

Ad hoc, post hoc and intrinsic-hoc in bioinformatics



Bioinformatics has undergone a transformation over the past decades. Early methods were often designed to meet the immediate analytical needs of biological data, offering accessible solutions built on established computational models. These tools delivered clear outputs and rapid analysis, becoming staples of exploratory data analysis. As omics data (for example, genomics, transcriptomics, proteomics and metabolomics) grow more complex, challenges such as parameter sensitivity and reproducibility issues arising from inconsistent results are becoming increasingly prominent. In response, researchers incorporate more advanced structures such as graph neural networks, transformers and large language models, and cross-domain synthesis into modeling approaches to enhance generalizability, robustness, reproducibility, interpretability and actionable insight. Yet few strategies can satisfy all these demands simultaneously, and such adaptations often induce more computing complexity and technical burdens. Considering that the data and computational frameworks are becoming complex, there is usually no single best solution for a fundamental biological question.

Researchers have adopted three methodological strategies – ad hoc, post hoc and intrinsic-hoc – that reflect recurring design patterns in bioinformatics. Revisiting these three strategies will help answer questions such as how to weigh trade-offs among robustness, interpretability and computational cost, when to use one of them or integrate them, which methods to apply in a particular scenario, and how to adapt them to different scenarios. It is important to recognize that the terms ‘ad hoc’, ‘post hoc’ and ‘intrinsic-hoc’ carry distinct, but related, meanings across disciplines. For instance, explainable AI applies these terms in the context of model interpretability, focusing on methods for understanding and explaining the behavior of machine learning models. We aim to synthesize these concepts in the context of bioinformatics and computational biology to emphasize the unique challenges posed

by high-dimensional, heterogeneous and domain-dependent big data, which differ substantially from most conventional machine learning benchmark settings.

Ad hoc methods in bioinformatics are characterized by their straightforward, problem-driven pattern recognition (for example, biomarkers, pathways and biological networks) for hypothesis testing and/or **early-stage hypothesis generation** (Fig. 1a). They are often deployed to **answer specific questions**, such as using rigorous statistical models for identifying differentially expressed genes (DEGs) between case and control groups. These tools are often efficient, easy to implement and adequate for **preliminary** or **exploratory** analyses where rapid turnaround is essential. However, they are sensitive to preprocessing steps, hypothesis formulations and parameter choices. Ad hoc methods often **lack built-in mechanisms to handle variability or uncertainty**, making them **prone to inconsistencies across parameter combinations or various datasets**. A typical example is using fold-change calculation and Student’s *t*-test to predict DEGs (Supplementary Information and Supplementary Fig. 1a). Such a method generates statistically significant results for testing whether a gene is differentially expressed under a certain biological condition in a given dataset, while the same gene may not be recognized by the Wilcoxon rank-sum test or in another dataset with the same biological condition. The significance (*P*-value) may also vary as a result of the sample size and data processing methods (for example, gene expression value normalization), sometimes making the differential comparison results inconsistent, even with the same fold change detected. Mostly, it is hard to discern the best or the most appropriate ad hoc tool for a specific analysis, except for some specific fields with a dominant tool – for example, AlphaFold¹ for protein 3D structure prediction and BLAST² for efficient sequence alignment.

To address the challenges of data variation, method variation and **high signal-to-noise ratio in real-world problems**, the rationale for moving from ad hoc to post hoc methods

is to enhance robustness, increase reliability in predictions, improve reproducibility and reconcile variability across analyses. **Post hoc methods** overcome the limitations of direct, single-target analyses by addressing additional questions, such as whether a result is robust or biased according to the experimental design or dataset selection. These methods **integrate outputs from multiple analyses**, through varying parameters of a given method, applying the same method to different datasets, or combining different analytical methods in one study. Results are then reconciled via **meta-analysis, consensus ranking, ensemble learning or belief theory** (Fig. 1b). The post hoc strategy introduces more layers of computational complexity and loses strong interpretability compared to ad hoc methods. The cost of post hoc methods is a higher requirement of computing resources and software engineering of multisystem implementations. Still, in the DEG analysis example, a consensus analysis by combining results from multiple datasets, parameter settings or tools can yield a more robust gene ranking for differential expression in a specific condition³. However, **biases introduced in individual analyses may persist in the aggregated results** (Supplementary Information and Supplementary Fig. 1b). It is worth noting that ‘post hoc’ in our framework carries a somewhat different meaning from its usage in explainable AI, where ‘post hoc’ typically refers to interpretability techniques applied after model training (for example, SHAP and LIME) to explain a black box model’s behavior. Here, we use ‘post hoc’ to denote strategies in bioinformatics that integrate and reconcile outputs from multiple analyses.

Finally, with the fast development of technologies, datasets and analytical methods, even post hoc strategies **may fail to yield robust or consensus results**. An example is the identification of hallmark genes for cellular senescence: among nine gene lists published over the years, only 39 out of 2,966 genes appeared in at least five, highlighting the lack of agreement⁴. **Intrinsic-hoc strategies** are adapted from intrinsic interpretability in explainable AI, where a model is inherently

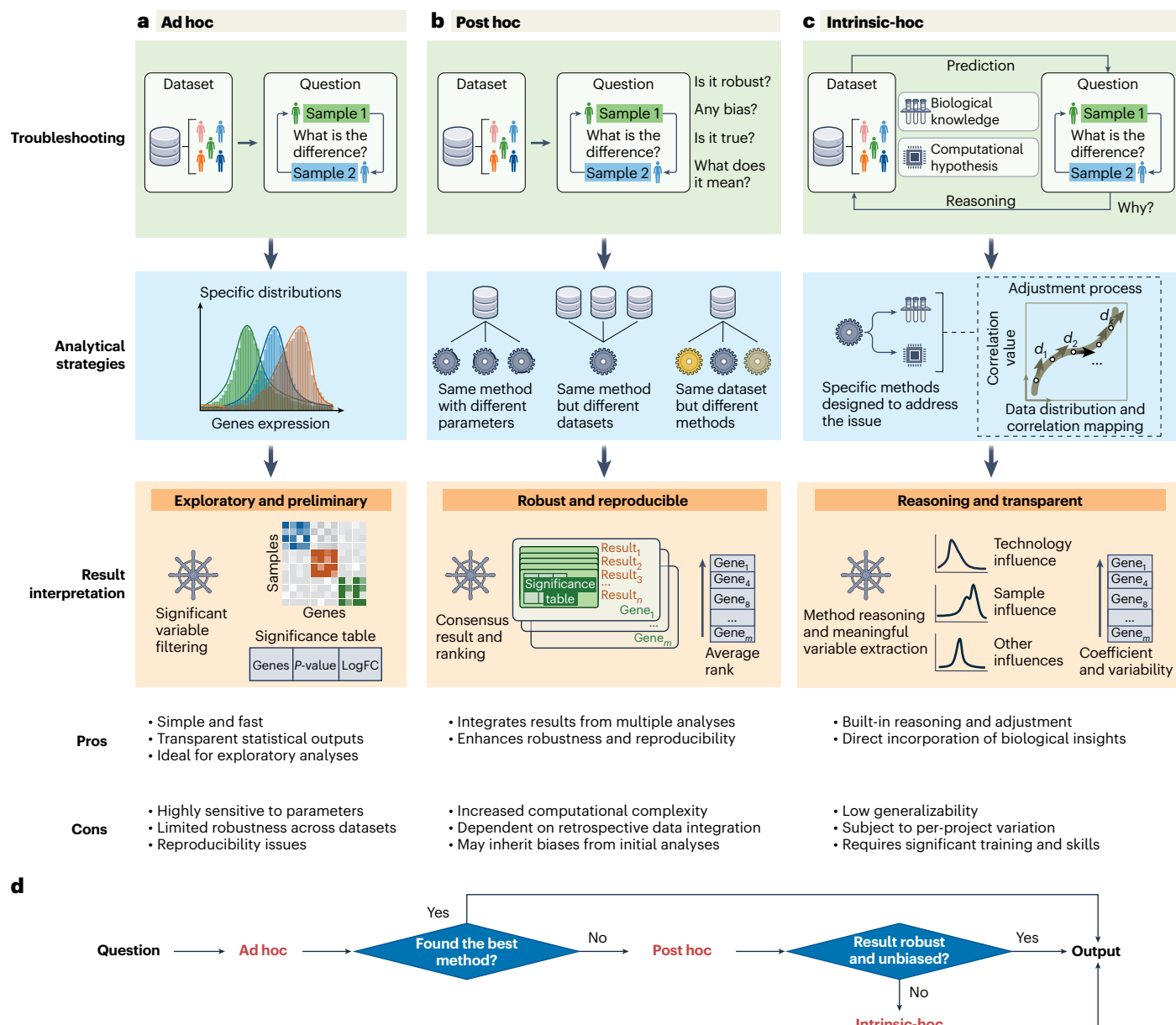


Fig. 1 | Comparison among ad hoc, post hoc and intrinsic-hoc strategies in bioinformatics. **a**, Ad hoc refers to a straightforward analytical pipeline that relies on statistical distributions, such as those used by edgeR and DESeq2, to identify DEGs. The results have good interpretability through significance tables and basic statistical metrics, making it well-suited for preliminary analyses. However, its outcomes can be highly sensitive to parameter choices and dataset variations, potentially affecting reproducibility. **b**, Post hoc refers to a more complex analytical strategy that organizes data into single or multiple datasets, applying varied parameters or multiple analytical methods. The results from these independent analyses are integrated through meta-analysis or consensus ranking, enhancing robustness and comprehensiveness. This panel highlights the interpretive challenges typical of post hoc analyses, which require additional interpretative steps to extract meaningful insights from computational outputs. **c**, Intrinsic-hoc refers to an advanced analytical design that inherently integrates

biological and computational constraints from the beginning. It explicitly models correlation structures and dynamically adjusts analytical processes. Unlike ad hoc and post hoc strategies, intrinsic-hoc methods embed biological insights and domain knowledge within the modeling process itself, transparently reasoning the prediction process. **d**, A flow chart for selection among three methodologies. Ad hoc methods are often the first line of approach when addressing a research question. If the selected tool or method is well-suited to the task with good benchmarking results, the outcome can be considered reliable. However, if uncertainty or variability remains, post hoc strategies should be used to test multiple hypotheses and assess the robustness and reproducibility of the findings. When results still appear biased or unstable, it suggests that the task is inherently complex, requiring intrinsic-hoc methods that incorporate biological knowledge, experimental design considerations or data distribution awareness directly into the algorithmic framework.

interpretable by design. Such models are built from the ground up to be understandable by humans – for example, linear regression, decision trees or neural networks whose architecture encodes domain knowledge (for example, layers corresponding to biological pathways). Interpretability is not added after the fact, but embedded directly into the model's assumptions and structure, via knowledge-guided or interpretable frameworks. In the context of bioinformatics, we define intrinsic-hoc as **biology-aware modeling** – approaches that integrate biological knowledge, experimental design or domain constraints directly into the algorithmic structure (Fig. 1c). For example, a recent DEG method, Memento⁵, characterizes how experimental factors affect the distribution of gene expression and tests their DEG performance specifically in interferon responses, perturbation and cell-type-specific quantitative trait loci (QTLs). It leverages the underlying molecular biology with statistical model training and provides predictions of how important and dynamic the gene's influence is across the dataset, as well as why it is picked or not picked as a DEG (Supplementary Information and Supplementary Fig. 1c). Other examples include a biologically informed deep neural network for prostate cancer discovery (P-NET)⁶, cell type classification using cell ontology graph (OnClass)⁷, and mutation rate prediction at the base pair resolution involving epigenomic, transcriptomic and regional variation of rare single-nucleotide variant rate (Roulette)⁸. Intrinsic-hoc methods are particularly powerful for tasks where model transparency and scientific validity are paramount, particularly when post hoc methods do not meet the needs. Nevertheless, because the deep layer of reasoning requires specialized training and fine-tuning, intrinsic-hoc methods are **difficult to generalize** and are often **more task-specific for biological contexts, data modalities or experimental assumptions**.

Ad hoc, post hoc and intrinsic-hoc strategies offer distinct advantages and trade-offs shaped by their assumptions, goals and interpretive depth (Supplementary Fig. 1d–f). To demonstrate that the trade-offs among the three strategy designs are generalizable, we added evidence from analytical tasks beyond DEG analysis in the Supplementary Information. There is no universally superior approach among the three categories for biological and biomedical data analysis; what matters is selecting the most appropriate method for the specific data, research questions and interpretive needs at hand. In general (Fig. 1d as a prototype), when carrying out a bioinformatics

project, we should always consider ad hoc methods first for exploratory analyses, for rapid assessments or when computational efficiency is prioritized. If a recognized ad hoc tool or method cannot be picked and further validation is needed to reconcile inconsistent findings, a post hoc strategy should be employed to assess robustness and potential bias. If the aggregated results are found to be robust and unbiased (where domain experts are usually involved), they can be finalized as output; otherwise, the analysis could move towards an intrinsic-hoc strategy, which embeds available domain knowledge directly into the model architecture. There are two main scenarios where an intrinsic-hoc approach becomes especially useful. First, when multiple methods used in a post hoc integration produce a wide range of solutions with no clear consensus, intrinsic-hoc strategies can help by embedding further constraints to stabilize the analysis. Second, when post hoc results conflict with well-established biological knowledge, intrinsic-hoc modeling is necessary to enforce those domain-specific rules. For example, in protein docking, if a homodimer is expected to exhibit symmetry but consensus-based selection yields asymmetric results between the two subunits, incorporating symmetrical constraints directly into the docking algorithm provides a biology-aware, intrinsic-hoc solution.

Method selection is not a linear decision, but a dynamic process based on confidence in results, robustness and interpretability. In practice, there are potentially other circumstances – researchers may directly jump to the development of intrinsic-hoc methods when existing ad hoc tools are not suitable and the models are expected to be aligned with underlying biological mechanisms and/or precision medicine. Moreover, emerging developments in AI and deep learning, especially those models emphasizing interpretability, are increasingly aligned with the goals of intrinsic-hoc strategies and may accelerate their adoption in bioinformatics research⁹. For example, multimodal AI approaches that integrate metabolic models with omics and imaging data offer a promising path towards biologically grounded, interpretable insights, which may improve the precision and impact of therapeutic decision-making¹⁰.

Just as linear and logistic regression methods represent fundamental paradigms in statistics, we believe ad hoc, post hoc and intrinsic-hoc strategies should be understood as core methodological classes in bioinformatics, each with its logic, structure and scope of

applicability. By articulating and elevating these categories into a broader methodological discourse, we hope to provide a more deliberate framework to guide method selection. A clear understanding of these strategies and their relationships will not only address inefficiencies and inconsistencies in current practices but also help bioinformaticians select and design appropriate analytical strategies for individual projects and large-scale consortia. We believe the information and discussion in this Correspondence can offer best practices to biologists for testing and formulating hypotheses; enhance interdisciplinary collaboration across biology, computer science and engineering; and improve the training of the next generation of scientists by embedding methodological thinking into bioinformatics education. In the long run, we hope this perspective contributes to shape bioinformatics as a more cohesive and theory-informed scientific discipline.

Data availability

All data used in this study are publicly available without restrictions. The 10k PBMCs dataset from a healthy donor can be downloaded from <https://www.10xgenomics.com/datasets/10k-human-pbmc3-v3-1-chromium-controller-3-1-high>. The acute kidney injury single-cell RNA-seq data can be downloaded from <https://cellxgene.cziscience.com/collections/bcb61471-2a44-4d00-a0af-ff085512674c>.

Qin Ma ^{1,2}, Hao Cheng¹, Yi Jiang¹, Anjun Ma ^{1,2} & Dong Xu ^{3,4}

¹Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA.

²Pelotonia Institute for Immuno-Oncology, James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. ³Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.

⁴C.S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA.

✉ e-mail: qin.ma@osumc.edu

Published online: 14 October 2025

References

1. Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. Mol. Biol.* **215**, 403–410 (1990).
3. Berger, E. et al. *Nat. Commun.* **10**, 773 (2019).
4. Qu, Y. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.11.21.568150> (2023).
5. Kim, M. C. et al. *Cell* **187**, 6393–6410.e6316 (2024).
6. Elmarakeby, H. A. et al. *Nature* **598**, 348–352 (2021).
7. Wang, S. et al. *Nat. Commun.* **12**, 5556 (2021).

8. Seplyarskiy, V. et al. *Nat. Genet.* **55**, 2235–2242 (2023).
9. Du, M., Liu, N. & Hu, X. *Commun. ACM* **63**, 68–77 (2019).
10. Occhipinti, A., Verma, S., Doan, L. M. T. & Angione, C. *Trends Cell Biol.* **34**, 85–89 (2024).

Acknowledgements

This work was supported by awards R01GM152585 (Q.M.), R01DK138504 (Q.M.), P01AI177687 (Q.M. and A.M.), P01CA278732 (Q.M. and A.M.), U54AG075931 (Q.M. and

A.M.), R21DK140693 (A.M.) and R35GM126985 (D.X.) from the US National Institutes of Health (NIH). This work was also supported by the Pelotonia Institute of Immuno-Oncology (PIIO) at The Ohio State University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and the PIIO.

Author contributions

Q.M., A.M. and D.X. jointly conceived and drafted the manuscript. Y.J. and H.C. implemented the case study in the Supplementary Information.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-025-02852-0>.

Peer review information *Nature Biotechnology* thanks Rong Fan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.