Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

How Much Does a Benchmark Weight Move a Leaderboard? An Empirical Analysis of Single-Cell Integration Rankings

Liu Chen

weidt1606@gmail.com

University of Calicut

Research Article

Keywords:

Posted Date: July 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10462090/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

How Much Does a Benchmark Weight Move a Leaderboard? An Empirical Analysis of Single-Cell Integration Rankings

Liu Chen
Independent Researcher
weidt1606@gmail.com

June 2026

Abstract

Biomedical benchmark leaderboards often combine several desirable properties into one score. The weights in that score express a scientific preference, but their effect on the reported ranking is rarely measured. We reanalyzed the public scIB single-cell data-integration benchmark to ask a narrow question: does the winning method change when the weight assigned to batch correction, relative to biological conservation, changes? The analysis covered 327 task-pipeline evaluations from five real atlas tasks, 17 method entries including an unintegrated baseline, and 14 metrics. We reproduced the benchmark's within-task scaling and swept the batch-correction weight over 1,001 values from 0 to 1. At the published weight of 0.4, scANVI ranked first with a mean task rank of 3.2; its task ranks ranged from 1 to 6. The winner was unchanged in a centered sensitivity band from 0.3 to 0.5, although as many as 6.0% of strictly ordered method pairs reversed. Across the full sweep, scANVI won through weight 0.641, Scanorama won from 0.642 to 0.644, and Harmony won from 0.645 onward. At the most discordant weight, Kendall's τ_b with the reference ranking was 0.494 and 25.2% of comparable method pairs were reversed. Resampling the five atlas tasks showed a separate source of uncertainty: at the reference weight, scANVI was the bootstrap winner in only 56.0% of 10,000 resamples. The result is therefore qualified. The published weight is locally stable in this reanalysis, but neither the winner nor the ordering is invariant to defensible changes in scientific priorities or task composition. Benchmark reports should include weight-sensitivity curves and task-resampling uncertainty rather than a single rank table.

1 Introduction

Computational biology uses benchmarks to choose among methods that cannot be compared by inspection. In single-cell analysis, an integration method may be judged on its removal of technical batch effects and its preservation of biological structure. These goals can conflict. A leaderboard resolves the conflict by combining several metrics, often with fixed weights, and then sorting the resulting score. This is convenient, but it turns a scientific value judgment into an apparently objective rank.

Good benchmarking guidance already treats dataset selection, preprocessing, metric choice, and aggregation as parts of the scientific design [39]. Yet the usual published artifact is still a single

ordering. The broader machine-learning literature calls this dependence on evaluation choices the "benchmark lottery" [8] and questions whether broad claims can be supported by narrow, composite benchmarks [30]. The metascientific concern is not that all leaderboards are wrong. It is that their uncertainty is hidden.

This paper measures one component of that uncertainty in a concrete case. We reanalyze the metric matrix released with scIB, a benchmark of atlas-level single-cell integration methods [20]. The empirical question is deliberately narrow:

> *How sensitive are the winning method and the full method ordering to the weight assigned to batch correction rather than biological conservation?*

The scIB score assigns weight 0.4 to batch correction and 0.6 to biological conservation. We treat 0.4 as the reference, examine a centered band from 0.3 to 0.5, and then map the full range from pure biological conservation to pure batch correction. The analysis does not try to identify a universally best integration method. It estimates how much one published leaderboard moves when its stated preference moves.

The paper makes three contributions. First, it gives a reproducible weight-sensitivity analysis using public benchmark outputs rather than simulated scores. Second, it separates weight sensitivity from uncertainty due to the small set of atlas tasks through bootstrap and leave-one-task-out analyses. Third, it tests two consequential design alternatives: retaining method-configuration pairs as separate entries and excluding the two label-assisted methods. These checks expose both a stable local conclusion and an unstable global one.

2 Previous Research

2.1 Reproducibility and benchmark-driven evidence

Concern about benchmark reliability belongs to a larger literature on reproducibility. Flexible analysis and selective reporting can make published findings more optimistic than their evidence warrants [15]. Survey and replication evidence has documented the practical scale of the problem across science [2, 3, 27]. Machine learning has parallel reporting deficits: many studies omit information needed for independent reproduction [13], while community interventions show that checklists, code, and structured reporting can improve practice [28]. In biomedicine, demands for transparent code and data are especially direct because a high score cannot be audited without them [14].

Benchmark papers introduce an additional structural bias. Developers choose datasets, tune their own method, and decide what to report. Neutral comparison studies were proposed to separate evaluation from method advocacy [5]. The "self-assessment trap" predicts that most developers will appear better than average under self-chosen evaluations [26]; empirical work has demonstrated over-optimistic performance estimates [16] and an advantage for newly introduced bioinformatics methods [6]. A systematic survey of omics tools likewise found that reproducibility and fair comparison are often limited by missing artifacts and inconsistent evaluation [22].

Some errors occur before aggregation. Test information can leak into training or model selection [18]; genomics has specific leakage risks when train and test sequences are not independent [40]. A model may also exploit a shortcut that raises the metric without learning the intended construct [11]. Repeated reuse of a public test set creates community-level adaptation, as illustrated by performance

changes on a new ImageNet test set [31]. After scores have been produced, inappropriate metrics [4] and invalid cross-dataset statistical comparisons [9] can still distort the conclusion. Our analysis begins later in this chain. It assumes the released scIB metric values and asks how the aggregation preference changes their interpretation.

2.2 Challenges and neutral comparisons in computational biology

Prospective challenges reduce some sources of bias by withholding reference answers. CASP established this model for protein structure prediction [25]; the same framework later made AlphaFold's improvement measurable against blinded targets [17]. DREAM applied challenge-based evaluation to pathway and network inference [36], supported by synthetic systems with known networks [34]. Those studies found that methods have condition-specific strengths [23] and that aggregated predictions can outperform an individual winner [24]. Challenge organizers have consequently argued for crowdsourced evaluation as a general biomedical research instrument [7, 33].

Neutral comparisons repeatedly find that rankings depend on the evaluation regime. Benchmark construction affects multiple-sequence-alignment accuracy [1]. In single-cell analysis, conclusions vary with data properties for differential expression [35], clustering [10], trajectory inference [32], and complete RNA-sequencing pipelines [37]. BEELINE showed that single-cell gene-regulatory network inference remains difficult on realistic references [29]. More directly, biomedical imaging challenge ranks can change under alternative test samples and aggregation rules [21], and a recent gene-regulatory-network reanalysis quantified pairwise rank reversals across several protocol axes [19].

The scIB benchmark is a useful empirical case because it makes the tradeoff explicit: the overall score is 40% batch correction and 60% biological conservation [20]. Its public metric matrix and analysis code permit an audit without rerunning computationally expensive integration pipelines. Reporting frameworks such as DOME [38] and practical guidance for biological machine learning [12] emphasize validation design and transparent reporting. Weight-sensitivity analysis is a direct extension of those principles to composite leaderboards.

3 Methods

3.1 Study design and data

This was a secondary analysis of published benchmark outputs. We used the commit-pinned `metrics.csv` file in the scIB reproducibility repository (commit `3afbffd`) and verified its SHA-256 digest before analysis. The input contains scores for 14 metrics. We restricted the primary analysis to the five real RNA atlas tasks and excluded the two simulated tasks. Table 1 describes the retained tasks. The resulting dataset contained 327 task-pipeline rows, 69 distinct method/preprocessing/output configurations, and 17 method entries when the unintegrated baseline was counted.

The five batch-correction metrics were principal-component-regression batch effect, batch silhouette width, integration local inverse Simpson's index, graph connectivity, and kBET. The nine biological-conservation metrics were normalized mutual information, adjusted Rand index, cell-type silhouette width, isolated-label F1, isolated-label silhouette, cell-cycle conservation, highly-variable-gene conservation, trajectory conservation, and cell-type local inverse Simpson's index. Some metrics are undefined for some outputs or tasks. As in the released scIB aggregation code, group means

Table 1: Real atlas tasks included in the reanalysis. Cell, batch, and label counts are from the scIB task metadata [20]. Rows are evaluated pipeline configurations in the released metric matrix.

Task	Cells	Batches	Cell labels	Rows
Human pancreas	16,382	9	14	69
Human lung atlas	32,472	16	17	69
Human immune cells	33,506	10	16	69
Human/mouse immune cells	97,952	23	19	69
Mouse brain	978,734	5	10	51
Total	1,158,546			327

used the available metrics for each row.

3.2 Score reconstruction

Let x_{dim} be the value of metric m for configuration i on task d. For each task and metric, we reproduced the published min-max transformation,

$$z_{dim} = \frac{x_{dim} - \min_j x_{djm}}{\max_j x_{djm} - \min_j x_{djm}}. \tag{1}$$

All metrics in the released matrix are oriented so that larger values are better. We then calculated the batch-correction component B_{di} as the mean of the available scaled batch metrics and the biological-conservation component C_{di} as the mean of the available scaled biology metrics. For batch weight w, the composite score was

$$S_{di}(w) = wB_{di} + (1 - w)C_{di}. \tag{2}$$

The published preference is $w = 0.4$.

A method can have several output, feature-selection, and scaling configurations. In the primary method-level analysis, the method score within each task was the largest observed $S_{di}(w)$ among its configurations. This answers the practical question "how well can this method perform under an evaluated setup?" but favors methods with more evaluated configurations; we therefore retained every configuration as a separate entry in a robustness analysis. If a method was unavailable for a task, we assigned that task's unintegrated score before ranking, matching the missing-value rule in the released cross-task aggregation code.

Methods were ranked separately within each task, with average ranks for ties. The cross-task statistic was the unweighted mean of the five task ranks. The winner was the method with the smallest mean rank. Thus, no large task dominated merely because it contained more cells.

3.3 Sensitivity outcomes

We evaluated 1,001 values of w from 0 to 1 in steps of 0.001. The primary outcome was winner identity. Secondary outcomes compared the complete ordering at each weight with the reference ordering at $w = 0.4$:

1. Kendall's τ_b, which accounts for tied mean ranks; and
2. the pairwise reversal rate, defined as the fraction of method pairs with opposite strict orderings at the candidate and reference weights. Pairs tied in either ordering were omitted from the reversal denominator.

Table 2: Top six methods at the reference batch weight $w = 0.4$. Each task rank uses the best evaluated configuration of that method; lower values are better. Ties receive average ranks.

Method	Human immune	Human/mouse immune	Lung	Mouse brain	Pancreas	Mean
scANVI*	3	4	2	1	6	3.2
Scanorama	1	2	4	6	9	4.4
fastMNN	2	6	5	8	4	5.0
scVI	6	3	3	5	8	5.0
Harmony	4	10	8	3	1	5.2
scGen*	5	1	1	14.5	5	5.3

We summarized a centered sensitivity band, $0.3 \leq w \leq 0.5$, as well as the complete range. The centered band is a diagnostic interval, not a claim that all users share those preferences. Transition points are reported to the resolution of the grid.

3.4 Task uncertainty and robustness analyses

To measure sensitivity to task composition, we resampled the five atlas tasks with replacement 10,000 times at five selected weights. The random seed was 20260723. When several methods tied for best mean rank in a resample, each received an equal fraction of that replicate. These are descriptive bootstrap frequencies, not probabilities that a method is scientifically best. We also omitted each task once and recalculated the reference ranking.

Two design checks were performed. First, all 69 configurations were ranked as separate leaderboard entries, avoiding within-method maximization. Second, we excluded scANVI and scGen because they use cell labels and were marked with an asterisk in the original benchmark. This check separates the score-weight question from the decision to compare label-assisted and unsupervised methods in one table.

No null-hypothesis test was used. The weight sweep covers the complete one-dimensional preference space for this score, and the number of atlas tasks is too small for a precise population model. Code, parameters, derived tables, and figure-generation scripts are included with the manuscript.

4 Results

4.1 The reference leaderboard was heterogeneous across tasks

At the published weight $w = 0.4$, scANVI had the smallest mean rank, 3.2, followed by Scanorama at 4.4 (Table 2). The aggregate hides substantial task variation. scANVI ranked first on mouse brain but sixth on pancreas. Scanorama ranked first on human immune cells but ninth on pancreas. scGen ranked first on two tasks yet tied for rank 14.5 on mouse brain. No method in the top six ranked in the top three on every task.

This heterogeneity matters for interpretation. The reference result is not a fivefold confirmation of one winner. It is a compromise among distinct task-level orderings.

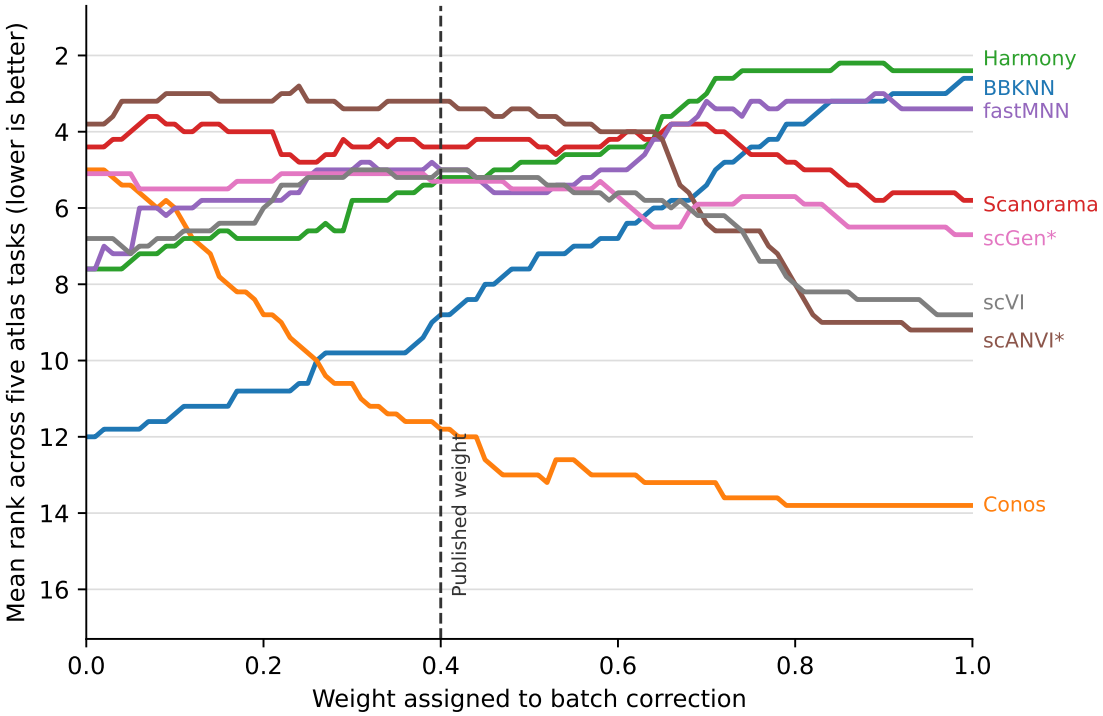


Figure 1: Mean task rank as the batch-correction weight changes. The figure shows methods that entered the top six at least once. The dashed line marks the published weight, $w = 0.4$. Direct labels identify trajectories at $w = 1$. Lower rank is better.

4.2 The winner was locally stable but globally weight-dependent

The rank trajectories are shown in Figure 1. scANVI remained the method-level winner from $w = 0$ through $w = 0.641$. Scanorama won for the narrow interval from 0.642 to 0.644, and Harmony won from 0.645 through 1. The brief Scanorama interval reflects changes in discrete task ranks: Scanorama is close to both scANVI and Harmony near the crossover, but it does not dominate either component.

The local sensitivity result was reassuring but not absolute. From $w = 0.3$ to $w = 0.5$, scANVI remained first. Within that band, Kendall's τ_b with the reference ordering never fell below 0.874, but the maximum pairwise reversal rate was 6.0%. The winner was stable while several lower-ranked pairs changed order.

The full sweep gave a different picture (Table 3). At pure biological conservation, scANVI led and 20.3% of strictly comparable pairs were reversed relative to the reference. At pure batch correction, Harmony led, BBKNN was second, and 23.9% of pairs were reversed. The largest discordance occurred at $w = 0.922$: $\tau_b = 0.494$ and 25.2% of comparable pairs were reversed.

Figure 2 shows why winner identity alone is incomplete. The full ordering begins to depart from the reference well before the winner changes. Conversely, the winner transition near 0.645 occurs after a region in which the rank correlation remains above 0.9.

Table 3: Selected points from the weight sweep. Kendall's τ_b and pairwise reversals compare each ordering with the reference at $w = 0.4$.

w_{batch}	w_{bio}	Winner rank	Winner	Runner-up rank	τ_b	Reversed pairs
0.0	1.0	3.8	scANVI*	Scanorama (4.4)	0.587	20.3%
0.2	0.8	3.2	scANVI*	Scanorama (4.0)	0.790	10.4%
0.4	0.6	3.2	scANVI*	Scanorama (4.4)	1.000	0.0%
0.5	0.5	3.4	scANVI*	Scanorama (4.2)	0.937	3.0%
0.6	0.4	4.0	scANVI*	Scanorama (4.2)	0.923	3.7%
0.8	0.2	2.4	Harmony	fastMNN (3.2)	0.578	20.9%
1.0	0.0	2.4	Harmony	BBKNN (2.6)	0.519	23.9%

4.3 The five-task sample added material uncertainty

At $w = 0.4$, scANVI was the bootstrap winner in 56.0% of task resamples. scGen won 19.9%, Harmony 11.7%, Scanorama 9.7%, and fastMNN 2.6%. Winner frequency shifted toward Harmony and BBKNN as the batch weight increased (Figure 3). At $w = 1$, Harmony and BBKNN had bootstrap winner frequencies of 48.6% and 40.5%, respectively, even though Harmony led in the observed five-task mean. Thus, a deterministic winner in the observed tasks did not imply a concentrated task-resampling distribution.

The leave-one-task-out analysis supported the same conclusion. scANVI remained first after omitting pancreas, lung, human immune, or human/mouse immune. When mouse brain was omitted, however, scGen became first with a mean rank of 3.0 and scANVI fell to second at 3.75. This follows directly from the task ranks in Table 2: mouse brain was scANVI's strongest task and scGen's weakest.

4.4 Design checks changed names but not the main pattern

When every method-configuration pair was retained as a separate leaderboard entry, four configurations won somewhere on the weight grid. Unscaled, highly-variable-gene scANVI led through $w = 0.556$; scaled, highly-variable-gene Scanorama led from 0.557 to 0.616; scaled, highly-variable-gene fastMNN led from 0.617 to 0.656 and again from 0.658 to 0.668; and scaled, highly-variable-gene Harmony led at 0.657 and from 0.669 onward. The one-point Harmony interval at 0.657 is a consequence of discrete average ranks and should not be interpreted as a precise scientific threshold. The configuration analysis nonetheless confirms that the identity of the top entry depends on the score preference.

Excluding the two label-assisted methods produced a simpler method-level result. Scanorama led from $w = 0$ through 0.647, and Harmony led from 0.648 through 1. At the reference weight, Scanorama's mean rank was 3.2 and scVI's was 3.4. This check changes the name of the reference winner, as expected, but places the main crossover at nearly the same batch weight as the full analysis. The choice to include label-assisted methods is therefore consequential for the headline name, while the tradeoff between biological conservation and batch removal remains visible either way.

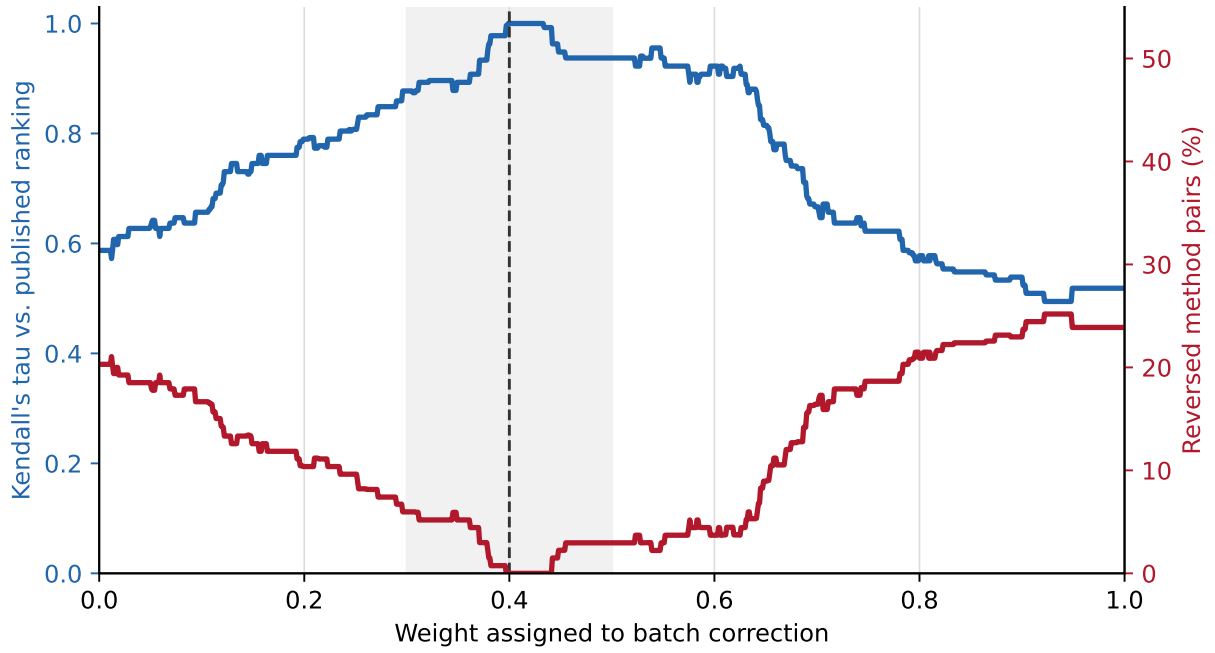


Figure 2: Stability of the full ordering relative to $w = 0.4$. Blue shows Kendall's τ_b (left axis); red shows the percentage of strictly ordered method pairs that reverse (right axis). The gray band is the centered 0.3–0.5 sensitivity interval.

5 Discussion

5.1 What the reanalysis establishes

The result is neither "the leaderboard is robust" nor "the leaderboard is arbitrary." It is more specific. Under the method-level aggregation used here, the scIB winner is stable to a moderate change around the published 40/60 weighting. A reader who prefers 30/70 or 50/50 would see the same winner. That local agreement is evidence in favor of the headline result.

The same result does not extend across the full preference space. A batch-focused user obtains Harmony, while a biology-focused user obtains scANVI when label-assisted methods are included and Scanorama when they are not. As many as one quarter of strictly ordered method pairs reverse relative to the reference. These are not numerical accidents: the methods occupy different positions on the batch-removal/biology-preservation tradeoff. Compressing that tradeoff to one score conceals the preference that selects the winner.

Task composition was at least as important as moderate weight variation. The reference winner appeared in 56.0% of task bootstraps, and one leave-one-out analysis changed the winner. This is not a conventional confidence statement; five named atlas tasks are not a random sample from a well-defined population. It is a stress test showing that the observed mean rank depends on which biological settings are represented.

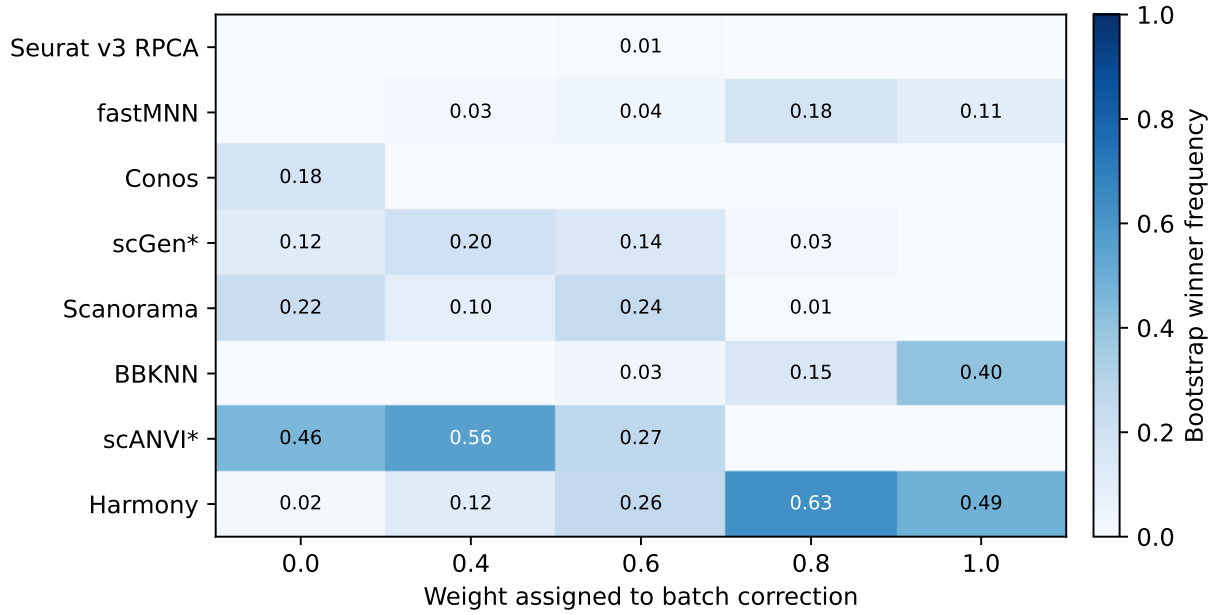


Figure 3: Winner frequencies from 10,000 bootstrap resamples of the five atlas tasks. Values below 0.01 are not printed. Tied winners divide one replicate equally. The frequencies describe sensitivity to the observed task set; they are not probabilities of universal superiority.

5.2 Implications for benchmark reporting

A useful leaderboard should expose, rather than erase, this structure. For a score with two components, the additional analysis is inexpensive:

1. report the two component scores as well as the composite;
2. sweep the weight over its full range and mark the published choice;
3. show winner intervals and full-ranking agreement, not only the top name;
4. resample or omit tasks to show dependence on benchmark composition; and
5. distinguish methods that use different information, such as cell labels, before combining them in one ordering.

These steps complement, rather than replace, neutral design, leakage audits, and reproducible artifacts. A beautifully stable aggregation cannot rescue a leaky split or an invalid construct. Conversely, a valid multi-objective benchmark can have several reasonable winners because users value its objectives differently. The benchmark should make that fact legible.

5.3 Limitations

This study has six main limitations. First, it analyzes one benchmark in one domain. The numerical transition near $w = 0.645$ is a property of these scIB scores, not a general constant. Second, only five real atlas tasks were available, so task-resampling results are coarse and strongly affected by individual atlases. Third, the primary analysis selects the best evaluated configuration of each method within each task. This reflects common use but can favor methods with more configurations; the configuration-level check partly addresses the issue.

Fourth, min-max scaling depends on the set of competitors. Adding a method with an extreme

score could change normalized components even if all existing raw metrics stayed fixed. Fifth, assigning the unintegrated baseline to unavailable method-task combinations follows the released analysis but is not the only defensible missing-data rule. Sixth, we reanalyzed released metric outputs rather than rerunning integration methods from raw data. We therefore do not test software reproducibility, leakage, metric implementation, or the biological validity of the reference labels.

The analysis is exploratory and was not preregistered. Grid transition points are descriptive. Most importantly, the study does not establish that scANVI, Scanorama, or Harmony is generally superior. It establishes where each becomes the top-ranked entry under a stated scoring and task-aggregation procedure.

6 Conclusion

A single leaderboard rank is the end of a chain of choices. In the scIB case, the published 40/60 weighting is locally stable: the winner survives a centered 0.3–0.5 sensitivity analysis. But the broader ordering still moves, the winner changes when batch correction receives roughly two-thirds of the weight, and the five-task sample leaves substantial winner uncertainty. Reporting only the reference table would reveal none of this.

The practical lesson is simple. Composite benchmark scores should be accompanied by weight-sensitivity and task-sensitivity analyses. These do not weaken a leaderboard. They show which conclusions the evidence supports, and which conclusions depend on a preference that should remain visible.

Data and Code Availability

The benchmark metric matrix analyzed in this study is publicly available from the scIB reproducibility repository [20]. We used the file https://github.com/theislab/scib-reproducibility/blob/3afbffd3674726e5146797be21cf6bd7470a2c5f/data/metrics.csv. No new primary biological data were generated. The complete analysis code, environment specification, and derived tables and figures will be deposited in a permanent public repository and made openly available no later than acceptance of this paper. The repository URL and archival identifier will be included in the accepted version. Until then, the code is available from the author upon reasonable request.

References

[1] Mohamed Radhouene Aniba, Olivier Poch, and Julie D. Thompson. Issues in bioinformatics benchmarking: The case study of multiple sequence alignment. *Nucleic Acids Research*, 38(21): 7353–7363, 2010. doi: 10.1093/nar/gkq625.

[2] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016. doi: 10.1038/533452a.

[3] C. Glenn Begley and Lee M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012. doi: 10.1038/483531a.

[4] Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. A critical analysis of

metrics used for measuring progress in artificial intelligence. *arXiv preprint arXiv:2008.02577*, 2020.

[5] Anne-Laure Boulesteix, Sabine Lauer, and Manuel J. A. Eugster. A plea for neutral comparison studies in computational sciences. *PLoS ONE*, 8(4):e61562, 2013. doi: 10.1371/journal.pone.0061562.

[6] Stefan Buchka, Alexander Hapfelmeier, Paul P. Gardner, Rory Wilson, and Anne-Laure Boulesteix. On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology*, 22(1):152, 2021. doi: 10.1186/s13059-021-02365-4.

[7] James C. Costello and Gustavo Stolovitzky. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clinical Pharmacology & Therapeutics*, 93(5): 396–398, 2013. doi: 10.1038/clpt.2013.36.

[8] Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

[9] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[10] Angelo Duò, Mark D. Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, 2018. doi: 10.12688/f1000research.15666.2.

[11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

[12] Joe G. Greener, Shaun M. Kandathil, Lewis Moffat, and David T. Jones. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55, 2022. doi: 10.1038/s41580-021-00407-0.

[13] Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11503.

[14] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush, and Hugo J. W. L. Aerts. Transparency and reproducibility in artificial intelligence. *Nature*, 586 (7829):E14–E16, 2020. doi: 10.1038/s41586-020-2766-y.

[15] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8): e124, 2005. doi: 10.1371/journal.pmed.0020124.

[16] Monika Jelizarow, Vincent Guillemot, Arthur Tenenhaus, Korbinian Strimmer, and Anne-Laure Boulesteix. Over-optimism in bioinformatics: An illustration. *Bioinformatics*, 26(16):1990–1998, 2010. doi: 10.1093/bioinformatics/btq323.

[17] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

[18] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

[19] Ihor Kendiukhov. Quantifying ranking instability across evaluation protocol axes in gene regulatory network benchmarking. *arXiv preprint arXiv:2603.03493*, 2026.

[20] Malte D. Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F. Mueller, Daniel C. Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022. doi: 10.1038/s41592-021-01336-8.

[21] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):5217, 2018. doi: 10.1038/s41467-018-07619-7.

[22] Serghei Mangul, Lana S. Martin, Brian L. Hill, Angela Ka-Mei Lam, Margaret G. Distler, Alex Zelikovsky, Eleazar Eskin, and Jonathan Flint. Systematic benchmarking of omics computational tools. *Nature Communications*, 10(1):1393, 2019. doi: 10.1038/s41467-019-09406-4.

[23] Daniel Marbach, Robert J. Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14):6286–6291, 2010. doi: 10.1073/pnas.0913357107.

[24] Daniel Marbach, James C. Costello, Robert Küffner, Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, The DREAM5 Consortium, Manolis Kellis, James J. Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012. doi: 10.1038/nmeth.2016.

[25] John Moult, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–v, 1995. doi: 10.1002/prot.340230303.

[26] Raquel Norel, John Jeremy Rice, and Gustavo Stolovitzky. The self-assessment trap: Can we all be better than average? *Molecular Systems Biology*, 7(1):537, 2011. doi: 10.1038/msb.2011.70.

[27] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716.

[28] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

[29] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, 2020. doi: 10.1038/s41592-019-0690-6.

[30] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, 2021.

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019.

[32] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, 2019. doi: 10.1038/s41587-019-0071-9.

[33] Julio Saez-Rodriguez, James C. Costello, Stephen H. Friend, Michael R. Kellen, Lara Mangravite, Pablo Meyer, Thea Norman, and Gustavo Stolovitzky. Crowdsourcing biomedical research: Leveraging communities as innovation engines. *Nature Reviews Genetics*, 17(8):470–486, 2016. doi: 10.1038/nrg.2016.69.

[34] Thomas Schaffter, Daniel Marbach, and Dario Floreano. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16): 2263–2270, 2011. doi: 10.1093/bioinformatics/btr373.

[35] Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, 2018. doi: 10.1038/nmeth.4612.

[36] Gustavo Stolovitzky, Don Monroe, and Andrea Califano. Dialogue on reverse-engineering assessment and methods: The DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences*, 1115(1):1–22, 2007. doi: 10.1196/annals.1407.021.

[37] Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S. Weber, Azadeh Seidi, Jafar S. Jabbari, Shalin H. Naik, and Matthew E. Ritchie. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, 16(6):479–487, 2019. doi: 10.1038/s41592-019-0425-8.

[38] Ian Walsh, Dmytro Fishman, Dario Garcia-Gasulla, Tiina Titma, Gianluca Pollastri, ELIXIR Machine Learning Focus Group, Jennifer Harrow, Fotis E. Psomopoulos, and Silvio C. E. Tosatto. DOME: Recommendations for supervised machine learning validation in biology. *Nature Methods*, 18(10):1122–1127, 2021. doi: 10.1038/s41592-021-01205-4.

[39] Lukas M. Weber, Wouter Saelens, Robrecht Cannoodt, Charlotte Soneson, Alexander Hapfelmeier, Paul P. Gardner, Anne-Laure Boulesteix, Yvan Saeys, and Mark D. Robinson. Essential guidelines for computational method benchmarking. *Genome Biology*, 20(1):125, 2019. doi: 10.1186/s13059-019-1738-8.

[40] Sean Whalen, Jacob Schreiber, William S. Noble, and Katherine S. Pollard. Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3):169–181, 2022. doi: 10.1038/s41576-021-00434-9.