FROM FAILED EXPERIMENTS TO MACHINE-ACTIONABLE KNOWLEDGE

Berend Smit

[*] *Laboratory of Molecular Simulation (LSMO), École Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland*

My view of the present state of research on The Lab of the Future: AI, Robotics, Automation

For most chemists, the way we think about data is deeply rooted in a tradition shaped by the history of scientific publishing. We were trained to believe that the goal of a publication is to communicate the *essential* results, that is, to focus on the successful experiments that lead to clear conclusions. Reporting failed or partially successful attempts would only confuse readers and obscure the key findings. The practical realities of publishing reinforced this mindset: journals used to charge authors per printed page, so including extra data literally cost more money. Moreover, articles were physically printed and submitted by mail, and sometimes even by *surface* mail. As Prof. Ben Feringa recounted, his university refused to pay for airmail postage, so his first manuscript made its way to the journal by sea. In this context, sharing only the essential results was not just a scientific convention but also an economic and logistical necessity.

Not only were there practical barriers—such as the impossibility of publishing all our data in a single article—but we also lacked the tools to handle such a large volume of information. Even if we had attempted to document every experiment, managing or interpreting the resulting data overload would have been unmanageable. We relied entirely on human expertise: the experienced chemist who could recognize patterns, filter out noise, and publish what truly mattered. We were, quite rightly, very grateful to these experts, whose intuition and judgment transformed the chaos of experimental data into a clear scientific article.

What has changed is that the very thing that once overwhelmed us—data—has become the fuel that drives progress. Where humans quickly reach their cognitive limits when faced with too much information, AI systems thrive. The more data they are given, the better they become at recognizing patterns, identifying correlations, and making predictions. In fact, for AI, a little data is often worse than no data at all; what it truly needs is massive, diverse, and well-structured datasets to learn effectively. This marks a fundamental shift in how we think about scientific knowledge. What was once seen as redundant, confusing, or unpublishable information—failed experiments, intermediate results, and incomplete series—has now become invaluable training material for AI, enabling it to extract insights that no human could discern alone.

It is not only about publishing raw data in a machine-readable format but also about adhering to well-defined and agreed standards. To illustrate this point, let's consider a typical workflow: a chemist synthesizes a new metal-organic framework (MOF) and measures an adsorption isotherm. The isotherm data are saved in an Excel sheet and uploaded as supplementary information, while the underlying crystal structure is deposited in the Cambridge Structural Database (CSD).[1] The National Institute of Standards and Technology (NIST) maintains a database of isotherms (the NIST-ISODB[2]), aiming to create a vast resource of adsorption data. NIST finds our isotherm relevant, and adds its to their database.

We were interested in using the NIST database to validate a large-scale benchmark molecular simulations study that predicts isotherms of different gases in MOFs from crystal structures. This, however, required us to link an isotherm from the NIST database to the corresponding crystal structure. As there are over 4,000 isotherms, we aimed to do this without reading the corresponding articles. Ongari et al[3] attempted to address this challenge by developing an automated method to connect adsorption isotherms from the NIST-ISODB with corresponding crystal structures in the CSD. They showed that inconsistent naming conventions ('Cu-BTC' vs. 'HKUST-1'), missing metal identities, and unstandardized metadata hinder this connection. This effort revealed a critical bottleneck: data without standardized metadata (see figure 1).

Ideally, the Excel file with adsorption data and its metadata should match the data deposited in the CSD. Since this metadata is often missing, NIST had to add it manually, for example, by using the caption of the corresponding isotherm. Frequently, the name used in the caption was an abbreviation (e.g., MOF1, MOF2, MOFA, MOFB, or any other abbreviation) to avoid the lengthy names in the CSD.

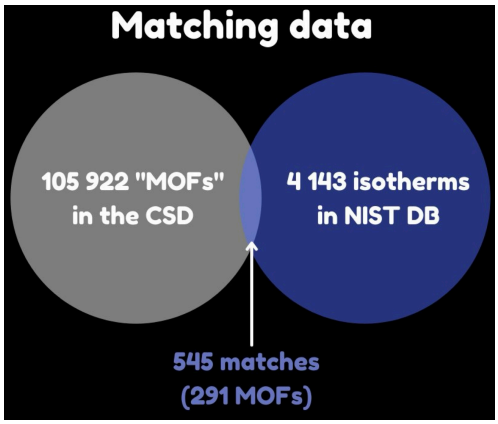

Figure 1: The number of matches of isotherms with their corresponding crystal structures by Ongari et al that could be obtained by just looking for the name reported of the MOF for which the isotherm was measured in the Cambridge Structural Database.

My recent research contributions to The Lab of the Future: AI, Robotics, Automation

Chemists have always learned from failure. Each unsuccessful synthesis, every amorphous powder or unexpected phase informs our intuition about what might work next. Yet, this tacit knowledge rarely enters the scientific record.[4] The study by Moosavi et al.[5] demonstrates how such 'chemical intuition' can be quantified in the case of synthesizing metal-organic frameworks (MOFs). As an example, Moosavi et al. repeated the synthesis of HKUST-1.[6]

HKUST-1 has been synthesized HKUST-1, and these groups report very different BET surface areas. Moosavi et al. used robotic synthesis to identify the conditions that produce the highest surface area. By reconstructing both failed and partially successful experiments and applying machine learning, they revealed which experimental variables

influenced the success of the synthesis. The machine learning model enabled us to quantify the importance of these variables, a factor that is usually not documented but enhances chemical intuition. This work marks a fundamental shift: instead of discarding failed experiments, they become a rich source of information that teaches us how to navigate the complex, multidimensional synthesis space of metal–organic frameworks (MOFs).

Outlook to future developments of research on The Lab of the Future: AI, Robotics, Automation

As argued by Jablonka et al.,[7] data management in chemistry should not be an afterthought but an intrinsic part of the scientific process. FAIR principles (findable, accessible, interoperable, and reusable) must be extended to make data machine-actionable. To achieve this, data collection, processing, and publication should be seamlessly integrated through electronic lab notebooks (ELNs) that automatically annotate and standardize data and their metadata directly from instruments. Such ELNs would serve as the central hub of the research process, ensuring that all data are stored in structured, open formats (e.g., JSON-LD or JCAMP-DX) and linked to controlled vocabularies and ontologies. Only then can we avoid the situation that, after 99 failed and partially successful experiments, we finally got our desired product; we need to spend 99% of our time documenting all our failures. Only then can we build an ecosystem in which computational tools autonomously understand and reuse experimental data.

Acknowledgments

The Swiss Science Foundation supports this work through the SNSF's Project Funding Scheme (214872) and Advanced Grant (216165). This work is part of the USorb-DAC Project, which is supported by a grant from The Grantham Foundation for the Protection of the Environment to RMI's climate tech accelerator program, Third Derivative, which provided additional support.

References

[1] The Cambridge Crystallographic Data Centre (CCDC) (2021) https://www.ccdc.cam.ac.uk/

[2] D. W. Siderius and V. K. Shen *NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials* (2016) https://adsorption.nist.gov/isodb/index.php#home

[3] D. Ongari, L. Talirz, K. M. Jablonka, D. W. Siderius, and B. Smit, *Data-Driven Matching of Experimental Crystal Structures and Gas Adsorption Isotherms of Metal-Organic Frameworks* J. Chem. Eng. Data **67** (7), 1743 (2022) http://dx.doi.org/10.1021/acs.jced.1c00958

[4] R. Brazil, *Illuminating'the ugly side of science': fresh incentives for reporting negative results* (2024) http://dx.doi.org/10.1038/d41586-024-01389-7

[5] S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou, and B. Smit, *Capturing chemical intuition in synthesis of metal-organic frameworks* Nat. Commun. **10** (1), 539 (2019) http://dx.doi.org/10.1038/s41467-019-08483-9

[6] S. S.-Y. Chui, S. M.-F. Lo, J. P. H. Charmant, A. G. Orpen, and I. D. Williams, *A Chemically Functionalizable Nanoporous Material* Science **283**, 1148 (1999) http://dx.doi.org/10.1126/science.283.5405.1148

[7] K. M. Jablonka, L. Patiny, and B. Smit, *Making the collective knowledge of chemistry open and machine actionable* Nat Chem **14** (4), 365 (2022) http://dx.doi.org/10.1038/s41557-022-00910-7