Position Paper on Dataset Engineering to Accelerate Science

Emilio Vital Brazil, Eduardo Soares, Lucas Villa Real, Leonardo Azevedo,
Vinicius Segura, Luiz Zerkowski, and Renato Cerqueira
IBM Research
Rio de Janeiro, RJ, Brazil
evital@br.ibm.com, eduardo.soares@ibm.com, lucasvr@br.ibm.com, lga@br.ibm.com,
vboas@br.ibm.com, luisvz@ibm.com, rcerq@br.ibm.com

Abstract

Data is a critical element in any discovery process. In the last decades, we observed exponential growth in the volume of available data and the technology to manipulate it. However, data is only practical when one can structure it for a well-defined task. For instance, we need a corpus of text broken into sentences to train a natural language machine-learning model. In this work, we will use the token *dataset* to designate a structured set of data built to perform a well-defined task. Moreover, the dataset will be used in most cases as a blueprint of an entity that at any moment can be stored as a table. Specifically, in science, each area has unique forms to organize, gather and handle its datasets. We believe that datasets must be a first-class entity in any knowledge-intensive process, and all workflows should have exceptional attention to datasets' lifecycle, from their gathering to uses and evolution. We advocate that science and engineering discovery processes are extreme instances of the need for such organization on datasets, claiming for new approaches and tooling. Furthermore, these requirements are more evident when the discovery workflow uses artificial intelligence methods to empower the subject-matter expert. In this work, we discuss an approach to bringing datasets as a critical entity in the discovery process in science. We illustrate some concepts using material discovery as a use case. We chose this domain because it leverages many significant problems that can be generalized to other science fields.

Introduction

Data is a critical element in any discovery process – it appears at the beginning of the process as input for experimentation, and at the end, as evidence to support the results. In the last few decades, we observed exponential growth in the volume of available data and the technology to generate and manipulate it (according to IDC, about 64 zettabytes were created or copied in 2020 (Overberg and Hand 2021)). However, data is only practical when one can use it for a well-defined task.

Besides the sheer volume, data can be unstructured and get more complex according to the domain, particularly in science applications. Moreover, organizing data and taking care of their entire lifecycle is especially vital when Artificial Intelligence (AI) techniques start to be critical in scientific processes. For instance, recently, the chemical industry has augmented traditional human-intensive work with automated, parallel, and iterative processes driven by AI to accelerate the materials-discovery (Pyzer-Knapp et al. 2022). This incorporation of AI in the materials-discovery workflow brought a set of novel problems in handling data, for example, how to qualify and filter thousand of molecule candidates created by machine learning generative techniques (Hoffman et al. 2021; Tadesse et al. 2022).

Nowadays scenario, where there is a high demand to accelerate scientific discoveries, which depends on a massive quantity of data, we advocate that it is paramount to look at data from a new perspective, bringing together data and tasks to remodel the concept of the dataset. To be used in any discovery process, data must have a well-defined structure, associations with domain knowledge, and a set of operations and analytics to evolve it. However, many methods to give such effects to data will strongly depend on the task we will perform with the data. For instance, representing a molecule is an enormous issue in chemical applications, and its choice must consider the final task (O'Boyle et al. 2011). Another common problem for working with data is how to define its lifecycle, and again we notice good practices for using the task to define it. We must create a dataset guided by the task that we will perform. Furthermore, while the operations executed on it are still creating information related to the original task, we consider it a natural evolution of the dataset, keeping its versioning. In summary, we propose working with datasets as first-class entities in the discovery workflow, using the task as its main characteristic to guide its complete lifecycle.

Our View

We are proposing an approach to the problem of managing data in the discovery process, where the task must guide the data lifecycle. Then, from now on, a *dataset* is defined as a set of data structured to perform a well-defined task. Which can be viewed as a blueprint of the actual data, that can be stored as tables at any time. In this text, we do not split the concept of dataset blueprint and dataset tables, but for critical discussion about how to implement such concepts it is paramount. In Figure 1, we illustrate the main components of our approach and their relationship. Our proposed approach to dataset engineering splits it into three critical

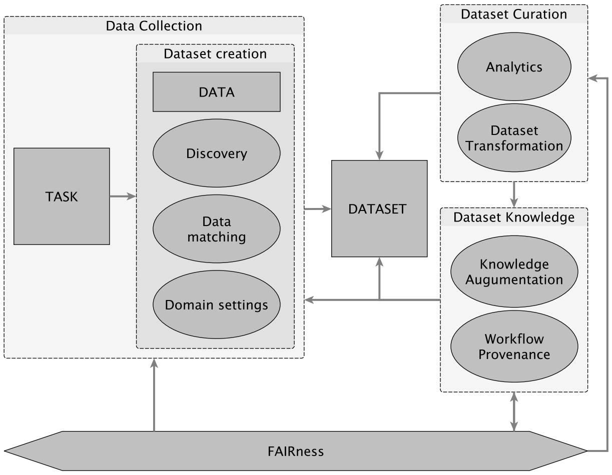

Figure 1: The relationship between the main components of our proposal to dataset engineering, the rectangles are instances, and ellipses represent sets of processes. FAIRness is a meta element that touches all components.

dimensions Data Collection, Dataset Curation, and Dataset knowledge. The data collection is responsible for collecting and organizing the data and creating the datasets from a task. It combines many challenges like data discovery and matching with domain settings, which is the set of processes responsible for connecting the data with the specificities of the task for scientific domains, such as material discovery. The dataset curation component is the set of processes to analyze datasets and enable their evolution. It focuses on the synergy between analytics and dataset transformation, supporting knowledge augmentation. The dataset knowledge piece is responsible for extending and connecting a dataset with a knowledge base, managing its lineage, and versioning. It is also vital feedback on the data collection dimension with the knowledge gathered in all components to improve the entire workflow closing the cycle. Furthermore, another critical element that we consider is the FAIR principles (Wilkinson et al. 2016), which touch all components helping to guide the dataset development. Although it is a meta part of our framework, in this work, we will discuss FAIRness in the context of the dataset knowledge.

The following sections present each main component from the material discovery perspective. This use case is rich in many aspects. For instance: (i) Computational chemistry is a vibrant field with critical issues that create significant challenges for dataset engineering, like how to represent a molecule (O'Boyle 2012). (ii) The amount of consumed and created data is enourmos (Ruddigkeit et al. 2012). (iii) The great potential to apply AI in many problems creates a new set of questions directly related to datasets (Suh et al. 2020).

Data Collection

Given that several organizations worldwide continuously produce new data as part of their discovery and engineering processes, collecting and aggregating such data comes with several challenges. Here we highlight some challenges that the scientific community has been trying to address.

Discovery Finding new repositories that host relevant data involves crawling the Internet. Such a task demands the use of an infrastructure with an efficient storage stack, a fast network bandwidth, and enough processing power to parse the retrieved pages. Once potential datasets have been found, one needs to identify the terms of use of such data and somehow assess the quality of that data. An initiative from bioschemas.org attempts to improve the website indexing process by defining a markup vocabulary for websites that host datasets from life sciences (Brickley, Burgess, and Noy 2019). On the other hand, many data providers keep falling back to manual curation steps when processing pipelines flag chemical structures with serious errors. Automatically assessing the quality of new data remains an open problem that we believe to be of critical importance for a high-quality data collection task.

Data matching Commonly, researchers attempt to improve upon results published by other organizations. This means that, once results of that new research are made public, the original data can be augmented by aggregating the new information. Unfortunately, data matching is not a straightforward task, as the following issues observed in material discovery indicate:

- Different notations to enumerate chemical structures: mapping between popular text-based notations such as *SMILES*, *SMARTS*, *InChI* and *InChIKey* is required, as publications are free to choose which notation to use (Saldivar-Gonzalez, Huerta-García, and Medina-Franco 2020);
- Multiple representation for the same molecular graph exist: *SMILES* strings are pervasive across data sources, yet one cannot simply resort to string comparison to tell if two *SMILES* represent the same molecule. A set of standardization rules allow the generation of canonical *SMILES* that converge to the same string. In practice, though, standardization rules differs between programs and, consequently, across data providers (Bento et al. 2020);
- Compounds may have different names and several synonyms: for instance, while a publication may refer to **[CH3][CH2][OH]** as *ethanol*, others may refer to that same compound as *alcohol*, *ethyl hydrate*, *anhydrol*, among several different alternative names (and possibly written in other languages);
- Typos and conversion errors: supplemental material, often shared via spreadsheets, are prone to conversion errors (such as gene names mistakenly converted to dates by Microsoft Excel (Abeysooriya et al. 2021)) and typographical errors that further difficult the task of automated data matching;
- Inaccurate cross-references among databases (Dashti et al. 2019).

Effectively, the data matching process becomes a pipeline where different techniques apply. Ongari et al. (2022), for instance, attempt matches by conventional name, by the publication venue and ID, by comparing the molecular graph's substructures, and even the pore volume of the crystal structures. It is clear that this is an open problem with several opportunities for improvement.

Domain settings To be useful, data retrieved from any given source repository must be transformed to meet the needs of the target domain. For instance, transformations like the normalization of relation units and standardization of compounds depend not only on predefined rules and conventions but also on ontologies that help establish a mapping from the source data to the desired output format. Automatically determining which functions provide this mapping is a key feature of a dataset engineering platform.

Dataset Curation

The dataset curation step is essential for the understanding of the data as well as the knowledge extraction from it. This step is composed by two different groups which are further detailed: Dataset Analytic, and Dataset Transformation.

Dataset Analytic

The dataset analytic process is composed by different methods which aims to uncover useful insights to experts. This step comprises the following analysis methods: Clustering Analysis, Covering Analysis, Causality Analysis, Data Visualization, Similarity Analysis, and Uncertainty quantification.

Clustering Analysis The clustering analysis is part of an unsupervised strategy used to discover existing patterns in a given dataset and group objects with similar characteristics given a context. According to (Hadipour et al. 2022), compounds clustering is vital to validate the diversity of the dataset, identify the similarity and heterogeneity among the objects contained in the dataset, and improve the challenging and costly process of establishing datasets for machine learning tasks (Elshawi et al. 2018). Understanding the categories of the compounds that need to be included in the dataset can significantly reduce the number of molecules that should be screened while, at the same time, ensuring the quality of the dataset. Different clustering techniques can be used at this step, including: CheS-Mapper (Gütlein, Karwath, and Kramer 2014), K-Means (Nugent and Meila 2010), Graph-based clustering (Tanemura, Das, and Merz Jr 2021), autocoencoders (Hadipour et al. 2022), and others.

Covering Analysis This step regards to the use of evaluation metrics and further insights to understand better the under- or over-generation of the data (Tadesse et al. 2022). It also helps to understand their characterizations at different levels of evaluation. Moreover, at this step the completeness of the data can be verified if necessary.

Therefore, such insights can benefit the quality of the data through improved interactions between machine learning researchers and domain experts in new molecules discovery (Tadesse et al. 2022).

Causality Analysis Understanding complicated interactions of chemical components is essential to new molecules discovery (Dang et al. 2015). Therefore, to detect the causal relations in the molecular structures play essential role for the description of molecular mechanisms and comprehend their functioning (Kelly et al. 2022).

The causality analysis favors the explainability of the dataset and may benefit the process of science discovering through machine learning models (Holzinger et al. 2021).

Data Visualization Data visualization is crucial for the dataset analysis and the advance of scientific researches (Fox and Hendler 2011). In terms of new molecules discovery field, data visualization enables decision-makers to discover design patterns, comprehend information, and form an opinion about potential new scientific discovery candidates(Ekins et al. 2016).

Designing new and better compounds requires understanding of the mechanism by which the molecules exert their biological effects. This also involves consideration of the uncertainty contained in the data, which data visualization helps to provide interpretability of it and allow researches to understand better the nature of the data (Rheingans and Joshi 1999).

Similarity Analysis Similarity-based methods are part of a feedforward reasoning process that relies on the proximity (in the feature space) of a target object to a given object (Angelov and Soares 2020). In terms of molecules analysis, molecular similarity implies that molecules of "similar" structure tend to have similar properties to their analogues (Samanta et al. 2020). Therefore, a common question that a researcher can make is: "Given a target molecule M which has a specific chemical activity, can I find the 30 molecules that are most similar to M so I can assess their behavior in a relevant quantitative-structure-activity (QSAR) analysis?".

The traditional strategy to calculate molecular similarity regards to encode the molecule as a vector of numbers. If fingerprints, for example Extended Connectivity FingerPrinting (ECFP) (Rogers and Hahn 2010), are used to produce a vector of bits which that describes molecule structure the Tanimoto similarity is commonly applied. However, as fingerprints just produce binary information regarding to the molecule structure, molecular descriptors as Mordred descriptor (Moriwaki et al. 2018) or PaDEL-descriptor (Yap 2011) richer information for the calculation of an improved similarity ranking. These interpretable features provided by molecular descriptors also favors human-in-the-loop as experts can incorporate their knowledge in the data to obtain a set of similar molecules that best fits the context that they are working.

Ranking metrics also serve an important purpose in evaluating the similarity of compounds. Distinct chemical domains are typically best described by different molecular features. Therefore, to provide experts with guided decision-making capabilities, various techniques can be used to evaluate the quality of any given (encoding, similarity function) pair for some class of compounds (Wassenaar et al. 2022).

Uncertainty Quantification Uncertainty quantification is a key component to provide measures of confidence necessary to complex decisions (Begoli, Bhattacharya, and Kusnezov 2019). Uncertainty estimation in predictions can have the potential to save considerable time and effort during the decision-making process. Additionally, applications and advances in the field of new molecules discovery has additional safety requirements that are demanding from scientists (Hirschfeld et al. 2020).

Indeed, model interpretability including associated confidence in output predictions is recognised as a principal shortcoming of current approaches for new molecules discovery (Wan, Sinclair, and Coveney 2021). Therefore, better communication of uncertainty positively contributes to the adoption of machine learning systems to accelerate scientific discoveries.

Dataset Transformation

Data transformation is a key step which embraces the processes of changing the format, structure, or values of data through Interactive AI and Batch Operations.

Interactive AI This step encompass solutions and algorithms where experts influences AI systems and vice-versa (Çelikok et al. 2019). This includes decision support solutions, recommender systems, and dialogue systems with focus on explainability, interpretability, and interaction leveraging the user experience with the dataset (Bellamy et al. 2019). This includes the operations of adjudication (Schaekermann 2020) and data aggregation (Edge, Larson, and White 2018). Interactive AI benefits the human-in-the-loop as it allows experts to incorporate useful, meaningful human interaction into the dataset (Zanzotto 2019). This task is specifically important for high stake applications as the dataset engineering to advance science.

Batch Operations Batch operations for data transformation includes processes for data cleaning, reduction, expansion, and generation (Fink 2009). Such operations are essential to guarantee the quality of the data and the best usability of them by machine learning approaches and experts. Batch operations gives a sense of how data is distributed, both from visual or quantitative perspectives (Yu 2010). Therefore, we can consider that data transformations of variables to ease both interpretation of data analyses and the application statistical and machine learning models to the dataset.

Dataset Knowledge

Knowledge Augmentation concerns the enhancement of the current knowledge base by acquiring/ingesting new data from other sources, and creating new knowledge by reasoning over the current base and over the new ingested data. To achieve Knowledge Augmentation, it is required to perform several tasks, like knowledge acquisition, formalization, storage/retrieval, learning, and reasoning (Silva de Oliveira, Sanin, and Szczerbicki 2022).

The knowledge base dataset may reside on disparate locations in heterogeneous data stores represented by different data models. A middleware that provides a seamless interface with an independent data model and data schemes is required to access heterogeneous data stores, like polystore systems (Stonebraker 2015; Özsu and Valduriez 2020).

Although exists well-curated, deeply-integrated, special-purpose repositories, many important datasets emerging from traditional, low-throughput bench science do not fit in the data models of these special-purpose repositories. It results in a diverse, less integrated, data ecosystem, exacerbating the discovery and re-usability of datasets for both humans and computation stakeholders. As an example, if a researcher wants to compare a dataset resulting from his/her experiment with other datasets, several questions should be answered, such as: (i) Where might the existing dataset have been published? (ii) How to start the search and using what search tools? (iii) Which characteristics should be used to filter the datasets? (iv) Are the datasets described with metadata, and metadata in what formats? After the dataset is found, other questions arise, like: (i) Can the dataset be downloaded? (ii) What is the data format? (iii) What are the requirements to integrate the data with local data? (iv) Can the data be automatically integrated? (v) Does the researcher have permission to use the data? Under what license conditions? Therefore, it is a grand challenge of data-intensive science to improve knowledge discovery for humans and computational agents in the discovery, access, integration and analysis of task-appropriate scientific data (Wilkinson et al. 2016). In 2016, Wilkinson *et al.* (Wilkinson et al. 2016) published the FAIR principles, a set of 15 recommendations for improving Findability, Accessibility, Interoperability, and Reusability of digital resources (Jacobsen et al. 2020). The principles are domain-independent and aim to facilitate reuse by humans and machines (Trojahn et al. 2022).

In another perspective, scientists struggle in performing comprehensive data analyses over their experiments if there is no information collected during the experiment workflow executions. To overcome this issue, they embrace provenance techniques on their experiments. Provenance (also referred to as lineage) data management techniques help reproduce, trace, assess, understand, and explain data, models, and their transformation processes (Herschel, Diestelkämper, and Lahmar 2017; Moreau et al. 2008; Buneman and Tan 2019). The provenance research community has evolved significantly to provide for several strategic capabilities, including experiment reproducibility (Thavasimani and Missier 2016), user steering (*i.e.,* runtime monitoring, interactive data analysis, runtime fine-tuning) (Souza et al. 2019b), raw data analysis (Sousa et al. 2016), and data integration for multiple workflows generating data in a data lake (Souza et al. 2019a).

Final Remarks

We argue that datasets should have a central role in knowledge-intensive processes, especially in scientific discovery. Moreover, we define its lifecycle in a task-oriented way, creating a synergy between the natural dataset evolution and its uses. The tasks and data combination is a powerful driver. It helps us finding many solutions to reuse data workflows, empower experts in the dataset lifecycle, and create data tooling. There are still many questions about the practical issues of datasets. However, this dataset definition can significantly impact many real-world problems supporting the acceleration of scientific discovery.

References

Abeysooriya, M.; Soria, M.; Kasu, M. S.; and Ziemann, M. 2021. Gene name errors: Lessons not learned. *PLOS Computational Biology*, 17(7): 1–13.

Angelov, P.; and Soares, E. 2020. Towards explainable deep neural networks (xDNN). *Neural Networks*, 130: 185–194.

Begoli, E.; Bhattacharya, T.; and Kusnezov, D. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1): 20–23.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5): 4–1.

Bento, A. P.; Hersey, A.; Felix, E.; Landrum, G. A.; Gaulton, A.; Atkinson, F.; Bellis, L. J.; Veij, M. D.; and Leach, A. R. 2020. An open source chemical structure curation pipeline using RDKit. *J. Cheminformatics*, 12(1): 51.

Brickley, D.; Burgess, M.; and Noy, N. 2019. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In *The World Wide Web Conference*, WWW '19, 1365–1375. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Buneman, P.; and Tan, W.-C. 2019. Data provenance: What next? *SIGMOD Rec.*

Çelikok, M. M.; Peltola, T.; Daee, P.; and Kaski, S. 2019. Interactive AI with a Theory of Mind. *arXiv preprint arXiv:1912.05284*.

Dang, T. N.; Murray, P.; Aurisano, J.; and Forbes, A. G. 2015. ReactionFlow: an interactive visualization tool for causality analysis in biological pathways. In *BMC proceedings*, volume 9, 1–18. BioMed Central.

Dashti, H.; Wedell, J. R.; Westler, W. M.; Markley, J. L.; and Eghbalnia, H. R. 2019. Automated evaluation of consistency within the PubChem Compound database. *Sci. Data*, 6(1): 190023.

Edge, D.; Larson, J.; and White, C. 2018. Bringing AI to BI: enabling visual analytics of unstructured data in a modern Business Intelligence platform. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems*, 1–9.

Ekins, S.; Perryman, A. L.; Clark, A. M.; Reynolds, R. C.; and Freundlich, J. S. 2016. Machine learning model analysis and data visualization with small molecules tested in a mouse model of Mycobacterium tuberculosis infection (2014–2015). *Journal of chemical information and modeling*, 56(7): 1332–1343.

Elshawi, R.; Sakr, S.; Talia, D.; and Trunfio, P. 2018. Big data systems meet machine learning challenges: towards big data science as a service. *Big data research*, 14: 1–11.

Fink, E. L. 2009. The FAQs on data transformation. *Communication Monographs*, 76(4): 379–397.

Fox, P.; and Hendler, J. 2011. Changing the equation on scientific data visualization. *Science*, 331(6018): 705–708.

Gütlein, M.; Karwath, A.; and Kramer, S. 2014. CheS-Mapper 2.0 for visual validation of (Q) SAR models. *Journal of cheminformatics*, 6(1): 1–18.

Hadipour, H.; Liu, C.; Davis, R.; Cardona, S. T.; and Hu, P. 2022. Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. *BMC bioinformatics*, 23(4): 1–22.

Herschel, M.; Diestelkämper, R.; and Lahmar, H. B. 2017. A survey on provenance: What for? What form? What from? *VLDB*.

Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; and Coley, C. W. 2020. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8): 3770–3780.

Hoffman, S. C.; Chenthamarakshan, V.; Zubarev, D. Y.; Sanders, D. P.; and Das, P. 2021. Sample-Efficient Generation of Novel Photo-acid Generator Molecules Using a Deep Generative Model. *arxiv*, (arXiv:2112.01625).

Holzinger, A.; Malle, B.; Saranti, A.; and Pfeifer, B. 2021. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. *Information Fusion*, 71: 28–37.

Jacobsen, A.; de Miranda Azevedo, R.; Juty, N.; Batista, D.; Coles, S.; Cornet, R.; Courtot, M.; Crosas, M.; Dumontier, M.; Evelo, C. T.; Goble, C.; Guizzardi, G.; Hansen, K. K.; Hasnain, A.; Hettne, K.; Heringa, J.; Hooft, R. W.; Imming, M.; Jeffery, K. G.; Kaliyaperumal, R.; Kersloot, M. G.; Kirkpatrick, C. R.; Kuhn, T.; Labastida, I.; Magagna, B.; McQuilton, P.; Meyers, N.; Montesanti, A.; van Reisen, M.; Rocca-Serra, P.; Pergl, R.; Sansone, S.-A.; da Silva Santos, L. O. B.; Schneider, J.; Strawn, G.; Thompson, M.; Waagmeester, A.; Weigel, T.; Wilkinson, M. D.; Willighagen, E. L.; Wittenburg, P.; Roos, M.; Mons, B.; and Schultes, E. 2020. FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence*, 2(1-2): 10–29.

Kelly, J.; Berzuini, C.; Keavney, B.; Tomaszewski, M.; and Guo, H. 2022. A review of causal discovery methods for molecular network analysis. *Molecular Genetics & Genomic Medicine*, e2055.

Moreau, L.; Ludäscher, B.; Altintas, I.; Barga, R. S.; Bowers, S.; Callahan, S.; Chin JR., G.; Clifford, B.; Cohen, S.; Cohen-Boulakia, S.; Davidson, S.; Deelman, E.; Digiampietri, L.; Foster, I.; Freire, J.; Frew, J.; Futrelle, J.; Gibson, T.; Gil, Y.; Goble, C.; Golbeck, J.; Groth, P.; Holland, D. A.; Jiang, S.; Kim, J.; Koop, D.; Krenek, A.; McPhillips, T.; Mehta, G.; Miles, S.; Metzger, D.; Munroe, S.; Myers, J.; Plale, B.; Podhorszki, N.; Ratnakar, V.; Santos, E.; Scheidegger, C.; Schuchardt, K.; Seltzer, M.; Simmhan, Y. L.; Silva, C.; Slaughter, P.; Stephan, E.; Stevens, R.; Turi, D.; Vo, H.; Wilde, M.; Zhao, J.; and Zhao, Y. 2008. Special Issue: The first provenance challenge. *CCPE*.

Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; and Takagi, T. 2018. Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10(1): 1–14.

Nugent, R.; and Meila, M. 2010. An overview of clustering applied to molecular biology. *Statistical methods in molecular biology*, 369–404.

O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; and Hutchison, G. R. 2011. Open Babel: An Open Chemical Toolbox. *Journal of Cheminformatics*, 3(1): 33.

Ongari, D.; Talirz, L.; Jablonka, K. M.; Siderius, D. W.; and Smit, B. 2022. Data-Driven Matching of Experimental Crystal Structures and Gas Adsorption Isotherms of Metal–Organic Frameworks. *Journal of Chemical & Engineering Data*, 67(7): 1743–1756.

Overberg, P.; and Hand, K. 2021. How to Understand the Data Explosion - WSJ. *https://www.wsj.com/articles/how-to-understand-the-data-explosion-11638979214*. Accessed on Nov. 1st 2022.

Özsu, M. T.; and Valduriez, P. 2020. *Principles of distributed database systems*. Springer, 4th edition.

O'Boyle, N. M. 2012. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. 4(1): 22.

Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; and Curioni, A. 2022. Accelerating Materials Discovery Using Artificial Intelligence, High Performance Computing and Robotics. 8(1): 84.

Rheingans, P.; and Joshi, S. 1999. Visualization of molecules with positional uncertainty. In *Data Visualization'99*, 299–306. Springer.

Rogers, D.; and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5): 742–754.

Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; and Reymond, J.-L. 2012. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. 52(11): 2864–2875.

Saldivar-Gonzalez, F.; Huerta-García, C.; and Medina-Franco, J. 2020. Chemoinformatics-based enumeration of chemical libraries: a tutorial. *Journal of Cheminformatics*, 12: 64.

Samanta, S.; O'Hagan, S.; Swainston, N.; Roberts, T. J.; and Kell, D. B. 2020. VAE-Sim: a novel molecular similarity measure based on a variational autoencoder. *Molecules*, 25(15): 3446.

Schaekermann, M. 2020. Human-AI Interaction in the Presence of Ambiguity: From Deliberation-based Labeling to Ambiguity-aware AI.

Silva de Oliveira, C.; Sanin, C.; and Szczerbicki, E. 2022. Smart Knowledge Engineering for Cognitive Systems: A Brief Overview. *Cybernetics and Systems*, 53(5): 384–402.

Sousa, V.; Oliveira, D.; Valduriez, P.; and Mattoso, M. 2016. Analyzing related raw data files through dataflows. *CCPE*.

Souza, R.; Azevedo, L.; Thiago, R.; Soares, E.; Nery, M.; Netto, M. A. S.; Brazil, E. V.; Cerqueira, R.; Valduriez, P.; and Mattoso, M. 2019a. Efficient runtime capture of multiworkflow data using provenance. In *IEEE eScience*.

Souza, R.; Silva, V.; Camata, J. J.; Coutinho, A.; Valduriez, P.; and Mattoso, M. 2019b. Keeping track of user steering actions in dynamic workflows. *FGCS*.

Stonebraker, M. 2015. The Case for Polystore. https://wp.sigmod.org/?p=1629.

Suh, C.; Fare, C.; Warren, J. A.; and Pyzer-Knapp, E. O. 2020. Evolving the Materials Genome: How Machine Learning Is Fueling the next Generation of Materials Discovery. 50(1): 1–25.

Tadesse, G. A.; Born, J.; Cintas, C.; Manica, M.; and Weldemariam, K. 2022. MPEGO: A toolkit for multi-level performance evaluation of generative models for material discovery domains. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Tanemura, K. A.; Das, S.; and Merz Jr, K. M. 2021. AutoGraph: Autonomous graph-based clustering of small-molecule conformations. *Journal of Chemical Information and Modeling*, 61(4): 1647–1656.

Thavasimani, P.; and Missier, P. 2016. Facilitating reproducible research by investigating computational metadata. In *IEEE Big Data*.

Trojahn, C.; Kamel, M.; Annane, A.; Aussenac-Gilles, N.; and Nguyen, B. L. 2022. A FAIR Core Semantic Metadata Model for FAIR Multidimensional Tabular Datasets. In Corcho, O.; Hollink, L.; Kutz, O.; Troquard, N.; and Ekaputra, F. J., eds., *Knowledge Engineering and Knowledge Management*, Lecture Notes in Computer Science, 174–181. Cham: Springer International Publishing. ISBN 978-3-031-17105-5.

Wan, S.; Sinclair, R. C.; and Coveney, P. V. 2021. Uncertainty quantification in classical molecular dynamics. *Philosophical Transactions of the Royal Society A*, 379(2197): 20200082.

Wassenaar, P. N.; Rorije, E.; Vijver, M. G.; and Peijnenburg, W. J. 2022. ZZS similarity tool: The online tool for similarity screening to identify chemicals of potential concern. *Journal of Computational Chemistry*, 43(15): 1042–1052.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Yap, C. W. 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, 32(7): 1466–1474.

Yu, C. H. 2010. Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research*, 3(1): 9–22.

Zanzotto, F. M. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64: 243–252.