

Editorial

In Silico Drug Design and Discovery: Big Data for Small Molecule Design—2nd Edition

Carmen Cerchia *  and Antonio Lavecchia 

Drug Discovery Laboratory, Department of Pharmacy, University of Naples Federico II, 80131 Naples, Italy; antonio.lavecchia@unina.it

* Correspondence: carmen.cerchia@unina.it

The volume and heterogeneity of data available to drug discovery have grown at a pace that would have been difficult to predict even a few years ago. Public repositories of bioactivity measurements, crystallographic coordinates, and gene-expression profiles now collectively catalog billions of data points [1–12], while protein structure prediction has opened entire families of targets to computational interrogation [13–17]. At the same time, the way that this data is used has changed. Computational chemistry is no longer a discipline that operates in a silo, feeding one-off predictions into a linear pipeline; instead, it functions as the connective tissue of multidisciplinary discovery campaigns in which machine learning, physics-based simulation, and experimental feedback loops are interwoven from the outset [18–22]. This shift is reshaping the expectations placed on in silico methods [23]. Scale still matters, but it is no longer sufficient on its own. What distinguishes productive computational work today is the quality of the biological question asked, the rigor of the data curation underlying the models, and the willingness to close the loop with experiment. The contributions assembled in this second edition of the *Special Issue* “In Silico Drug Design and Discovery: Big Data for Small Molecule Design” illustrate this maturation across a remarkably diverse set of therapeutic areas and methodological strategies.

A first thematic thread running through the *Special Issue* concerns the use of structural modeling and simulation to illuminate poorly characterized targets. Adewumi and Mosebi tackled estrogen receptor- $\alpha 36$, a clinically relevant but structurally unresolved variant implicated in endocrine-resistant breast cancer [24]. By constructing a homology model [25] of the ligand-binding domain and performing molecular dynamics (MD) simulations coupled with MM/PBSA free-energy calculations, they provided a structural rationale for the preferential binding of brousoflavonol B to ER- $\alpha 36$ over ER- $\alpha 66$, offering a starting point for the design of selective inhibitors. In a different biological context, Giordano et al. combined sequence analysis, phylogenetics, homology modeling, and docking to characterize lipoxygenases from marine diatoms and to propose a classification linked to possible substrate preferences [26]. This study demonstrates how bioinformatic methods can convert sequence data into testable functional hypotheses. Rehman et al. applied a more classical discovery workflow to leucyl-tRNA synthetase from *Mycobacterium tuberculosis*, combining virtual screening with MD and then validating selected hits by isothermal titration calorimetry and antimycobacterial assays [27]. Their identification of Macimorelin as a promising binder of LeuRS underscores how computational repurposing screens, when reinforced by orthogonal experimental evidence, can yield viable chemical leads against drug-resistant pathogens. Yu et al. employed the SILCS methodology and MD to propose that bile acids and their conjugates act as negative allosteric modulators of the M1 muscarinic receptor, with potential implications for colon cancer biology [28]. This work



Received: 10 April 2026

Accepted: 14 April 2026

Published: 16 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

brings together membrane-context signaling, endogenous metabolite pharmacology, and receptor allostery, three dimensions that are often studied in isolation.

A second theme is the transition from single-modality prediction toward integrated, data-rich models that capture multiple dimensions of molecular behavior [29,30]. Umemori et al. addressed a difficult problem in preclinical safety: the prediction of reactive metabolite formation associated with drug-induced liver injury [31,32], by developing a binary classifier for cysteine-trapping outcomes using proprietary experimental data [33]. Their head-to-head comparison of a message-passing neural network and a random forest model is instructive: under a realistic time-split validation, the graph-based architecture outperformed the fingerprint-based model, yet the analysis of contributory substructures retained interpretability in medicinal chemistry terms. This interplay between predictive accuracy and mechanistic transparency reflects a growing expectation in the field. Debnath et al. reported GramSeq-DTA, a framework for drug–target affinity prediction that combines structural representations of compounds and proteins with chemical perturbation signatures derived from the L1000 transcriptomic resource [34]. This study suggests that purely structure-based representations may not capture all the information relevant to affinity prediction. Piazza et al. provide a complementary illustration of how machine learning can be embedded into a practical discovery pipeline [35]. After developing and benchmarking a large panel of predictive models, they used the best performer to drive a virtual screening campaign for CDK9 inhibitors, ultimately identifying compounds with low-micromolar biochemical activity and promising cellular profiles. This work exemplifies how ML can complement rather than replace medicinal chemistry expertise, helping to narrow the candidate pool and focus experimental effort.

A distinct and increasingly important strand of contemporary computational drug design concerns generative models that learn the distribution of chemical space and propose novel molecules rather than simply scoring existing ones [36–41]. Islam and Caulfield developed a ligase-conditioned junction tree variational autoencoder to generate synthesizable, drug-like molecular glues aimed at promoting the degradation of amyloid- β 42 through the ubiquitin–proteasome system [42]. This contribution is notable because it addresses targeted protein degradation in the context of neurodegeneration and treats generative design as a conditioned problem in which target biology, ligase specificity, and synthetic realism are jointly optimized.

Taken together, the contributions in this *Special Issue* suggest several directions that are likely to define the next phase of big-data-driven drug discovery. First, multimodality is becoming essential [43–45]. The most informative models will be those that integrate chemical structure, target structure, transcriptomic perturbation, safety endpoints, and real medicinal chemistry constraints within a single framework, rather than treating each data type separately [46,47]. Second, the coupling of data-driven and physics-based reasoning is gaining ground. Several papers in this *Special Issue* already combine predictive ranking with docking, MD, calorimetry, or cellular phenotyping, and such hybrid strategies are likely to become even more important as the community confronts sparse, noisy, and biased datasets [48]. Third, the standards by which computational work is evaluated are shifting. Retrospective benchmark performance alone is no longer a convincing endpoint; what matters is whether a model can deliver biologically meaningful insight and experimentally testable hypotheses with sufficient robustness [49–51]. Alongside these methodological trends, the community would benefit from greater attention to uncertainty quantification, data provenance, negative-result sharing, and model interpretability [52–55], elements that are still treated as optional refinements.

In conclusion, this second edition of “In Silico Drug Design and Discovery: Big Data for Small Molecule Design” offers a snapshot of a field that is maturing rapidly. The

studies collected here indicate that the real promise of big data for small-molecule design is ultimately not automation, but the construction of closed-loop discovery systems in which computation, synthesis, and experiment iteratively refine one another, turning data into decisions, and decisions into better medicines.

Funding: This research received no external funding.

Acknowledgments: C.C. acknowledges MUR (Ministero dell’Università e della Ricerca), PON R&I 2014–2020 Asse IV “Istruzione e Ricerca per il recupero—REACT-EU”, Azione IV.4—“Contratti di Ricerca su tematiche dell’Innovazione” and support from the University of Naples “Federico II” (Grant FRA 2022, Line C, INITIATIVE). During the preparation of this manuscript, the authors used ChatGPT (version 5.4, OpenAI) for the purposes of English language refinement. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MD Molecular Dynamics

SILCS Site identification by ligand competitive saturation

References

1. Zdrazil, B. Fifteen Years of ChEMBL and Its Role in Cheminformatics and Drug Discovery. *J. Cheminform.* **2025**, *17*, 32. [[CrossRef](#)] [[PubMed](#)]
2. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2025 Update. *Nucleic Acids Res.* **2025**, *53*, D1516–D1525. [[CrossRef](#)] [[PubMed](#)]
3. Burley, S.K.; Bhatt, R.; Bhikadiya, C.; Bi, C.; Biester, A.; Biswas, P.; Bittrich, S.; Blaumann, S.; Brown, R.; Chao, H.; et al. Updated Resources for Exploring Experimentally-Determined PDB Structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Res.* **2025**, *53*, D564–D574. [[CrossRef](#)] [[PubMed](#)]
4. Subramanian, A.; Narayan, R.; Corsello, S.M.; Peck, D.D.; Natoli, T.E.; Lu, X.; Gould, J.; Davis, J.F.; Tubelli, A.A.; Asiedu, J.K. A next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171*, 1437–1452. [[CrossRef](#)]
5. Zhou, Y.; Zhang, Y.; Zhao, D.; Yu, X.; Shen, X.; Zhou, Y.; Wang, S.; Qiu, Y.; Chen, Y.; Zhu, F. TTD: Therapeutic Target Database Describing Target Druggability Information. *Nucleic Acids Res.* **2024**, *52*, D1465–D1477. [[CrossRef](#)]
6. Knox, C.; Wilson, M.; Klinger, C.M.; Franklin, M.; Oler, E.; Wilson, A.; Pon, A.; Cox, J.; Chin, N.E.; Strawbridge, S.A.; et al. DrugBank 6.0: The DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* **2024**, *52*, D1265–D1275. [[CrossRef](#)]
7. Malik, A.; Arsalan, M.; Moreno, C.; Mosquera, J.; Félix, E.; Kizilören, T.; Muthukrishnan, V.; Zdrazil, B.; Leach, A.R.; O’Boyle, N.M. ChEBI: Re-Engineered for a Sustainable Future. *Nucleic Acids Res.* **2026**, *54*, D1768–D1778. [[CrossRef](#)]
8. The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53*, D609–D617. [[CrossRef](#)]
9. Liu, T.; Hwang, L.; Burley, S.K.; Nitsche, C.I.; Southan, C.; Walters, W.P.; Gilson, M.K. BindingDB in 2024: A FAIR Knowledgebase of Protein-Small Molecule Binding Data. *Nucleic Acids Res.* **2025**, *53*, D1633–D1644. [[CrossRef](#)]
10. Béquignon, O.J.M.; Bongers, B.J.; Jespers, W.; IJzerman, A.P.; van der Water, B.; van Westen, G.J.P. Papyrus: A Large-Scale Curated Dataset Aimed at Bioactivity Predictions. *J. Cheminform.* **2023**, *15*, 3. [[CrossRef](#)]
11. Jiang, P.; Sinha, S.; Aldape, K.; Hannenhalli, S.; Sahinalp, C.; Rupp, E. Big Data in Basic and Translational Cancer Research. *Nat. Rev. Cancer* **2022**, *22*, 625–639. [[CrossRef](#)] [[PubMed](#)]
12. Hunter, F.M.I.; Ioannidis, H.; Bento, A.P.; Bosc, N.; Corbett, S.; Felix, E.; Magarinos, M.P.; Manners, E.; Smit, I.A.; de Veij, M.; et al. Drug and Clinical Candidate Drug Data in ChEMBL. *J. Med. Chem.* **2025**, *68*, 19800–19827. [[CrossRef](#)] [[PubMed](#)]
13. Bertoni, D.; Tsenkov, M.; Magana, P.; Nair, S.; Pidruchna, I.; Querino Lima Afonso, M.; Midlik, A.; Paramval, U.; Lawal, D.; Tanweer, A.; et al. AlphaFold Protein Structure Database 2025: A Redesignated Interface and Updated Structural Coverage. *Nucleic Acids Res.* **2026**, *54*, D358–D362. [[CrossRef](#)] [[PubMed](#)]
14. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; et al. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500. [[CrossRef](#)]
15. Kovalevskiy, O.; Mateos-Garcia, J.; Tunyasuvunakool, K. AlphaFold Two Years on: Validation and Impact. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2315002121. [[CrossRef](#)]

16. Varadi, M.; Velankar, S. The Impact of AlphaFold Protein Structure Database on the Fields of Life Sciences. *Proteomics* **2023**, *23*, 2200128. [[CrossRef](#)]
17. Percudani, R.; De Rito, C. Predicting Protein Function in the AI and Big Data Era. *Biochemistry* **2025**, *64*, 2345–2352. [[CrossRef](#)]
18. Bienstock, R.J. AI/ML Methodologies and the Future-Will They Be Successful in Designing the next Generation of New Chemical Entities? *J. Cheminform.* **2025**, *17*, 46. [[CrossRef](#)]
19. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477. [[CrossRef](#)]
20. Klambauer, G.; Hochreiter, S.; Rarey, M. Machine Learning in Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59*, 945–946. [[CrossRef](#)]
21. Chen, W.; Liu, X.; Zhang, S.; Chen, S. Artificial Intelligence for Drug Discovery: Resources, Methods, and Applications. *Mol. Ther. Nucleic Acids* **2023**, *31*, 691–702. [[CrossRef](#)] [[PubMed](#)]
22. Deng, J.; Yang, Z.; Ojima, I.; Samaras, D.; Wang, F. Artificial Intelligence in Drug Discovery: Applications and Techniques. *Brief. Bioinform.* **2022**, *23*, bbab430. [[CrossRef](#)] [[PubMed](#)]
23. Zitnik, M. AI-Enabled Drug Discovery Reaches Clinical Milestone. *Nat. Med.* **2025**, *31*, 2490–2491. [[CrossRef](#)] [[PubMed](#)]
24. Adewumi, A.T.; Mosebi, S. Characteristic Binding Landscape of Estrogen Receptor-A36 Protein Enhances Promising Cancer Drug Design. *Biomolecules* **2023**, *13*, 1798. [[CrossRef](#)]
25. Nema, P.; Purohit, A.; Soni, V.; Kashaw, S.K. A Comprehensive Review on Homology Modelling in Drug Development: Methodologies, Tools, and Evaluation. *Mini Rev. Med. Chem.* **2026**, *26*. [[CrossRef](#)]
26. Giordano, D.; Bonora, S.; D’Orsi, I.; D’Alelio, D.; Facchiano, A. Structural and Functional Characterization of Lipxygenases from Diatoms by Bioinformatics and Modelling Studies. *Biomolecules* **2024**, *14*, 276. [[CrossRef](#)]
27. Rehman, Z.U.; Najmi, A.; Zoghebi, K. Insights into the Effects of Ligand Binding on Leucyl-tRNA Synthetase Inhibitors for Tuberculosis: In Silico Analysis and Isothermal Titration Calorimetry Validation. *Biomolecules* **2024**, *14*, 711. [[CrossRef](#)]
28. Yu, W.; MacKerell, A.D.; Weber, D.J.; Raufman, J.-P. Bile Acids Are Potential Negative Allosteric Modulators of M1 Muscarinic Receptors. *Biomolecules* **2025**, *15*, 1326. [[CrossRef](#)]
29. Angelo, J.S.; Guedes, I.A.; Barbosa, H.J.C.; Dardenne, L.E. Multi-and Many-Objective Optimization: Present and Future in de Novo Drug Design. *Front. Chem.* **2023**, *11*, 1288626. [[CrossRef](#)]
30. Luukkonen, S.; van den Maagdenberg, H.W.; Emmerich, M.T.M.; van Westen, G.J.P. Artificial Intelligence in Multi-Objective Drug Design. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102537. [[CrossRef](#)]
31. Park, B.K.; Boobis, A.; Clarke, S.; Goldring, C.E.P.; Jones, D.; Kenna, J.G.; Lambert, C.; Laverty, H.G.; Naisbitt, D.J.; Nelson, S.; et al. Managing the Challenge of Chemically Reactive Metabolites in Drug Development. *Nat. Rev. Drug Discov.* **2011**, *10*, 292–306. [[CrossRef](#)] [[PubMed](#)]
32. Ray, A.S. Practical Perspectives on Addressing Hepatotoxicity during Clinical Candidate Selection. *Expert Opin. Drug Metab. Toxicol.* **2025**, *21*, 751–753. [[CrossRef](#)] [[PubMed](#)]
33. Umemori, Y.; Handa, K.; Yoshimura, S.; Kageyama, M.; Iijima, T. Development of a Novel In Silico Classification Model to Assess Reactive Metabolite Formation in the Cysteine Trapping Assay and Investigation of Important Substructures. *Biomolecules* **2024**, *14*, 535. [[CrossRef](#)] [[PubMed](#)]
34. Debnath, K.; Rana, P.; Ghosh, P. GramSeq-DTA: A Grammar-Based Drug–Target Affinity Prediction Approach Fusing Gene Expression Information. *Biomolecules* **2025**, *15*, 405. [[CrossRef](#)]
35. Piazza, L.; Poles, C.; Bononi, G.; Granchi, C.; Di Stefano, M.; Poli, G.; Giordano, A.; Medugno, A.; Napolitano, G.M.; Tuccinardi, T.; et al. Machine Learning-Based Virtual Screening for the Identification of Novel CDK-9 Inhibitors. *Biomolecules* **2026**, *16*, 12. [[CrossRef](#)]
36. Cerchia, C.; Delre, P.; Falbo, E.; Lavecchia, A. Artificial Intelligence in De Novo Drug Design. In *Applied Artificial Intelligence for Drug Discovery: From Data-Driven Insights to Therapeutic Innovation*; Lavecchia, A., Ed.; Springer Nature: Cham, Switzerland, 2026; pp. 157–186, ISBN 978-3-031-98022-0.
37. Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K.F. Generative Models for Molecular Discovery: Recent Advances and Challenges. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, e1608. [[CrossRef](#)]
38. Cheng, Y.; Gong, Y.; Liu, Y.; Song, B.; Zou, Q. Molecular Design in Drug Discovery: A Comprehensive Review of Deep Generative Models. *Brief. Bioinform.* **2021**, *22*, bbab344. [[CrossRef](#)]
39. Tong, X.; Liu, X.; Tan, X.; Li, X.; Jiang, J.; Xiong, Z.; Xu, T.; Jiang, H.; Qiao, N.; Zheng, M. Generative Models for de Novo Drug Design. *J. Med. Chem.* **2021**, *64*, 14011–14027. [[CrossRef](#)]
40. Bajorath, J. Chemical Language Models for Molecular Design. *Mol. Inform.* **2024**, *43*, e202300288. [[CrossRef](#)]
41. Bhattacharya, D.; Cassady, H.J.; Hickner, M.A.; Reinhart, W.F. Large Language Models as Molecular Design Engines. *J. Chem. Inf. Model.* **2024**, *64*, 7086–7096. [[CrossRef](#)]
42. Islam, N.N.; Caulfield, T.R. Conditioned Generative Modeling of Molecular Glues: A Realistic AI Approach for Synthesizable Drug-like Molecules. *Biomolecules* **2025**, *15*, 849. [[CrossRef](#)] [[PubMed](#)]

43. Chakraborty, C.; Bhattacharya, M.; Pal, S.; Chatterjee, S.; Das, A.; Lee, S.-S. AI-Enabled Language Models (LMs) to Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) in Drug Discovery and Development. *J. Adv. Res.* **2025**, *78*, 377–389. [[CrossRef](#)] [[PubMed](#)]
44. Luo, Y.; Liu, X.Y.; Yang, K.; Huang, K.; Hong, M.; Zhang, J.; Wu, Y.; Nie, Z. Toward Unified Ai Drug Discovery with Multimodal Knowledge. *Health Data Sci.* **2024**, *4*, 0113. [[CrossRef](#)] [[PubMed](#)]
45. Elemento, O.; Leslie, C.; Lundin, J.; Tourassi, G. Artificial Intelligence in Cancer Research, Diagnosis and Therapy. *Nat. Rev. Cancer* **2021**, *21*, 747–752. [[CrossRef](#)]
46. Huynh, D.L.; Seal, S.; Reid, D.; Carpenter, A.E.; Bender, A.; Spjuth, O.; Seal, S.; Huynh, D.L.; Chelbi, M.; Patra, A.; et al. AI Agents in Drug Discovery: Applications and Case Studies. *Drug Discov. Today* **2026**, 104650. [[CrossRef](#)]
47. Markossian, S.; Baskir, H.M.; Grossman, A.C.; Chandran, S.; Rele, S.; Hall, M.D.; Amaro, R.E.; Tropsha, A.; Zakharov, A.V. In Silico Drug Discovery: Bridging the Gaps in Preclinical Translation. *Drug Discov. Today* **2026**, *31*, 104567. [[CrossRef](#)]
48. Fromer, J.C.; Coley, C.W. Computer-Aided Multi-Objective Optimization in Small Molecule Discovery. *Patterns* **2023**, *4*, 100678. [[CrossRef](#)]
49. Pitt, W.R.; Bentley, J.; Boldron, C.; Colliandre, L.; Esposito, C.; Frush, E.H.; Kopec, J.; Labouille, S.; Meneyrol, J.; Pardoe, D.A.; et al. Real-World Applications and Experiences of AI/ML Deployment for Drug Discovery. *J. Med. Chem.* **2025**, *68*, 851–859. [[CrossRef](#)]
50. Wang, Y.; Hu, Z.; Chang, J.; Yu, B. Thinking on the Use of Artificial Intelligence in Drug Discovery. *J. Med. Chem.* **2025**, *68*, 4996–4999. [[CrossRef](#)]
51. Doron, G.; Genway, S.; Roberts, M.; Jasti, S. Generative AI: Driving Productivity and Scientific Breakthroughs in Pharmaceutical R&D. *Drug Discov. Today* **2025**, *30*, 104272. [[CrossRef](#)]
52. Gangwal, A.; Lavecchia, A. IMPACT Framework: Establishing Global Standards for Artificial Intelligence Implementation, Methodology, and Translation in Drug Discovery. *WIREs Comput. Mol. Sci.* **2026**, *16*, e70072. [[CrossRef](#)]
53. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584. [[CrossRef](#)]
54. Lamens, A.; Bajorath, J. Explainable Artificial Intelligence for Molecular Design in Pharmaceutical Research. *Chem. Sci.* **2026**, *17*, 1411–1422. [[CrossRef](#)]
55. Bajorath, J. From Explainable Artificial Intelligence to Human Understanding. *Artif. Intell. Life Sci.* **2025**, *7*, 100131. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.