The role of computational results databases in accelerating the discovery of catalysts

Databases of computational results hold high promise for accelerating catalysis research. Still, many challenges remain and consensus on facets such as metadata, reliability and curation is crucial to transform the hype into an attractive technology.

Carles Bo, Feliu Maseras and Núria López

Science has the power to generate great volumes of data. Over the years, numerous collective efforts, including those made by the Manhattan Project, CERN, the Genomics Consortium and large research facilities in astronomy and climate, have had to collect, label, store and curate scientific digital outputs. Chemistry and materials science are different. Highly standardized thermodynamic data have been curated by organizations such as the National Institute of Standards and Technology (NIST) through Chemistry Handbooks[1], but probably the most successful stories in chemistry-related data archiving are the neat curation of a protein database[2] and the maintenance of crystallographic records for molecules[3] and materials[4]. In addition, private efforts made by chemical companies ensure that they can trace experiments performed more than half a century ago.

Computational techniques employed in the study of the atomistic processes occurring at the sub-nanometre scale in chemistry, biochemistry, physics and materials science have been based on solving the Schrödinger equation in different ways. It has been calculated that about 30% of the total use of supercomputers at the European level are devoted to various kinds of density functional theory (DFT) approximations. Now, the robustness of the implementations in the different quantum chemistry codes ensures that the data obtained are of the same quality irrespective of the computational codes, provided that high standards are used[5]. Multipurpose, Dropbox-like initiatives such as Zenodo, Figshare and Dryad allow the allocation of space for storage of scientific data. However, the massive simulations generate high volumes of data that are neither tagged nor syndicated, thus the efforts of computation are partially in vain.

Several initiatives have been developed to systematically collect information from calculations and build databases that can be easily mined and can serve as the first step in artificial intelligence models. One of the earliest examples is the Quantum Chemistry Literature Database, which collected data from published manuscripts[6]. This evolved into a website[7], but updates seem to have been discontinued in 2013. Some reference databases such as the Computational Chemistry Comparison and Benchmark Database by NIST[8], the Benchmark Energy and Geometry Database[9] or the Minnesota Databases[10] boosted the development of new DFT functionals, solvation models and empirical dispersion corrections. Others focus on specific issues, such as the Alexandria Library, dedicated to force field development[11] and the PubChemQC project, which took data encoded in the PubChem database and calculated the first 10 excited states for over 2 million molecules[12]. The software generates different kinds of inputs for molecular codes using Open Babel[13]. For instance, it can search for molecules where the HOMO–LUMO gap is smaller than a certain value (for example, 1.0 eV). Regarding structured data, the pioneering definition of the Chemical Markup Language[14] was followed by The Quixote Project[15], which put forward an ambitious programme for collaborative and open quantum chemistry data management. In turn, CatApp[16] was among the first attempts to recycle computational data in the field of heterogeneous catalysis.

Nowadays, stand-alone data extractors, such as ExcelAutomat[17], EsiGen[18] and the cclib library[19] are available. They search in output files for patterns, which are then parsed, summarized, uploaded to spreadsheets and/or assigned to variables for molecular codes. In addition, data repositories and specialized software platforms are currently under development. But it is in materials science where the developments have arguably been the most exciting. For instance the US-driven Materials Genome Initiative[20] has a specific Materials Project[21], which aims at employing supercomputers and the most advanced electronic structure analyses to provide an open web-based access to all the structures already computed (including those not yet synthetized), and the resources needed for the design of novel materials[22]. Important efforts have also been made at the European level through the European Materials Modelling Council[23] to achieve transferability of data and codes through the Modelling Data conceptual frameworks[24]. Although the core of the work has been related to transferability in multiscale modelling (data inheritance between scales), some excellent work on ontologies has been developed. In Europe, two initiatives have been working in parallel as centres of excellence (2014–2017), namely the Novel Materials Discovery (NoMaD) repository[25] and the Automated Interactive Infrastructure and Database (AiiDA)[26,27]. NoMaD creates, collects, stores and cleanses computational materials science data, and develops tools for data mining to find structures, correlations and novel results that do not appear within smaller datasets. The system is decentralized and can incorporate results from different sources. Particular areas of interest are heat-transport tensors for many materials, catalytic activation of CO_2 and thin coating films to protect novel hybrid perovskite solar cells from degradation in moist environments through high-throughput screening of potential transparent oxide semiconductors. In turn, AiiDA is a flexible and scalable infrastructure that allows the management, preservation and dissemination of computational results, but also works on the data ensuring its searchability and providing workflows. The core of their set-up is driven by automation, data, environment and sharing, adopting concepts and tools from computer science. Particularly, the Materials Cloud[28] has been developed to improve the viability of industrially adapted databases. However, the changes in the European policies with the new Virtual Materials

Market Place strategy might affect these initiatives. The Computational Materials Repository[29] is structured in different projects — for instance two-dimensional materials — and extracts the data through the combined use of the Atomic Simulation Environment (ASE) and Python scripts. In turn, ioChem-BD (Input/Output Chemistry Big Data)[30,31] focuses on the chemical properties and links data generation and open-access publication. In addition to providing tools for data curation and post-processing, ioChem-BD builds up a distributed network of independent nodes around a central server where data and metadata are indexed. In any of these databases, the transferability of the information between the molecular and periodic approaches is a challenge, particularly when it comes to addressing chemical and catalytic problems where the recognition of the structural and electronic patterns that build the active site is a major goal. Merging the best properties of homogeneous and heterogeneous catalysts, for instance in the area of single-atom catalysis[32], requires this pattern recognition. However, the representation of crystalline materials in a format that allows the comparison to their molecular counterparts is a major hurdle[33].

While the path is well traced, efforts are still needed to give the proper semantics to data[34] in order to allow transferability between the different fields where first-principles results can be relevant. Among the challenges ahead, it is mandatory to transform data available in plain websites into structured data, accompanied by valuable metadata, which follows the findable, accessible, interoperable, recyclable (FAIR) principles. Code interoperability and standard data formats are motivating international research actions[35,36], such as the Molecular Sciences Software Institute[37] in the US and the European Materials Modelling Consortium[24], but much effort worldwide is still needed. This will allow the integration in multiscale methodologies that can go from the atomistic perspective to the device level. The integration with experimental results is yet another long-term challenge. Finally, the combination of databases with machine learning is having a big impact on the field by allowing an increasing number of applications[33,38–42] and is seen as a major opportunity in industry[43]. However, this blooming computational field also demands massive curated data and robust algorithm benchmarks[44], as there is a risk of the hype masking the real advantages of these powerful tools[45,46].

All of these advances will ensure that artificial intelligence technologies based on computational or mixed computational/experimental databases will accelerate the discovery of active, selective and stable catalysts that, for instance, are able to break the standard linear scaling relationships, follow the principles of green chemistry and provide the tools for a circular economy in the areas of chemistry, catalysis and materials science. ❐

Carles Bo[1,2], Feliu Maseras[1,3] and Núria López[1]*

[1]Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology, Tarragona, Spain. [2]Universitat Rovira i Virgili, Departament de Química Física i Inorgànica, Tarragona, Spain. [3]Universitat Autònoma de Barcelona, Departament de Química, Bellaterra, Spain.

**e-mail: nlopez@iciq.es*

References

1. NIST Chemistry WebBook. *NIST Standard Reference Database Number 69* (NIST, accessed 26 May 2018); https://webbook.nist.gov/chemistry
2. *Protein Data Bank* (RCSB, accessed 26 May 2018); https://www.rcsb.org
3. *The Cambridge Structural Database* (CCDC, accessed 26 May 2018); https://www.ccdc.cam.ac.uk
4. *Inorganic Crystal Structure Database* (FIZ Karlsruhe, accessed 26 May 2018); http://www2.fiz-karlsruhe.de/icsd_home.html
5. Lejaeghere, K. et al. *Science* **351**, aad3000 (2016).
6. Ohno, K. & Morokuma, K. *Quantum Chemistry Literature Data Base—Bibliography of Ab Initio Calculations for 1978–1980* (Elsevier, Amsterdam, 1982).
7. *QCLDB II* (QCDB Group, accessed 25 May 2018); http://qcldb2.ims.ac.jp
8. *Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101 Release 19* (NIST, accessed 26 May 2018); https://cccbdb.nist.gov
9. Hobza, P. *Benchmark Energy and Geometry Database* (Institute of Organic Chemistry and Biochemistry, Prague, accessed 26 May 2018); http://www.begdb.com
10. *Databases Truhlar Research Group* (accessed 26 May 2018); http://truhlar.chem.umn.edu/content/databases
11. Ghahremanpour, M. M., van Maaren, P. J. & van der Spoel, D. *Sci. Data* **5**, 180062 (2018).
12. Nakata, M. & Shimazaki, T. J. *Chem. Inf. Model* **57**, 1300–1308 (2017).
13. *Open Babel: The Open Source Chemistry Toolbox* (accessed 26 May 2018); http://openbabel.org/wiki/Main_Page
14. Murray-Rust, P. & Rzepa, H. S. *J. Cheminformatics* **3**, 44 (2011).
15. Adams, S. et al. *J. Cheminformatics* **3**, 38 (2011).
16. Hummelshøj, J. S., Abild-Pedersen, F., Studt, F., Bligaard, T. & Nørskov, J. K. *Angew. Chem. Int. Ed.* **51**, 272–274 (2011).
17. Laloo, J. Z. A., Laloo, N., Rhyman, L. & Ramassami, P. *J. Comput. Aided Mol. Des.* **31**, 667–673 (2017).
18. Rodríguez-Guerra Pedregal, J., Gómez-Orellana, P. & Maréchal, J.-D. *J. Chem. Inf. Model.* **58**, 561–564 (2018).
19. O'Boyle, N. M., Tenderholt, A. L. & Langner, K. M. *J. Comput. Chem.* **29**, 839–845 (2008).
20. Materials Genome Initiative (accessed 29 May 2018); https://mgi.gov
21. The Materials Project (accessed 30 August 2018); https://www.materialsproject.org
22. Tabor, D. P. et al. *Nat. Rev. Mater.* **3**, 5 (2018).
23. The European Materials Modelling Council (accessed 30 August 2018); https://emmc.info
24. de Bass, A. F. *What Makes a Material Function* (EU, 2017)
25. *NOMAD Repository* (NOMAD Laboratory, accessed 23 May 2018); https://nomad-repository.eu
26. *Automated Interactive Infrastructure and Database for Computational Science* (accessed 25 May 2018); http://www.aiida.net
27. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. *Comput. Mater. Sci.* **111**, 218–230 (2016).
28. Web platform "Materials Cloud" could help industry streamline research efforts. *Marvel* http://nccr-marvel.ch/highlights/2018-05-web-platform-materials-cloud-could-help-industry (30 May 2018).
29. *Computational Materials Repository* (CAMd, accessed 14 September 2018); https://cmr.fysik.dtu.dk
30. Álvarez-Moreno, M. et al. *J. Chem. Inf. Model.* **55**, 95 (2015).
31. *ioChem-BD* (accessed 29 May 2018); http://www.iochem-bd.org
32. Chen, Z. *Nat. Nanotech* **13**, 702–707 (2018).
33. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. *Nature* **559**, 547–555 (2018).
34. Wang, B., Dobosh, P. A., Chalk, S., Sopek, M. & Ostlund, N. S. *J. Phys. Chem. A* **121**, 298–307 (2016).
35. Rossi, E. et al. *J. Comput. Chem.* **35**, 611–621 (2014).
36. Ghiringhelli, L. M. *npj Comput. Mater.* **3**, 46 (2017).
37. The Molecular Sciences Software Institute (accessed 30 August 2018); https://molssi.org
38. Schütt, K. T., Arbabzadah, F., Chmiela, S. & Müller, K.-R. *Nat. Commun.* **8**, 13890 (2017).
39. Janet, J. P. & Kulik, H. J. *Chem. Sci.* **8**, 5137–5152 (2017).
40. Ferguson, A. L. *ACS Cent. Sci.* **4**, 938–941 (2018).
41. Gómez-Bombarelli, R. et al. *ACS Cent. Sci.* **4**, 268–276 (2018).
42. Nandy, A., Duan, C., Janet, J. P., Gugler, S. & Kulik, H. Preprint at https://doi.org/10.26434/chemrxiv.6987074.v1 (2018).
43. Jones, G. *Nat. Catal.* **1**, 311–313 (2018).
44. Wu, Z. et al. *Chem. Sci.* **9**, 513–530 (2018).
45. Lemonick, S. Is machine learning overhyped? *Chem. Eng. News* https://cen.acs.org/physical-chemistry/computational-chemistry/machine-learning-overhyped/96/i34 (2018).
46. PASC18 panel discussion. Is HPC facing a game change? *YouTube* https://www.youtube.com/watch?v=mTqzCvm0G5c (16 July 2018).