

Blind Challenges Let Us See the Path Forward for Predictive Models

Published as part of *Journal of Chemical Information and Modeling* special issue “Open Science and Blind Data: The Antiviral Discovery Challenge”.

John D. Chodera, W. Patrick Walters,* Sriram Kosuri, and James S. Fraser

Cite This: <https://doi.org/10.1021/acs.jcim.6c00205>

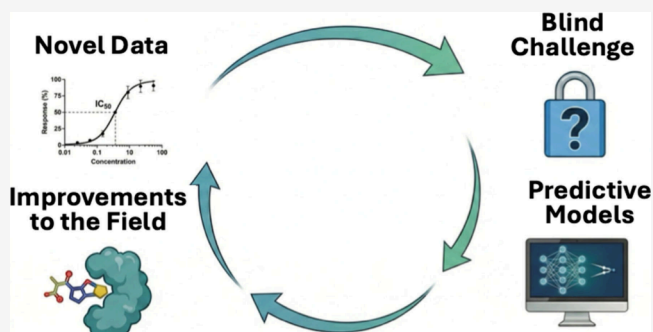
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: The rapid proliferation of AI/ML models in drug discovery heralds an era of extraordinary progress but also raises urgent questions about whether the true predictive performance is as good as advertised. On-target prediction models often benefit from high-resolution structural or atomistic representations that capture the subtleties of binding affinity and pose. In contrast, off-target and ADMET liabilities have typically relied on more implicit representations of molecular interactions. Retrospective benchmarks often provide a misleading picture of how successful these diverse representations are at predicting properties, and the community lacks standardized, prospective comparisons. Blind challenges, such as the OpenADMET × ASAP × PolarisHub Challenge featured in this issue, are crucial for realistically evaluating progress, encouraging iterations, and directing collective efforts toward major accuracy barriers. With ongoing investment in large-scale, open data creation, and community-led challenges, predictive modeling is poised to rapidly transform drug discovery by enabling accurate, multiparameter optimization.



■ DRUG DISCOVERY IS COSTLY

Drug discovery and development is costly, time-consuming, and subject to failure.¹ While clinical phases are individually the most costly, the sequential nature of drug discovery and high cumulative failure rates mean that the majority of cost per approved drug—where total costs per approved drug now exceed \$1B^{2,3}—is due to accumulated failures in the discovery phase and associated issues that likely could have been addressed earlier in discovery. The success of a drug discovery program hinges on multiparameter optimization: the empirical balancing of on-target potency, off-target specificity, and ADMET properties. This balancing act is where predictive models have the greatest potential to reduce costly cycles of design, synthesis, and testing.

■ COMPUTER-AIDED DRUG DISCOVERY (CADD) HOLDS ENORMOUS POTENTIAL TO ACCELERATE PROGRESS

Computer-aided drug discovery (CADD) has long sought to guide molecular design decisions with predictive models, aiming to save time and reduce attrition.⁴ Even modest improvements in model accuracy can yield superlinear returns by reducing the number of compounds that need to be synthesized or advanced.^{5,6}

A surge of enthusiasm for new drug discovery methods driven by artificial intelligence (AI) and machine learning

(ML) has brought significant new talent, techniques, and energy to the field. This has increased expectations for major breakthroughs in performance, similar to what AlphaFold achieved with protein structure prediction (as well as the expansion into related areas, such as nucleic acid and small-molecule prediction, using software inspired by AlphaFold).⁷

The history of blind challenges in related fields of structural biology and computational chemistry underscores their value. The CASP experiment in protein structure prediction—supported by decades of systematic data curation from the Protein Data Bank (PDB)—galvanized progress and culminated in AlphaFold’s breakthrough. Blind challenges drive method development by providing essential experimental feedback, sharing valuable benchmarks with the community, and maintaining focus on aspirational goals for models that deliver real utility.

As CADD looks for its “AlphaFold” moment, it remains difficult to evaluate how well these models actually perform in practice. Retrospective benchmarks are plagued by data

Received: January 21, 2026

leakage, inconsistent curation, and a lack of standardized data sets.^{8,9} Moreover, practical AI/ML models often demand large, high-quality data sets that remain scarce in many critical domains of drug discovery. Blind challenges provide an essential solution. By assessing models prospectively on common, well-designed data sets unavailable during training, blind challenges create a level playing field, generate realistic estimates of predictive utility, focus the field on critical problems in need of solutions, and foster rapid community-wide iteration and learning to accelerate progress. Similarly, SAMPL,¹⁰ D3R¹¹/CELPP,¹² and CACHE¹³ have advanced free energy calculations, docking, and hit identification, respectively. Without careful prospective assessment, it is easy to fool ourselves into overestimating practical performance—a risk Richard Feynman famously warned against (“The first principle is that you must not fool yourself and you are the easiest person to fool.^{14*}”). Currently, the CACHE initiative has begun to fill this gap for *on-target* hit identification, providing a template for how prospective evaluation can sharpen models and align community efforts.

■ ADMET PROPERTIES AND ANTITARGET AS A FOCUS FOR BLIND CHALLENGES

While on-target binding is often the first focus of predictive modeling, the ultimate fate of a drug candidate is usually determined by ADMET properties and interactions with antitargets, proteins where drug binding alters toxicity and pharmacokinetics. Poor solubility, metabolic instability, transporter efflux, or unexpected channel binding (e.g., hERG¹⁵) are among the leading causes of drug discovery failure. While factors such as poor solubility are primarily physicochemical, other critical bottlenecks, such as metabolic instability, transporter-mediated efflux, or unexpected ion channel binding, are often mediated by a relatively limited repertoire of proteins. While many of these liabilities are now screened for in the preclinical phase,¹⁶ they remain primary drivers of attrition in the discovery pipeline. We dub this specific set of proteins “antitargets”. This set includes targets that mediate toxicity and metabolizing enzymes (such as CYPs¹⁷). While interactions with the latter are sometimes optimized for specific profiles, they represent a finite landscape of interactions that, if better predicted, would benefit the entire field. The limited size of this set suggests that understanding these liabilities could be essential to improving success rates. Incorporating ADMET and antitarget data into predictive frameworks ensures that the next generation of AI/ML models does not simply identify binders but also identifies compounds with a realistic chance of becoming safe and effective medicines. Moreover, any gains made in the ability to predict compound interactions with these antitarget proteins are likely to benefit *all* drug discovery efforts. In contrast, gains in the predictive ability against a specific target may not generalize.

■ BLIND CHALLENGES REQUIRE EVERGREEN DATA GENERATION EFFORTS

For blind challenges to succeed in transforming the field, they require an evergreen source of new data. Retrospective repositories such as ChEMBL aggregate valuable information from the literature but often resemble “dumpster diving” for data: large numbers of small heterogeneous assay data sets, inconsistent conditions, correlated or biased data sets generated for an orthogonal purpose, and mixed measurement

types complicate the ability to both build accurate models and assess predictive utility with appropriate statistical power.

Centralized, large-scale initiatives can overcome these limitations by generating large, robust, high-quality, consistent data sets tailored to predictive modeling needs. Economies of scale, advanced technologies, and active learning can reduce costs, increase scale, and ensure the data generated is highly informative and fit-for-purpose for building and assessing predictive models.

In the near term, we anticipate that individual academic groups will continue to generate valuable public data sets on specific targets, providing a critical testing ground for new models. However, the most pressing ADMET data sets remain locked behind closed doors in industry, limiting broad community impact. Initiatives like OpenADMET (currently funded by ARPA-H Avoid-ome, the Gates Foundation, and the Astera Institute), in partnership with ASAP (the AI-driven Structure-enabled Antiviral Platform) and PolarisHub, aim to break this barrier by generating open data sets that capture both structural and functional information on key antitargets. OpenADMET breaks down these barriers by leveraging high-throughput experimentation and structural biology to provide insights into how diverse small molecules bind to the “Avoid-ome”. By pairing X-ray and cryoEM structures with comprehensive biochemical and cellular assay data generated under consistent conditions, the initiative provides the community with the “ground truth” needed to move beyond reliance on heterogeneous literature databases. These collaborations create a continuous pipeline of high-throughput experimental data that can be used in future challenges and benchmarking efforts, ensuring that the challenges featured in this issue have a sustainable source of high-quality data. Sustained blind challenges on individual antitarget data sets will sharpen models for well-defined liabilities, while complementary ADMET challenges from individual target-based campaigns, such as the pan-coronavirus study highlighted here, will test whether models can handle the multiparameter trade-offs that ultimately determine success in drug development. Together, these dual-challenge formats will be essential to ensuring that predictive modeling keeps pace with the complex realities of drug discovery.

■ AUTHOR INFORMATION

Corresponding Author

W. Patrick Walters — OpenADMET, Boston, Massachusetts 01581, United States; orcid.org/0000-0003-2860-7958; Email: pat.walters@omsf.io

Authors

John D. Chodera — Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York 10065, United States

Sriram Kosuri — Octant, Inc., Emeryville, California 94608-1016, United States

James S. Fraser — Department of Bioengineering and Therapeutic Sciences and Quantitative Biosciences Institute, University of California San Francisco, San Francisco, California 94143, United States; orcid.org/0000-0002-5080-2859

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jcim.6c00205>

Funding

Research reported in this publication was partially supported by the Advanced Research Projects Agency for Health (ARPA-H) under AVOID-OME: Structurally enabling the “avoid-ome” to accelerate drug discovery and Award Number 1AY1AX000035-01.

Notes

The authors declare no competing financial interest.

REFERENCES

- (1) Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B. Diagnosing the Decline in Pharmaceutical R&D Efficiency. *Nat. Rev. Drug Discovery* **2012**, *11* (3), 191–200.
- (2) Ringel, M. S.; Scannell, J. W.; Baedeker, M.; Schulze, U. Breaking Eroom's Law. *Nat. Rev. Drug Discovery* **2020**, *19* (12), 833–834.
- (3) *Erooms_law: Eroom's Law*; Github.
- (4) Brown, F. K.; Sherer, E. C.; Johnson, S. A.; Holloway, M. K.; Sherborne, B. S. The Evolution of Drug Design at Merck Research Laboratories. *J. Comput. Aided Mol. Des.* **2017**, *31* (3), 255–266.
- (5) Shirts, M. R.; Mobley, D. L.; Brown, S. P. Free-Energy Calculations in Structure-Based Drug Design. *Drug Design* **2010**, *1*, 61–86.
- (6) Retchin, M.; Wang, Y.; Takaba, K.; Chodera, J. D. DrugGym: A Testbed for the Economics of Autonomous Drug Discovery. *bioRxiv* **2024**, No. 2024.05.28.596296.
- (7) Ahdritz, G.; Bouatta, N.; Floristean, C.; Kadyan, S.; Xia, Q.; Gerecke, W.; O'Donnell, T. J.; Berenberg, D.; Fisk, I.; Zanichelli, N.; Zhang, B.; Nowaczynski, A.; Wang, B.; Stepniewska-Dziubinska, M. M.; Zhang, S.; Ojewole, A.; Guney, M. E.; Biderman, S.; Watkins, A. M.; Ra, S.; Lorenzo, P. R.; Nivon, L.; Weitzner, B.; Ban, Y.-E. A.; Chen, S.; Zhang, M.; Li, C.; Song, S. L.; He, Y.; Sorger, P. K.; Mostaque, E.; Zhang, Z.; Bonneau, R.; AlQuraishi, M. OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization. *Nat. Methods* **2024**, *21* (8), 1514–1524.
- (8) Graber, D.; Stockinger, P.; Meyer, F.; Mishra, S.; Horn, C.; Buller, R. Resolving Data Bias Improves Generalization in Binding Affinity Prediction. *Nat. Mach. Intell.* **2025**, *7* (10), 1713–1725.
- (9) Bernett, J.; Blumenthal, D. B.; Grimm, D. G.; Haselbeck, F.; Joeres, R.; Kalinina, O. V.; List, M. Guiding Questions to Avoid Data Leakage in Biological Machine Learning Applications. *Nat. Methods* **2024**, *21* (8), 1444–1453.
- (10) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K. Overview of the SAMPL5 Host-Guest Challenge: Are We Doing Better? *J. Comput. Aided Mol. Des.* **2017**, *31* (1), 1–19.
- (11) Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Walters, W. P.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. D3R Grand Challenge 2: Blind Prediction of Protein-Ligand Poses, Affinity Rankings, and Relative Binding Free Energies. *J. Comput. Aided Mol. Des.* **2018**, *32* (1), 1–20.
- (12) Wagner, J. R.; Churas, C. P.; Liu, S.; Swift, R. V.; Chiu, M.; Shao, C.; Feher, V. A.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *Structure* **2019**, *27* (8), 1326–1335.e4.
- (13) Ackloo, S.; Al-Awar, R.; Amaro, R. E.; Arrowsmith, C. H.; Azevedo, H.; Batey, R. A.; Bengio, Y.; Betz, U. A. K.; Bologna, C. G.; Chodera, J. D.; Cornell, W. D.; Dunham, I.; Ecker, G. F.; Edfeldt, K.; Edwards, A. M.; Gilson, M. K.; Gordijo, C. R.; Hessler, G.; Hillisch, A.; Hogner, A.; Irwin, J. J.; Jansen, J. M.; Kuhn, D.; Leach, A. R.; Lee, A. A.; Lessel, U.; Morgan, M. R.; Moulton, J.; Muegge, I.; Oprea, T. I.; Perry, B. G.; Riley, P.; Rousseaux, S. A. L.; Saikatendu, K. S.; Santhakumar, V.; Schapira, M.; Scholten, C.; Todd, M. H.; Vedadi, M.; Volkamer, A.; Willson, T. M. CACHE (Critical Assessment of Computational Hit-Finding Experiments): A Public-Private Partnership Benchmarking Initiative to Enable the Development of Computational Methods for Hit-Finding. *Nat. Rev. Chem.* **2022**, *6* (4), 287–295.
- (14) Feynman, R. P. Cargo Cult Science. *Engineering and Science* **1974**, *37* (7), 10–13.
- (15) Garrido, A.; Lepailleur, A.; Mignani, S. M.; Dallemagne, P.; Rochais, C. HERG Toxicity Assessment: Useful Guidelines for Drug Design. *Eur. J. Med. Chem.* **2020**, *195*, 112290.
- (16) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing Safety-Related Drug Attrition: The Use of in Vitro Pharmacological Profiling. *Nat. Rev. Drug Discovery* **2012**, *11* (12), 909–922.
- (17) Denisov, I. G.; Makris, T. M.; Sligar, S. G.; Schlichting, I. Structure and Chemistry of Cytochrome P450. *Chem. Rev.* **2005**, *105* (6), 2253–2277.