Advancing the ColabFit Exchange towards a Web-scale Data Source for Machine Learning Interatomic Potentials

Eric G. Fuemmeler
Department of Aerospace Engineering and Mechanics, University of Minnesota

Gregory P. Wolfe
Center for Soft Matter Research, Department of Physics, New York University

Amit Gupta
Department of Aerospace Engineering and Mechanics, University of Minnesota

Joshua A. Vita
Materials Science Division, Lawrence Livermore National Laboratory

Ellad B. Tadmor
Department of Aerospace Engineering and Mechanics, University of Minnesota

Stefano Martiniani
Center for Soft Matter Research, Department of Physics, New York University
Simons Center for Computational Physical Chemistry, Department of Chemistry
Courant Institute of Mathematical Sciences

Abstract

Data-driven (DD) interatomic potentials (IPs) trained on large collections of first principles calculations are rapidly becoming essential tools in the fields of computational materials science and chemistry for discovery pipelines and performing atomic-scale simulations. Despite this, apart from a few notable exceptions, there is a distinct lack of well-organized, public datasets in common formats available for use with IP development. This deficiency precludes the research community from implementing widespread benchmarking, which is essential for gaining insight into model performance and transferability, and also limits the development of more general universal (perhaps even multi-source) IPs. To address this issue, last year we introduced the ColabFit Exchange, the first database providing open access to a large collection of systematically organized datasets from multiple domains that is especially designed for IP development. It has now grown to contain 369 datasets spanning nearly 400,000 unique chemistries. Here we discuss recent updates to the ColabFit Exchange, including data statistics for the ever-growing database, modifications to the data standard and database backend, and new tools to utilize the data for machine learning (ML) applications.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Leveraging modern computing infrastructures, high-throughput pipelines for density functional theory (DFT) calculations have been able to produce results for millions of atomic configurations spanning a wide range of chemistries and applications [1, 2, 3, 4, 5, 6]. These methods have led to the creation of a number of massive datasets of first principles calculations, such as Materials Project [7], OpenCatalyst Project [8, 9], and Alexandria [10], among others [11, 12, 13, 14], which have served as critical resources for materials discovery and interatomic potential (IP) development. While these repositories have proven extremely useful, there still exist opportunities for continued development and dissemination of datasets specifically tailored to fit the needs of developers of data-driven (DD) IPs. In particular, datasets intended for use with IP development typically include a variety of non-equilibrium atomic configurations or hand-selected structures depending on the target application. Furthermore, datasets intended for fitting DDIPs are often carefully pruned and refined to enable the models to efficiently learn the physical behaviors relevant for the accurate prediction of a given material property, and to achieve stable simulations. This is in contrast to many of the largest and most used datasets that are commonly derived from (relaxation) trajectories. Conversely, existing databases of quantum mechanical (QM) calculations focus predominantly on stable equilibrium structures relevant to material discovery. Even in the case of databases that do contain portions of the data that may be suitable for use in DDIP fitting, they are rarely organized in a way that facilitates model benchmarking or targeted analysis of model behavior across chemical compound space.

In addition to the issues of content and structure of existing QM calculation databases, common methods for organizing and distributing DDIP training datasets, such as the use of personal GitHub repositories [9, 15, 16, 17, 18], uploads to Figshare [19, 20, 21, 22, 23], Zenodo [24, 25, 26, 27], or other file sharing methods are inconsistent and not conducive to interpretability and interoperability of the datasets. Datasets stored in this manner often use custom formats (Extended XYZ, HDF5, VASP OUTCARs, CSV, JSON) depending upon the specific research group that generated them, and despite government insistence [28, 29] typically lack metadata necessary for interpretability and reproducibility of the data (missing units, unspecified DFT settings, undocumented inconsistencies in data structure). Unfortunately, even this limited approach for sharing data is pursued by only a handful of researchers, with the vast majority of DDIP datasets being entirely inaccessible to the general public or made available through private correspondence “upon reasonable request,” without always honoring such requests. The end result is a significant decrease in reproducibility of published results and the effective loss of non-trivial amounts of effort and computational time spent on data generation, inevitably hindering scientific progress.

Therefore, to address these concerns, the ColabFit Exchange was developed [30]. It serves as a FAIR[31] (findable, accessible, interoperable, and reusable) exchange of datasets designed for DDIP training to help to facilitate collaboration and drive innovation by: 1) defining a consistent, efficient, and standardized method for storing the data; 2) enabling the organization of the data into meaningful, well-documented groupings; and 3) providing tools for easily accessing and contributing to the database in order to promote community engagement. In this work, we describe recent updates to the ColabFit Exchange, including the manner in which the data is represented, changes to the database backend infrastructure, and new ways to utilize the data.

2 Database Structure

The original storage backend used in the ColabFit Exchange was a non-relational database (MongoDB), chosen for the potential flexibility of data point representation. To increase speed of data ingestion and retrieval and facilitate long-term growth of the database, the ColabFit Exchange has moved to a relational (SQL-style) database, Vast DB [32], which is designed as a high speed, flash-based data lake. For full details on data components that remain largely unchanged, please refer to the original ColabFit Exchange publication [30]. More significant changes are detailed below.

2.1 Changes to low-level components (COs and POs)

Briefly, the two fundamental building blocks of the ColabFit Data Standard are Configurations (COs) and Property Objects (POs). Each CO stores a representation of an atomistic object of interest and typically serves as input ($\mathbf{x}$) to a DD pipeline. POs, on the other hand, store instances of property

Table 1: Counts of objects of interest in the ColabFit Exchange, excluding the data from the OpenCatalyst datasets. These values do *not* double count in the case where there exist duplicates of a given object (e.g., when an identical configuration was uploaded in multiple datasets, or an author is credited on multiple publications). Here, a “chemical system” refers to a set of unique constituent atom types.

Objects	Count
Datasets	369
Configuration sets	1,327
Property objects	125,291,927
Configurations	86,888,666
Atoms	3,385,190,598
Chemical systems	395,338
Publications	168
Authors	724

values associated with COs and typically serve as predictive targets ($\mathbf{y}$). To streamline the process of data selection and use, properties belonging to the same calculation, *e.g.* energy and atomic forces from one snapshot of a relaxation trajectory, are now stored together. This greatly improves data retrieval and export, which is essential for efficient downstream machine learning (ML) applications. Additionally, calculated values have been converted to standard units, rather than being documented in their original units, as was done previously. This simplifies multi-source training, an increasingly important strategy towards the development of foundation models.

As a primary target for DDIP tasks, representation of appropriate energy calculations is critical. The ColabFit Exchange now stores a single value for the energy that would be considered conjugate with atomic forces (those forces representing the negative gradient of the calculated energy with respect to changes in atomic positions). Due to differences in terminology between different calculation software, this energy value, regardless of its designation in the software or data files, has been stored under a single designation, allowing the user to choose the appropriate target without confusion caused by inconsistent property naming.

2.2 Changes to other components (MDs and DSs)

Metadata was previously stored as a separate object of nested attribute names and values. To remain true to the mission of enabling data reproducibility, with the resulting potential size and complexity of metadata content, metadata is now stored on disk as a high performance dual NFS/object, with a direct file address stored in the corresponding PO or CO.

3 Data Overview

Table 1 provides a summary of the contents of the ColabFit Exchange, which is currently (September 2024) composed of 369 unique datasets contributed by their authors or gathered from the literature. These datasets are further broken down into 1,327 configuration sets, which can be readily combined, split, or grouped in order to define new datasets based on the needs of the community. In total, the ColabFit Exchange contains over 125 million POs, corresponding to over 218 million computed properties. Note that the OpenCatalyst datasets (which are included in the ColabFit Exchange) are not included in these summary statistics, as they are already well-documented elsewhere in the literature [8, 9] and their large sizes ($\sim$114 million COs for OC20) would obscure the results from the other datasets. As the ColabFit Exchange continues to grow, updated statistics summarizing its contents can be found at `https://colabfit.org`.

The $\sim$86 million atomic configurations (for a total of $\sim$3.4 billion atoms) spanning nearly 400,000 chemical systems can be further analyzed based on their chemical composition, as shown in Fig. S1. Here, a “chemical system” is defined as a set of unique constituent atom types, e.g., C, C-H, C-H-N, . . . , and is indicative of the types of chemistries explored within the ColabFit Exchange. Though single element datasets are the most common (see Fig. S2), 95% of the configurations in the ColabFit Exchange include at least two elements, meaning the ColabFit Exchange may be used as a starting point for the development of many multi-element models. Much of the multi-element data comes from larger datasets designed for the construction of "universal" IPs intended to model all relevant types of atomic interactions [33, 34, 35], such as the Materials Project trajectory dataset [35], and others from the literature [21, 34, 36]. By providing access to all of these datasets within a unified framework, the ColabFit Exchange will simplify the process of constructing training datasets for new chemical systems that have not yet been explicitly sampled by the datasets currently in the ColabFit Exchange.

The values in Table 2 provide a further breakdown of the most prevalent computed properties stored within the ColabFit Exchange that are available for supervised training. Energies are the most commonly computed property, followed by forces. Note that the energy counts in Table 2 represent the energy conjugate with atomic forces, as discussed above. The force property count in Table 2 represents the number of POs with calculated forces. Each PO typically represents a multi-atom system, and may therefore contain multiple individual force vectors. Stresses are available for about 11% of the POs in the ColabFit Exchange. The ColabFit Exchange also includes additional properties that are well-defined within the schema but are generally less relevant to DDIP development, *e.g*, band gap.

Table 2: Counts of property instances in the ColabFit Exchange, excluding the data from the OpenCatalyst datasets. These values *do* double count in the case where two identical copies of a property exist (e.g., two distinct configurations were uploaded with identical potential energies) in order to accurately reflect the number of target values in the ColabFit Exchange. Though many of the datasets currently in the ColabFit Exchange contain more computed properties than the three shown here, energies, forces, and stresses are the three that are predominantly used for training DDIPs.

Property Instance (PI)	Count
Energy	123,096,727
Atomic forces	75,200,861
Stress	13,804,217

At the dataset level, Fig. S2 shows that the ColabFit Exchange has a wide range of dataset sizes, both in terms of the total number of atoms and the number of unique atom types contained within a given dataset. Since the original ColabFit publication, the number of datasets with greater than 20 atom types has grown from 3 to 42. The increased number of multi-element and larger datasets reflects our effort to gather such datasets, as well as a broader trend within the community to generate such datasets.

4 Data Integration

In the original ColabFit publication [30], we highlighted a streamlined ML workflow taking data from the ColabFit Exchange, using that data for training within the KLIFF [37] package, and storing and deploying the resulting model via the OpenKIM [38] platform. Since then we have developed new and integrated with existing tools to help facilitate the use of data on the ColabFit exchange for the development of DDIPs and other advanced material/chemical ML models. For example, we have integrated with the Open MatSci ML Toolkit [39] for the training of state-of-the-art graph neural networks. Within this framework, models in a variety of common architectures can be trained from data downloaded from the ColabFit Exchange in pre-formatted Lightning Memory-Mapped Database (LMDB) files and benchmarked against existing models. In addition, improved query and data fetching performance enabled by changes to the data structure and database backend has allowed for the integration of a streaming PyTorch `DataLoader` within KLIFF. During training, batches of data are efficiently fetched and processed on-the-fly directly from the ColabFit exchange. As datasets become larger, and multi-source training becomes more common, we envision this becoming an increasingly important avenue for utilizing data on the ColabFit Exchange.

5 Conclusion

In this work we have provided an overview of the key updates made to the ColabFit infrastructure. The database has undergone significant growth in the past year. This growth can be summarized by the following counts: datasets more than doubled from 139 to over 360; distinct systems more than quadrupled from $\sim$70,000 to $\sim$400,000; total properties increased from $\sim$28 million to over 200 million. In addition, as detailed, improvements to the data standard and the database backend will accelerate further sustained growth and allow for the development of new features and tools. Of particular interest is the development of high-throughput data transformation pipelines, improved inter-dataset analytics, *e.g.* similarity metrics, and new ways to interface and utilize the data for large-scale ML applications. Along these lines, we have developed several new ways to integrate ColabFit data into fitting pipelines, including integration with MatSciML [39] and the addition of a streaming dataloader into KLIFF [37]. We hope that these and future integrations will enable and simplify workflows for leveraging the full power of data maintained on the ColabFit Exchange. We invite the community to upload data by visiting `https://colabfit.org` or the GitHub repository `https://github.com/colabfit/data-lake`, which in turn strengthens our mission to make DDIP data findable, accessible, interoperable, and reusable.

Acknowledgements

This research was supported through the National Science Foundation (NSF) under grant OAC-2039575. S.M. acknowledges the Simons Center for Computational Physical Chemistry for financial support. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

[1] Anubhav Jain, Geoffroy Hautier, Charles J. Moore, Shyue Ping Ong, Christopher C. Fischer, Tim Mueller, Kristin A. Persson, and Gerbrand Ceder. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science*, 50(8):2295–2310, June 2011.

[2] Rickard Armiento, Boris Kozinsky, Marco Fornari, and Gerbrand Ceder. Screening for high-performance piezoelectrics using high-throughput density functional theory. *Physical Review B*, 84(1), July 2011.

[3] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *JOM*, 65(11):1501–1509, 11 2013.

[4] Antoine A. Emery and Chris Wolverton. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites. *Scientific Data*, 4(1), October 2017.

[5] Aini Palizhati, Wen Zhong, Kevin Tran, Seoin Back, and Zachary W. Ulissi. Toward predicting intermetallics surface properties with high-throughput DFT and convolutional neural networks. *Journal of Chemical Information and Modeling*, 59(11):4742–4749, October 2019.

[6] Daniel Wines, Kamal Choudhary, Adam J. Biacchi, Kevin F. Garrity, and Francesca Tavazza. High-throughput DFT-based discovery of next generation two-dimensional (2d) superconductors. *Nano Letters*, 23(3):969–978, January 2023.

[7] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.

[8] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. The open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10), 2021.

[9] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Felix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence Zitnick. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5), 2023.

[10] Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. Machine-learning-assisted determination of the global zero-temperature phase diagram of materials. *Advanced Materials*, 35(22):2210788, 2023.

[11] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, Gus L.W. Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy. Aflowlib.org: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, 6 2012.

[12] Claudia Draxl and Matthias Scheffler. Nomad: The fair concept for big data-driven materials science. *MRS Bulletin*, 43(9):676–682, 9 2018.

[13] Dipendra Jha, Kamal Choudhary, Francesca Tavazza, Wei keng Liao, Alok Choudhary, Carelyn Campbell, and Ankit Agrawal. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature Communications*, 10(1):1–12, 12 2019.

[14] Kamal Choudhary, Kevin F. Garrity, Andrew C. E. Reid, Brian DeCost, Adam J. Biacchi, Angela R. Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A. Gilad Kusne, Andrea Centrone, Albert Davydov, Jie Jiang, Ruth Pachter, Gowoon Cheon, Evan Reed, Ankit Agrawal, Xiaofeng Qian, Vinit Sharma, Houlong Zhuang, Sergei V. Kalinin, Bobby G. Sumpter, Ghanshyam Pilania, Pinar Acar, Subhasish Mandal, Kristjan Haule, David Vanderbilt, Karin Rabe, and Francesca Tavazza. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6(1), November 2020.

[15] Fan-Shun Meng, Jun-Ping Du, Shuhei Shinzato, Hideki Mori, Peijun Yu, Kazuki Matsubara, Nobuyuki Ishikawa, and Shigenobu Ogata. General-purpose neural network interatomic potential for the α-iron and hydrogen binary system: Toward atomic-scale understanding of hydrogen embrittlement. *Physical Review Materials*, 5(11), November 2021.

[16] A. Rohskopf, C. Sievers, N. Lubbers, M.a. Cusentino, J. Goff, J. Janssen, M. McCarthy, D. Montes Oca de Zapiain, S. Nikolov, K. Sargsyan, D. Sema, E. Sikorski, L. Williams, A.p. Thompson, and M.a. Wood. Fitsnap: Atomistic machine learning with lammps. *Journal of Open Source Software*, 8(84):5118, 2023.

[17] Rasha Atwi, Matthew Bliss, Maxim Makeev, and Nav Nidhi Rajput. MISPR: an open-source package for high-throughput multiscale molecular simulations. *Scientific Reports*, 12(1), September 2022.

[18] John L. A. Gardner, Zoé Faure Beaulieu, and Volker L. Deringer. Synthetic data enable experiments in atomistic machine learning, 2022.

[19] Anders S. Christensen and O. Anatole von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces, 2020.

[20] Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9(1), December 2022.

[21] So Takamoto, Chikashi Shinagawa, Daisuke Motoki, Kosuke Nakago, Wenwen Li, Iori Kurata, Taku Watanabe, Yoshihiro Yayama, Hiroki Iriguchi, Yusuke Asano, Tasuku Onodera, Takafumi Ishii, Takao Kudo, Hideki Ono, Ryohto Sawada, Ryuichiro Ishitani, Marc Ong, Taiki Yamaguchi, Toshiki Kataoka, Akihide Hayashi, Nontawat Charoenphakdee, and Takeshi Ibuka. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nature Communications*, 13(1), May 2022.

[22] Xingyi Guan, Akshaya Das, Christopher J. Stein, Farnaz Heidar-Zadeh, Luke Bertels, Meili Liu, Mojtaba Haghighatlari, Jie Li, Oufan Zhang, Hongxia Hao, Itai Leven, Martin Head-Gordon, and Teresa Head-Gordon. A benchmark dataset for hydrogen combustion. *Scientific Data*, 9(1), May 2022.

[23] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1), August 2014.

[24] Yury Lysogorskiy, Cas van der Oord, Anton Bochkarev, Sarath Menon, Matteo Rinaldi, Thomas Hammerschmidt, Matous Mrovec, Aidan Thompson, Gábor Csányi, Christoph Ortner, and Ralf Drautz. Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon. *npj Computational Materials*, 7(1), June 2021.

[25] Penghua Ying, Haikuan Dong, Ting Liang, Zheyong Fan, Zheng Zhong, and Jin Zhang. Atomistic insights into the mechanical anisotropy and fragility of monolayer fullerene networks using quantum mechanical calculations and machine-learning molecular dynamics simulations. *Extreme Mechanics Letters*, 58:101929, January 2023.

[26] Alex M. Maldonado, Igor Poltavsky, Valentin Vassilev-Galindo, Alexandre Tkatchenko, and John A. Keith. Modeling molecular ensembles with gradient-domain machine learning force fields. *Digital Discovery*, 2, May 2023.

[27] Pandu Wisesa, Christopher M. Andolina, and Wissam A. Saidi. Development and validation of versatile deep atomistic potentials for metal oxides. *The Journal of Physical Chemistry Letters*, 14(2):468–475, January 2023.

[28] Office of Science and Technology Policy, Executive Office of the President. Increasing access to the results of federally funded scientific research. `https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf`, February 22, 2013 2013.

[29] Office of Science and Technology Policy, Executive Office of the President. Ensuring free, immediate, and equitable access to federally funded research. `https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf`, August 25 2022.

[30] Joshua A. Vita, Eric G. Fuemmeler, Amit Gupta, Gregory P. Wolfe, Alexander Quanming Tao, Ryan S. Elliott, Stefano Martiniani, and Ellad B. Tadmor. ColabFit exchange: Open-access datasets for data-driven interatomic potentials. *The Journal of Chemical Physics*, 159(15):154802, 10 2023.

[31] Matthias Scheffler, Martin Aeschlimann, Martin Albrecht, Tristan Bereau, Claudia Felser, Mark Greiner, Axel Groß, Christoph Koch, Kurt Kremer, E Wolfgang, Markus Scheidgen, Christof Wöll, and Claudia Draxl. Fair data – new horizons for materials research. *Nature*, 604(April):1–20, 2022.

[32] The vast data platform. `https://www.vastdata.com/whitepaper`. Accessed: 2024-09-06.

[33] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148(24):241733, June 2018.

[34] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, Nov 2022.

[35] Bowen Deng, Peichen Zhong, KyuJung Jun, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. Chgnet: Pretrained universal neural network potential for charge-informed atomistic modeling, 2023.

[36] Leonid Komissarov and Toon Verstraelen. Zeo-1, a computational data set of zeolite structures. *Scientific Data*, 9(1), February 2022.

[37] Mingjian Wen, Yaser Afshar, Ryan S. Elliott, and Ellad B. Tadmor. KLIFF: A framework to develop physics-based and machine learning interatomic potentials. *Computer Physics Communications*, 272:108218, 2022.

[38] E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller, and C. A. Becker. The potential of atomistic simulations and the Knowledgebase of Interatomic Models. *JOM*, 63(7):17, 2011.

[39] Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, Matthew Spellings, Mikhail Galkin, and Santiago Miret. Matsciml: A broad, multi-task benchmark for solid-state materials modeling. *arXiv preprint arXiv:2309.05934*, 2023.

Supplemental Information

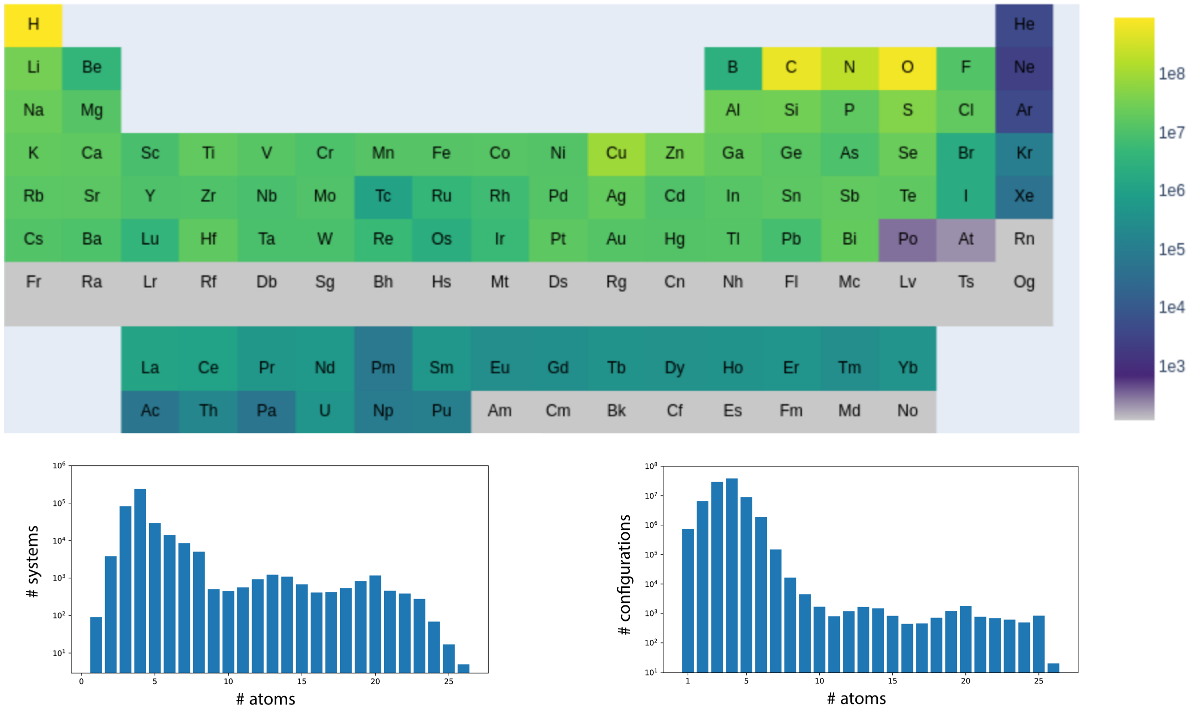

Fig. S1: Chemical composition of the ColabFit Exchange, spanning 91 of the 118 elements on the periodic table, for a total of 395,338 unique chemical systems. After excluding the OpenCatalyst data (which is not represented in this figure), the majority of the database is composed of organic molecules (C, H, and O alone make up ∼67% of the data shown in this figure) due to the relative popularity and availability of molecular datasets. There is currently no data for elements with atomic numbers between 86 and 88, or greater than 94. The bottom panel shows histograms of the number of unique chemical systems (left) or configurations (right) present in the ColabFit Exchange for different numbers of atomic types (i.e., the number of unary/binary/ternary/... systems or configurations). Four datasets account for all data with greater than 10 atom types.

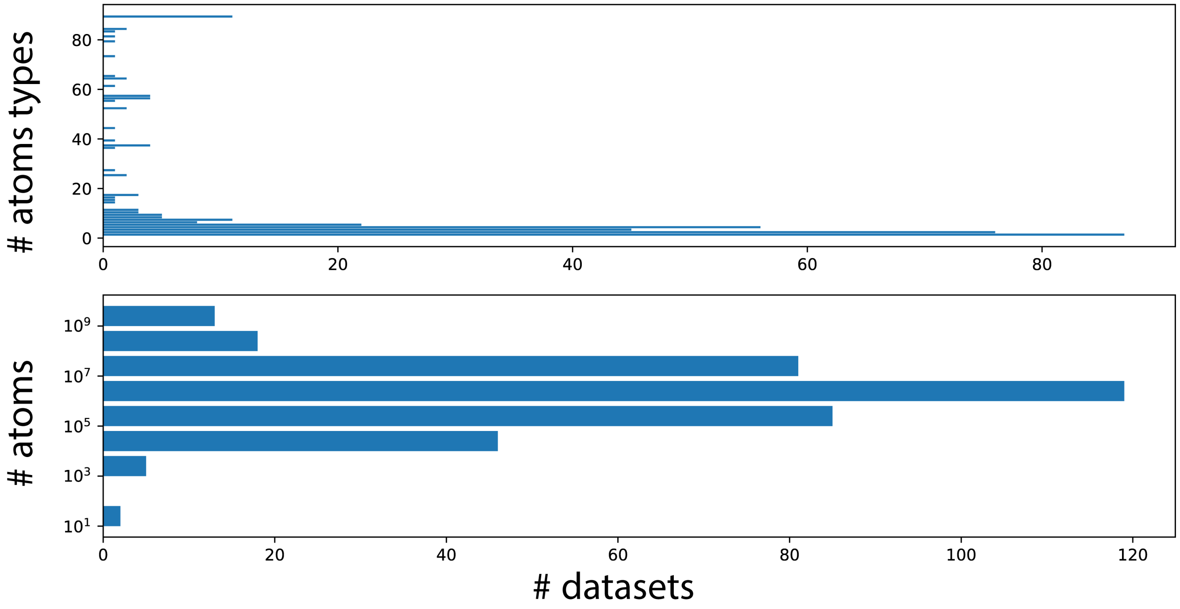

Fig. S2: Histogram showing the sizes of the datasets currently in the ColabFit Exchange. The distribution of the total number of atoms summed over all COs in a given dataset is Gaussian-like, centered about a mean of 10^6.