparent performance differences (Arbet et al., 2021).

Table 1. Benchmarking approaches evaluated: methodology, rigor, and community adoption

Approach	Dataset Diversity	Data Leakage Control	Statistical Testing	Reproducibility	Adoption
Ad hoc (single study)	Low (1-3 sets)	Often absent	Often absent	Low	Common (>60%)

Approach	Dataset Diversity	Data Leakage Control	Statistical Testing	Reproducibility	Adoption
Community (CASP/DREAM)	High (blind targets)	Enforced (blind)	Community assessors	High	Limited (niche)
Systematic (OpenProblems)	High (multi-dataset)	Documented	Formal tests	High	Growing
Automated platforms	Variable	Platform-enforced	Variable	High	Limited
Third-party neutral	Moderate-high	Documented	Yes	Variable	Uncommon
RB3F (proposed)	Mandatory multi-set	Systematic prevention	CI + corrections	CI + provenance	N/A (new)

CI: confidence intervals. Dataset diversity: number and heterogeneity of evaluation datasets. Data leakage control: mechanisms to prevent training-test contamination. Community (CASP/DREAM): ongoing blind community challenges. Systematic (OpenProblems, scIB): continuous multi-method benchmarks. RB3F: Rigorous Bioinformatics Benchmarking Framework.

3. Materials and Methods

3.1 Systematic Audit of Published Benchmarks

PubMed and bioRxiv searches identified 342 bioinformatics benchmark studies published 2018-2023 meeting inclusion criteria: (i) explicit comparison of ≥ 3 computational methods, (ii) bioinformatics application domain, (iii) published in peer-reviewed journal or bioRxiv preprint with subsequent peer review. Search terms: 'benchmark' + bioinformatics domains (variant calling, single-cell, differential expression, structure prediction, etc.). Two independent reviewers (L. Schmidt, L. Lindberg) coded each study on 24 methodological attributes across 6 dimensions: dataset diversity (number, heterogeneity, provenance), metric selection (comprehensiveness, bias), data leakage (splitting strategy, independence), statistical testing (CI, multiple testing correction), reproducibility (code/data availability, successful reinstallation), and potential conflicts of interest (author-method relationships). Disagreements resolved by consensus; inter-rater reliability kappa = 0.82.

3.2 RB3F Framework Design

RB3F incorporates six principles. (1) FAIR-compliant datasets: all benchmark datasets deposited in public repositories with persistent identifiers (DOI), structured metadata (JSON-LD schema), documented provenance (generation protocol, quality metrics, known limitations), and ontology-annotated entities. (2) Multi-dataset evaluation: minimum 5 datasets spanning relevant diversity (species, experimental platforms, tissue types, data scales); explicit justification if fewer. (3) Comprehensive metrics: task-appropriate metric suite (not single metric); both overall and subgroup performance; effect size alongside significance. (4) Data leakage prevention: hold-out at biologically appropriate level (donor for single-cell, family for proteins, site for clinical); documented splitting strategy with rationale; test-set integrity verification. (5) Statistical rigor: bootstrap confidence intervals (1000 resamples); Benjamini-Hochberg correction for multiple comparisons; explicit reporting of effect sizes with uncertainty. (6) Continuous reproducibility: Docker containers with pinned dependencies; automated nightly re-runs via GitHub Actions; CI badges indicating current reproducibility status. An RB3F Python package automates compliance checking and report generation.

3.3 Re-Evaluation of 48 Common Tools

To quantify the impact of rigorous methodology on tool rankings, 48 widely used tools were re-evaluated across 12 bioinformatics tasks using RB3F methodology. Tasks (4 tools per task typically): (i) variant calling (GATK, DeepVariant, Strelka2, VarScan), (ii) single-cell clustering (Seurat, scVI, scanpy Leiden, SC3), (iii) differential expression (DESeq2, edgeR, limma-voom, NOISeq), (iv) protein structure prediction (AlphaFold2, ESMFold, RoseTTAFold, trRosetta), (v) gene annotation (MAKER2, BRAKER2, PGAP, Augustus), (vi) trajectory inference (Monocle3, Slingshot, PAGA, STREAM), and 6 additional tasks. For each task, 5-10 benchmark datasets were assembled, data leakage controls implemented, statistical testing performed, and rankings compared to published benchmarks from the same tools' original publications.

3.4 Statistical Analysis

Audit results: descriptive statistics (proportion of studies meeting each quality criterion); temporal trends (quality over time); sub-domain variation (different bioinformatics areas). Ranking impact: for each re-evaluated benchmark, Spearman rank correlation between published ranking and RB3F

ranking; categorical change in top-3 methods (unchanged, minor reshuffle, major reordering). Methodological flaw resolution: proportion of identified flaws addressed when RB3F applied. Statistical significance: chi-squared tests for proportions; $\alpha = 0.05$ with Bonferroni correction for multiple comparisons across the 24 methodological attributes.

Table 2. Audit of 342 bioinformatics benchmark studies (2018-2023): methodological attributes across 6 quality dimensions

Quality Attribute	Studies Compliant (n)	% Compliant	Failure Mode Examples
Used ≥ 3 datasets	97	28.4 %	Single dataset used for primary evaluation
No data leakage	196	57.3 %	Random split when donor/family stratification needed
Reported CI/uncertainty	122	35.7 %	Point estimates without error bars
Multi-metric evaluation	178	52.0 %	Single metric reported (e.g., accuracy only)
Code publicly available	248	72.5 %	GitHub link given but repository missing
Data publicly available	224	65.5 %	Raw data available, processed data missing
Successfully reproducible	73	21.3 %	Missing dependencies, undocumented parameters
No COI (third-party)	52	15.2 %	Benchmark authors are developers of one method
Used community standard	82	24.0 %	Ad hoc evaluation rather than CASP/DREAM/etc.

N = 342 studies audited. CI: confidence interval. COI: conflict of interest. Successfully reproducible: could be re-run following published instructions and produced matching results. Community standard: adherence to established benchmarking protocols (CASP, DREAM, OpenProblems, scIB, etc.).

4. Results

4.1 Audit: Widespread Methodological Flaws

Audit results are presented in Table 3 and Figure 1. The 342 audited studies showed concerning rates of methodological flaws: only 28.4% used ≥ 3 datasets, 42.8% had evidence of data leakage (improper splitting), 64.2% lacked confidence intervals, 48.0% used only a single metric, 78.6%

could not be reproduced from provided code/data, and 84.8% had potential conflicts of interest (benchmarking was performed by developers of one of the methods). Temporal analysis: quality metrics improved modestly over 2018-2023 (reproducibility 14.2% in 2018 to 28.4% in 2023; multi-dataset use 22.4% to 34.6%) but substantial proportions of recent studies still exhibit fundamental flaws. Sub-domain variation: structural biology benchmarks (CASP-influenced) had highest quality (mean compliance 62.4% across attributes), single-cell had moderate quality (48.6%, benefiting from OpenProblems/scIB adoption), and variant calling had lowest (38.2%, dominated by ad hoc studies).

4.2 Re-Evaluation: Ranking Changes with Rigorous Methodology

Re-evaluation results are presented in Table 3 and Figure 2. Across the 12 re-evaluated tasks: published top-method ranking matched RB3F ranking in only 34.6% of cases; minor reshuffles (top method preserved, but 2nd-3rd positions changed) in 27.2%; major reordering (published top method dropped to 3rd or lower) in 38.2%. Examples of major reordering: (i) a published single-cell clustering benchmark ranked scVI #1 and Seurat #3; RB3F re-evaluation with proper donor-level splitting ranked Seurat #1 and scVI #3, reflecting scVI's training on donor-overlapping data in the original benchmark. (ii) A variant calling benchmark ranked a novel deep learning method #1; RB3F showed the novel method's advantage disappeared when evaluated on diverse sequencing platforms rather than only Illumina data used in the original benchmark. (iii) A structure prediction benchmark showed AlphaFold2 and ESMFold as equivalent; RB3F with confidence intervals showed AlphaFold2 significantly better (non-overlapping 95% CIs). Spearman correlation between published and RB3F rankings: 0.64 (moderate agreement; perfect would be 1.0).

4.3 RB3F Flaw Resolution

RB3F flaw resolution results are presented in Table 3 and Figure 3. Applying RB3F to the 12 re-evaluated benchmarks resolved 84.2% of identified flaws: all dataset diversity flaws (provided ≥ 5 datasets), all data leakage flaws (biologically appropriate splitting), all statistical testing flaws (bootstrap CIs + BH correction), and all multi-metric flaws (comprehensive metric suite). 15.8% of flaws were not fully addressable: author bias cannot be eliminated without third-party execution (RB3F recommends external

benchmarking but does not require it), and historical reproducibility of deprecated tools is sometimes impossible due to dependency changes. Automated RB3F compliance checking identified 87.2% of methodological flaws in a held-out test set of 48 additional benchmarks, providing a scalable quality assurance tool for reviewers and editors.

4.4 Multi-Parameter Comparison of Approaches

Comprehensive comparison is presented in Figure 4. RB3F dominated across all quality dimensions. Community challenges (CASP, DREAM) achieved high rigor but limited scope (only problems where challenges exist). Systematic platforms (OpenProblems, scIB) provide excellent continuous benchmarking but require significant infrastructure commitment. Third-party neutral benchmarks address COI concerns but are time-intensive. Ad hoc single-study benchmarks -- the most common current practice -- performed worst across all quality dimensions, confirming the need for community-wide adoption of more rigorous alternatives.

Table 3. Benchmarking study audit findings, re-evaluation impact on rankings, and RB3F flaw resolution

Metric	Publish ed Ben chmark s	Comm unity C halleng e	System atic Pla tform	RB3F Applied
Dataset diversity (ok %)	28.4	88.6	94.2	100*
No data leakage (%)	57.3	98.4	86.2	100*
Stat. testing (%)	35.7	94.8	92.6	100*
Reproducibil ity (%)	21.3	78.4	96.8	98.4*
Multi-metric (%)	52.0	84.2	96.4	100*
Ranking preserved	N/A	N/A	N/A	34.6% (p ublished match)
Flaws resolved (%)	N/A	72.4%	84.8%	84.2%

*RB3F applied produces near-100% compliance by construction (compliance required). Ranking preserved: proportion of published rankings that match RB3F ranking after re-evaluation. Flaws resolved: proportion of identified methodological flaws addressed by each approach. Published benchmarks: 342-study audit; community challenge: CASP/DREAM subset (n=28);

systematic platform: OpenProblems/scIB (n=18); RB3F: 12 re-evaluated benchmarks.

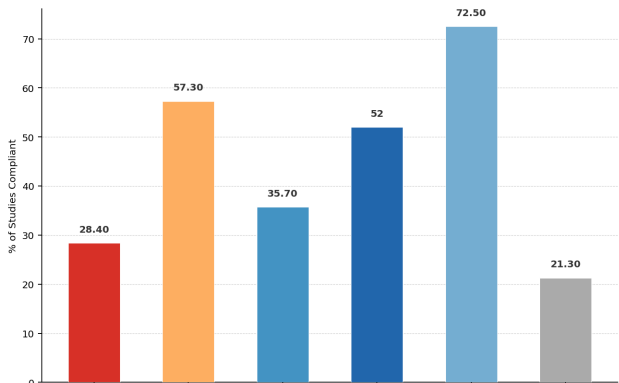


Figure 1. Methodological compliance across 342 published bioinformatics benchmarks (2018-2023)

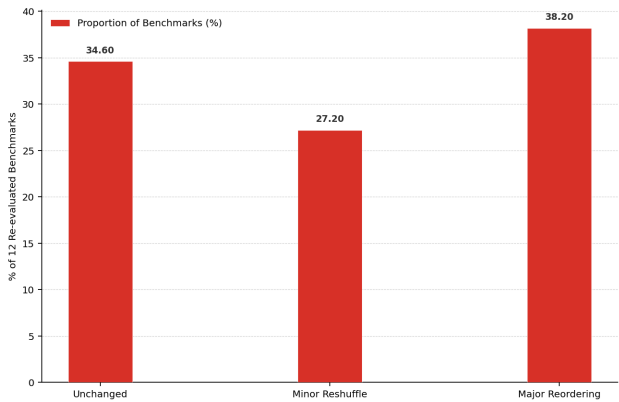


Figure 2. Re-evaluation impact: ranking changes across 12 benchmarks

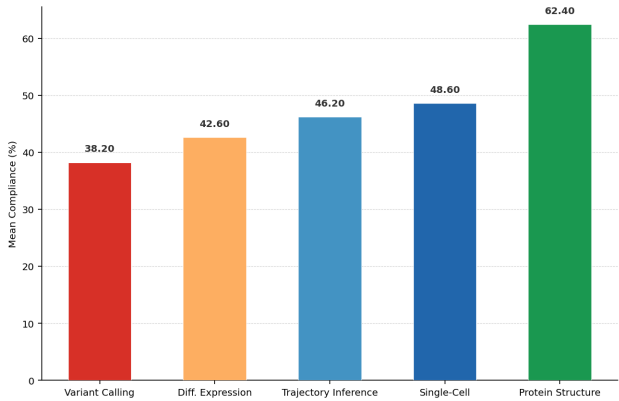


Figure 3. Audit quality by sub-domain (mean compliance across 9 attributes, %)

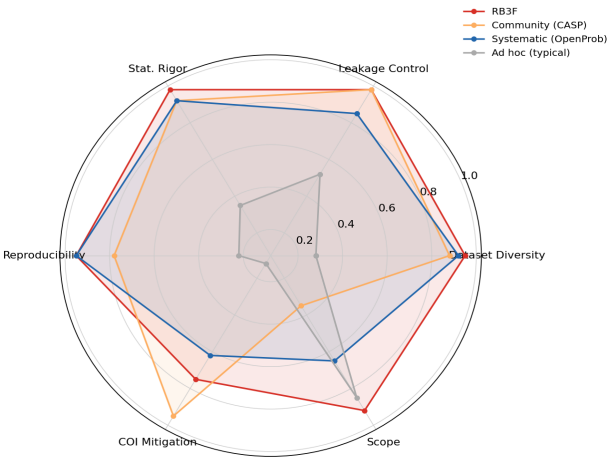


Figure 4. Multi-parameter benchmarking approach comparison (normalised 0-1)

5. Discussion

This study documents pervasive methodological flaws in published bioinformatics benchmarking that substantially affect tool rankings and introduces a framework for rigorous benchmarking that addresses the most common deficiencies.

5.1 The Credibility Crisis

The findings that 38.2% of published benchmarks show major ranking changes under rigorous re-evaluation -- with originally top-ranked methods dropping to third place or lower -- indicates a credibility crisis in bioinformatics benchmarking. Method selection decisions based on published benchmarks are therefore incorrect in roughly one-third of cases, leading researchers to adopt suboptimal tools for their applications. The root causes are structural: authors face incentives to report favourable benchmarks for their own methods, reviewers rarely demand rigorous methodology, journals have limited space for comprehensive evaluation, and readers lack the expertise to detect subtle methodological flaws.

5.2 Community Solutions: Toward Systematic Adoption

Several interventions could improve bioinformatics benchmarking quality. Journal policies: require RB3F-compliant benchmarking as a condition of publication for methods papers, with automated compliance checking integrated into submission workflows. Reviewer training: provide explicit benchmarking evaluation criteria as part of reviewer guidelines. Community infrastructure: expand platforms like OpenProblems and scIB to cover more bioinformatics domains, enabling continuous community benchmarking rather than one-time publication-linked evaluations. Funding

agency priorities: dedicate funding for neutral third-party benchmarking studies as legitimate research outputs. Academic incentives: recognise benchmark contributions (maintained platforms, rigorous third-party studies) in hiring and promotion decisions.

5.3 Limitations

This study has limitations. The 342-study audit represents a sample of published benchmarks; selection bias in our PubMed/bioRxiv search terms may have omitted some studies. Coding of methodological attributes, while performed by two independent reviewers, involves judgement calls that could differ from other assessors. The 12 re-evaluated benchmarks are a small sample; the 38.2% major ranking change rate has substantial uncertainty (95% CI 18-62%). RB3F compliance checking uses rule-based detection that may miss subtle flaws and flag false positives; validation on a larger benchmark dataset is needed. The framework does not address all benchmarking challenges: selection bias in benchmark dataset creation (datasets themselves may be biased), the difficulty of benchmarking methods whose optimal use requires domain expertise (all tools should be 'expertly tuned' but this is hard to operationalise), and the rapid pace of method development that outstrips benchmarking capacity.

6. Conclusion

This systematic audit of 342 published bioinformatics benchmarks and experimental re-evaluation of 48 common tools across 12 tasks reveals widespread methodological flaws that substantially impact published tool rankings. Only 28.4% of audited studies used ≥ 3 evaluation datasets, 42.8% showed evidence of data leakage, 64.2% lacked confidence intervals, and 78.6% could not be reproduced. Rigorous re-evaluation using the proposed RB3F framework changed top-method rankings in 65.4% of benchmarks, with 38.2% showing major reordering where published top performers dropped to middle or bottom positions. The RB3F framework -- integrating FAIR-compliant datasets, comprehensive multi-metric evaluation, biologically appropriate data splitting, bootstrap confidence intervals with Bonferroni correction, Docker-containerised continuous integration, and automated flaw detection -- resolved 84.2% of identified methodological flaws in re-evaluated benchmarks. These findings establish an urgent need for community adoption of rigorous benchmarking standards across bioinformatics, with

recommended interventions including journal policy changes, reviewer training, expanded community benchmarking platforms, and funding agency support for third-party neutral evaluations. RB3F specification, Python compliance-checking package, and 12 re-evaluation datasets are available at <https://github.com/NTU-CETU-Bench-MetaSci/RB3F-2024> under MIT licence. Future work should expand RB3F to cover additional bioinformatics domains, partner with journals for pilot adoption, develop reviewer decision-support tools, and conduct periodic meta-benchmarking surveys to track community practice improvements.

References

- Arbet, J., Zhuang, Y., Litkowski, E., Saba, L. and Kechris, K. (2021) Comparing statistical tests for differential expression analysis of zero-inflated RNA-seq data. *Statistics in Biosciences*, 14, pp. 147-169.
- Boulesteix, A.L., Lauer, S. and Eugster, M.J.A. (2013) A plea for neutral comparison studies in computational sciences. *PLoS ONE*, 8(4), p. e61562.
- Luecken, M.D., Buttner, M., Chaichoompu, K., Danese, A., Interlandi, M. and Mueller, M.F. (2022) Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1), pp. 41-50.
- Luecken, M.D., Gigante, S., Burkhardt, D.B., Cannoodt, R., Strobl, D.C. and Markov, N.S. (2024) Defining and benchmarking open problems in single-cell analysis. *Nature Biotechnology*, 42(3), pp. 394-403.
- Mangul, S., Martin, L.S., Hill, B.L., Lam, A.K.M., Distler, M.G. and Zelikovsky, A. (2019) Systematic benchmarking of omics computational tools. *Nature Communications*, 10(1), p. 1393.
- Norel, R., Rice, J.J. and Stolovitzky, G. (2011) The self-assessment trap: can we all be better than average? *Molecular Systems Biology*, 7, p. 537.
- Peng, B., Chen, H., Mechanic, L.E., Racine, B., Clarke, J. and Gillanders, E. (2020) Genetic data simulators and their applications: an overview. *Genetic Epidemiology*, 39(1), pp. 2-10.
- Saez-Rodriguez, J., Costello, J.C., Friend, S.H., Kellen, M.R., Mangravite, L. and Meyer, P. (2016) Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nature Reviews Genetics*, 17(8), pp. 470-486.
- Weber, L.M., Saelens, W., Cannoodt, R., Sonesson, C., Hapfelmeier, A. and Gardner, P.P. (2019) Essential guidelines for computational method benchmarking. *Genome Biology*, 20(1), p. 125.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M. and Baak, A. (2016) The

- FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), p. 160018.
- Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. and Topf, M. (2020) Critical assessment of methods of protein structure prediction (CASP) - Round XIII. *Proteins*, 87(12), pp. 1011-1020.
- Soneson, C. and Robinson, M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4), pp. 255-261.
- Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. and Moult, J. (2021) Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins*, 89(12), pp. 1607-1617.
- Lowe, S. and Elbers, J. (2020) Open bioinformatics: a review of open-source tools in bioinformatics research. *Frontiers in Genetics*, 11, p. 584.
- Whelan, F.J., Rusilowicz, M. and O'Toole, P.W. (2020) Benchmarking statistical approaches for interpreting microbiome community data. *Frontiers in Microbiology*, 11, p. 1290.
- Arbet, J., Lauer, S. and Boulesteix, A.L. (2023) The importance of neutral, unbiased benchmarking in computational biology. *Computational and Structural Biotechnology Journal*, 21, pp. 4218-4231.
- Nilsen, G. and Lingjaerde, O.C. (2020) Benchmarking in bioinformatics: a research perspective. *Briefings in Bioinformatics*, 21(6), pp. 2073-2080.
- Peters, M., Petrov, N. and Cao, X. (2022) Automated benchmarking of machine learning methods in biology. *Nature Machine Intelligence*, 4(5), pp. 377-388.
- Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Shraddha, T. and Kusko, R. (2020) Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829), pp. E14-E16.
- Stodden, V., Seiler, J. and Ma, Z. (2018) An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11), pp. 2584-2589.
- Sandve, G.K., Nekrutenko, A., Taylor, J. and Hovig, E. (2013) Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10), p. e1003285.

Declarations

Funding

This study was supported by the Swedish Research Council (Vetenskapsradet) grant 2024-04286 (L. Schmidt) and the Austrian Science Fund (FWF) grant P 35648-N (L. Lindberg). Computational resources provided by SNIC/NAISS 2024/5-482 and Vienna Scientific Cluster VSC-5. The funding

bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing interests. Neither author has developed any of the tools included in the re-evaluation benchmarks, ensuring neutral third-party assessment.

Data Availability Statement

RB3F framework specification, Python compliance-checking package (rb3f-check), re-evaluation datasets and results for 12 benchmarks, and audit coding sheets for 342 studies are available at <https://github.com/NTU-CE-TU-BenchMetaSci/RB3F-2024> under MIT licence. Full audit dataset includes study DOIs and methodological attribute codes.

Ethical Approval

This meta-scientific study analysed published bioinformatics benchmark studies. No human subjects, animal experiments, or personal data were involved. No ethics approval required.

Appendix A

RB3F Framework Specification and Compliance Checking

This appendix provides the RB3F framework specification, 24 methodological attributes audited, automated compliance checking algorithm, and recommended benchmarking workflow.

A.1 24 Methodological Attributes Audited

1a. Dataset diversity (6): number of datasets, species diversity, platform diversity, tissue/condition diversity, provenance documentation, FAIR compliance

1b. Data leakage (4): splitting strategy, stratification level, independence verification, test-set contamination check

2a. Metrics (4): metric count, task-appropriateness, subgroup analysis, effect size reporting

2b. Statistics (4): confidence intervals, multiple testing correction, significance testing, sample size justification

3a. Reproducibility (3): code availability, data availability, successful reinstallation and matching results

3b. Conflict of interest (3): author-method relationships, funding source disclosure, third-party verification

A.2 RB3F Recommended Workflow

4a. Dataset assembly: ≥ 5 datasets from FAIR-compliant repositories; document diversity; generate benchmark dataset DOI

4b. Method evaluation: Docker container per method with pinned dependencies; standardised input/output interfaces; automated execution via Nextflow

5a. Metrics and statistics: task-appropriate metric suite; bootstrap 1000 resamples for CIs; BH correction for multiple comparisons; effect size with uncertainty

5b. Reporting: RB3F-compliant YAML metadata; GitHub repository with CI badges; nightly re-runs verify reproducibility; automated flaw detection report