JARVIS-Leaderboard: A Large Scale Benchmark of Materials Design Methods

Kamal Choudhary,* Daniel Wines, Kevin F. Garrity, Maureen Williams, and Francesca Tavazza
Material Measurement Laboratory, National Institute of Standards and Technology, Maryland, 20899, USA.

Kangming Li and Jason Hattrick-Simpers
Department of Materials Science and Engineering, University of Toronto, 27 King's College Cir, Toronto, ON, Canada.

Vishu Gupta
Department of Electrical and Computer Engineering, Northwestern University, Evanston, Illinois, 60208, USA
Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, 08544, USA and
Ludwig Institute for Cancer Research, Princeton University, Princeton, New Jersey, 08544, USA

Aldo H. Romero
Department of Physics and Astronomy, West Virginia University, Morgantown, WV 26506, USA

Jaron T. Krogel and Kayahan Saritas
Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge,Tennessee 37831,USA

Addis Fuhr and Panchapakesan Ganesh
Center for Nanophase Materials Science, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

Paul R. C. Kent
Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

Keqiang Yan, Yuchao Lin, and Shuiwang Ji
Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA

Ben Blaiszik
Globus, University of Chicago, Illinois, 60637, USA. Data Science and Learning Division, Argonne National Lab, Illinois, 60439, USA.

Patrick Reiser
Institute of Nanotechnology, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany

Pascal Friederich
Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany and
Institute of Nanotechnology, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany

Ankit Agrawal
Department of Electrical and Computer Engineering, Northwestern University, Evanston, Illinois, 60208, USA

Pratyush Tiwary and Eric Beyerle
Department of Chemistry and Biochemistry and Institute for Physical Science and Technology,
University of Maryland, College Park, MD 20742, United States

Peter Minch and Trevor David Rhone
Department of Physics, Applied Physics and Astronomy,
Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Ichiro Takeuchi
Department of Materials Science and Engineering, University of Maryland, College Park, MD 20742, USA

Robert B. Wexler
Department of Chemistry and Institute of Materials Science and Engineering,
Washington University in St. Louis, St. Louis, MO 63130, USA

Arun Mannodi-Kanakkithodi
School of Materials Engineering, Purdue University, West Lafayette, IN, 47907, USA

Elif Ertekin
Department of Mechanical Science and Engineering, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA and Materials Research Laboratory, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA

Avanish Mishra and Nithin Mathew
Theoretical Division (T-1), Los Alamos National Laboratory, Los Alamos, NM, 87545, USA

Mitchell Wood and Andrew Dale Rohskopf
Center for Computing Research, Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

Shih-Han Wang, Luke E. K. Achenie, and Hongliang Xin
Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Adam J. Biacchi
Physical Measurement Laboratory, National Institute of Standards and Technology, Maryland, 20899, USA.

Lack of rigorous reproducibility and validation are significant hurdles for scientific development across many fields. Materials science, in particular, encompasses a variety of experimental and theoretical approaches that require careful benchmarking. Leaderboard efforts have been developed previously to mitigate these issues. However, a comprehensive comparison and benchmarking on an integrated platform with multiple data modalities with perfect and defect materials data is still lacking. This work introduces JARVIS-Leaderboard, an open-source and community-driven platform that facilitates benchmarking and enhances reproducibility. The platform allows users to set up benchmarks with custom tasks and enables contributions in the form of dataset, code, and meta-data submissions. We cover the following materials design categories: Artificial Intelligence (AI), Electronic Structure (ES), Force-fields (FF), Quantum Computation (QC) and Experiments (EXP). For AI, we cover several types of input data, including atomic structures, atomistic images, spectra, and text. For ES, we consider multiple ES approaches, software packages, pseudopotentials, materials, and properties, comparing results to experiment. For FF, we compare multiple approaches for material property predictions. For QC, we benchmark Hamiltonian simulations using various quantum algorithms and circuits. Finally, for experiments, we use the inter-laboratory approach to establish benchmarks. There are 1281 contributions to 274 benchmarks using 152 methods with more than 8 million data-points, and the leaderboard is continuously expanding. The JARVIS-Leaderboard is available at the website: https://pages.nist.gov/jarvis_leaderboard/

* kamal.choudhary@nist.gov

I. INTRODUCTION

The accelerated design and characterization of materials of technological interest has been a rapidly evolving area of research in the last few decades [1]. Materials design requires approaches spanning a variety of length and time scales[2]. For atomistic design, the methods employed may include computational approaches such as density functional theory, tight-binding, force-fields, and highly accurate approaches such as quantum Monte Carlo or quantum computations. A wide range of approaches are employed above the purely atomistic level, such as mesoscale and finite-element methods. Similarly, experimental characterization approaches include X-ray diffraction, vibroscopy, manometry, scanning electron microscopy, and magnetic susceptibility measurements.

Moreover, data produced from these techniques can be of various types: chemical formulae, atomic/micro-structures, images, spectra, and text-documents[4–6]. The data analysis and curation methods add further complexity to benchmarking efforts, which are extremely important [7–18]. For example, more than 70 % of research works were shown to be non-reproducible [19–21], and this number could be much higher depending upon the field of investigation. Although there have been significant advances in individual fields, there is an urgent need to establish a large-scale benchmark for systematic, reproducible, transparent, and unbiased scientific development .

Developing such metrology is a highly challenging task, even for one of these methods, let alone the entire galaxy of available methods. Projects and approaches such as the materials genome and FAIR initiatives [1, 22], have resulted in several well-curated datasets and benchmarks. These, in turn, have led to several materials informatics applications[23–26]. Although electronic structure approaches such as density functional theory (DFT) tend to be more reproducible than other categories [16, 27], a systematic effort must be made to validate these methods and estimate the error in predictions. Hence, it is highly desirable to have a large-scale benchmarking platform in the materials science field for reproducibility and method validation.

Massive progress in fields such as image recognition/image classification (ImageNet [28]), protein structure prediction (AlphaFold [29]), large language modeling (Generative pretrained transformers (GPT)) [30]) has been possible primarily because of well-defined benchmarks in respective fields. With regards to AI methods for structure-to-property predictions [31], benchmarking efforts have enabled drastic improvements in the accuracy of predicted properties (i.e., moving away from descriptor-based predictions and including graph neural networks in the model architectures to improve accuracy).

For deterministic electronic structure methods such as DFT, extensive benchmarking of software and different DFT approximations (functionals, pseudopotentials, etc.) has led to increased reproducibility and precision in individual results and workflows [27, 32]. Such benchmarks allow a wide community to solve problems collectively and systematically. In addition, since there already exists highly accurate models for specific tasks (i.e., energy prediction), more comprehensive evaluations of the models are required so that the performance ranking is not overfitted to one biased data source. We believe that such a universal and large-scale set of benchmarks for materials science will significantly benefit the scientific community.

To this date, several benchmarks of individual methods have already been developed. For artificial intelligence (AI) methods, there have been several benchmarks and leaderboards such as MatBench [33]. MatBench provides a leaderboard for machine learned structure-based property predictions of inorganic materials using 13 supervised machine learning tasks (thermodynamic, tensile, optical, thermal, elastic, and electronic properties) from 10 datasets (including DFT and experiment) [33]. Similar AI benchmarking and leaderboard platforms include MoleculeNet [34], OpenCatalystProject[35], sGDML [36, 37], mLEARN [38], MatScholar [39], and AtomAI [40]. For electronic structure methods, some of the notable benchmarks include the work by Lejaeghere et al.[27], Borlido et al.[41], Huber et al. [42], Zhang et al.[43], Tran et al.[44] and several other projects [45–48]. Other method benchmarks include phase-field benchmarks by Wheeler et al.[49], Lindsay et al. [50], and microscopy benchmarks such as by Wei et al.[51]. A few additional benchmarking studies in materials science include Refs. 52–64. More details on some of these benchmarking efforts are provided in later sections.

The goal of this project is to provide a more comprehensive framework for materials benchmarking than previous works. In particular, most existing efforts: 1) lack the flexibility to readily incorporate new tasks or benchmarks, which is a limitation given the continuous discovery of new materials and quantities in science, 2) are specialized towards a single modality, such as electronic structure, rather than providing a comprehensive framework that can accommodate multiple modalities, 3) offer only a limited set of tasks or properties, 4) are primarily focused on computational methods, overlooking the importance of experimental benchmarking, and 5) make adding contributions to existing platforms rather complex, creating a barrier to entry. In general, there is a need to simplify the process of user contributions to leaderboards to foster broader community engagement.

In this work, we present a user-friendly, comprehensive approach to integrate the benchmarking of both computational, experimental and data-analytics methods. The JARVIS-Leaderboard framework (`https://pages.nist.gov/jarvis_leaderboard/`) covers a variety of categories: Artificial Intelligence (AI), Electronic Structure (ES), Force-field (FF), Quantum Computation (QC), and Experiments (EXP). It also covers various data types, including atomic structures, spectra, images, and text. This project can be used to: (1) check the state-of-the-art methods in respective fields, (2) add a contribution model on an existing benchmark, (3) add a new benchmark, (4) compare new ideas and approaches to well-known approaches. To enhance reproducibility, we encourage each contribution to (1) be from peer-reviewed articles with an associated DOI for all contributions, models, and tools, (2) include a run script to exactly reproduce the results (especially for computational tools), (3) include a metadata file with details such as team name, contact information, computational timing and software (with software

version)/hardware used in order to enhance transparency.

It is important to note differences between a typical data-repository and a benchmarking platform. Some of the key distinguishing factors between a usual large data-repository (such as JARVIS-DFT) and the present leaderboard effort are: 1) the leaderboard contains well-characterized/well-known samples/tasks (i.e., with digital object identifier/peer-reviewed article links) with all the scripts/metadata easily available to reproduce the results rather than just being a look-up table to find data, 2) large data repositories usually contain more variation in materials chemistry/structure and less variation of methods while the leaderboard focuses on a larger number of method comparisons.

For example, the JARVIS-DFT contains DFT data for more than 80,000 materials and millions of material properties with a few specific ES methods and hence there are only a few entries for, say, the electronic bandgap of Silicon from different methods, while the leaderboard contains electronic bandgaps for Silicon using more than 17 ES methods from various contributors. Similarly, JARVIS-ALIGNN project contains AI models for more than 80 properties/tasks of materials, i.e., just one model for a well-known property such as formation energy, while there are more than 12 methods for formation energy task in the leaderboard (as discussed later).

Furthermore, the JARVIS-leaderboard attempts to bridge together multiple categories of methods (AI, ES, FF, QC, EXP) and types of data (single properties, structure, spectra, text, etc.) with the goal of broadening benchmarking efforts across several fields of study. What differentiates the JARVIS-Leaderboard from platforms such as MatBench [33], is that MatBench [33] provides a handful of tasks to evaluate ML methods on larger datasets (i.e. 10^4 entries, most of which are from the Materials Project [65]). A potential drawback of this approach is that the resulting performance rankings could be biased towards the data distribution of a single source. In contrast, the JARVIS-Leaderboard covers a broader range of datasets and properties and provides a better overview of model performance.

Recently in the field of machine learning in materials science, there has been a fixation on performance metrics for newly developed models. This begs the question of whether or not benchmarking can be destructive to the development of new methods if these new methods cannot immediately outperform the previous state-of-the-art approaches. This also begs the question of whether or not benchmarking can lead to overfitting or poor generalization [66, 67].

Therefore, we outline how the leaderboard can also be used to identify and focus on some of the major challenges in different fields, such as: (1) how to evaluate extrapolation capability[68]? (2) why is it difficult to develop a reasonably good AI model with similar accuracy to electronic structure methods?, (3) how can we reduce the computational cost of higher accuracy electronic structure predictions (such as bandgaps and bandoffsets)?, (4) how do we identify examples of materials that require high-fidelity methods (beyond DFT accuracy)?, (5) how can we identify material space where methodological improvements need to be targeted?, (6) how can we establish figures of merits for mesoscale models such as phase field?, (7) how can we make atomistic image analysis quantitative rather than qualitative?, (8) and how do we develop and benchmark multi-modal models (such as text, image, video, atomic structures etc.) [69]?

The JARVIS-Leaderboard is seamlessly integrated into the existing and well-established NIST-JARVIS infrastructure [70, 71], which hosts several datasets, tools, applications, and tutorials for materials design, motivated by the materials genome initiative [1]. The framework is open access to the entire materials science community for progressing the field collectively and systematically. JARVIS (Joint Automated Repository for Various Integrated Simulations) [70, 71] is a repository designed to automate materials discovery and optimization using classical force-field, density functional theory, machine learning calculations, and experiments. Nevertheless, the leaderboard is not limited to NIST-JARVIS infrastructure and can be linked with other external projects as well.

Since its creation in 2017, JARVIS has had over 50,000 users worldwide, over 45 JARVIS-associated articles have been published, and over 80,000 materials currently reside in the database. As these numbers continue to multiply, significant effort on external outreach to the materials science community has been an additional goal of JARVIS, with several events (https://jarvis.nist.gov/events/) such as the Artificial Intelligence for Materials Science (AIMS) and Quantum Matters in Materials Science (QMMS) workshops and hands-on JARVIS-Schools, which have had hundreds of participants throughout the last few years. Based on the level of success and support from the community with regard to the existing JARVIS infrastructure, we believe that the integration of the JARVIS-Leaderboard will have a similar level of engagement and success, with a growing number of contributors from all over the world (in government, academia and industry) and in different sub-fields of materials science.

II. RESULTS AND DISCUSSION

A. Leaderboard overview

At the homepage, information regarding the number of methods, benchmarks, contributions, and datapoints are provided. A snapshot of the homepage with various categories is shown in Fig. 1a. Clicking on one of the entries (or searching in the 'Search' box) such as "formation_energy_peratom" opens a new tab with available contributions. This new tab consists of 1) a description of the benchmark, 2) a plot of various available contributions (as shown in Fig. 1b), 3) explicit table for the plot

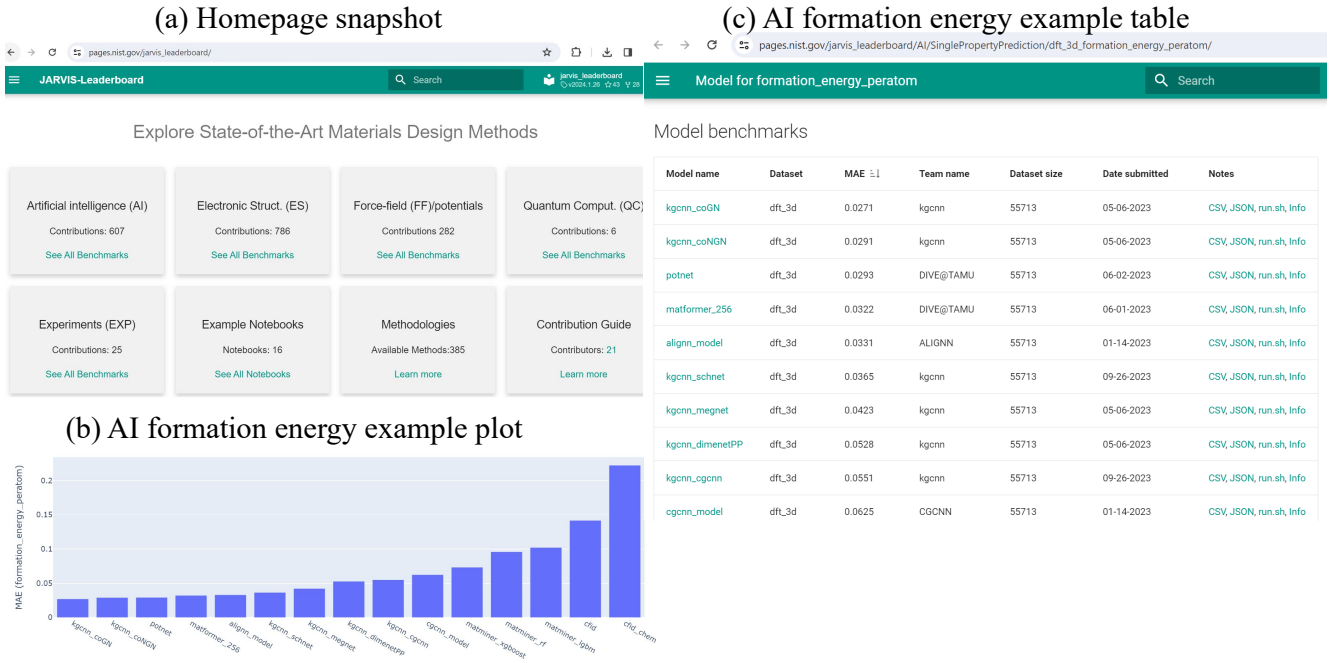

FIG. 1. Leaderboard snapshot with an example output for AI based formation energy per atom model on the JARVIS-DFT (dft_3d) dataset. a) homepage sanpshot showing list of categories and number of available contributions at the time of writing, b) an example AI regression model benchmark for formation energy with several contributions. The methods are sorted based on the mean absolute error (MAE) values. Lower MAE values indicate higher accuracy, c) explicit table for the plot in panel b.Links to individual csv.zip (AI-SinglePropertyPrediction-formation_energy_peratom-dft_3d-test-mae.csv.zip), json.zip (dft_3d_formation_energy_peratom.json.zip), shell script (run.sh) and detailed info (metadata.json) files are provided to help enhance reproducibility. Such results plots and tables are available for each benchmark in the leaderboard.

(as shown in Fig. 1c). For each contribution, links are provided to the submitted data (in .csv.zip format), reference benchmark data (in JSON file), a shell script to reproduce the contribution (run.sh file) and metadata file (metadata.json). The metadata file contains details about the team name, electronic mail address of the contributor(s), DOI number, software (with software version), hardware, instrument, computational timing and other relevant details of a benchmark.

There are several categories for the benchmarks including AI, ES, QC, FF and EXP and their combinations. Some example contributions and a summary table are also provided on the webpage to help a user navigate through the project. The summary table breaks down the available information into categories and sub-categories of different methodologies.

JARVIS-Leaderboard is an evolving project, so additions to the project are anticipated, welcome, and easy to make. We show a general flowchart for adding a new benchmark to the leaderboard in Fig. 2. The user can populate the reference dataset (with well-defined data splits) used for a specific benchmark (e.g. for 2D exfoliation energies in JARVIS-DFT dataset using an AI method: "AI-SinglePropertyPrediction-exfoliation_energy-dft_3d-test"). AI benchmarks have pre-defined training/validation/test identifiers and target data in a corresponding json.zip file, while other methods have only reference test set for evaluation because they do not require model training like an AI method does. For most benchmarks in the leaderboard, experimental data is used as the reference data.

There is a helper script jarvis_populate_data.py to generate a benchmark dataset. A user can apply their method, train models, or run experiments on that dataset and prepare a csv.zip file, a metadata.json file, and also if possible, a conda environment.yaml/Nix/Dockerfile and a run.sh file. This step helps to reproduce the benchmark. These files are kept in a folder with the name of the folder as the team name and can be uploaded to a user's GitHub account by the automated jarvis_upload.py script. This script automatically forks the parent usnistgov/jarvis_leaderboard repo for the user, adds the team-name folder with its files in that forked repo, runs a few minimal sanity checks on the new contribution, and then makes a pull request to the parent repo. The contribution addition and automated testings are carried out using GitHub actions. The administrators of the JARVIS-Leaderboard at NIST will verify the contributions and then finally, it will become part of the leaderboard website.

This project is available on GitHub at: https://github.com/usnistgov/jarvis_leaderboard. The administrators of the JARVIS-Leaderboard at NIST will fully oversee the upload of contributions and benchmarks. A tree structure of the repo is

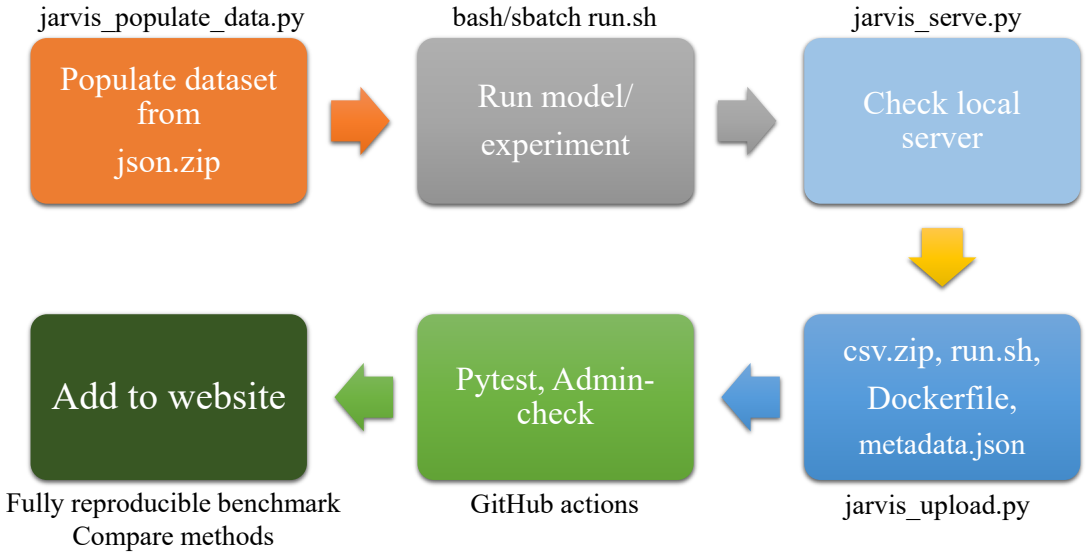

FIG. 2. A flow-chart showing the processes involved in uploading a new contribution to the leaderboard. The jarvis_populate_data.py scripts generate a benchmark dataset. A user can apply their method, train models, or run experiments on that dataset and prepare a csv.zip, a metadata.json file, and other files in a new folder in the contributions directory. The contributions can be locally checked by the user using jarvis_server.py script. Then the folder can be uploaded to a user's GitHub account by the automated jarvis_upload.py script involving several GitHub uploading steps. The administrators of the JARVIS-Leaderboard at NIST will verify the contributions and then finally, it will become part of the leaderboard website.

shown in Fig. 3. There are two main directories in the repo: (1) benchmarks (reference) and (2) leaderboard contributions (for various leaderboard entries), as shown by the green highlighted boxes in Fig. 3.

The "benchmarks" directory has folders for the AI, ES, QC, FF, and EXP categories. Within them, there are sub-folders for specific sub-categories such as (1) SinglePropertyPrediction (where the output of a model/experiment is one single number for an entry), (2) SinglePropertyClass (where the output is class-ids, i.e., 0,1,.. instead of floating values), (3) ImageClass (for multi-class image classification), (4) TextClass (for multi-label text classification), (5) MLFF (machine learning force-field), (6) Spectra (for multi-value data) and (7) EigenSolver (for Hamiltonian simulation). In each of these sub-folders, there are .json.zip files with well-defined reference datasets and available properties as also available in the JARVIS-Tools package `https://jarvis-tools.readthedocs.io/en/master/databases.html`. To avoid storage of large files in the GitHub repo, the actual datasets are part of JARVIS-Tools and are stored in the Figshare repository with specific DOIs and version numbers.

Next, in the "contributions" directory, there is a collection of folders that consist of .csv.zip, metadata.json files, and optionally a Dockerfile and run.sh file. The csv.zip file contains identifier (id) entries and corresponding prediction values obtained by the corresponding model/method. These test identifiers (such as JVASP-1408 in Fig. (3)) must match the test set IDs in the json.zip file in the benchmarks folder for the metric measurements to work. Each of the csv.zip files must contain six components in the filename to place the contribution in the appropriate webpage. The components are the categories (such as AI), sub-categories (such as ImageClass), property (such as bravais_lattice), dataset-name (such as stem_2d_image as available in the JARVIS-Tools database page), and data-split. For entries in the AI category, the data is in train-validation-test splits (using a fixed random number generator). For the current leaderboard format, we report the performance accuracy in the test set only. These files can be easily edited with common text editors. Each contribution folder (e.g. alignn-model) consists of one or several csv.zip files corresponding to each benchmark (such as for formation energies, bandgap, etc.).

Model-specific details are kept in the metadata.json file with *required* keys such as model_name, project_url, team_name and an email address. Users can keep other data such as the uncertainty, time taken, and instrument/software/hardware used in the metadata file as well. For computational models, the run.sh script can be used to reproduce the contributions completely as a single command line script or job submission script. If a method requires additional steps or details beyond a simple command

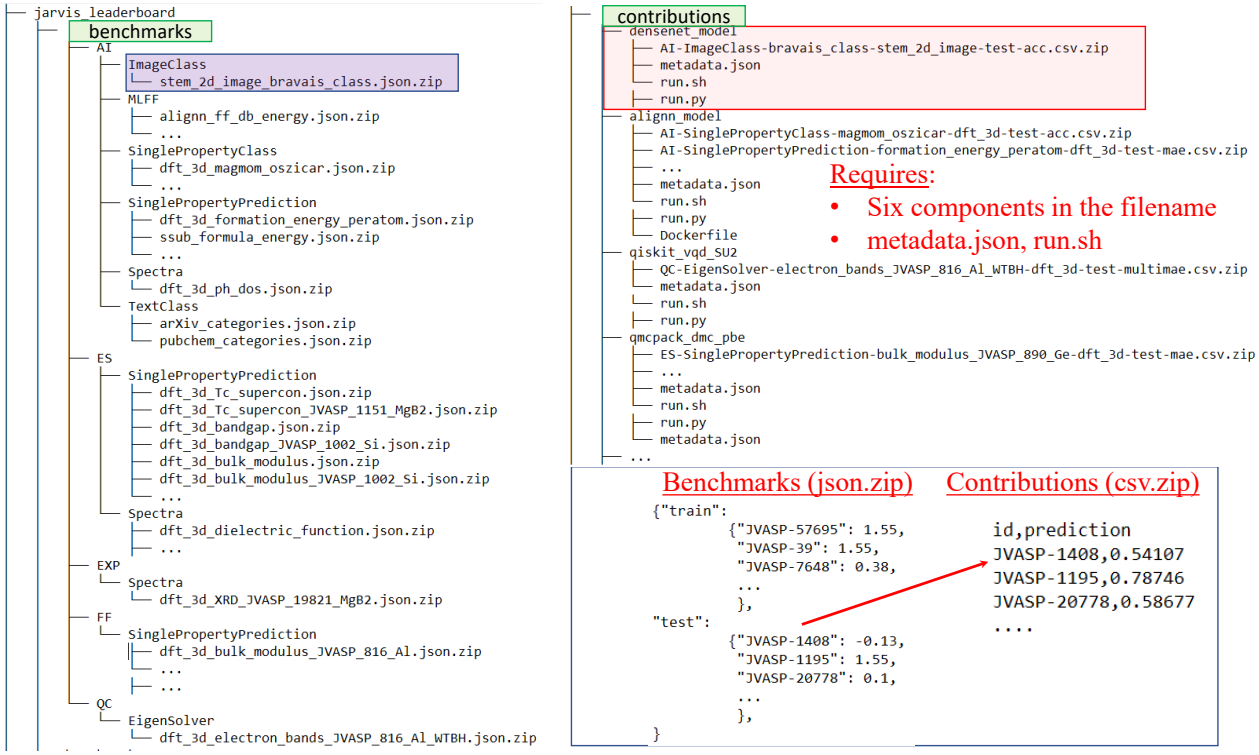

FIG. 3. A tree diagram for directory and file-structure in the leaderboard. There are two main directories in the repo: (1) benchmarks (reference) and (2) leaderboard contributions (for various leaderboard entries). In the "benchmarks" directory, there are folders for the AI, ES, QC, FF, and EXP categories. Within them, there are sub-folders for specific sub-categories. In the "contributions" directory there is a collection of folders that consists of .csv.zip, metadata.json files, and optionally a Dockerfile and run.sh file for available contributions from each method. The csv.zip file contains entries of identifier (id) and corresponding prediction values as obtained by the corresponding model/method. These test identifiers (such as JVASP-1408) must match the test set ids in the json.zip file in the benchmarks folder for the metric measurements to work.

line script, a user can upload a README file containing the additional details. For enhanced reproducibility, we also optionally allow users to include a Dockerfile and an ipython/Google-colab notebook for each benchmark. These notebooks can be used to run the contributions in the Google-cloud without downloading anything locally.

In addition, there is a "docs" directory in the JARVIS-leaderboard. The docs folder consists of a directory structure that is similar to the benchmarks folder with categories names (AI, ES, etc.), and sub-categories (such as SinglePropertyPrediction, ImageClass etc.) with markdown (.md) files that will be converted automatically into corresponding html pages for the website. For each benchmark (i.e., json.zip file), a corresponding docs entry (i.e., md file) should be present. A new benchmark must be associated with a peer-reviewed article and a DOI, in order to have trust in the reference benchmark data. A new benchmark must also be verified by the JARVIS-Leaderboard administrators.

As mentioned above, there already exist several other materials science-specific benchmarks. We compare some of these benchmarks in Table 1 based on the categories that are included. We find that there is no single, large-scale benchmark encompassing the various fields as in the JARVIS-Leaderboard. Also, the data format, metadata, and website for these different leaderboards vary significantly. Hence, having a uniform way to compare different methods would greatly help the materials community.

B. Benchmarks

The benchmarks consist of experimental data, density functional theory, or numerical solutions that are well-known and have already been published in peer-reviewed articles or books. A benchmark should be considered the "ground truth" for a particular task. Therefore, it is mandatory to have a digital object identifier (DOI) for each benchmark from a peer-reviewed article. There can be multiple contributions from different models or experiments for a benchmark, e.g., contributions from various DFT functionals in predicting the electronic bandgap of silicon with respect to experimental data. Typically, for electronic structure (ES) method based contributions, the benchmarks are experimental data; for artificial intelligence (AI) methods, they are the test

TABLE 1. Comparison of benchmark infrastructure available for materials design methods for several categories.

Projects	AI	ES	FF	QC	EXP
MoleculeNet[34]	✓	-	-	-	-
MatBench[33]	✓	-	-	-	-
OpenCatalystProject[35]	✓	-	-	-	-
SciML[72]	✓	-	-	-	-
SGDML[37]	✓	-	-	-	-
GuacaMol[73]	✓	-	-	-	-
Alchemy[74]	✓	-	-	-	-
ML4Chem[75]	✓	-	-	-	-
DGL-LifeSci[76]	✓	-	-	-	-
CCCBDB[77]	-	✓	-	-	✓
Delta-DFT[27]	-	✓	-	-	-
SSSP[78]	-	✓	-	-	-
OpenKIM[79]	-	-	✓	-	-
IPR[80]	-	-	✓	-	-
JARVIS-FF[81]	-	-	✓	-	-
Mlearn[38]	-	-	✓	-	-
QuantumVolume[82]	-	-	-	✓	-
SupermarQ[83]	-	-	-	✓	-
Olympus[84]	-	-	-	-	✓
Golem[85]	-	-	-	-	✓
HTE-MC[86]	-	-	-	-	✓
JARVIS-LB	✓	✓	✓	✓	✓

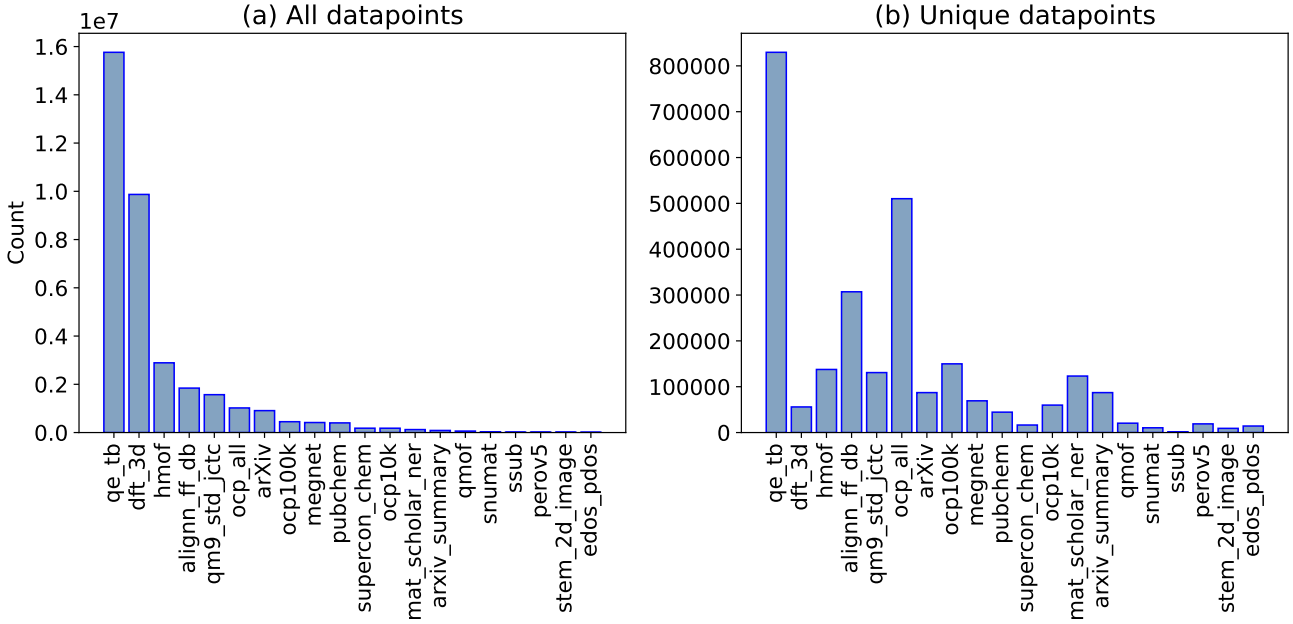

FIG. 4. Distribution of data in each dataset. (a) all entries in leaderboard, (b) entries with unique identifiers. Note that one identifier (such as JVASP-1002 for silicon) can have multiple properties (such as bandgap, bulk modulus etc.). A script to generate this figure is also provided on the leaderboard website as the leaderboard is continuously evolving.

split; for force-field (FF) methods, they are electronic structure data; for quantum computation (QC), they are analytical results; and for experiments (EXP), they are other experiments. Currently, we have more than 270 benchmarks in the leaderboard. The JARVIS-Leaderboard flexible and dynamic nature allows addition of new benchmarks as well.

Each entry in the benchmark dataset consists of a unique identifier. Most of these datasets are integrated into JARVIS-Tools databases page already (but not limited by it), with an associated JARVIS ID number (JID) and are backed up in Figshare,

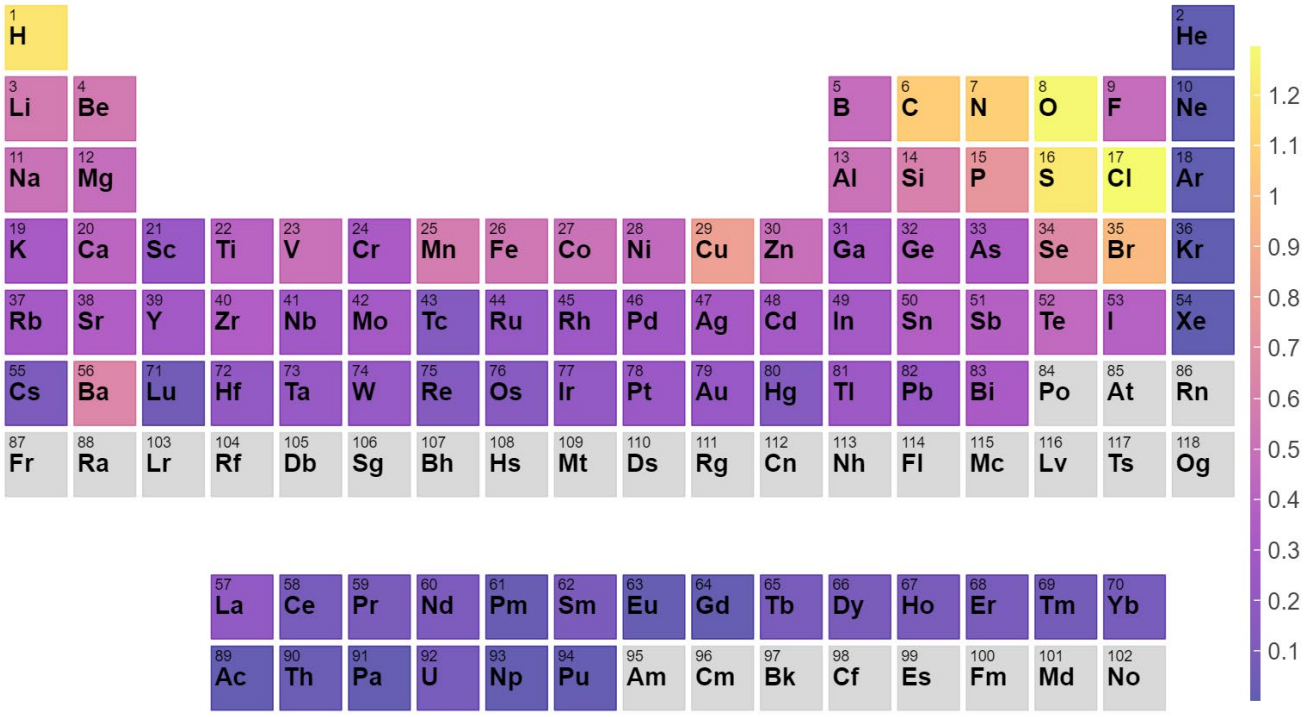

FIG. 5. Periodic table element distribution for entries in all the datasets. This is calculated by taking into account all the element specific entries normalized by total entries i.e. these are percentage probabilities.

Google Drive and NIST-internal storage systems. The number of entries can vary from a few (which is especially applicable for experimental and high-accuracy computational methods, where generating a very large dataset is not feasible in terms of time and resources) to hundreds of thousands of entries in a dataset.

An overview of the dataset can be found in Fig. 4. Considering all possible entries in the dataset, we have close to 7 million datapoints. For example, an atomic structure can have multiple properties calculated, such as bandgaps and formation energies, among other properties. We find the JARVIS-DFT-3D dataset to have the largest number of entries. Considering unique systems, we can find the distribution in Fig. 4b). In this case, qe-tb (fitting dataset for ThreeBodyTB.jl [94]) is one of the largest datasets available in the leaderboard. Note that these datasets contain all varieties of data modalities such as atomic structure, images, spectra and text. In Fig. 5, we show the fractional distribution of periodic table elements in the entire dataset. We find that the most common elements are C, N, O, Cu which is similar to the natural abundance of these elements.

Experimental results are uploaded as benchmarks (i.e. what is regarded as the reference). In the absence of experimental data, high-fidelity computational methods can be used as a reference. If there are multiple experimental measurements available in the literature, each can be individually added as separate benchmarks (i.e., different json.zip files to distinguish one benchmark from another) and users can submit contributions for each of them. As time and the materials science field progresses, certain experimental data may need to be revisited (i.e. more accurate measurements in the future or results are reported that contradict previous experimental data). As a response to this, separate reference (experimental) benchmarks can be added, and users will be able to plot and compare the evolution of these benchmarks over time.

In addition, Leaderboard users can raise an issue on GitHub pertaining to reference benchmarks. The administrators will also upload a README file which contains additional information about the experiments conducted, including associated DOI, experimental conditions and provide details if additional experiments conducted on the same material/property exist in the literature. The experimental conditions described in the README file can be important when comparing the reference benchmark to calculated results, which may be in different conditions than the experiment (i.e. the bandgap of a material is never measured at 0 K, as DFT predicts).

Contributions to the leaderboard in the form of user-submitted experimental data can be compared with previous experiments, electronic structure methods or other numerical results. ES-based contributions are benchmarked against experimental results and can be compared with other ES methods. QC data can be compared with classical computation data or exact analytical results. For FF, contributions can be compared to DFT (or other ES data) or high-level interatomic potential benchmark suites (specifically for MLFFs) [198]. For AI, a test dataset is used. Unlike other methods, AI methods can have both “train” and “test” datasets, while others have only “test” sets in the corresponding dataset. For AI methods, if the “train” dataset is not

provided and only "test" is given, the benchmark can be used for checking extrapolation behavior such as vacancy formation energy benchmarks.

C. Analysis of Benchmarks

Presently, the leaderboard has 5 categories, 10 sub-categories, 152 methods, 274 benchmarks, 1281 contributions and 8714228 datapoints. In this section, we show a few of the hundreds of example analyses that can be carried out using the available benchmarks and contributions. In Fig. 6, we show the MAE of the AI computed formation energy and ES computed bandgap for Si for a variety of contributions in the leaderboard. In Fig. 6(a) we see the comparison of 12 AI models (each AI model had a well-defined 80:10:10 split for training, validation and testing respectively from the JARVIS-3D database) and find the kgcnn_coGN [140] has the highest accuracy/lowest error, followed by Potnet[141], Matformer[142] and ALIGNN[130, 131] models. This can be attributed to the fact that as we include more structural information and use deep-learning methods rather than descriptor methods, we get improvement in accuracy.

Similarly, in Fig. 6(b) we compare the bandgap of Si using several methods and find GLLB-sc [113] calculated with GPAW [112] to yield the lowest error, while G_0W_0 [98] (VASP [87, 88]), GW_0 [98] (VASP [87, 88]), TBmBJ [102, 103] (VASP), and DMC [96] (QMCPACK [114]) methods follow. This can be attributed to the inclusion of the discontinuity potential (GLLB-sc [113]) or kinetic energy density (TBmBJ [102, 103]) in the density functional or incorporating many-body physics (G_0W_0 [98], GW_0 [98], DMC [96]) into the methodology, which can lead to improved accuracy for bandgap prediction. Also, similar methods such as PBE[89] data from Open Quantum Materials Database (OQMD) [199, 200], AFLOW [201] and Materials Project [65] compare well with each other.

In Fig. 6(c) we compare how several classical FFs compute the Voigt bulk modulus of Si. In Fig 6(d) we compare several MLFF models for the forces of Si. We compare various pretrained MLFFs and other MLFFs we specifically trained on the MLEARN [38] dataset (PBE-based DFT data). We see that ALIGNN-FF [170] and MatGL [139] perform similarly for prediction of forces. Fig. 6(d) provides a comprehensive comparison of MLFFs that are trained and tested on the same dataset and pretrained models that were trained elsewhere. The comparisons are presented in tabular form for all the benchmarks on the leaderboard website. We have provided tools and notebooks in the leaderboard GitHub repository that can be used for making such plots for all the available benchmarks and contributions. A collection of such figures for method comparison is available in the supplementary information (Supplementary Figures 1-298). We have also added interactive plots for such comparisons on the website. These tools can aid in identifying examples of materials that require high-fidelity methods beyond the accuracy of DFT in order to understand their underlying properties. In addition, these tools can be used to validate electronic structure methods and provide insight for error estimation.

The leaderboard has a large number of benchmarks and can enable a more comprehensive comparison of different methods for better revealing their respective advantages and limitations. For instance, neural networks outperform descriptor-based models by a large degree in all of the 10 regression tasks in the latest Matbench [33] leaderboard. To check if this is also the case for 44 regression benchmarks in the current JARVIS leaderboard, we compare the performance of the best descriptor-based model to that of the best neural network. As shown in Fig. 7, the best neural network outperform the best descriptor-based model in 34 tasks, but only 14 out of 44 (32 %) tasks see a performance difference by more than 20 %. This indicates that descriptor-based models are still competitive with respect to neural networks, especially considering their better interpretability and orders of magnitude lower training cost [66, 67]. Notably, the best descriptor-based model is found to outperform the best neural network in 10 tasks including those with 10^4-10^5 training data, opening up interesting questions and potential direction to further model improvement. For instance, the inferior performance of neural networks in the regression tasks for the heat capacity and hMOF data may be related to the recently revealed incapability of graph neural networks in capturing periodicity [202].

D. Analysis of Error Metrics

Although a metric such as the MAE can be useful to compare methods for a specific benchmark, it is difficult to compare across different methods, since MAE values can differ substantially. Hence, we use the mean absolute deviation (MAD, computed with respect to the average value of the training data as a baseline/random-guess model) to MAE ratio for both AI and ES single-property-prediction categories. Mean absolute deviation values act as a baseline/random-guessing model for the benchmark and contributed models should have MAE performance better than MAD values. We show the MAD/MAE ratios for AI and ES benchmarks in Fig. 8. We find that the MAD/MAE values range from 2 to 50. MAD/MAE values close to 1 suggest low predictive power. We observe that quantum properties such as the bandgap have lower MAD/MAE than classical quantities (quantities that do not require quantum mechanical simulations) such as total energy or bulk modulus. Interestingly, such trends for classical vs. quantum quantities are observed for both the AI and ES approaches.

E. Interactive View of Benchmarks and Contributions

In addition to making bar plots as shown in Fig. 6 and Fig. 8, the raw data available in benchmarks and contributions can be presented in various other forms such as scatter plots, bandstructures, adsorption spectra, and diffraction spectra. In Fig. 9, we show example comparisons of different methods for AI, ES, QC and EXP categories including (a) formation-energy-per atom model using AI, (b) bulk modulus predictions using ES, (c) electronic bandstructure of Al using QE with different quantum circuits [186], (d) CO_2 capture for zeolite at several labs in round-robin fashion [157]. In Fig. 9a), we find that formation energy is one of the easiest quantities to train AI models and even simple chemistry only-based models can perform reasonably well (i.e., cfid_chem). Including more structural features (such as bond angles and dihedral angles) and using deep learning models (such as graph neural network vs descriptor based models) further helps improve accuracy. Similarly, for ES example for predicting bulk modulus, we find irrespective of DFT based method used, they are in relatively close agreement with experimental bulk modulus data as shown in Fig. 9b). In Fig 9c), we find that the selection of a quantum circuit is critically important for predicting electronic band structures well. Here, we used 6 different quantum [186] circuits and found the SU(2) [152] circuit to compare well with classical computer-based electronic bandstructures. This can be attributed to various entanglements captured in the SU(2) [152] circuits that may be missing in other circuits. Finally, for experimental inter-laboratory/round-robin type measurements of the zeolite CO_2 isotherm, we find excellent agreement across different labs [157].

III. METHODS

The JARVIS-Leaderboard aims to provide a comprehensive framework covering a variety of length and time-scale approaches [2] to enable realistic materials design. In this section, we provide a brief overview of the methods that are currently available in the leaderboard. In this work we use the terms categories, sub-categories, methods, benchmarks, and contributions often, so we define them as follows.

Currently, there are five main “categories” in the leaderboard: Artificial Intelligence (AI), Electronic Structure (ES), Force-field (FF), Quantum Computation (QC), and Experiments (EXP). Each category is divided into “sub-categories", a list of which is provided on the website. These sub-categories include single-property-prediction, single-property-classification, atomic force prediction, text classification, text-token classification, text generation, image classification, image segmentation, image generation, spectra-prediction, and eigensolver. These sub-categories are highly flexible and new categories can be easily added. “Benchmarks" are the reference data (in the form of json.zip file, discussed later) used to calculate performance metrics for each specific contribution. “Methods" are a set of precise specifications for evaluation against a benchmark. For example, within the ES category, density functional theory (DFT) performed with the specifications of the Vienna Ab initio Simulation Package (VASP)[87, 88], Perdew-Burke-Ernzerhof (PBE) [89] functional and PAW [87, 88] pseudopotentials (VASP-PBE-PAW) is a method. Similarly, within the AI category, descriptor/feature-based models with specifications of MatMiner [90] chemical features and the LightGBM [91]software is a method. “Contributions" are individual data (in the form of csv.zip files) for each benchmark computed with a specific method. Each contribution files consist of six components: category (e.g. AI), sub-category (e.g. SinglePropertyPrediction), property (e.g. formation energy), dataset (e.g. dft_3d), data-split (e.g. test), metric (e.g. mae).

A. Electronic structure

Electronic structure approaches cover short length scales and short time scales with high-fidelity. There are a variety of ES methodologies such as such as tight-binding [92–94], density functional theory (DFT)[95], quantum Monte Carlo [96], dynamical mean field theory [97] and many-body perturbation theory (Green’s function with screened Coulomb potential, GW methods)[98]. For each of the methodologies, there are a number of specifications to completely describe a method including the exact software, exchange-correlation functional, pseudopotential, and other relevant parameters. Example methods used in this work are given in Table 2.

Each method in the ES category can have a variety of contributions. For example, using a specific method, one can calculate various properties such as bandgaps, formation energies, bulk moduli, solar cell efficiencies, and superconducting transition temperatures as well as spectral quantities such as dielectric functions. While there are more than 400 approximate exchange-correlation functionals proposed in DFT literature [159], currently, we have OptB88vdW [100], Opt86BvdW [101], LDA [99], PBE[89], PBEsol [108], GLLB-sc [113], TBmBJ [102, 103], SCAN [104], r2SCAN [105], HSE06 [106], in the leaderboard. We use converged k-points and cut-offs as available in the JARVIS-DFT database [160]. We have used the Vienna Ab initio Simulation Package (VASP) [87, 88], ABINIT [109–111], GPAW [112] and Quantum Espresso (QE) [107] as DFT software packages, but other packages can be easily added as well. In addition, we use VASP [87, 88] to perform GW calculations including “single-shot" G_0W_0 and self-consistent GW_0 methods [98]. Other ES approaches include tight-binding (TB) [92] and quantum Monte Carlo (QMC) [96]. For TB, we use the recently developed ThreeBodyTB.jl code[94] along with the Wannier90 [115] code, while the QMCPACK [114] code is used for diffusion Monte Carlo (DMC) [96] calculations.

TABLE 2. Summary of current benchmark categories and methods available in the JARVIS-Leaderboard at the time of writing. More details can be found in the individual metadata.json file. Note that the number of methods is continuously growing.

Category	General name	Method Specification
ES	DFT [99]	VASP[87, 88] (PBE[89], LDA [99], OptB88vdW [100], Opt86BvdW [101], TBmBJ [102, 103], SCAN [104], r2SCAN [105], HSE06 [106])
		QE [107] (PBE[89], PBEsol [108])
		ABINIT [109–111] (PBE[89])
		GPAW [112] (PBE[89], LDA [99], GLLB-sc [113])
	QMC[96]	QMCPACK[114] (DMC[96])
	GW[98]	VASP[87, 88] (G_0W_0 [98], GW_0 [98])
	TB [92]	ThreeBodyTB.jl [94] (Wannier90 [115])
AI	Descriptor	CFID [116], MagPie [117], MatMiner [67, 90], crystal feature model [118], ElemNet [119–122], IRNet[123–125], BRNet[126–128], SNAP [129]
,	Graph-based	ALIGNN [130, 131], CGCNN [132], SchNet [133], AtomVision [134], ChemNLP [135], DimeNet+ [136, 137], CHGNet [138], M3GNET [139] kgcnn_coGN [140], Potnet[141], Matformer[142]
	Transformers	OPT [143], GPT [30], T5 [144]
FF	LJ [145]	LAMMPS [146] (2D-Liquid)
	EAM [147]	LAMMPS [146] (FCC-Al)
	REBO [148]	LAMMPS [146] (Diamond-Si)
	AMBER99sb-ildn [149]	GROMACS [150] (Alanine dipeptide)
	CHARMM36m [151]	GROMACS [150] (α-aminoisobutyric acid)
QC	Algorithms	Qiskit [152] (VQE [153], VQD [154])
		PennyLane [155, 156] (VQE [153], VQD [154])
	Circuits	Qiskit [152] (PauliTwo Design [152], SU(2) [152])
EXP	Diffraction	XRD (Bruker D8)
	Manometry	CO_2 adsorption FACT lab [157]
	Vibroscopy	Kevlar FAVIMAT [158]
	Magnetometry	Susceptibility (PPMS) [158]

B. Force-field

Force fields can be used in molecular dynamics and Monte Carlo simulations for studying larger time and length scales compared to electronic structure methods. Traditional force fields are developed for specific chemical systems and applications and may not be transferable to other uses. It is important to check the validity of an FF before using it in a particular application. Moreover, the development of FFs is a cumbersome task. Examples of typical FFs include embedded-atom method (EAM) potentials [147] (i.e. Al099.eam.alloy for aluminum system [161]), Lennard Jones (LJ)[145] for 2D liquids, reactive empirical bond order (REBO)[148] for Si, and classical, atomistic force fields for biomolecular systems[162, 163]. Recently, machine learning force fields (MLFF)[164–169] have become popular because of their higher accuracy and ease of development (such as SNAP [129] FFs). Nevertheless, early generations of MLFFs were also developed for specific types of chemistry and applications. Very recently, several MLFFs have been developed that can be used to simulate any combination of periodic table elements. Some of these FFs include M3GNET [139], ALIGNN-FF [170], and CHGNet [138]. In the leaderboard, we include benchmarks for energies, forces, and stress tensors for both specific systems and universal datasets.

Traditional FFs are available in LAMMPS [146], while MLFFs are integrated into the Atomic Simulation Environment (ASE) [171] package. Some of these MLFFs are now available in LAMMPS and other large-scale MD codes. In addition to static quantities, FFs can be used for Monte Carlo simulations, such as CO_2 adsorption in metal-organic frameworks (MOFs) [172] using the RASPA [173] code. In addition to energy, force, and stress, we also have FF benchmarks for classical properties such as the bulk modulus. For biomolecular systems, GROMACS[174] is commonly used, and we present here free energy differences and conformational state population benchmarks for three model peptides[175–177].

C. Artificial intelligence

Recently artificial intelligence methods have become popular for materials prediction across all lengths and time scales. We currently have benchmarks for four types of data used as input for the AI models: (1) atomic structure, (2) spectra, (3) images, and 4) text. AI techniques can be used for both forward prediction and inverse design. For atomic structure datasets, we use DFT datasets such as JARVIS-DFT[70, 71], Materials Project (MP) [65], Tight binding three-body dataset (TB3) [94], Quantum-Machine 9 (QM9) [178, 179]. For spectral data, we use either DFT-based spectra of, for example, electron or phonon density

of states (DOS), Eliashberg functions, or numerical XRD spectra. For images, we have simulated and experimental scanning transmission electron microscope (STEM) and scanning tunneling microscopy (STM) images for 2D materials. For text data, we have used the publicly available arXiv dataset.

Currently, we have models for feature-based/tabular models (such as RandomForest [180], Gradient boosting [180], Linear regression [180]), graph based models (such as ALIGNN [130, 131], SchNet [133], CGCNN [132], M3GNET [139], AtomVision [134], ChemNLP [135]) as well as transformers (such as OPT [143], GPT [30], and T5 [144]). These models use popular AI code bases including PyTorch [181], scikit-learn [180], TensorFlow [182], LightGBM [91], JAX [183], and HuggingFace [184]. These models are used for a variety of properties such as formation energies, electron bandgaps, phonon spectra, forces, text data etc.

D. Quantum computation

Quantum chemistry is one of the most promising applications of quantum computations [185]. Quantum computers with relatively few logical qubits can potentially exceed the performance of much larger classical computers because the size of Hilbert space increases exponentially with the number of electrons in the system. Predicting the energy levels of a Hamiltonian is a typical and fundamentally important problem in quantum chemistry. We use Hamiltonian simulations with quantum algorithms and compare it with classical solvers. Determination of appropriate quantum circuit for a specific QC problem is a challenging task. For example, we use the tight-binding Hamiltonians for electrons and phonons in JARVIS-DFT and evaluate the electron bandstructures using quantum algorithms (such as variational quantum eigen solver (VQE) [153] and variational quantum deflation (VQD) [154]) and with different quantum circuits (such as PauliTwo design [152] and SU(2) [152] circuits). We primarily use the Qiskit [152] software in this work through the JARVIS-Tools/AtomQC [186] interface, but other packages such as Tequila [187], Circq [188], and Pennylane [155, 156] can also be easily integrated. In addition to studying algorithm and circuit architecture dependence, the leaderboard can be used for studying the noise-levels in quantum circuits across different quantum computers, which is a key issue hindering quantum computer commercialization. Currently, we are only using statevector simulators for the quantum algorithms available in the Qiskit [152] library.

E. Experiments

Although experimental results for material properties and spectra are referenced in comparison to computational methods (within the JARVIS-Leaderboard and other leaderboards such as MatBench [33]), we dedicated a portion of the JARVIS-Leaderboard to experimental benchmarking. Benchmarking experiments essentially boils down to the comparison of different experiments for the same desired result/s. A systematic way to perform this benchmarking is through round-robin testing [189]. This is an inter-laboratory test performed independently several times, which can involve multiple scientists and a variety of methods and equipment. This approach has been applied successfully for a range of materials science applications [157, 190–193], but many more of such experiments are still needed. Specifically in the JARVIS-Leaderboard, we include experimental round-robin results for manometric measurements of CO_2 adsorption [157]. It is important to note that the experimental results included in the leaderboard are for well-characterized materials with well-defined properties and phenomena that can be easily reproduced (in contrast to replication attempts of variable experiments, such as the recent attempt to synthesize room temperature superconductors [194–197]). Some of the experiments we used for benchmarking purposes are XRD, magnetometry, vibroscopy, and scanning electron microscopy (SEM) and transition electron microscopy (TEM). We purchase the samples from industrial vendors with available identifiers such as CAS-number. We also carried out XRD for MgB_2 (a superconducting material) to verify its crystal structure before carrying out magnetometry measurements to determine the transition temperature. This measurement was compared with numerical XRD data. Magnetometry measurements for superconductors were also conducted to compare their superconducting transition temperatures with respect to predicted or experimentally available values [158]. Strain-stress measurements were done for Kevlar for failure analysis [158]. We have several instruments such as Bruker D8, Titan, Quantum design PPMS and FAVIMAT in the leaderboard currently.

F. Metrics used

We use several metrics in the leaderboard depending on the "sub-categories" mentioned above. We use mean absolute error (MAE), accuracy (acc), multi-mae (L1 norm of multi dimensional data), recall-oriented understudy for gisting evaluation (ROUGE) for the singlepropertyprediction, singlepropertyclassification, spectra/eigensolver/atomic forces and textGen/textsummary subcategories respectively. As the user contributes their data to compare against the reference data (benchmarks), other complementary metrics (such as those available in the sklearn.metrics library) can be easily calculated as the raw contribution data is also made available through the website. For the sake of readability and ease of use, we primarily

employ the metrics mentioned above. For single property prediction, there is only scalar values per column in the csv.zip file with id and prediction separate by comma., For spectra, force-prediction and other multi-value quantities (i.e.with multiple prediction values per id) we concatenate the array and separate by semicolon (to avoid comma convention in csv files). The benchmark data is also stored in a similar format. We provide tools to convert these csv.zip files into json or other file formats if needed. We also provided notebooks to visualize the data through Jupyter/Colab notebooks. In addition, we plan to eventually add metrics for timing, uncertainty, development cost and other details.

ACKNOWLEDGEMENTS

K.C., D.W., K.F.G., A.F., A.J.B., M.W., and F.T. thank the National Institute of Standards and Technology for funding, computational, and data-management resources. This work was performed with funding from the CHIPS Metrology Program, part of CHIPS for America, National Institute of Standards and Technology, U.S. Department of Commerce. K.C. thanks the computational support from XSEDE (Extreme Science and Engineering Discovery Environment) computational resources under allocation number TG-DMR 190095. Contributions from K.C. were supported by the financial assistance award 70NANB19H117 from the U.S. Department of Commerce, National Institute of Standards and Technology. J.T.K., K.S., P.G. and P.R.C.K. were supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, as part of the Computational Materials Sciences Program and Center for Predictive Simulation of Functional Materials. A.F. and P. G. were supported by the Center for Nanophase Materials Sciences, which is a US Department of Energy, Office of Science User Facility at Oak Ridge National Laboratory. AHR thanks the Supercomputer Center and San Diego Supercomputer Center through allocation DMR140031 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. AHR also recognizes the support of West Virginia Research under the call Research Challenge Grand Program 2022 and NASA EPSCoR Award 80NSSC22M0173. N.M. and A.M. acknowledge support from the U.S. Department of Energy through the LANL LDRD Programs under grant no. 20210036DR and 20220814PRD4, respectively. V.G. and A.A. were supported by NIST award 70NANB19H005 and NSF award CMMI-2053929. S.H.W. especially thanks to the NSF Non-Academic Research Internships for Graduate Students (INTERN) program (CBET-1845531) for supporting part of the work in NIST under the guidance of K.C. A.M.K. acknowledges support from the School of Materials Engineering at Purdue University under startup account F.10023800.05.002. P.F. acknowledges support by the Federal Ministry of Education and Research (BMBF) under Grant No. 01DM21001B (German-Canadian Materials Acceleration Center).

Please note certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply the recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan). The Los Alamos National Laboratory is operated by the Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

IV. DATA AVAILABILITY STATEMENT

Multiple datasets used in this work are available at the Figshare repository: https://figshare.com/authors/Kamal_Choudhary/4445539. Index and usage guidelines are provided at https://pages.nist.gov/jarvis/databases/.

V. CODE AVAILABILITY STATEMENT

JARVIS-Leaderboard package mentioned in the article can be found at https://github.com/usnistgov/jarvis_leaderboard.

VI. COMPETING INTERESTS

The authors declare no competing financial or non-financial interests.

VII. AUTHOR CONTRIBUTIONS

K.C. conceived of the idea, developed the workflow, designed several benchmarks and oversaw the project. K.C., D.W., K.L., K.F.G. and V.G. wrote the first draft of the manuscript. K.C., D.W., K.L., K.F.G., V.G., A.H.R., J.T.K., K.S., A.F., R.W., A.M.K., K.Y., Y.L., P.R., A.M., S.H.W., E.B., A.D.R., T.D.R., A.J.B., F.T. uploaded benchmarks and contributions to the leaderboard. All authors contributed in editing and revising the manuscript. List of contributors to the GitHub repository for this work is also available at: https://github.com/usnistgov/jarvis_leaderboard/graphs/contributors.

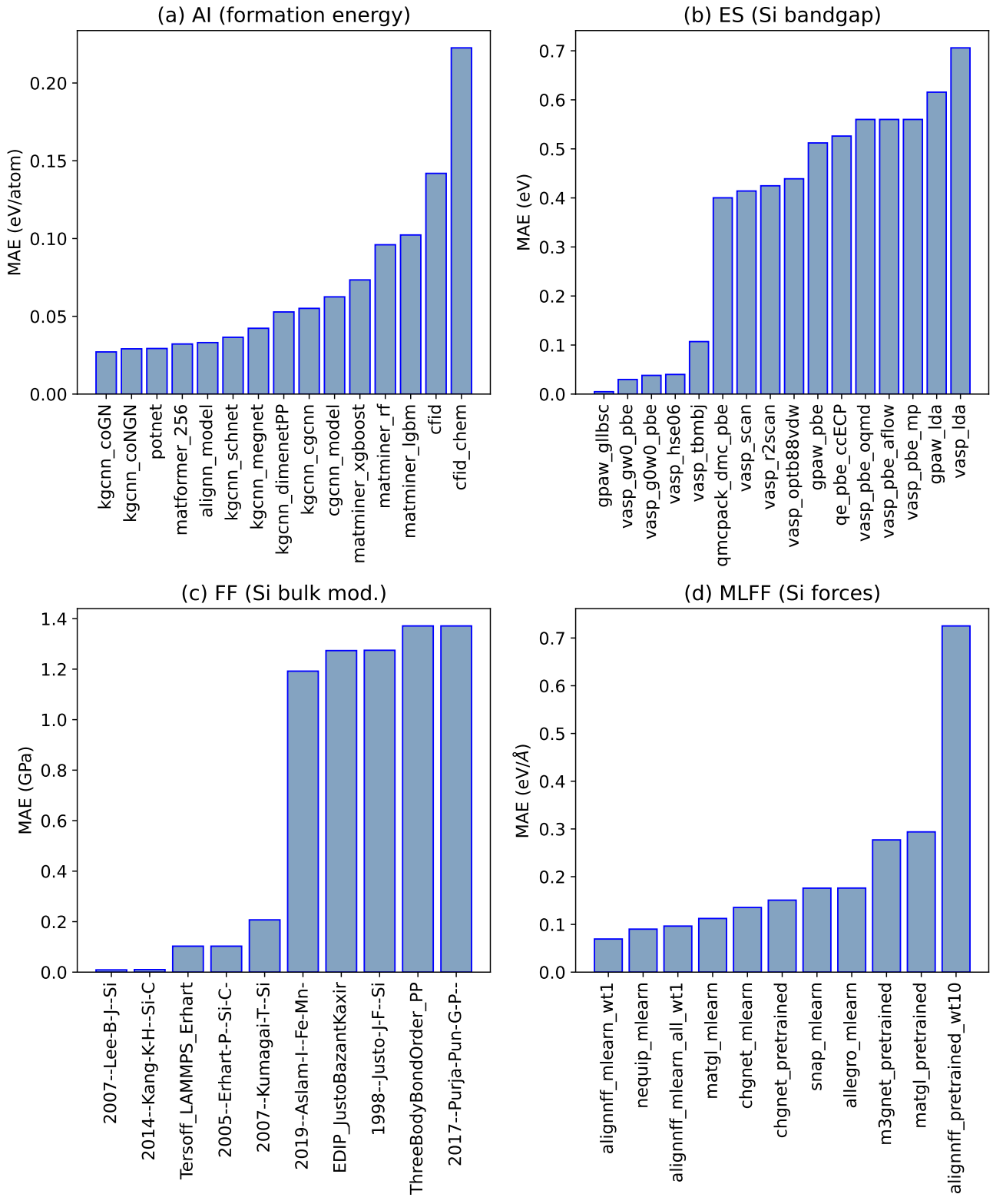

FIG. 6. Example mean absolute errors for benchmarks including (a) artificial intelligence (AI) formation energy for test set with 5572 materials in JARVIS-DFT 3D dataset, (b) electronic structure (ES) Si (JARVIS-DFT ID: JVASP-1002) bandgap, (c) classical force-field (FF) based Voigt bulk modulus of Si and (d) machine learning force-field (MLFF) based forces for Si. We provide Jupyter/Google colab notebooks to easily plot such comparisons for all available benchmarks. Also, similar analysis figures for all the available benchmarks are available in the supplementary information (Supplementary Figures 1-298). As a note, these plots are a current snapshot of the leaderboard, and it is possible that new and more accurate models will be developed and added here in the future.

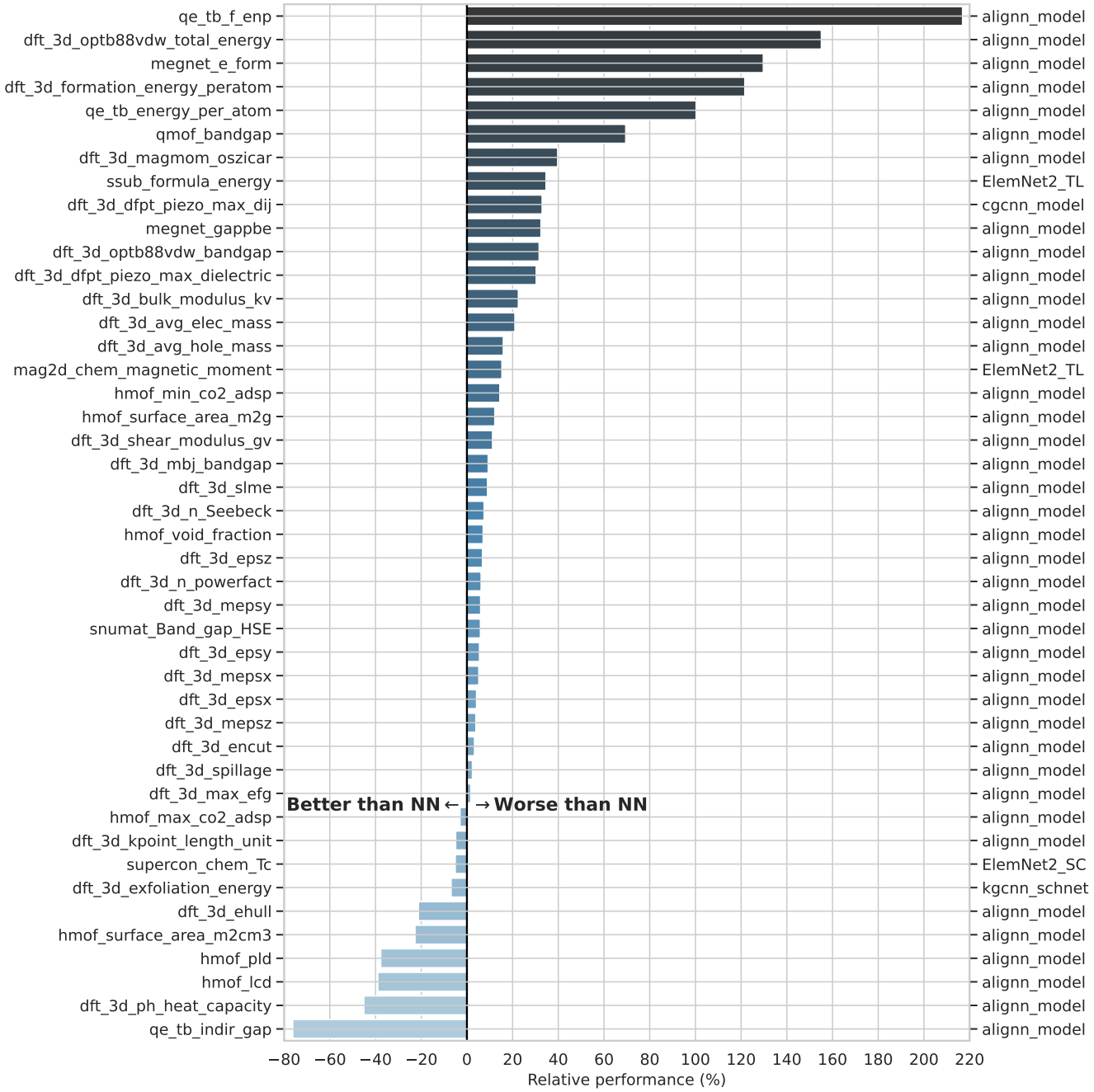

FIG. 7. Relative performance computed as the ratio of the MAE difference between the best descriptor-based model and the best neural networks to the MAE of the best neural networks in the AI regression benchmarks. The benchmark name and the corresponding best performing neural network are indicated in the left and right y axis, respectively. For all the considered AI benchmarks, the best descriptor-based model is the tree-based model using Magpie [117] and Voronoi-tessellation [203] features. As a disclaimer, these plots are a current snapshot of the leaderboard, and it is possible that new and more accurate models will be developed in the future.

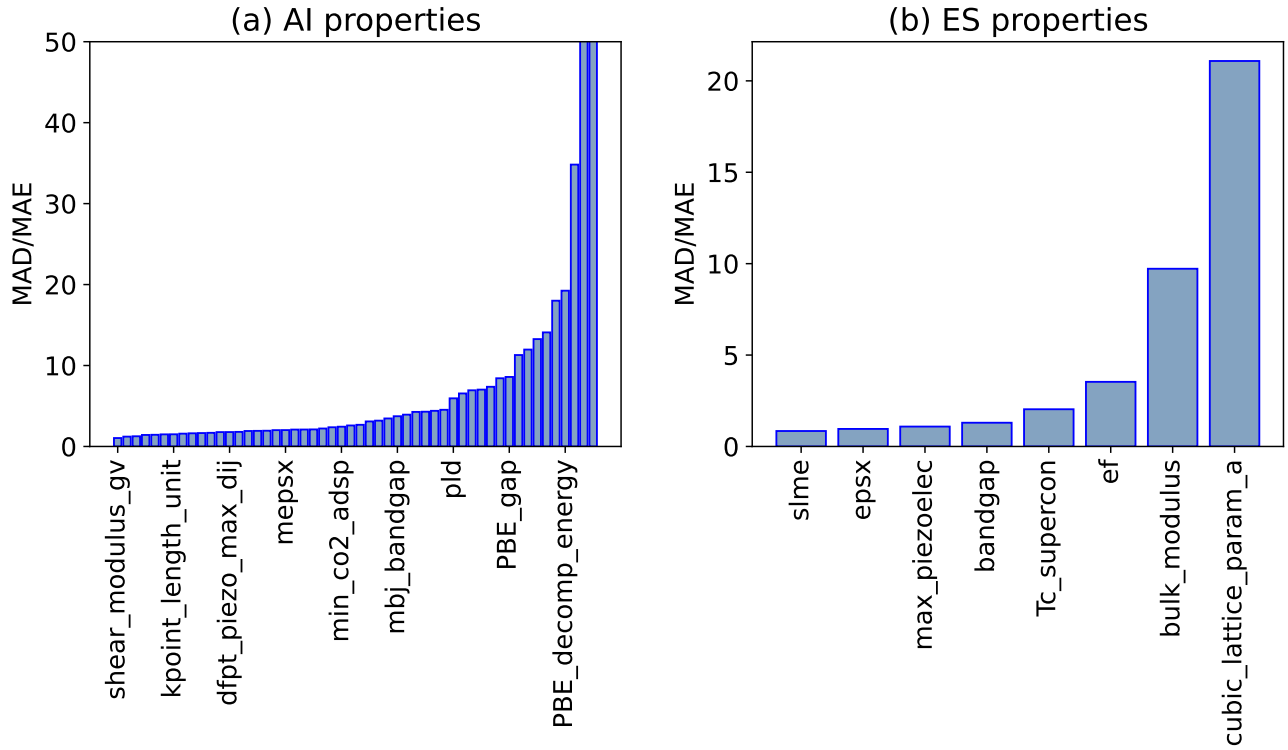

FIG. 8. Mean absolute deviation (MAD) to mean absolute error (MAE) ratio for (a) AI and (b) electronic structure methods. MAD:MAE serves as uniform criteria for comparing performances of models.

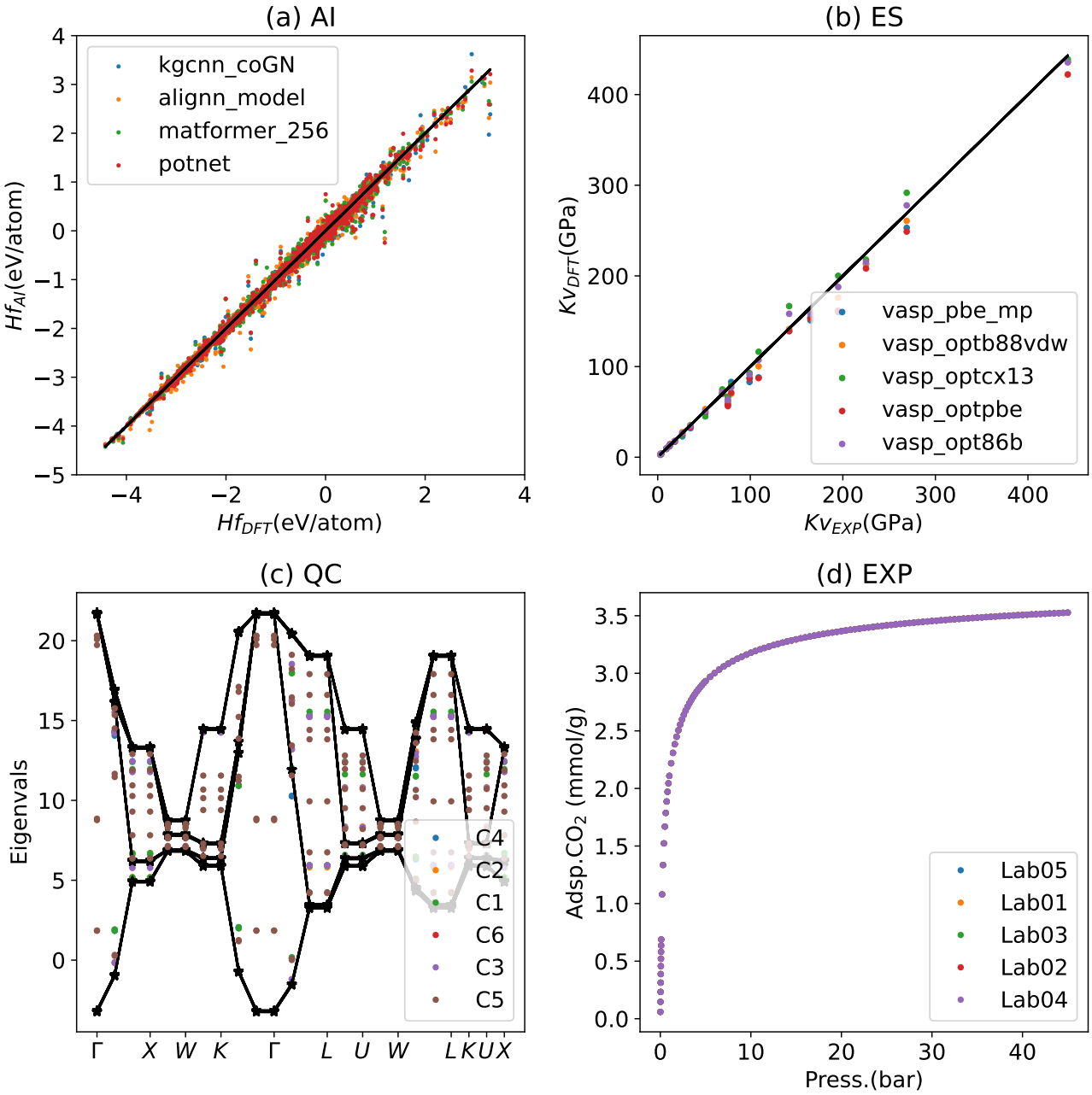

FIG. 9. Example results for AI, ES, QC and EXP results. (a) formation-energy-per atom model using AI for JARVIS-DFT 3D dataset with 5572 materials in the test set, (b) bulk modulus predictions using ES methods for 21 materials, (c) electronic bandstructure of Aluminum using QC methods with different quantum circuits on a coarse k-point mesh, (d) CO_2 capture for zeolite (ZSM-5) at several labs in inter-laboratory/round-robin fashion.

VIII. REFERENCES

[1] C. H. Ward and J. A. Warren, *Materials genome initiative: materials data* (US Department of Commerce, National Institute of Standards and Technology, 2015).

[2] W. D. Callister *et al.*, *Fundamentals of materials science and engineering*, Vol. 471660817 (Wiley London, 2000)
.

[3] L.-Q. Chen, "Phase-field models for microstructure evolution," Annu. Rev. Mat. Res. **32**, 113–140 (2002),

[4] A. Agrawal, K. Gopalakrishnan, and A. Choudhary, "Materials image informatics using deep learning," in *Handbook on Big Data and Machine Learning in the Physical Sciences: Volume 1. Big Data Methods in Experimental Materials Discovery*, World Scientific Series on Emerging Technologies, edited by "" ("WorldScientific, 2020) pp. 205–230.

[5] K. Choudhary *et al.*, "Recent advances and applications of deep learning methods in materials science," npj Comp. Mat. **8**, 59 (2022).

[6] D. J. Audus, K. Choudhary, B. L. DeCost, A. G. Kusne, F. Tavazza, and J. A. Warren, "Artificial intelligence for materials," in *Artificial Intelligence for Science*, Chap. Chapter 23, pp. 413–430.

[7] C. F. Camerer, et al., "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," Nat. Hum. Behav. **2**, 637–644 (2018).

[8] Daniele Fanelli, "Is science really facing a reproducibility crisis, and do we need it to?," Proc. Nat. Acad. Sci. **115**, 2628–2631 (2018).

[9] Zhu Sun, et al., "Are we evaluating rigorously? Benchmarking recommendation for reproducible evaluation and fair comparison," Proc. 14th ACM Conf. on Recomm. Sys. (2020).

[10] Valentin Amrhein, Fränzi Korner-Nievergelt, Tobias Roth, "The earth is flat (p> 0.05): significance thresholds and the crisis of unreplicable research," PeerJ **5**, e3544 (2017).

[11] David Robert Grimes, Chris T. Bauch, John P.A. Ioannidis, "Modelling science trustworthiness under publish or perish pressure," Roy. Soc. Open Sci. **5.1**, 171511 (2018).

[12] Eric M. Prager, et al., "Improving transparency and scientific rigor in academic publishing," J. Neuro. Res. **97.4**, 377–390 (2019).

[13] Anastasios G. Papadiamantis, et al., "Metadata stewardship in nanosafety research: Community-driven organisation of metadata schemas to support FAIR nanoscience data," Nanomat. **10.10**, 2033 (2020).

[14] Hao-Nan Zhu, Cindy Rubio-González, "On the reproducibility of software defect datasets," 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE) IEEE, 2023.

[15] Franklin Sayre, Amy Riegelman, "The reproducibility crisis and academic libraries," Coll. and Res. Lib. **79.1**, 2 (2018).

[16] Susi Lehtola, Miguel A. L. Marques, "Reproducibility of density functional approximations: How new functionals should be reported," J. Chem. Phys. **159**, 114116 (2023).

[17] Anastasios G. Papadiamantis, Logan Ward, Jason Hattrick-Simpers , "Metadata stewardship in nanosafety research: Community-driven organisation of metadata schemas to support FAIR nanoscience data," Dig. Disc. **3**, 281-286 (2024).

[18] Genevera I. Allen, Luqin Gan, Lili Zheng, "Interpretable Machine Learning for Discovery: Statistical Challenges and Opportunities," Ann. Rev. Stat. and App. **11** (2023).

[19] J. Park, J. D. Howe, and D. S. Sholl, "How reproducible are isotherm measurements in metal–organic frameworks?" Chem. Mat. **29**, 10487–10495 (2017).

[20] M. Baker, "1,500 scientists lift the lid on reproducibility," Nature **533**, 452–454 (2016)
.

[21] M. Hutson, "Artificial intelligence faces reproducibility crisis," Science **359**, 725–726 (2018)
.

[22] M. D. Wilkinson *et al.*, "The fair guiding principles for scientific data management and stewardship," Sci. Data **3**, 1–9 (2016).

[23] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," APL Mat. **4**, 053208 (2016).

[24] J. Rickman, T. Lookman, and S. Kalinin, "Materials informatics: From the atomic-level to the continuum," Acta Mat. **168**, 473–510 (2019).

[25] A. Agrawal and A. Choudhary, "Deep materials informatics: Applications of deep learning in materials science," MRS Comm. **9**, 779–792 (2019).

[26] V. Gupta, W.-k. Liao, A. Choudhary, and A. Agrawal, "Evolution of artificial intelligence for application in contemporary materials science," MRS Comm. **13**, 754–763 (2023).

[27] K. Lejaeghere *et al.*, "Reproducibility in density functional theory calculations of solids," Science **351**, aad3000 (2016).

[28] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," Int. J. Comp. Vis. **115**, 211–252 (2015).

[29] J. Jumper *et al.*, "Highly accurate protein structure prediction with alphafold," Nature **596**, 583–589 (2021).

[30] T. Brown *et al.*, "Language models are few-shot learners," Adv. in Neur. Info. Proc. Sys. **33**, 1877–1901 (2020).

[31] X. Zhang *et al.*, "Artificial intelligence for science in quantum, atomistic, and continuum systems." Preprint at https://arxiv.org/abs/2307.08423 (2023).

[32] E. Bosoni *et al.*, "How to verify the precision of density-functional-theory implementations via reproducible and universal workflows," Nat. Rev. Phys. **6**, 45–58 (2024).

[33] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm," npj Comp. Mat. **6**, 138 (2020).

[34] Z. Wu *et al.*, "Moleculenet: a benchmark for molecular machine learning," Chem. Sci. **9**, 513–530 (2018).
[35] L. Chanussot *et al.*, "Open catalyst 2020 (oc20) dataset and community challenges," ACS Catal. **11**, 6059–6072 (2021).
[36] S. Chmiela *et al.*,"Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, e1603015 (2017).
[37] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, "sgdml: Constructing accurate and data efficient molecular force fields using machine learning," Comp. Phys. Comm. **240**, 38–45 (2019).
[38] Y. Zuo *et al.*, "Performance and cost assessment of machine learning interatomic potentials," J. Phys. Chem. A **124**, 731–745 (2020).
[39] L. Weston *et al.*,"Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," J. Chem. Inf. Model. **59**, 3692–3702 (2019)
.
[40] Maxim Ziatdinov, Ayana Ghosh, Chun Yin (Tommy) Wong, and Sergei V. Kalinin, "AtomAI framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy," Nat. Mach. Intel. **4**, 1101–1112 (2022).
[41] P. Borlido, T. Aull, A. W. Huran, F. Tran, M. A. Marques, and S. Botti, "Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids," J. Chem. Theor. Comp. **15**, 5069–5079 (2019).
[42] S. P. Huber *et al.*, "Common workflows for computing material properties using different quantum engines," npj Comp. Mat. **7**, 136 (2021).
[43] G.-X. Zhang, A. M. Reilly, A. Tkatchenko, and M. Scheffler, "Performance of various density-functional approximations for cohesive properties of 64 bulk solids," New J. Phys. **20**, 063020 (2018)
.
[44] Richard Tran *et al.*, "The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysts," ACS Catal. **13**, 3066–3084 (2023).
[45] P. Jurečka, J. Šponer, J. Černỳ, and P. Hobza, "Benchmark database of accurate (mp2 and ccsd (t) complete basis set limit) interaction energies of small model complexes, dna base pairs, and amino acid pairs," Phy. Chem. Chem. Phys. **8**, 1985–1993 (2006).
[46] B. Brauer, M. K. Kesharwani, S. Kozuch, and J. M. Martin, "The s66x8 benchmark for noncovalent interactions revisited: Explicitly correlated ab initio methods and density functional theory," Phys. Chem. Chem. Phys. **18**, 20905–20925 (2016).
[47] R. A. Mata and M. A. Suhm, "Benchmarking quantum chemical methods: Are we heading in the right direction?" Angewandte Chemie Int. Ed. **56**, 11011–11018 (2017).
[48] D. E. Taylor *et al.*, "Blind test of density-functional-based methods on intermolecular interaction energies," J. Chem Phys. **145**, 124105 (2016).
[49] D. Wheeler *et al.*, "Pfhub: the phase-field community hub," J. Open Res. Soft. **7** (2019).
[50] A. D. Lindsay *et al.*, "2.0 - MOOSE: Enabling massively parallel multiphysics simulation," SoftwareX **20**, 101202 (2022).
[51] Jingrui Wei *et al.*, "Benchmark Tests of Atom Segmentation Deep Learning Models with a Consistent Dataset," Micro. and Microanalys. **29**, 552–562 (2023).
[52] J. Ren, F. Wang, J. Zhang, Q. Zheng, M. Ren, and B. Shi, "Diligent102: A photometric stereo benchmark dataset with controlled shape and material variation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022) pp. 12581–12590.
[53] M. Li, Z. Zhou, Z. Wu, B. Shi, C. Diao, and P. Tan, "Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials," IEEE Trans. on Im. Proc. **29**, 4159–4173 (2020).
[54] A. N. Henderson, S. K. Kauwe, and T. D. Sparks, "Benchmark datasets incorporating diverse tasks, sample sizes, material systems, and data heterogeneity for materials informatics," Data in Brief **37**, 107262 (2021).
[55] V. Fung, J. Zhang, E. Juarez, and B. G. Sumpter, "Benchmarking graph neural networks for materials chemistry," npj Comp. Mat. **7**, 84 (2021).
[56] A. Cecen, H. Dai, Y. C. Yabansu, S. R. Kalidindi, and L. Song, "Material structure-property linkages using three-dimensional convolutional neural networks," Acta Mat. **146**, 76–84 (2018).
[57] S. G. Baird, R. Issa, and T. D. Sparks, "Materials science optimization benchmark dataset for multi-objective, multi-fidelity optimization of hard-sphere packing simulations," Data in Brief **50**, 109487 (2023).
[58] L. Chen, H. Tran, R. Batra, C. Kim, and R. Ramprasad, "Machine learning models for the lattice thermal conductivity prediction of inorganic materials," Comp. Mat. Sci. **170**, 109155 (2019).
[59] S. Tian *et al.*,"Quartet protein reference materials and datasets for multi-platform assessment of label-free proteomics," Genome Bio. **24**, 202 (2023).
[60] N. Fu *et al.*, "Materials transformers language models for generative materials design: a benchmark study." Preprint at https://arxiv.org/abs/2206.13578 (2022).
[61] B. Meredig *et al.*, "Can machine learning identify the next high-temperature superconductor? examining extrapolation performance for materials discovery," Mol. Syst. Des. Eng. **3**, 819–825 (2018).
[62] E. Lejeune, "Mechanical mnist: A benchmark dataset for mechanical metamodels," Ext. Mech. Lett. **36**, 100659 (2020).
[63] C. L. Clement, S. K. Kauwe, and T. D. Sparks, "Benchmark aflow data sets for machine learning," Int. Mat. Manufact. Innov. **9**, 153–156 (2020).
[64] D. Varivoda, R. Dong, S. S. Omee, and J. Hu, "Materials property prediction with uncertainty quantification: A benchmark study," App. Phys. Rev. **10**, 021409 (2023),
[65] A. Jain *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mat. **1**, 011002 (2013),
[66] K. Li, B. DeCost, K. Choudhary, M. Greenwood, and J. Hattrick-Simpers, "A critical examination of robustness and generalizability of machine learning prediction of materials properties," npj Comp. Mat. **9**, 55 (2023).
[67] Kangming Li, Daniel Persaud, Kamal Choudhary, Brian DeCost, Michael Greenwood, and Jason Hattrick-Simpers, "Exploiting redundancy in large materials datasets for efficient machine learning with less data," Nature Communications **14**, 7283 (2023).

[68] K. Choudhary and B. G. Sumpter, "Can a deep-learning model make fast predictions of vacancy formation in diverse materials?" AIP Adv. **13** (2023).

[69] R. Vuorio, S.-H. Sun, H. Hu, and J. J. Lim, "Multimodal model-agnostic meta-learning via task-aware modulation," in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).

[70] K. Choudhary *et al.*, "The joint automated repository for various integrated simulations (jarvis) for data-driven materials design," npj Comp. Mat. **6**, 173 (2020).

[71] D. Wines *et al.*, "Recent progress in the JARVIS infrastructure for next-generation data-driven materials design," App. Phys. Rev. **10**, 041302 (2023),

[72] J. T. et al, "Scimlbench: A benchmarking suite for ai for science," (2021).

[73] N. Brown, M. Fiscato, M. H. Segler, and A. C. Vaucher, "Guacamol: benchmarking models for de novo molecular design," J. Chem. Info. Model. **59**, 1096–1108 (2019).

[74] G. Chen *et al.*, "Alchemy: A quantum chemistry dataset for benchmarking ai models." Preprint at https://arxiv.org/abs/1906.09427 (2019).

[75] M. E. Khatib and W. A. de Jong, "Ml4chem: A machine learning package for chemistry and materials science." Preprint at https://arxiv.org/abs/2003.13388 (2020).

[76] F. Broccatelli, R. Trager, M. Reutlinger, G. Karypis, and M. Li, "Benchmarking accuracy and generalizability of four graph neural networks using large in vitro adme datasets from different chemical spaces," Mol. Info. **41**, 2100321 (2022).

[77] R. D. Johnson *et al.*, "Nist computational chemistry comparison and benchmark database," http://srdata. nist. gov/cccbdb (2006).

[78] G. Prandini, A. Marrazzo, I. E. Castelli, N. Mounet, and N. Marzari, "Precision and efficiency in solid-state pseudopotential calculations," npj Comp. Mat. **4**, 72 (2018).

[79] D. S. Karls, M. Bierbaum, A. A. Alemi, R. S. Elliott, J. P. Sethna, and E. B. Tadmor, "The openkim processing pipeline: A cloud-based automatic material property computation engine," J. Chem. Phys. **153**, 064104 (2020).

[80] L. M. Hale, Z. T. Trautt, and C. A. Becker, "Evaluating variability with atomistic simulations: the effect of potential and calculation methodology on the modeling of lattice and elastic constants," Model. Sim. Mat. Sci. and Engineering **26**, 055003 (2018).

[81] K. Choudhary, A. J. Biacchi, S. Ghosh, L. Hale, A. R. H. Walker, and F. Tavazza, "High-throughput assessment of vacancy formation and surface energies of materials using classical force-fields," J. Phys.: Cond. Matt. **30**, 395901 (2018).

[82] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, "Validating quantum computers using randomized model circuits," Phys. Rev. A **100**, 032328 (2019).

[83] T. Tomesh *et al.*, "Supermarq: A scalable quantum benchmark suite," in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, 2022) pp. 587–603.

[84] Florian Häse *et al.*, "Olympus: a benchmarking framework for noisy optimization and experiment planning," Machine Learning: Science and Technology **2**, 035021 (2021).

[85] M. Aldeghi, F. Häse, R. J. Hickman, I. Tamblyn, and A. Aspuru-Guzik, "Golem: an algorithm for robust experiment and process optimization," Chem. Sci. **12**, 14792–14807 (2021).

[86] J. R. Hattrick-Simpers *et al.*, "An inter-laboratory study of zn–sn–ti–o thin films using high-throughput experimental methods," ACS Comb. Sci. **21**, 350–361 (2019).

[87] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169 (1996).

[88] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," Comp. Mat. Sci. **6**, 15–50 (1996).

[89] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865–3868 (1996).

[90] L. Ward *et al.*, "Matminer: An open source toolkit for materials data mining," Comp. Mat. Sci. **152**, 60–69 (2018).

[91] G. Ke *et al.* "Lightgbm: A highly efficient gradient boosting decision tree," Adv. Neur. Info. Proc. Sys. **30**, 3146–3154 (2017).

[92] W. A. Harrison, *Electronic structure and the properties of solids: the physics of the chemical bond* (Courier Corporation, 2012).

[93] K. F. Garrity and K. Choudhary, "Database of wannier tight-binding hamiltonians using high-throughput density functional theory," Sci. Data **8**, 106 (2021).

[94] K. F. Garrity and K. Choudhary, "Fast and accurate prediction of material properties with three-body tight-binding model for the periodic table," Phys. Rev. Mat. **7**, 044603 (2023).

[95] R. M. Martin, *Electronic structure: basic theory and practical methods* (Cambridge university press, 2020).

[96] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, "Quantum Monte Carlo simulations of solids," Rev. Mod. Phys. **73**, 33–83 (2001).

[97] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, and C. Marianetti, "Electronic structure calculations with dynamical mean-field theory," Rev. Mod. Phys. **78**, 865 (2006).

[98] G. Onida, L. Reining, and A. Rubio, "Electronic excitations: density-functional versus many-body green's-function approaches," Rev. Mod. Phys. **74**, 601–659 (2002).

[99] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864–B871 (1964).

[100] J. Klimeš, D. R. Bowler, and A. Michaelides, "Chemical accuracy for the van der waals density functional," J. Phys.: Cond. Matt. **22**, 022201 (2009).

[101] J. c. v. Klimeš, D. R. Bowler, and A. Michaelides, "Van der waals density functionals applied to solids," Phys. Rev. B **83**, 195131 (2011).

[102] F. Tran and P. Blaha, "Importance of the kinetic energy density for band gap calculations in solids with density functional theory," J. Phys. Chem. A **121**, 3318–3325 (2017).

[103] D. P. Rai, M. P. Ghimire, and R. K. Thapa, "A dft study of bex (x = s, se, te) semiconductor: Modified becke johnson (mbj) potential," Semicond. **48**, 1411–1422 (2014).

[104] J. Sun, A. Ruzsinszky, and J. P. Perdew, "Strongly constrained and appropriately normed semilocal density functional," Phys. Rev. Lett. **115**, 036402 (2015).

[105] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, "Accurate and numerically efficient r2scan meta-generalized gradient approximation," *The Journal of Physical Chemistry Letters*, J. Phys. Chem. Lett. **11**, 8208–8215 (2020).

[106] J. Heyd, G. E. Scuseria, and M. Ernzerhof, "Hybrid functionals based on a screened coulomb potential," J. Chem. Phys. **118**, 8207–8215 (2003),

[107] P. Giannozzi *et al.*, "Quantum espresso: a modular and open-source software project for quantum simulations of materials," J. Phys.: Cond. Matt. **21**, 395502 (2009).

[108] J. P. Perdew *et al.*, "Restoring the density-gradient expansion for exchange in solids and surfaces," Phys. Rev. Lett. **100**, 136406 (2008).

[109] X. Gonze *et al.*, "Recent developments in the abinit software package," Comp. Phys. Comm. **205**, 106–131 (2016).

[110] A. H. Romero *et al.*, "Abinit: Overview, and focus on selected capabilities," J. Chem. Phys. **152**, 124102 (2020).

[111] X. Gonze *et al.*, "The abinit project: Impact, environment and recent developments," Comp. Phys. Comm. **248**, 107042 (2020).

[112] J. Enkovaara *et al.*, "Electronic structure calculations with gpaw: a real-space implementation of the projector augmented-wave method," J. Phys.: Cond. Matt. **22**, 253202 (2010).

[113] M. Kuisma, J. Ojanen, J. Enkovaara, and T. T. Rantala, "Kohn-sham potential with discontinuity for band gap materials," Phys. Rev. B **82**, 115106 (2010).

[114] J. Kim *et al.*, "Qmcpack: an open source ab initio quantum monte carlo package for the electronic structure of atoms, molecules and solids," J. Phys.: Cond. Matt. **30**, 195901 (2018).

[115] A. A. Mostofi *et al.*, "An updated version of wannier90: A tool for obtaining maximally-localised wannier functions," Comp. Phys. Comm. **185**, 2309–2310 (2014).

[116] K. Choudhary, B. DeCost, and F. Tavazza, "Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape," Phys. Rev. Mat. **2**, 083801 (2018).

[117] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Comp. Mat. **2**, 16028 (2016).

[118] R. B. Wexler, G. S. Gautam, E. B. Stechel, and E. A. Carter, "Factors governing oxygen vacancy formation in oxide perovskites," J. Am. Chem. Soc. **143**, 13212–13227 (2021).

[119] D. Jha *et al.*, "Elemnet: Deep learning the chemistry of materials from only elemental composition," Sci. Rep. **8**, 17593 (2018).

[120] D. Jha *et al.*, "Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning," Nat. Comm. **10**, 1–12 (2019).

[121] V. Gupta *et al.*, "Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data," Nat. Comm. **12**, 6595 (2021).

[122] V. Gupta, W.-k. Liao, A. Choudhary, and A. Agrawal, "Pre-activation based representation learning to enhance predictive analytics on small materials data," in *2023 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2023) pp. 1–8.

[123] D. Jha *et al.*, "Irnet: A general purpose deep residual regression framework for materials discovery," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, edited by "" (2019) pp. 2385–2393.

[124] D. Jha *et al.*, "Enabling deeper learning on big data for materials informatics applications," Sci. Rep. **11**, 4244 (2021).

[125] D. Jha, V. Gupta, W.-k. Liao, A. Choudhary, and A. Agrawal, "Moving closer to experimental level materials property prediction using ai," Sci. Rep. **12**, 1–9 (2022).

[126] V. Gupta, W.-k. Liao, A. Choudhary, and A. Agrawal, "Brnet: Branched residual network for fast and accurate predictive modeling of materials properties," in *Proceedings of the 2022 SIAM international conference on data mining (SDM)* (SIAM, 2022) pp. 343–351.

[127] V. Gupta, A. Peltekian, W.-k. Liao, A. Choudhary, and A. Agrawal, "Improving deep learning model performance under parametric constraints for materials informatics applications," Sci. Rep. **13**, 9128 (2023)
.

[128] C. J. Bartel, *et al.*, "A critical examination of compound stability predictions from machine-learned formation energies," Preprint at https://arxiv.org/abs/2102.13090 (2021).
.

[129] C. Chen *et al.*, "Accurate force field for molybdenum by machine learning large materials data," Phys. Rev. Mat. **1**, 043603 (2017).

[130] K. Choudhary and B. DeCost, "Atomistic line graph neural network for improved materials property predictions," npj Comp. Mat. **7**, 185 (2021).

[131] V. Gupta *et al.*, "Structure-aware graph neural network based deep transfer learning framework for enhanced predictive analytics on diverse materials datasets," npj Comp. Mat. **10**, 1 (2024).

[132] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**, 145301 (2018).

[133] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet – a deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018),

[134] K. Choudhary, R. Gurunathan, B. DeCost, and A. J. Biacchi, "Atomvision: A machine vision library for atomistic images," J. Chem. Info. Model. (2023).

[135] K. Choudhary and M. L. Kelley, "ChemNLP: A Natural Language-Processing-Based Library for Materials Chemistry Text Data," J. Phys. Chem. C **127**, 17545–17555 (2023).

[136] J. Gasteiger, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," in *International Conference on Learning Representations (ICLR)* (2020).

[137] J. Gasteiger, S. Giri, J. T. Margraf, and S. Günnemann, "Fast and uncertainty-aware directional message passing for non-equilibrium molecules," in *Machine Learning for Molecules Workshop, NeurIPS* (2020)
.

[138] B. Deng, P. Zhong, K. Jun, K. Han, C. J. Bartel, and G. Ceder, "CHGNet: Pretrained universal neural network potential for charge-informed atomistic modeling." Preprint at https://arxiv.org/abs/2302.14231 (2023)
.

[139] C. Chen and S. P. Ong, "A universal graph deep learning interatomic potential for the periodic table," Nat. Comp. Sci. **2**, 718–728 (2022).

[140] P. Reiser, A. Eberhard, and P. Friederich, "Graph neural networks in tensorflow-keras with raggedtensor representation (kgcnn)," Soft. Imp. , 100095 (2021).

[141] Y. Lin, K. Yan, Y. Luo, Y. Liu, X. Qian, and S. Ji, "Efficient approximations of complete interatomic potentials for crystal property prediction," in *Proceedings of the 40th International Conference on Machine Learning* (2023).

[142] K. Yan, Y. Liu, Y. Lin, and S. Ji, "Periodic graph transformers for crystal material property prediction," in *The 36th Annual Conference on Neural Information Processing Systems* (2022) pp. 15066–15080.

[143] S. Zhang *et al.*, "Opt: Open pre-trained transformer language models." Preprint at http://arxiv.org/abs/2205.01068 (2022).

[144] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." Preprint at http://arxiv.org/abs/1910.10683 (2020).

[145] J. E. Jones and S. Chapman, "On the determination of molecular fields.—i. from the variation of the viscosity of a gas with temperature," Proc. Roy. Soc. London. Series A, Containing Papers of a Mathematical and Physical Character **106**, 441–462 (1924),

[146] A. P. Thompson *et al.*, "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," Comp. Phys. Comm. **271**, 108171 (2022).

[147] M. S. Daw and M. I. Baskes, "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals," Phys. Rev. B **29**, 6443–6453 (1984).

[148] J. Tersoff, "New empirical approach for the structure and energy of covalent systems," Phys. Rev. B **37**, 6991–7000 (1988).

[149] K. Lindorff-Larsen *et al.*, "Improved side-chain torsion potentials for the amber ff99sb protein force field," Proteins: Struct., Func., and Bioinfo. **78**, 1950–1958 (2010).

[150] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "Gromacs: fast, flexible, and free," J. Comp. Chem. **26**, 1701–1718 (2005).

[151] S. Mehdi, D. Wang, S. Pant, and P. Tiwary, "Accelerating all-atom simulations and gaining mechanistic understanding of biophysical systems through state predictive information bottleneck," J. Chem. Theor. Comp. **18**, 3231–3238 (2022).

[152] "IBM Quantum," `https://quantum-computing.ibm.com` (2021).

[153] A. Peruzzo *et al.*, "A variational eigenvalue solver on a photonic quantum processor," Nat. Comm. **5**, 4213 (2014).

[154] O. Higgott, D. Wang, and S. Brierley, "Variational Quantum Computation of Excited States," Quantum **3**, 156 (2019).

[155] V. Bergholm *et al.*, "Pennylane: Automatic differentiation of hybrid quantum-classical computations." Preprint at http://arxiv.org/abs/1811.04968 (2022).

[156] J. M. Arrazola *et al.*, "Differentiable quantum computational chemistry with pennylane." Preprint at http://arxiv.org/abs/2111.09967 (2023).

[157] H. G. T. Nguyen *et al.*, "A reference high-pressure co2 adsorption isotherm for ammonium zsm-5 zeolite: results of an interlaboratory study," Adsorption **24**, 531–539 (2018).

[158] A. Engelbrecht-Wiggans *et al.*, "Effects of temperature and humidity on high-strength p-aramid fibers used in body armor," Textile Res.Journ. **90**, 2428–2440 (2020),

[159] S. Lehtola, C. Steigemann, M. J. Oliveira, and M. A. Marques, "Recent developments in libxc—a comprehensive library of functionals for density functional theory," SoftwareX **7**, 1–5 (2018).

[160] K. Choudhary and F. Tavazza, "Convergence and machine learning predictions of monkhorst-pack k-points and plane-wave cut-off in high-throughput dft calculations," Comp. Mat. Sci. **161**, 300–308 (2019).

[161] K. Choudhary, F. Y. P. Congo, T. Liang, C. Becker, R. G. Hennig, and F. Tavazza, "Evaluation and comparison of classical interatomic potentials through a user-friendly interactive web-interface," Sci. Data **4**, 160125 (2017).

[162] D. A. Case *et al.*, "The amber biomolecular simulation programs," J. Comp. Chem. **26**, 1668–1688 (2005).

[163] J. Huang *et al.*, "Charmm36m: an improved force field for folded and intrinsically disordered proteins," Nat. Methods **14**, 71–73 (2017).

[164] I. Novoselov, A. Yanilkin, A. Shapeev, and E. Podryabinkin, "Moment tensor potentials as a promising tool to study diffusion processes," Comp. Mat. Sci. **164**, 46–56 (2019).

[165] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).

[166] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[167] L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).

[168] V. Botu and R. Ramprasad, "Adaptive machine learning framework to accelerate ab initio molecular dynamics," Int. J. Quant. Chem. **115**, 1074–1083 (2015),

[169] J. S. Smith *et al.*, "Automated discovery of a robust interatomic potential for aluminum," Nat. Comm. **12** (2021).

[170] K. Choudhary, B. DeCost, L. Major, K. Butler, J. Thiyagalingam, and F. Tavazza, "Unified graph neural network force-field for the periodic table: solid state applications," Dig. Disc. (2023).

[171] A. H. Larsen *et al.*, "The atomic simulation environment—a python library for working with atoms," J. Phys.: Cond. Matt. **29**, 273002 (2017).

[172] K. Choudhary, T. Yildirim, D. W. Siderius, A. G. Kusne, A. McDannald, and D. L. Ortiz-Montalvo, "Graph neural network predictions of metal organic framework co2 adsorption properties," Comp. Mat. Sci. **210**, 111388 (2022).

[173] D. Dubbeldam, S. Calero, D. E. Ellis, and R. Q. Snurr, "Raspa: molecular simulation software for adsorption and diffusion in flexible nanoporous materials," Mol. Sim. **42**, 81–101 (2016),
[174] S. Páll *et al.*, "Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS," J. Chem. Phys. **153** (2020),
[175] S.-T. Tsai, Z. Smith, and P. Tiwary, "Sgoop-d: Estimating kinetic distances and reaction coordinate dimensionality for rare event systems from biased/unbiased simulations," J. Chem. Theor. Comp. **17**, 6757–6765 (2021).
[176] S. Mehdi, D. Wang, S. Pant, and P. Tiwary, "Accelerating all-atom simulations and gaining mechanistic understanding of biophysical systems through state predictive information bottleneck," J. Chem. Theor. Comp. **18**, 3231–3238 (2022).
[177] D. Wang and P. Tiwary, "State predictive information bottleneck," J. Chem. Phys. **154**, 134111 (2021).
[178] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17," J. Chem. Info. Model. **52**, 2864–2875 (2012).
[179] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1** (2014).
[180] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[181] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library." Preprint at http://arxiv.org/abs/1912.01703 (2019).
[182] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015), software available from tensorflow.org.
[183] J. Bradbury *et al.*, "JAX: composable transformations of Python+NumPy programs," (2018).
[184] T. Wolf *et al.*,"Huggingface's transformers: State-of-the-art natural language processing." Preprint at http://arxiv.org/abs/1910.03771 (2020).
[185] M. A. Nielsen and I. L. Chuang, "Quantum computation and quantum information," Phys. Today **54**, 60 (2001).
[186] K. Choudhary, "Quantum computation for predicting electron and phonon properties of solids," J. Phys.: Cond. Matt. **33**, 385501 (2021).
[187] J. S. Kottmann *et al.*, "Tequila: a platform for rapid development of quantum algorithms," Quantum Sci. Tech. **6**, 024009 (2021).
[188] C. Developers, "Cirq," (2022), See full list of authors on Github: https://github .com/quantumlib/Cirq/graphs/contributors.
[189] R. H. Pierson and E. A. Fay, "Guidelines for interlaboratory testing programs," Analyt. Chem. **31**, 25A–49A (1959).
[190] N. D. Lowhorn *et al.*, "Round-robin studies of two potential seebeck coefficient standard reference materials," in *2007 26th International Conference on Thermoelectrics* (2007) pp. 361–365.
[191] S. Moylan, C. U. Brown, and J. Slotwinski, "Recommended protocol for round-robin studies in additive manufacturing," J. Test. and Eval. **44**, 1009–1018 (2016).
[192] C. U. Brown, G. Jacob, M. Stoudt, S. Moylan, J. Slotwinski, and A. Donmez, "Interlaboratory study for nickel alloy 625 made by laser powder bed fusion to quantify mechanical property variability," J. Mat. Eng. and Perf. **25**, 3390–3397 (2016).
[193] E. Alleno *et al.*, "Invited Article: A round robin test of the uncertainty on the measurement of the thermoelectric dimensionless figure of merit of Co0.97Ni0.03Sb3," Rev. Sci.Inst. **86**, 011301 (2015),
[194] Y. Jiang *et al.*, "$Pb_9Cu(PO_4)_6(OH)_2$: Phonon bands, localized flat-band magnetism, models, and chemical analysis," Phys. Rev. B **108**, 235127 (2023).
[195] S. Lee, J.-H. Kim, and Y.-W. Kwon, "The first room-temperature ambient-pressure superconductor," Preprint at http://arxiv.org/abs/2307.12008 (2023).
[196] K. Guo, Y. Li, and S. Jia, "Ferromagnetic half levitation of lk-99-like synthetic samples," Sci. China Phys., Mech. &; Astro. **66** (2023), 10.1007/s11433-023-2201-9.
[197] K. Kumar, N. K. Karn, Y. Kumar, and V. P. S. Awana, "Absence of superconductivity in LK-99 at ambient conditions." Preprint at http://arxiv.org/abs/2308.03544 (2023).
[198] Y. Zuo *et al.*, "Performance and cost assessment of machine learning interatomic potentials," J. Phys. Chem. A **124**, 731–745 (2020).
[199] J. E. Saal *et al.*, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd)," JOM **65**, 1501–1509 (2013).
[200] S. Kirklin *et al.*, "The open quantum materials database (oqmd): assessing the accuracy of dft formation energies," npj Comp. Mat. **1**, 15010 (2015).
[201] S. Curtarolo *et al.*, "Aflow: An automatic framework for high-throughput materials discovery," Comp. Mat. Sci. **58**, 218–226 (2012).
[202] S. Gong, T. Xie, Y. Shao-Horn, R. Gomez-Bombarelli, and J. C. Grossman, "Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity." Preprint at https://arxiv.org/abs/2208.05039 (2022).
[203] L. Ward *et al.*, "Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations," Phys. Rev. B **96**, 024104 (2017).

IX. FIGURE LEGENDS

1. Leaderboard snapshot with an example output for AI based formation energy per atom model on the JARVIS-DFT (dft_3d) dataset. a) homepage sanpshot showing list of categories and number of available contributions at the time of writing, b) an example AI regression model benchmark for formation energy with several contributions. The methods are sorted based on the mean absolute error (MAE) values. Lower MAE values indicate higher accuracy, c) explicit table for the plot in panel b.Links to individual csv.zip (AI-SinglePropertyPrediction-formation_energy_peratom-dft_3d-test-mae.csv.zip), json.zip (dft_3d_formation_energy_peratom.json.zip), shell script (run.sh) and detailed info (metadata.json) files are provided to help enhance reproducibility. Such results plots and tables are available for each benchmark in the leaderboard.

2. A tree diagram for directory and file-structure in the leaderboard. There are two main directories in the repo: (1) benchmarks (reference) and (2) leaderboard contributions (for various leaderboard entries). In the "benchmarks" directory, there are folders for the AI, ES, QC, FF, and EXP categories. Within them, there are sub-folders for specific sub-categories. In the "contributions" directory there is a collection of folders that consists of .csv.zip, metadata.json files, and optionally a Dockerfile and run.sh file for available contributions from each method. The csv.zip file contains entries of identifier (id) and corresponding prediction values as obtained by the corresponding model/method. These test identifiers (such as JVASP-1408) must match the test set ids in the json.zip file in the benchmarks folder for the metric measurements to work.

3. A tree diagram for directory and file-structure in the leaderboard. There are two main directories in the repo: (1) benchmarks (reference) and (2) leaderboard contributions (for various leaderboard entries). In the "benchmarks" directory, there are folders for the AI, ES, QC, FF, and EXP categories. Within them, there are sub-folders for specific sub-categories. In the "contributions" directory there is a collection of folders that consists of .csv.zip, metadata.json files, and optionally a Dockerfile and run.sh file for available contributions from each method. The csv.zip file contains entries of identifier (id) and corresponding prediction values as obtained by the corresponding model/method. These test identifiers (such as JVASP-1408) must match the test set ids in the json.zip file in the benchmarks folder for the metric measurements to work.

4. Distribution of data in each dataset. (a) all entries in leaderboard, (b) entries with unique identifiers. Note that one identifier (such as JVASP-1002 for silicon) can have multiple properties (such as bandgap, bulk modulus etc.). A script to generate this figure is also provided on the leaderboard website as the leaderboard is continuously evolving.

5. Periodic table element distribution for entries in all the datasets. This is calculated by taking into account all the element specific entries normalized by total entries i.e. these are percentage probabilities.

6. Example mean absolute errors for benchmarks including (a) artificial intelligence (AI) formation energy for test set with 5572 materials in JARVIS-DFT 3D dataset, (b) electronic structure (ES) Si (JARVIS-DFT ID: JVASP-1002) bandgap, (c) classical force-field (FF) based Voigt bulk modulus of Si and (d) machine learning force-field (MLFF) based forces for Si. We provide Jupyter/Google colab notebooks to easily plot such comparisons for all available benchmarks. Also, similar analysis figures for all the available benchmarks are available in the supplementary information (Supplementary Figures 1-298). As a note, these plots are a current snapshot of the leaderboard, and it is possible that new and more accurate models will be developed and added here in the future.

7. Relative performance computed as the ratio of the MAE difference between the best descriptor-based model and the best neural networks to the MAE of the best neural networks in the AI regression benchmarks. The benchmark name and the corresponding best performing neural network are indicated in the left and right y axis, respectively. For all the considered AI benchmarks, the best descriptor-based model is the tree-based model using Magpie [117] and Voronoi-tessellation [203] features. As a disclaimer, these plots are a current snapshot of the leaderboard, and it is possible that new and more accurate models will be developed in the future.

8. Mean absolute deviation (MAD) to mean absolute error (MAE) ratio for (a) AI and (b) electronic structure methods. MAD:MAE serves as uniform criteria for comparing performances of models.

9. Example results for AI, ES, QC and EXP results. (a) formation-energy-per atom model using AI for JARVIS-DFT 3D dataset with 5572 materials in the test set, (b) bulk modulus predictions using ES methods for 21 materials, (c) electronic bandstructure of Aluminum using QC methods with different quantum circuits on a coarse k-point mesh, (d) CO_2 capture for zeolite (ZSM-5) at several labs in inter-laboratory/round-robin fashion.

X. TABLE LEGENDS

1. Comparison of benchmark infrastructure available for materials design methods for several categories.

2. Summary of current benchmark categories and methods available in the JARVIS-Leaderboard at the time of writing. More details can be found in the individual metadata.json file. Note that the number of methods is continuously growing.