A call to incorporate early-stage drug discovery priorities into multiomics AI benchmarking

Kristina Ulicna[1,2*†], Oren Kraus[2], Anne E. Carpenter[3], Fabian J. Theis[4], Tommaso Biancalani[5], Paula Andrea Marin Zapata[6], William J. Godinez[7], Rob Moccia[8], Cuong Q. Nguyen[9], Nathaniel Robichaud[10], Alexandra Pettet[11], Djork-Arné Clevert[12], Judith Mueller[13], Ian Barrett[14], Alisandra K. Denton[1,2*]

[1]Valence Labs, Montréal, Québec, Canada.
[2]Recursion, Salt Lake City, UT, USA.
[3]Broad Institute of MIT & Harvard, Cambridge, MA, USA.
[4]Helmholtz Munich & Technical University of Munich (TUM), Germany.
[5]Genentech, San Francisco, CA, USA.
[6]Bayer, Berlin, Germany.
[7]Novartis Biomedical Research, Emeryville, CA, USA.
[8]Valid Inc., Natick, MA, USA.
[9]Genesis Molecular AI, San Francisco, CA, USA.
[10]Nomic Bio, Montréal, Québec, Canada.
[11]GlaxoSmithKline (GSK), San Francisco, CA, USA.
[12]Pfizer Research & Development, Berlin, Germany.
[13]Merck / Merck Sharp & Dohme (MSD), Cambridge, MA, USA.
[14]AstraZeneca, Cambridge, UK.

*Corresponding authors: kristina.smith.ulicna@gmail.com; ali@valencelabs.com;
†The author has moved to a new affiliation since the completion of this work.

Correspondence Letter

The integration of complementary high-dimensional cellular data types, known as *multiomics*, offers unprecedented potential to decode complex biology and accelerate therapeutic discovery. Combined with artificial intelligence (AI), these approaches

promise to enable a more holistic understanding of cellular states, both healthy and diseased, and their modulation by candidate drugs. When strategically embedded into pharmaceutical pipelines, multiomics AI holds transformative potential to accelerate and de-risk the path to novel medicines. Yet, a critical disconnect persists between the perceived technical progress in multiomics AI research and its tangible impact on drug development. This gap is partly anchored in limited, and often misaligned, benchmarking standards that prioritize leaderboard-style metrics over translational relevance. In this correspondence, leading experts drawing on academic and industry expertise call for a rethinking of benchmarking strategies to better reflect the realities of early-stage drug discovery. We scope broader discussions of AI benchmarking in biology [1, 2] by focusing on priorities specific to drug discovery. Our aim is to catalyze the development of systematic datasets and evaluative frameworks that provide practical, decision-informing value tailored to therapeutic development.

Although the term *multiomics* remains broad and often ambiguous, its relevance to biology-driven stages of drug discovery is becoming increasingly evident across both target- and phenotype-based approaches (Figure 1A). By integrating molecular layers (*e.g.*, genomics, transcriptomics, proteomics) with phenotypic readouts (*e.g.*, high-content imaging, morphological profiling, viability assays), multiomics is uniquely suited to capture the underlying structure of biological systems and deepen our understanding of drug-induced biological mechanisms across diverse disease areas and evolving drug programs. However, multiomics AI models must gain more flexibility and adaptability to shifting assay constraints and decision points along the pipeline to reliably and transparently surface promising drug candidates early, while flagging likely failures before they become costly downstream (Figure 1B). This could be achieved by learning **perturbation effect fingerprints**, *i.e.* relatable biological knowledge representations which reliably distil relevant information from high-dimensional and noisy data. The goal of such fingerprints is to translate and complement the signal across bioassays of different levels of overlap, complexity, throughput, and species of the model organism and comprehend their interplay with candidate drugs (Figure 1C).

The promise of these fingerprints hinges on systematic and rigorous evaluation, which needs to move beyond overfitting to narrow, over-specified datasets focused solely on optimising technical, leaderboard-style metrics such as accuracy. When used in isolation, such inadequate scoring risks favoring models that simply recapitulate previously known relationships, such as annotated gene-gene or gene-compound interactions [3], rather than discovering novel insights (Figure 1D). This creates a fundamental tension: benchmarks may reward **familiarity over discovery**. In early-stage drug discovery, this leads to cycling through known targets and optimizing best-in-class drug designs rooted in well-characterized pathways, while overlooking bold and vast first-in-class opportunities in uncharted biology. In later stages, this could risk flagging only obvious, well-understood toxicity links, while missing more subtle yet critical side-effect signals that, if undetected early, can lead to costly failures downstream. While confirming the validity of yet-unknown biological concepts

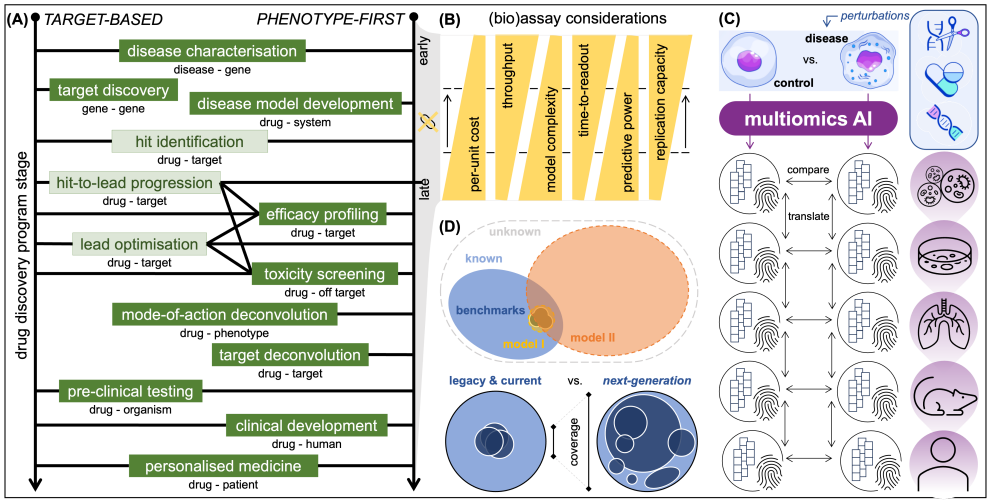

Fig. 1 Opportunities and challenges for multiomics AI in drug discovery. (A) Multiomics AI has broad relevance (green shading) across both target-based (left) and phenotype-first (right) discovery strategies, supporting nearly all stages from early biology through feedback-informed chemistry to clinical trials via improved integration and benchmarking. **(B)** Multiomics AI must flexibly navigate the evolving landscape of (bio)assay and data readout constraints as therapeutic programs advance to prioritize informative experiments and discontinue failing candidates as early as possible (dashed lines). **(C)** Central to this effort is learning biologically meaningful *perturbation effect fingerprints* that reliably compare between control (*e.g.*, modelling healthy) and perturbation (*e.g.*, modelling diseased) states, supporting insights into their modulation by candidate drugs. These fingerprints must adapt to increasing system complexity, spanning noisy single-cell data, heterogeneous cell lines, organoid-like structures, animal models, and patient samples. **(D)** Yet, robust benchmarking remains a barrier as current evaluations often rely on narrow, highly-overlapping datasets that reward models for recovering known biology (model I), rather than those capable of generalizing and uncovering novel insights (model II). The purpose-built next generation of benchmarks should therefore increase the coverage of the biological areas, systems, and species and include datasets of varying size, overlap, and distributions.

is challenging at first, pharmaceutical companies are uniquely positioned to address this through feedback loops between computational and experimental teams. With the resources to run experiments, plausible hypotheses can and should get tested and drive iterative refinement of multiomics AI models, implementing the so-called lab-in-the-loop approach [4, 5].

However, current multiomics AI models must overcome much broader challenges of benchmarking robustness to meaningfully support drug discovery. The limited scope of current benchmarks is understandable: the multiomics field is still emerging and faces unique and significant hurdles. Generating high-quality, large-scale, paired datasets is financially demanding, and their annotation is even more technically complex and resource-intensive due to lack of consensus or feasibility of human expert interpretation. As a result, many datasets lack a ground truth reference, making it hard to define tasks and metrics that measure real progress in representing multiomics

data. Separating meaningful model advances from small, incremental gains is often further worsened by the absence of meaningful comparisons to alternative methods, especially to simple, clear baselines like random rankings, statistical heuristics, linear models or intuitive data analysis. Without such context, demonstrating added value of complex multiomics AI models is difficult, and has proven to become a major oversight to date [6–8].

In this letter, we advocate for designing a set of multiomic AI benchmarking tasks tailored to real-world drug discovery goals, such as grouping genes or compounds by functional similarity, forecasting perturbation effect fingerprints into specific biological contexts, translating insights across modalities or species, and tracking cellular state transitions over time. This would liberate current benchmarks, often relying on legacy datasets from early studies, which have become *de facto* standards for subsequent methods evaluation, often despite their inappropriateness for the task. Rather than discarding these default evaluation tools designed without the drug discovery goals in mind, we should clearly communicate their scope and limitations, and complement them with next-generation, purpose-built tasks that better reflect the biological and translational demands. Such benchmarks should be routinely positioned within well-defined lower baselines and, where possible, upper bounds to critically assess model performance and to enable fast, convenient and reproducible evaluation of new and existing models against them. Establishing and adhering to such standards will drive consistency across academia and industry, empowering both interpretability-focused and action-oriented use cases.

In summary, we recommend the following core principles for developing a set of actionable, next-generation multiomics AI benchmarks:

- **Open, transparent datasets:** Public release of multiomic data with detailed quality-control metrics, standardized annotations, and transparent reporting of dataset biases.

- **Biology-inspired tasks:** Development of evaluation tasks and performance measures that reflect true drug discovery impact beyond leaderboard-style metrics, and ideally cross-validating the findings experimentally in the real world.

- **Meticulous baseline comparisons:** Systematic evaluation of existing models against simple, intuitive baselines informed by the dataset biases, distributions and coverage.

- **Shared benchmarking platforms:** Community-hosted repositories and leaderboards that integrate datasets, tasks, and results to accelerate method development, as inspired by other pharmaceutics-adjacent fields [9, 10].

We recognise that these recommendations are ambitious, but their implementation is urgent, necessary, and achievable by deliberately aligning the multiomics AI community toward real-world impact in drug discovery.

Acknowledgements. We thank Cas Wognum (Valence Labs) for his early guidance, ongoing support, and leadership in helping establish the multiomics steering committee and shaping this correspondence letter from its inception. We also thank Emmanuel Noutahi (Valence Labs) for his thoughtful feedback on the final versions of the manuscript and for providing support throughout the writing phase that enabled this work.

Declarations. F.J.T. consults for Immunai Inc., CytoReason Ltd., Cellarity, Bio-Turing Inc., and Genbio.AI Inc., and has an ownership interest in Dermagnostix GmbH and Cellarity. A.E.C. serves as scientific advisor for companies that use image-based profiling and Cell Painting (Recursion, SyzOnc, Quiver Bioscience) and receives honoraria for occasional scientific visits to pharmaceutical and biotechnology companies. Except for A.E.C. and F.J.T., all other authors are employees of for-profit companies. The authors declare no additional competing interests.

References

[1] Mahmood, F.: A benchmarking crisis in biomedical machine learning. Nature Medicine **31**, 1060 (2025) https://doi.org/10.1038/s41591-025-03637-3

[2] Brooks, T.G., Lahens, N.F., Mrčela, A., Grant, G.R.: Challenges and best practices in omics benchmarking. Nature Reviews Genetics **25**, 326–339 (2024) https://doi.org/10.1038/s41576-023-00679-6

[3] Celik, S., Hütter, J.-C., Melo Carlos, S., Lazar, N.H., Mohan, R., Tillinghast, C., Biancalani, T., Fay, M.M., Earnshaw, B.A., Haque, I.S.: Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. PLOS Computational Biology **20**(10), 1012463 (2024) https://doi.org/10.1371/journal.pcbi.1012463

[4] Noutahi, E., Hartford, J., Tossou, P., Whitfield, S., Denton, A.K., Wognum, C., Ulicna, K., Craig, M., Hsu, J., Cuccarese, M., Bengio, E., Beaini, D., Gibson, C., Cohen, D., Earnshaw, B.: Virtual Cells: Predict, Explain, Discover (arXiv:2505.14613) (2025). https://doi.org/10.48550/arXiv.2505.14613

[5] Frey, N.C., Hötzel, I., Stanton, S.D., Kelly, R., Alberstein, R.G., Makowski, E., Martinkus, K., Berenberg, D., Bevers, J.I., Bryson, T., Chan, P., Czubaty, A., D'Souza, T., Dwyer, H., Dziewulska, A., Fairman, J.W., Goodman, A., Hofmann, J., Isaacson, H., Gligorijević, V.: Lab-in-the-loop therapeutic antibody design with deep learning. bioRxiv (2025) https://doi.org/10.1101/2025.02.19.639050

[6] Boiarsky, R., Singh, N.M., Buendia, A., Amini, A.P., Getz, G., Sontag, D.: Deeper evaluation of a single-cell foundation model. Nature Machine Intelligence **6**(12), 1443–1446 (2024) https://doi.org/10.1038/s42256-024-00949-w

[7] Ahlmann-Eltze, C., Huber, W., Anders, S.: Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear baselines. Nature Methods (2025) https://doi.org/10.1038/s41592-025-02772-6

[8] Seal, S., Dee, W., Shah, A., Zhang, A., Titterton, K., Cabrera, A.A., Boiko, D., Beatson, A., Carreras Puigvert, J., Singh, S., Spjuth, O., Bender, A., Carpenter, A.E.: Small molecule bioactivity benchmarks are often well-predicted by counting cells. bioRxiv (2025) https://doi.org/10.1101/2025.04.27.650853

[9] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Topf, M.: Critical assessment of techniques for protein structure prediction, fourteenth round. CASP14 Abstract Book (2020). https://doi.org/10.1002/prot.26237

[10] Wognum, C., Ash, J.R., Aldeghi, M., *et al.*: A call for an industry-led initiative to critically assess machine learning for real-world drug discovery. Nature Machine Intelligence **6**, 1120–1121 (2024) https://doi.org/10.1038/s42256-024-00911-w