Trust, Reproducibility, and Progress: The Roles of Independent Blind Prediction and Assessment and Benchmarking in Computational Biology

Gaia Andreoletti, Serghei Mangul,
Sage Bionetworks
2901 3rd Ave, suite 330, Seattle (WA), United States of America
Email: gaia.andreoeltti@sagebase.org

Predrag Radivojac
Northeastern University
177 Huntington Avenue, Boston, MA 02115, United States of America
Email: *predrag@northeastern.edu*

Steven E. Brenner
University of California Berkeley
461A Koshland Hall #3102, CA 94720-3102, United States of America
Email: brenner@compbio.berkeley.edu

Abstract

When evaluations aren't trustworthy, entire research programs can chase mirages. Objective benchmarks and independent assessment have repeatedly catalyzed progress across computational biology, from protein structure prediction to variant interpretation and single-cell analysis. This workshop gathers leaders of community challenges and benchmarking infrastructures together with domain experts to provide a contemporary view of how to design trustworthy evaluations, why blind prediction matters, and where standards, infrastructure, and policy must evolve to meet the demands of AI-driven biology. We summarize the motivation and scope of the workshop; provide background on methodological and infrastructural advances that enable rigorous benchmarking; highlight invited speakers' contributions; and outline anticipated outcomes and community calls-to-action.

Keywords: benchmarking; blind prediction; reproducibility; evaluation; community challenges; standards; FAIR; open science; foundation models; genomics; proteomics; structural biology; biomedicine

1. Background and Motivation

Benchmarking, the systematic, reproducible evaluation of computational methods on well-characterized tasks, has been a primary engine of progress in computational biology. Community experiments built on blind prediction and independent assessment have delivered clear snapshots of the state-of-the-art while revealing gaps, biases, and practical limitations.

Structured challenges have established trust by separating training from evaluation via held-out or "model-to-data" test sets and by publishing transparent rules and metrics. In the "model-to-data" infrastructure, algorithms are sent to the data in a secure environment; test data never leave the enclave. These approaches have improved reproducibility through standardized data, well-defined tasks, and shared tooling such as containers, workflows, and leaderboards. They have accelerated translation by focusing evaluation on end-user-relevant criteria, including pathogenicity classification accuracy, clinical utility, and robustness across cohorts. Moreover, they have focused innovation by clarifying failure modes, encouraging ablation studies, and highlighting generalizable approaches.

The advent of large neural models and multi-omics assays makes rigorous, forward-compatible benchmarking both more urgent and more challenging. Data heterogeneity, potential leakage, shifting distributions, and compute-intensive models complicate fair comparisons. At the same time, funders, journals, and regulators increasingly look to community benchmarks to establish standards of evidence for methods intended for biomedical research and clinical use.

2. Principles of Trustworthy Benchmarking

1. Clear problem framing and intended use: precisely define the task and scope, with explicit target populations and contexts of use.

2. Data stewardship and curation: representative, consent-appropriate datasets; provenance; management of confounders; FAIR data practices; model-to-data evaluation when needed.

3. Separation of concerns: strict separation of training/validation/test; preregistered evaluation protocols where feasible.

4. Transparent, meaningful metrics: align with real-world utility; calibration and uncertainty; confidence intervals and significance tests.

5. Baselines, ablations, and error analysis: strong baselines; ablation studies; post-challenge error analyses.

6. Reproducible workflows: containerized submissions, version-controlled pipelines, archival of code/models, reference implementations and persistent identifiers.

7. Sustained infrastructure and governance: long-lived, community-maintained platforms and steering committees; open licenses; clear conflict-of-interest policies.

8. Equity and generalizability: evaluate across demographic, technical, and laboratory strata; robustness under dataset shift and rare-event conditions.

3. Landscape of Community Experiments and Benchmarking Infrastructures

The computational biology community has developed a rich ecosystem of benchmarking initiatives, each addressing distinct aspects of method evaluation and validation. Long-standing challenges such as CASP for protein structure prediction and CAFA for function annotation have demonstrated the power of iterative, multi-year assessments to drive algorithmic innovation and establish performance baselines. These initiatives have set precedents for blind prediction protocols, independent assessment frameworks, and community engagement that continue to influence new benchmarking efforts.

Recent years have witnessed a proliferation of domain-specific benchmarks spanning genomics, transcriptomics, proteomics, and integrative multi-omics analysis. Initiatives focused on variant interpretation, single-cell analysis, spatial transcriptomics, and clinical genomics have emerged to address the unique challenges of their respective domains. Simultaneously, infrastructure projects have developed platforms for secure model evaluation, including federated learning environments, containerized submission systems, and continuous benchmarking frameworks that enable ongoing assessment as new methods and data become available.

Despite this progress, significant gaps remain. Many benchmarks lack the infrastructure to prevent data leakage from large pre-trained models, struggle to evaluate generalization across diverse populations and experimental conditions or fail to capture clinically relevant performance metrics. The tension between open science principles and the need for truly held-out test data continues to challenge benchmark design. Furthermore, the computational resources required to evaluate modern deep learning models create barriers to participation and comprehensive comparison, while the rapid pace of methodological innovation often outstrips the capacity of static benchmarks to remain relevant.

4. Speakers

- Steven Brenner, University of California Berkeley, USA
- Predrag Radivojac, Northeastern University
- Brett Beaulieu-Jones, University of Chicago, USA
- Lana Garmire, University of Michigan, USA
- Panel Q&A

5. Workshop Format and Participation

This workshop begins with a framing introduction, followed by invited-talk blocks followed by a panel discussion with open Q&A and a live demo. Interactive polling will solicit audience views on priorities (e.g., preregistration, dataset access models, and required reporting) and capture community-endorsed recommendations.

5.1 Live Demo

Finding the right biomedical challenge or benchmark for your research can be frustratingly difficult. Information is scattered across different websites and platforms, making it hard to discover relevant datasets and tasks. OpenChallenges (OC) solves this problem by bringing together information about over 2,000 biomedical challenges from sources like Synapse and Kaggle into one searchable platform.

We will demonstrate how OC works with AI chatbots to make finding challenges as easy as having a conversation. Instead of manually searching through websites, you can simply ask questions like "show me genomics challenges with public baseline results" and get immediate, relevant answers. During the live demo, participants will see how to search for challenges, save useful queries, and export results in formats ready for analysis.

This session is designed for anyone interested in benchmarking—no technical expertise required. You'll leave with practical tools to discover challenges relevant to your work, whether you're looking for datasets to test new methods, teaching materials for courses, or opportunities to participate in community experiments.

6. Expected Outcomes

- A concise checklist for designing and reporting trustworthy benchmarks in computational biology.
- A community “wish list” for infrastructure and policy (e.g., preregistration; FAIR test sets; standardized container interfaces; uncertainty and error reporting).
- A shared bibliography of landmark challenges, infrastructure papers, and best-practice guidance to support new evaluations.

7. Acknowledgments

We thank the invited speakers for their contributions to community benchmarking and the organizers and sponsors of the Pacific Symposium on Biocomputing for supporting this workshop. We also acknowledge community contributors and maintainers of benchmarking platforms, dataset providers, and assessors for their essential roles in rigorous, open science.