Data-Centric Path to Autonomous Chemical Discovery

FLORIAN BOSER,[a] FRANK GLORIUS[a*] and ANDREW I. COOPER[b*]

[a]*Organisch-Chemisches Institut, Universität Münster, Corrensstraße 40, Münster 48149, Germany*

[b]*Materials Innovation Factory and Department of Chemistry and Leverhulme Research Centre for Functional Materials Design, University of Liverpool, Liverpool L69 7ZD, United Kingdom*

For many years, chemistry was advanced by knowledge-driven experimental practice, theory, and serendipitous findings, placing the chemist in a role more akin to an artisan.[1,2] This model has been hugely successful, supplying the world with important inventions such as drugs and plastics. However, the field of chemistry now faces two interlinked and foundational challenges: a reproducibility crisis that undermines trust in published results, and the high-dimensionality of chemical (reaction) space, which vastly exceeds the capacity of established, manual exploration.[3–6] The latter issue is not new—there has always been a huge gulf between the size of chemical space and our ability to explore it—but global challenges such as resource scarcity, increased energy use, and population growth have created an urgency to accelerate our rate of innovation in chemistry.

At the heart of these challenges lies an 'information gap'—a divide between the fragmentary, selectively reported results found in the literature and the comprehensive, structured data required for practical and generalisable predictive models and reliable knowledge transfer.[7–9] The roots of the information gap are both systemic and cultural. Reporting and selection biases favour the publication of successful, "positive" outcomes while obscuring failed or negative results, leading to a distorted representation of chemical reactivity and an overestimation of reliability. As a result, existing datasets usually lack the comprehensive, unbiased data required for generalisable, predictive Machine Learning (ML) models.[3,7–9]

Self-driving laboratories (SDLs) could address this information gap, if adopted widely enough, and could constitute a paradigm shift in chemical research. SDLs seek to close the loop of the scientific method—autonomously generating hypotheses, designing experiments, executing them with robotic precision, and analysing outcomes through fully integrated ML workflows. By integrating every stage of the experimental pipeline within modular, data-centric platforms, SDLs can produce large-scale, high-quality, machine-readable datasets—including vital metadata and negative results. In doing so, they have the potential to overcome the physical and operational disconnects that impede reproducibility, slow the pace of innovation, and contribute to inefficient resource use.[4,5,10–13] Nevertheless, their full transformative potential is contingent upon adherence to FAIR data principles—Findable, Accessible, Interoperable, and Reusable.[7,14,15]

If SDLs become a standard mode of operation for chemistry research, then this might fundamentally change the human role within the research cycle, liberating researchers

from repetitive, physically demanding and potentially toxic and hazardous manual procedures, opening the opportunity to focus on higher-level creative, strategic and interdisciplinary challenges, such as complex hypothesis formulation, integrative data analysis, and campaign design. There are already fully agentic Artificial Intelligence (AI) systems that can support the entire discovery process, integrating literature search, hypothesis generation, experimental design and data analysis, representing the emergence of AI co-scientists.[16,17] While deep scientific intuition remains indispensable, the day-to-day execution of the scientific method is increasingly delegated to autonomous agents and robots, which accelerates discovery, enhances sustainability and enables efficient and profound exploration of chemical space.[10,11,18,19]

The Glorius Group pursues a comprehensive strategy for driving the Labs of the Future by constructing a holistic, data-centric chemical ecosystem: from automated high-quality data generation and informative molecular representation (featurisation), through the application of ML models to accelerate discovery, and towards reconstruction of missing information from biased historical data.

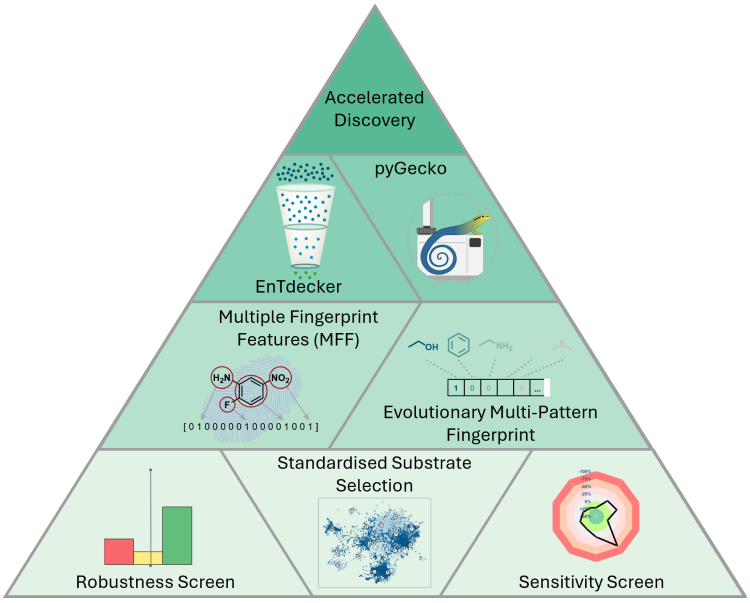

Fig. 1. Recent contributions of the Glorius Group to the data-driven Lab of the Future. The foundational layer is based on systematic screening tools and standardisation, and is designed to generate high-quality, information-rich experimental data.[20–22] This information is preserved within powerful and interpretable molecular representations.[23,24] At the apex, high-quality experimental data from HTE and EnTdecker's ML predictions lead to accelerated discovery.[25–27]

Our early work focused on enriching the informativeness and reproducibility of chemical data through systematic tools for reaction evaluation: The Robustness Screen, for example, employs potentially interfering, fragment-based probes to systematically map a reaction's functional group tolerance, identifying potential reaction inhibitors and decreased product formation without the need for extensive substrate scopes.[20] Similarly, the Sensitivity Screen evaluates how sensitive a reaction is to minor, unintentional deviations from its optimal published conditions, directly probing a

common source of irreproducibility.[22,28] Building on this principle, our work on Standardised Substrate Selection leverages unsupervised ML to guide the experimental selection of substrates for a newly discovered reaction. This approach—publicly available as a web application—is designed to eliminate human selection bias and achieve a representative, objective coverage of chemical space.[21,29]

To accelerate the generation of standardised, high-quality data, we developed a workflow for quantitative, calibration-free High-throughput Experimentation (HTE). Leveraging our open-source pyGecko library—enabling rapid and automated processing of GC-MS and GC-FID data—and a commercial microreactor, we eliminate the laborious bottleneck of creating product-specific calibration standards. This enables the rapid parallel analysis and visualisation of 96 reactions in minutes, generating standardised, high-quality information that is highly sought after for the training of predictive ML models.[25]

Complementing data generation, our work in molecular representation seeks to create descriptors that are both powerful and intuitive. We first developed Multiple Fingerprint Features (MFF), a versatile descriptor based on the simple concatenation of diverse, structure-based fingerprints, and more recently, the Evolutionary Multi-Pattern Fingerprint (EvoMPF), which uses an evolutionary algorithm to autonomously generate dense, problem-specific, and highly interpretable representations directly from the data.[23,24]

The true power of this data-centric ecosystem is realised when high-quality datasets and informative representations are deployed in predictive models that directly guide experimental campaigns. A prime example is our EnTdecker platform, a ML tool—publicly available as a web application—that provides chemists with near-instantaneous predictions of excited-state properties crucial for energy transfer catalysis.[26,30] This predictive engine is integrated into a data-driven, three-layer screening strategy, which synergistically combines *in silico* substrate mapping, luminescence quenching, and high-throughput reaction screening to dramatically accelerate the discovery of novel dearomative cycloadditions.[27] Such integrated digital-physical workflows, where predictive models serve as the cognitive core for experimental design, exemplify how data-driven chemistry is evolving to become the intelligent 'brain' for the autonomous laboratories of the future.

The Cooper Group has work in automated or “high-throughput” chemistry since the mid-2000’s,[31–33] but more recently our focus has shifted toward computational design of materials,[34,35] autonomous ‘mobile robotic chemists’,[10,11] and human-in-the-loop[36] and machine reasoning methods,[37] working toward a concept of ‘hybrid intelligence’ for chemistry laboratories (Fig. 2). Our long-term goal is to unlock the power of automation and robotics by building a hybrid reasoning platform that can avoid the decision-making bottlenecks that are otherwise imposed by the mismatch between automated experiments and human working patterns (Fig. 2a). We see mobile robots (Fig. 2b) as an enabling technology here because they are inherently modular, and they can carry out experiments in any order, like a human researcher.[10,11] We also believe that Large Language Models (LLM), notwithstanding their clear limitations, will play an important role in this hybrid intelligence paradigm. We have already demonstrated[37] that LLMs can greatly enhance the performance of searches in high-dimensional chemical space, for example for catalysis problems (Fig. 2c).

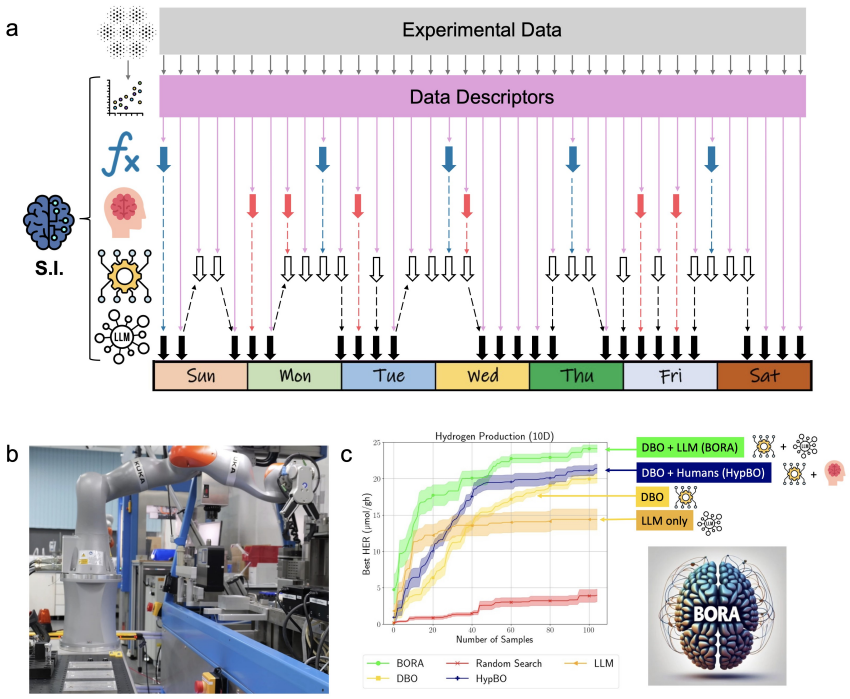

Fig. 2. a, Hybrid intelligence for autonomous laboratories. A supervisory intelligence (S.I.) coordinates (from top to bottom) experimental data, physics-based models (f_x),[34,35] human reasoning,[36] algorithmic searches, and machine reasoning,[37] leveraging the advantages and mitigating the weaknesses of these various inputs. b, Mobile robots[10,11] are an enabling technology for modular, flexible laboratories where conditional decisions are made using hybrid intelligence. c, 'LLM-in-the-loop' reasoning—BORA, a hybrid of LLMs and Bayesian optimisation (BO)[37] can outperform human-in-the-loop (HypBO), BO only (DBO) and LLM-only search strategies, as illustrated here for a 10-dimensional photocatalysis problem. A key advantage is that LLM reasoning, unlike human reasoning, can be deployed as frequently as needed in a closed-loop experiment (a).

Looking to the future, the field is advancing toward universal, human-compatible robotics and generalist, reasoning-capable AI. The prospect of modular robots, able to seamlessly operate existing laboratory infrastructure, foreshadows the liberation of automation from bespoke, platform-specific constraints. In parallel, 'foundation' models and agentic AI—endowed with broad chemical intuition and reasoning capacities—are positioned to become the cognitive core of the autonomous laboratory, coordinating complex workflows, adapting dynamically, and integrating virtual and physical experimentation.[16,19,38,39]

The synergy of modular robotics and agentic AI promises a virtuous cycle: improved automation generates richer data, which empowers better AI, in turn enabling more ambitious automation and accelerating discovery. However, this vision depends fundamentally on the universal, rigorous application of FAIR data standards.[15] Only with high-quality, machine-readable, openly shared data can the potential of automation

and AI be fully realised, thus closing the information gap, ensuring reproducibility, and ushering in a new era of collaborative, autonomous discovery.

Acknowledgments

Generous financial support from the ERC Advanced Grant (Agreement No. 101098156, HighEnT) and the Deutsche Forschungsgemeinschaft (Priority Program SPP 2363, "Molecular Machine Learning") is gratefully acknowledged. We also acknowledge funding from the AI for Chemistry: AIchemy hub (EPSRC grants EP/Y028775/1 and EP/Y028759/1), the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design, and the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 856405). AIC thanks the Royal Society for a Research Professorship (RSRP\S2\232003).

References

1. A. McNally, C. K. Prier, D. W. C. MacMillan, *Science* **334**, 1114 (2011).
2. A. Y. Rulev, *New J. Chem.* **41**, 4262 (2017).
3. F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen, *et al.*, *Angew. Chem. Int. Ed.* **61**, e202204647 (2022).
4. G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, *et al.*, *Chem. Rev.* **124**, 9633 (2024).
5. M. Abolhasani, E. Kumacheva, *Nat. Synth.* **2**, 483 (2023).
6. O. Bayley, E. Savino, A. Slattery, T. Noël, *Matter* **7**, 2382 (2024).
7. M. L. Schrader, F. R. Schäfer, F. Schäfers, F. Glorius, *Nat. Chem.* **16**, 491 (2024).
8. N. S. Eyke, B. A. Koscher, K. F. Jensen, *Trends Chem.* **3**, 120 (2021).
9. R. Mercado, S. M. Kearnes, C. W. Coley, *J. Chem. Inf. Model.* **63**, 4253 (2023).
10. B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, *et al.*, *Nature* **583**, 237 (2020).
11. T. Dai, S. Vijayakrishnan, F. T. Szczypiński, J.-F. Ayme, E. Simaei, *et al.*, *Nature* **635**, 890 (2024).
12. Y. Shi, P. L. Prieto, T. Zepel, S. Grunert, J. E. Hein, *Acc. Chem. Res.* **54**, 546 (2021).
13. M. Seifrid, R. Pollice, A. Aguilar-Granda, Z. Morgan Chan, K. Hotta, *et al.*, *Acc. Chem. Res.* **55**, 2454 (2022).
14. M. A. Butakova, A. V. Chernov, O. O. Kartashov, A. V. Soldatov, *Nanomaterials* **12**, 12 (2022).
15. M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, *et al.*, *Sci. Data* **3**, 160018 (2016).
16. A. E. Ghareeb, B. Chang, L. Mitchener, A. Yiu, C. J. Szostkiewicz, *et al.*, arXiv, ID: 2505.13400, (2025).
17. M. D. Skarlinski, S. Cox, J. M. Laurent, J. D. Braza, M. Hinks, *et al.*, arXiv, ID: 2409.13740, (2024).
18. F. Häse, L. M. Roch, A. Aspuru-Guzik, *Trends Chem.* **1**, 282 (2019).
19. H. Hysmith, E. Foadian, S. P. Padhy, S. V. Kalinin, R. G. Moore, *et al.*, ChemRxiv, doi: 10.26434/chemrxiv-2024-3xq9z, (2024).

20. K. D. Collins, F. Glorius, *Nat. Chem.* **5**, 597 (2013).
21. D. Rana, P. M. Pflüger, N. P. Hölter, G. Tan, F. Glorius, *ACS Cent. Sci.* **10**, 899 (2024).
22. L. Pitzer, F. Schäfers, F. Glorius, *Angew. Chem. Int. Ed.* **58**, 8572 (2019).
23. P. M. Pflüger, M. Kühnemund, F. Katzenburg, H. Kuchen, F. Glorius, *Chem* **10**, 1391 (2024).
24. F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, *Chem* **6**, 1379 (2020).
25. F. Katzenburg, F. Boser, F. R. Schäfer, P. M. Pflüger, F. Glorius, *Digit. Discov.* **4**, 384 (2025).
26. L. Schlosser, D. Rana, P. Pflüger, F. Katzenburg, F. Glorius, *J. Am. Chem. Soc.* **146**, 13266 (2024).
27. D. Rana, C. Hümpel, R. Laskar, L. Schlosser, S. Korgitzsch, *et al.*, *J. Am. Chem. Soc.* **147**, 28359 (2025).
28. F. Schäfer, L. Lückemeier, F. Glorius, *Chem. Sci.* **15**, 14548 (2024).
29. *Substrate Selection - Evaluating Pharmaceutical Scope Relevance*. https://pharmascope.uni-muenster.de/ (accessed August 4, 2025).
30. *EnTdecker - Accelerating Substrate Discovery for EnT Catalysis*. https://entdecker.wwu.de/ (accessed August 4, 2025).
31. C. L. Bray, B. Tan, C. D. Wood, A. I. Cooper, *J. Mater. Chem.* **15**, 456 (2005).
32. S. Jana, S. P. Rannard, A. I. Cooper, *Chem. Commun.* **28**, 2962 (2007).
33. C. L. Bray, B. Tan, S. Higgins, A. I. Cooper, *Macromolecules* **43**, 9426 (2010).
34. A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, *et al.*, *Nature* **543**, 657 (2017).
35. M. O'Shaughnessy, J. Glover, R. Hafizi, M. Barhi, R. Clowes, *et al.*, *Nature* **630**, 102 (2024).
36. A. Cissé, X. Evangelopoulos, S. Carruthers, V. V. Gusev, A. I. Cooper, *Proc. 33rd Int. Joint Conf. Artificial Intel. (IJCAI-24)*, 3881–3889, (2024).
37. A. Cissé, X. Evangelopoulos, V. V. Gusev, A. I. Cooper, *Proc. 34th Int. Joint Conf. Artificial Intel. (IJCAI-25)*, in press, (2025).
38. Q. Zhu, F. Zhang, Y. Huang, H. Xiao, L. Zhao, *et al.*, *Natl. Sci. Rev.* **9**, nwac190 (2022).
39. S. Gao, A. Fang, Y. Huang, V. Giunchiglia, A. Noori, *et al.*, *Cell* **187**, 6125 (2024).