AI-Driven Molecular Discovery through Automated Dataset Generation and Execution

Sergi Vela
[1]*Institut de Química Avançada de Catalunya (IQAC-CSIC), Carrer de Jordi Girona 18, Barcelona, Spain*
sergi.vela@iqac.csic.es

The discovery and optimization of molecules and materials increasingly rely on computational chemistry explore the chemical space. In modern discovery pipelines, this exploration is increasingly driven by machine learning (ML) [1] models, whose accuracy and scope are partially determined by the quality and diversity of the datasets on which they are trained.[2] Generally, these datasets are obtained in two ways: (i) generated combinatorially from a pool of pre-defined fragments, which inherently restricts the resulting chemical space, or (ii) mined from existing repositories. The former approach is simpler, while the latter offers much greater diversity. Experimental databases such as the Cambridge Structural Database (CSD) contain vast structural information reflecting the creativity and synthetic efforts of thousands of researchers over decades. However, their direct exploitation for QC-driven discovery is often complicated by structural errors and missing electronic information, particularly in the case of transition metal complexes.

In recent work,[3] this issue was tackled through the development of cell2mol, a tool that automates the chemical interpretation of crystallographic data. It recovers molecular connectivity, total charge, and metal oxidation states directly from CIF files. Building upon this foundation, the authors further enabled the automated ground-state spin assignment using an ML model trained on more than 2000 complexes, achieving 98% accuracy.[4] These AI-powered tools facilitate the mining of experimental crystal structures into QC-ready molecular datasets with exceptional chemical diversity, as was demonstrated in the FORMED dataset for the exploration of organic electronic materials.[5]

Despite these advances in dataset generation and chemical space exploration, the practical execution of such studies poses significant operational challenges. Generally, computational workflows involve structure preparation, input generation, job submission and monitoring, post-processing, and long-term storage of results, which demands translational skills. To reduce human effort, and eliminate random errors, workflows often rely on improvised scripts that are seldom maintained, leading to reproducibility issues.

To address these challenges, we introduce SCOPE, an open-source Python package created as an end-to-end platform for automated computational chemistry workflows. SCOPE is built around a modular, object-oriented architecture that unifies chemical representation, workflow definition, execution on HPC systems, and structured data management within one framework. It supports dynamic workflows that respond to computation outcomes, and interfaces directly with SLURM-managed HPC environments for job submission, monitoring, and results retrieval. At present, SCOPE provides support for Gaussian and Quantum Espresso, covering a broad range of electronic structure applications.

SCOPE shares the goal of other similar tools like AiiDA[6] or QMFlows[7], but stands out in that it is built upon a hierarchy of chemistry-aware objects capable of representing molecules and molecular crystals, including systems with transition metal complexes. By focusing on chemical awareness, SCOPE simplifies the use of common techniques in computational chemistry for the generation and curation of structures, and for the analysis of results, which reduces user overhead and enhances reproducibility, particularly in high-throughput screening contexts. Overall, SCOPE consolidates years of work in data curation and dataset creation for ML-ready computational discovery projects in molecule-based materials.

References

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh. *Nature* 559 (2018) 547
[2] C. Bo, F. Maseras, and N. López, *Nature Catalysis* 1 (2018) 809.
[3] S. Vela, R. Laplaza, Y. Cho, C. Corminboeuf. *npj Comput Mater* **8** (2022) 188.
[4] Y. Cho, R. Laplaza, S. Vela, C. Corminboeuf. Digital Discovery 3 (2024), 1638.
[5] T. J. Blaskovits, R. Laplaza, S. Vela, C. Corminboeuf. Advanced Materials, 36 (2024) 2305602.
[6] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari and B. Kozinsky. Computational Materials Science, 111 (2016) 218.
[7] F. Zapata, L. Ridder, J. Hidding, C. R. Jacob, I. Infante and L. Visscher, Journal of Chemical Information and Modeling 59 (2019) 3191.