npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design

Kamal Choudhary[1,2,3]✉, Kevin F. Garrity[1], Andrew C. E. Reid[1], Brian DeCost[1], Adam J. Biacchi[4], Angela R. Hight Walker[4], Zachary Trautt[1], Jason Hattrick-Simpers[1], A. Gilad Kusne[1], Andrea Centrone[4], Albert Davydov[1], Jie Jiang[5], Ruth Pachter[5], Gowoon Cheon[6], Evan Reed[6], Ankit Agrawal[7], Xiaofeng Qian[8], Vinit Sharma[9,10], Houlong Zhuang[11], Sergei V. Kalinin[12], Bobby G. Sumpter[12], Ghanshyam Pilania[13], Pinar Acar[14], Subhasish Mandal[15], Kristjan Haule[15], David Vanderbilt[15], Karin Rabe[15] and Francesca Tavazza[1]

The Joint Automated Repository for Various Integrated Simulations (JARVIS) is an integrated infrastructure to accelerate materials discovery and design using density functional theory (DFT), classical force-fields (FF), and machine learning (ML) techniques. JARVIS is motivated by the Materials Genome Initiative (MGI) principles of developing open-access databases and tools to reduce the cost and development time of materials discovery, optimization, and deployment. The major features of JARVIS are: JARVIS-DFT, JARVIS-FF, JARVIS-ML, and JARVIS-tools. To date, JARVIS consists of ≈40,000 materials and ≈1 million calculated properties in JARVIS-DFT, ≈500 materials and ≈110 force-fields in JARVIS-FF, and ≈25 ML models for material-property predictions in JARVIS-ML, all of which are continuously expanding. JARVIS-tools provides scripts and workflows for running and analyzing various simulations. We compare our computational data to experiments or high-fidelity computational methods wherever applicable to evaluate error/uncertainty in predictions. In addition to the existing workflows, the infrastructure can support a wide variety of other technologically important applications as part of the data-driven materials design paradigm. The JARVIS datasets and tools are publicly available at the website: https://jarvis.nist.gov.

npj Computational Materials (2020)6:173 ; https://doi.org/10.1038/s41524-020-00440-1

INTRODUCTION

The Materials Genome Initiative (MGI) (https://mgi.gov/, The website provides information about several activities and events under the Materials Genome Initiative (MGI); https://www.nist.gov/mgi, The website provides information about various projects under the National Institute of Standards and Technology (NIST)'s Materials Genome Initiative (MGI) chapter) was introduced in 2011 to accelerate materials discovery using computational[1–7], experimental[8–11] and data analytics[12–14] approaches. The MGI has revolutionized several fields for materials-applications, such as batteries[15], thermoelectrics[16], and alloy-design[17], thorough open-access public database and tool development[18]. The MGI encourages systematic Process-Structure-Property-Performance (PSPP)[19]-based efficient design-approaches rather than Edisonian trial-error methods[20].

Especially in the field of computational materials design, quantum mechanics-based density functional theory (DFT)[21] has proven to be an immensely successful technique, and several databases of automated DFT calculations are widely used in materials design applications. Despite their successes, existing DFT databases face limitations due to issues intrinsic to conventional DFT approaches, e.g., the generalized gradient approximation of Perdew-Burke-Ernzerhof (GGA-PBE)[21,22]. Drawbacks of the existing DFT databases include non-inclusion of van der Waals (vdW) interactions[6], bandgap underestimations[23], non-inclusion of spin-orbit coupling[5], overly simplifying magnetic ordering[24], neglecting defects[25] (point, line, surface and volume), unconverged computational parameters such as k-points[26], ignoring temperature effects[27] (generally DFT calculations are performed at 0 K), lack of layer/thickness-dependent properties of low dimensional materials[28], and lacking interfaces/heterostructures of materials[29], all of which can be critical for realistic material-applications. In addition, there are several other computational approaches, such as classical force-field (FF)[30], computational microscopy, phase-field (PF), CALculation of PHAse Diagrams (CALPHAD)[31], and Orientation Distribution Functions (ODF)[32] which lack the integrated tools and databases that have been developed for DFT-based computational approaches. Finally, the integration of computational approaches with experiments, the application of statistical uncertainty analysis, and the implementation of data analytics and artificial intelligence (AI) techniques require significant developments to meet the goals set forth by the MGI.

Some of the notable materials databases are: Automatic-FLOW for Materials Discovery (AFLOW)[1], Materials-project[2], Khazana[15], Open Quantum Materials Database (OQMD)[3], Novel Materials Discovery (NOMAD)[7], Computational Materials Repository (CMR)[33],

[1]Materials Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. [2]Theiss Research, La Jolla, CA 92037, USA. [3]Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, USA. [4]Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. [5]Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright–Patterson Air Force Base, Dayton, OH 45433, USA. [6]Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305, USA. [7]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA. [8]Department of Materials Science and Engineering, Texas A&M University, Texas, TX 77843, USA. [9]Joint Institute for Computational Sciences, University of Tennessee, Knoxville, TN 37996, USA. [10]National Institute for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [11]School for Engineering of Matter Transport and Energy, Arizona State University, Tempe, AZ 85287, USA. [12]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [13]Materials Science and Technology Division, Los Alamos National Lab, Los Alamos, NM 87545, USA. [14]Department of Mechanical Engineering, Virginia Tech, Blacksburg, VA 24061, USA. [15]Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08901, USA. ✉email: kamal.choudhary@nist.gov

NIMS-MatNavi(NIMS-MatNavi database. https://mits.nims.go.jp/. This website host information about several material classes and their properties), NREL-MatDB[34], Inorganic Crystal Structure Database (ICSD)[35], Materials-Cloud[36], Citrine(Ctrine informatics. https://citrine.io. This website hosts several tools for accelerated materials design), OpenKIM[37], Predictive Integrated Structural Materials Science (PRISMS)[38], and Phase-Field hub (PFhub)[39]. Some of the commonly used computational-tools are Python Materials Genomics (PYMATGEN)[40], Atomic Simulation Environment (ASE)[41], Automated Interactive Infrastructure and Database (AIIDA)[4] and MPinterfaces[42]. The data most commonly included in these databases consists of crystal structures, formation energies, bandgaps, elastic constants, Poisson ratios, piezoelectric constants, and dielectric constants. These material properties can be used directly to screen for potentially interesting materials for a given application as candidates for experimental synthesis and characterization, as well as part of a PSPP design approach to better understand the factors driving material performance. Beyond the directly calculated material properties mentioned above, several selection metrics are also being developed to aid materials design, such as scintillation attenuation length[43], thermoelectric complexity factor[44], spectroscopy limited maximum efficiency[45,46], exfoliation energy[6], and spin-orbit spillage[5,24,47]. Akin to DFT-like standard computational approaches that are used as screening tools for experiments, machine learning (ML)[12–14,48] models for materials design are being developed as pre-screening tools for other conventional computational methods such as DFT. In addition, ML tools are proposed to accelerate experimental methods directly based on computational data[49]. All of the above developments show immense promise for accelerating materials design.

The principles mentioned above constitute the foundations of the Joint Automated Repository for Various Integrated Simulations (JARVIS) (https://jarvis.nist.gov) infrastructure, a set of databases and tools to meet some of the current material-design challenges. The main components of JARVIS are: JARVIS-DFT, JARVIS-FF, JARVIS-ML, and JARVIS-tools. JARVIS is developed and hosted at the National Institute of Standards and Technology (NIST) (Please note that commercial software is identified to specify procedures. Such identification does not imply recommendation by the National Institute of Standards and Technology) as part of the MGI. A detailed documentation webpage for the database is available at: https://jarvis-materials-design.github.io/dbdocs/.

Started in 2017, JARVIS-DFT[5,6,23–25,28,29,45,49,50] is a repository based on DFT calculations that mainly uses the vdW-DF-OptB88 van der Waals functional[51]. The database also uses beyond-GGA approaches for a subset of materials, including the Tran-Blaha modified Becke-Johnson (TBmBJ) meta-GGA[52], the hybrid functional PBE0, the hybrid range-separated functional Heyd-Scuseria-Ernzerhof (HSE06), Dynamical Mean Field Theory (DMFT), and G_0W_0. In addition to hosting conventional properties such as formation energies, bandgaps, elastic constants, piezoelectric constants, dielectric constants, and magnetic moments, it also contains previously unavailable datasets, such as exfoliation energies for van der Waals bonded materials, the spin-orbit coupling (SOC) spillage, improved meta-GGA bandgaps, frequency-dependent dielectric functions, the spectroscopy limited maximum efficiency (SLME), infrared (IR) intensities, electric field gradients (EFG), heterojunction classifications, and Wannier tight-binding Hamiltonians. These datasets are compared to experimental results wherever possible to evaluate their accuracy as predictive tools. JARVIS-DFT also introduced protocols such as automatic k-point convergence, which can be critical for obtaining precise and accurate results. JARVIS-DFT is distributed through the website: https://jarvis.nist.gov/jarvisdft/.

The JARVIS-FF[25,53] database, also started in 2017, is a repository of classical force-field/potential computational data intended to help a user select the most appropriate force-field for a specific application. Many classical force-fields are developed for a particular set of properties (such as energies), and may not have been tested for properties not included in training (such as elastic constants, or defect formation energies). JARVIS-FF provides an automatic framework to consistently calculate and compare basic properties, such as the bulk modulus, defect formation energies, phonons, etc., that may be critical for specific molecular-dynamics simulations. JARVIS-FF relies on DFT and experimental data to evaluate accuracy. JARVIS-FF is distributed through the website: https://jarvis.nist.gov/jarvisff/.

The JARVIS-ML[45,49,50,54,55] is a repository of machine learning (ML) model parameters, descriptors, and ML-related input and target data. JARVIS-ML introduced Classical Force-field Inspired Descriptors (CFID) in 2018 as a universal framework to represent a material's chemistry-structure-charge related data. With the help of CFID and JARVIS-DFT data, several high-accuracy classification and regression ML models were developed, with applications to fast materials-screening and energy-landscape mapping. Some of the trained property models include formation energies, exfoliation energies, bandgaps, magnetic moments, refractive indexes, dielectric constants, thermoelectric performance, and maximum piezoelectric and infrared modes. Also, several ML interpretability analyses have provided physical-insights beyond intuitive materials-science knowledge[54]. These models, the workflow, the datasets, etc. are disseminated to enhance the transparency of the work. Recently, JARVIS-ML was expanded to include ML models to analyze STM-images in order to directly accelerate the interpretation of experimental images. Graph convolution neural network models are currently being developed for automated handling of images and crystal-structure analysis in materials science. JARVIS-ML is distributed through the website: https://jarvis.nist.gov/jarvisml/.

JARVIS-tools is the underlying computational framework used for automation, data-generation, data-handling, analysis and dissemination of all the above repositories. JARVIS-tools uses cloud-based continuous integration, low-software dependency, auto-documentation, Jupyter and Google-Colab notebook integration, pip installation and related strategies to make the software robust and easy to use. JARVIS-tools also hosts several examples to enable a user to reproduce the data in the above repositories or to apply the tools for their own applications. JARVIS-tools are provided through the GitHub page: https://github.com/usnistgov/jarvis.

While JARVIS has some features in common with existing DFT-based computational databases, we note that there are several features currently unique to the JARVIS framework. First, JARVIS has a tight integration between FF and DFT techniques. Second, JARVIS includes CFID ML learning descriptors and several ML models based on those descriptors, including solar-cell efficiency, thermoelectrics, exfoliation energies, infrared active modes, and refractive index etc. Finally, JARVIS-DFT itself features heavy use of a van der Waals functional, a 2D materials database, a STM image database, spin-orbit calculations, spin-orbit spillage, solar cell efficiency, meta-GGA functional calculations, other post-GGA electronic structure calculations, 2D heterostructure design app and a Wannier function database. We also provide REST-API framework for users to download and upload materials data using JARVIS-API.

This paper is organized as follows: (1) we introduce the main computational techniques, organized by the time and length scales, (2) we illustrate JARVIS-tools and its functionalities, (3) we discuss the contents of the major JARVIS databases, (4) we demonstrate some of the derived applications, and (5) we discuss outstanding challenges and future work.

RESULTS AND DISCUSSION

Overview of computational techniques

There are many computational tools for simulating realistic materials depending on the time and length scales of interest[56]. Before we discuss the details of JARVIS, we will provide a brief list of these techniques and highlight their range of applicability, as summarized in Fig. 1. Relevant techniques include quantum mechanical computations, classical/molecular mechanics, mesoscale modeling, finite element analysis, and engineering design. Each of these methodologies has its own ontology and semantics for describing themselves and the PSPP relationship. For example, 'structure' may imply electronic configurations in the quantum regime, atomic arrangement in molecular mechanics, microstructure, segments in phase field-based mesoscale modeling, and mesh-structure in finite element analysis. Material properties are calculated using corresponding physical laws such as the Schrödinger equation in the quantum regime, or Newton's laws of motion for classical regimes. For realistic material design, it is important to integrate these methods. A major challenge for multiscale modeling is propagating the results of one simulation into another while capturing the relevant physics. Artificial Intelligence (AI) techniques have been applied in each of these domains and can be used to integrate the methods to a certain extent[12]. In JARVIS, we primarily focus on atomistic-based classical and quantum simulations and machine-learning, but we also attempt to integrate other simulation methods with our atomistic data for a few specific applications such as using DFT based elastic constants in orientation distribution function based finite element simulations.

Software and databases

The JARVIS infrastructure (Fig. 2) is a combination of databases and tools for running and integrating some of the computational methods mentioned above. The general procedure for adding a dataset to JARVIS is as follows. We start with the goal of finding or designing a material to display or optimize a given property. Then, we decide on an appropriate computational method, as well as a computationally efficient way to screen for the best candidate materials. The screening process can proceed in several steps, with computationally inexpensive methods applied first, followed by more computationally intensive methods on the remaining materials. Whenever possible, the data is compared with available experiments to evaluate the accuracy and quality of the database. Once a large enough dataset is generated, machine learning techniques can be utilized to accelerate the traditional computational approaches.

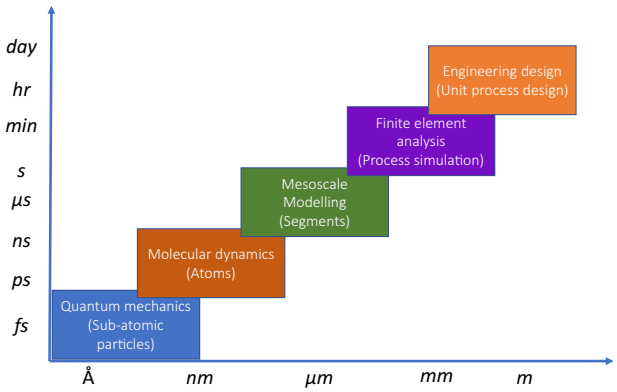

Fig. 1 Length and time-scale based computational materials design techniques. We primarily focus on the lowest two levels of the computational methodologies, DFT and MD, but we integrate with other simulation methods for specific applications.

As an example of making use of multiple computational tools within the same framework, we consider finding materials to maximize solar-cell efficiency. We develop a screening criterion (Spectroscopic Limited Maximum Efficiency, SLME, a part of *JARVIS-tools*) and calculate the necessary properties (dielectric function and band gap, a part of *JARVIS-DFT*). We test the method by comparing known materials to experiment (*precision and accuracy assessment*), and we perform more accurate meta-GGA and GW calculations (*JARVIS-Beyond DFT*) as additional screening and validation steps. Finally, we develop a machine learning model (*JARVIS-ML*) to accelerate future materials design. Details of this example can be found in refs. [45,46]. Similar case-studies for thermoelectrics, dielectrics, and infrared-phonon modes are available in ref. [50] and ref. [55].

The database component of JARVIS consists of JARVIS-DFT for DFT calculations and JARVIS-FF for molecular dynamics simulations. JARVIS-ML hosts several machine learning models based on our datasets. JARVIS-tools contains tools for automating, post-processing and disseminating generated data, as well as several derived applications such as JARVIS-Heterostructure. We also include precision and accuracy analyses of the generated data, which consists of comparing DFT data with experiments, comparing FF data with DFT, comparing ML models with DFT, etc. As a lower-level technique (see Fig. 1), JARVIS-DFT data can be fed into JARVIS-FF and JARVIS-ML models, but not vice versa. We use JARVIS-ML to accelerate both JARVIS-DFT and JARVIS-FF. In this way, the JARVIS-infrastructure establishes a joint integration for automation and generation of repositories. We provide several social-media platforms to build a community of interest. Some of the key resources for the JARVIS-infrastructure are shown in Table 1.

JARVIS-tools

JARVIS-tools is a python-based software package with ≈20,000 lines of code and consisting of several python-classes and functions. JARVIS-tools can be used for (a) the automation of simulations and data-generation, (b) post-processing and analysis of generated data, and (c) the dissemination of data and methods, as shown in Fig. 3. It uses cloud-based continuous integration checking including GitHubAction, CircleCI, TravisCI, CodeCov, and PEP8 linter to maintain consistency in the code and its functionalities. The JARVIS-tools is distributed through an open GitHub repository: https://github.com/usnistgov/jarvis.

An example python class in JARVIS-tools is 'Atoms'. It uses atomic coordinates, element types and lattice vectors to build an 'Atoms' object from which several properties, such as density and chemical formula, can be calculated. This 'Atoms' class, along with several other modules (discussed later), can be used for setting up

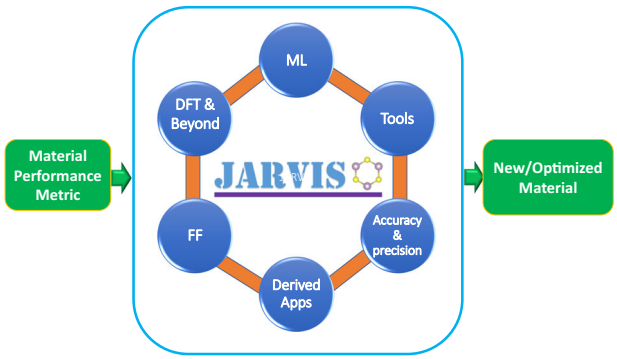

Fig. 2 An overview of the JARVIS infrastructure. For a given materials performance metric, several JARVIS components can work together to design optimized or completely new materials.

Table 1. An overview of resources available in the JARVIS infrastructure.

Resource	Website	Brief description
Homepage	https://jarvis.nist.gov/	Description and API
FF	https://jarvis.nist.gov/jarvisff	Evaluation of classical force field
DFT	https://jarvis.nist.gov/jarvisdft	Density functional theory data
ML	https://jarvis.nist.gov/jarvisml	Machine learning models
Tools	https://github.com/usnistgov/jarvis	Scripts for running simulations
Downloads	https://www.ctcms.nist.gov/~knc6/downloads.html	Downloadable metadata
Notebooks	https://github.com/JARVIS-Materials-Design/jarvis-tools-notebooks	Jupyter/Google-Colab notebooks
Heterostruct	https://jarvis.nist.gov/jarvish	2D heterostructure properties
WannierTB	https://jarvis.nist.gov/jarviswtb	Wannier tight binding models
BeyondDFT	https://jarvis.nist.gov/jarvisbdft	High-level ab-initio methods
Publications	https://www.ctcms.nist.gov/~knc6/pubs.html	JARVIS-related publication
Tools docs	https://jarvis-tools.readthedocs.io/en/latest/	Documentation (docs) of tools
DB docs	https://jarvis-materials-design.github.io/dbdocs/	Documentation on the database
Tools pypi	https://pypi.org/project/jarvis-tools/	Pypi repository of tools
Workshops	https://www.ctcms.nist.gov/~knc6/workshops.html	JARVIS-related workshops
ResearchG.	https://www.researchgate.net/project/NIST-JARVIS	Social media researchgate page
Twitter	https://twitter.com/jarvisnist	Social media twitter page
Facebook	https://www.facebook.com/jarvisnist/	Social media Facebook page
Linkedin	https://www.linkedin.com/company/jarvisnist	Social media Linkedin page
YouTube	https://www.youtube.com/channel/UClChK_t7kmVx_QMStQH_T9g	Social media Youtube page
Google group	https://groups.google.com/forum/#!forum/jarvis-nist	Social media google-group

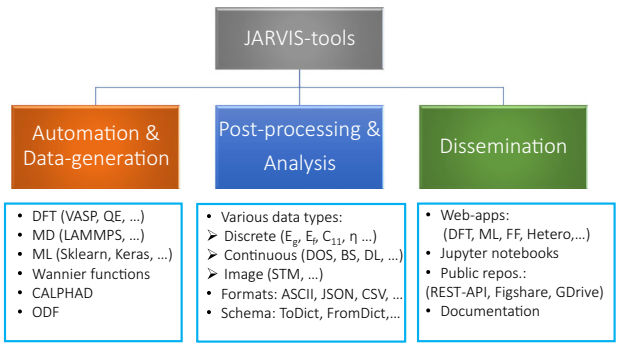

Fig. 3 Three main components of the JARVIS-tools package and their capabilities.

```
>>> from jarvis.core.atoms import Atoms
>>> box = [[2.715, 2.715, 0], [0, 2.715, 2.715], [2.715, 0, 2.715]]
>>> coords = [[0, 0, 0], [0.25, 0.25, 0.25]]
>>> elements = ["Si", "Si"]
>>> Si = Atoms(lattice_mat=box, coords=coords, elements=elements)
>>> density = round(Si.density,2)
>>> print (density)
2.33
>>>
>>> from jarvis.db.figshare import data
>>> dft_3d = data(dataset='dft_3d')
>>> print (len(dft_3d))
36099
>>> from jarvis.io.vasp.inputs import Poscar
>>> for i in dft_3d:
...     atoms = Atoms.from_dict(i['atoms'])
...     poscar = Poscar(atoms)
...     jid = i['jid']
...     filename = 'POSCAR-'+jid+'.vasp'
...     poscar.write_file(filename)
>>> dft_2d = data(dataset='dft_2d')
>>> print (len(dft_2d))
1070
>>> for i in dft_2d:
...     atoms = Atoms.from_dict(i['atoms'])
...     poscar = Poscar(atoms)
...     jid = i['jid']
...     filename = 'POSCAR-'+jid+'.vasp'
...     poscar.write_file(filename)
```

Fig. 4 Examples of using python classes in JARVIS-tools for constructing 'Atoms' class and downloading data. More tutorial-based examples are available on the documentation pages.

calculations with external software packages. An example of the 'Atoms' class is shown in Fig. 4.

The 'Atoms' class along with many other modules in JARVIS-tools are used to generate input files for automating software codes. Currently, JARVIS-tools can be used to automate DFT calculations with packages such as Vienna Ab-initio simulation package (VASP)[57,58], Quantum Espresso (QE)[59]; MD with Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)[60]; ML with Scikit-learn[61], Keras[62], and LightGBM[63]; Wannier calculations with Wannier90[64] and Wanniertools[65]. A number of predefined workflows are available in JARVIS-tools that are continuously being used to calculate properties of uncharacterized or existing materials in the database. Three workflows are shown in Fig. 5. For DFT calculations, an input Atoms class is used to generate input files for VASP (Fig. 5a) with the 'VaspJob' class in order to calculate the desired properties, such as the energy. We automatically perform calculations to converge numerical parameters like the k-points and plane-wave cut-off for individual materials. Geometry optimization is then carried out with energy, force, and stress relaxation. We have chosen a particular set of pseudopotentials or PAWs as tested and recommended by the software developers of various codes. Subsequent properties, such as band structure, dielectric function, elastic constants, piezoelectric constants or spin-orbit spillage are computed on the relaxed structure. Later, custom jobs can also be run on the optimized structure using 'VaspJob', such as Wannier90 calculations using the 'Wannier90-Win' class, which generates the input files for an Atom class and a

chosen set of pseudopotentials, disentanglement window and other controlling parameters. All of these steps produce a JavaScript Object Notation (JSON) file once the calculations are done as a signature of their completion. The workflows can be restarted from intermediate computations, making the calculations robust to interruptions due to computer failure, etc. We also add several error-handlers in the workflows to automatically re-submit a calculation if a typical error is encountered.

A similar workflow is shown for an example of FF based on LAMMPS calculations in Fig. 5b. Here, for a particular force-field such as Ni-Al[53], for example, all the structures related to Ni, Al, and Ni–Al are obtained from the DFT database and converted into a LAMMPS input format using 'Atoms', 'LammpsData' and 'LammpsJob' objects. Then a series of geometry optimization, vacancy formation energy, surface energy, and phonon-related calculations are run, based on the symmetry of the structure. All of these steps use a set of ".mod" module files with input parameters that control respective LAMMPS calculations. The obtained results are compared with corresponding DFT data, to evaluate the quality of an FF for a particular system or simulation.

In machine learning calculations, the input materials-data is transformed into several machine-readable descriptors[66] such as CFID dataset or STM image 'numpy' arrays. As we are not going to generate another set of data for testing ML models, we split the dataset into training and testing sets in a 90:10 or similar split. Using k-fold cross-validation, we obtain hyperparameters for the chosen algorithm, for example, the number of trees, learning rate, etc. in the case of Gradient Boosting Decision Tree (GBDT). We choose the optimized parameters and train on 90% train data and test on the 10% test data to evaluate the truly predictive performance on unseen data. We also carry out k-fold cross-validation using the finalized model to get model uncertainty. Later, we can analyze interpretability with techniques such as feature importance in tree-based algorithms or filters in neural networks. These models are saved in Pickle, cPickle and Joblib modules for model persistency. We also carry out uncertainty analysis using methods such as prediction interval and Monte-Carlo dropouts[67]. A few examples and Jupyter notebooks are provided on the GitHub page to illustrate the above-mentioned methods. More details about the individual python modules mentioned above can be found in the JARVIS-tools documentation (https://jarvis-tools.readthedocs.io/en/latest/). A documentation on integrating JARVIS-tools with the database is available at (https://jarvis-materials-design.github.io/dbdocs/).

After running the automated calculations, the data is post-processed to predict various material properties (such as bandgap, formation energy, spin-orbit spillage, SLME, density of states, phonons, dielectric function, or STM image). Many of the python classes use 'ToDict' and 'FromDict' methods that help store the metadata. These metadata are then used with HTML[68], Javascript, Flask[69] and other related software to make web-pages and web-apps. The metadata is also shared in public repositories such as Figshare (https://figshare.com/authors/Kamal_Choudhary/4445539), and JARVIS-Representational state transfer (REST) API, based on the MGI philosophy of creating and using interoperable datasets. Note that through the JARVIS-REST API, a user can download JARVIS data and can also upload/store their own data. If the stored data follows the schema (in XSD format), then the API automatically generates HTML pages for the user's data. The data generated in JARVIS is mainly stored in Extensible Markup Language (XML), JavaScript Object Notation (JSON), Comma-Separated Values (CSV) or American Standard Code for Information Interchange (ASCII) format and, again, JARVIS-tools can be used to analyze the pre-calculated data for materials design. A wrapper-code for the REST-API upload and download is available at (https://github.com/usnistgov/jarvis/blob/master/jarvis/db/restapi.py). An example of downloading precalculated dataset with JARVIS-tools is shown in Fig. 4. JARVIS-tools, along with the various software shown in Fig. 3, has led to several databases shown in Fig. 6.

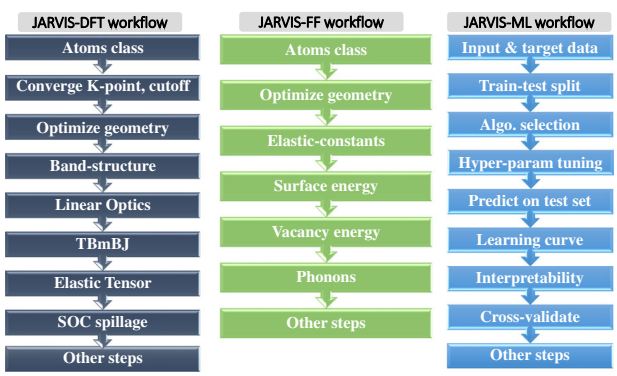

Fig. 5 Flowcharts showing some of the main steps used in most-commons calculations. a JARVIS-DFT, (**b**) JARVIS-FF, and (**c**) JARVIS-ML workflows.

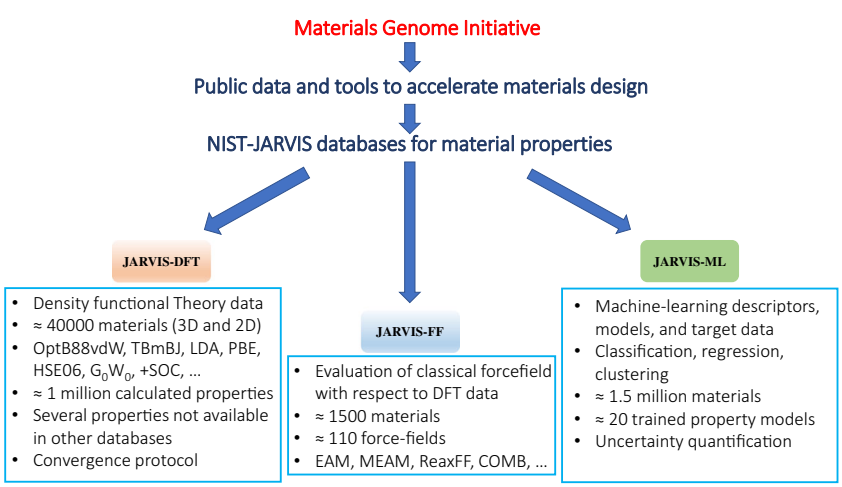

Fig. 6 Three main databases in JARVIS and a summary of their contents.

JARVIS-DFT

Density functional theory is one of the most commonly used techniques in condensed-matter physics to solve real-world materials problems. In DFT, instead of solving the fully interacting Schrödinger equation, we solve the Kohn-Sham equations, which describe an effective non-interacting problem, greatly improving computational efficiency. Although exact in principle, DFT requires several approximations in practice. In particular, various levels of approximation to the exchange-correlation functional are possible, which require different computational effort. Most existing DFT databases use the common GGA-PBE throughout all the material-classes. JARVIS-DFT can be viewed as an attempt to build a repository beyond existing DFT databases. JARVIS-DFT[5,6,23–25,28,29,45,49,50] was started in 2017 and contains data for ≈40,000 materials, with ≈1 million calculated properties, mainly based on the VASP package. Although there are several DFT-functionals adopted in JARVIS-DFT, we use vdW-DF-OptB88 consistently for all the 3D, 2D, 1D, and 0D materials. This functional has been shown to provide accurate predictions for lattice-parameters and energetics for both vdW and non-vdW bonded materials[28]. In addition to hosting 3D bulk materials, the database consists of 2D monolayer, 1D-nanowire, and 0D-molecular materials (as shown in Table 2). However, to date, 3D and 2D materials have primarily been distributed publicly. Moreover, other exchange-correlation functionals are considered (as shown in Table 3), which can help estimate the prediction uncertainty. While vdW-DF-OptB88 can predict accurate lattice parameters and formation energies, bandgaps are still underestimated. Calculations with hybrid functionals (such as range-separated HSE06 and PBE0) and many-body approaches (such as G_0W_0) remain too computationally expensive[21] to use in a high-throughput methodology for thousands of materials. Hence, a meta-GGA Tran-Blaha-modified Becke-Johnson (TBmBJ) potential is used to provide a good balance between computational expense and accuracy. The TBmBJ accuracy is shown to be close enough to the high-level methods such as HSE06 at up to ten times lower computational expense[52]. Accurate prediction of optical gaps by calculation of the frequency-dependent dielectric function is important for several applications, for example, solar-cell efficiency calculations. Accurate prediction of bandgaps also helps in obtaining accurate frequency-dependent dielectric functions, which can be critical for solar-cell efficiency calculations; however, TBmBJ cannot describe the excitonic nature of electron-hole pairs in low-dimensional materials. In addition to TBmBJ, we are generating HSE06, PBE0, G_0W_0, and DMFT datasets, which can be considered as beyond-DFT methods discussed in the next section. Next, SOC is varied to analyze the differences introduced by this coupling. These differences are used to discover 3D and 2D topological materials. In addition, several DFT databases are developed including properties such as frequency-dependent dielectric function and electric field gradient. A few important protocols such as k-point automatic convergence are also introduced. A snapshot of the JARVIS-DFT website along with a list of properties that are available is shown in Fig. 7. JARVIS-DFT has several filtering options on the website to screen candidate materials. We provide the input files as downloadable .zip files, especially for the users who do not have much expertize in using python-based codes. Raw input and output files (on the order of 1 terabyte) will soon be made publicly available through the Figshare repository, NIST-Materials data repository, and Materials Data Facility (MDF). A summary table, with the number of data available with vdW-DF-OptB88 and other methods, is shown in Tables 2 through 4. Table 2, Table 3, and Table 4 provides a summary of available materials classes, DFT functionals used and materials properties available in the JARVIS-DFT database, respectively.

Table 2. A brief summary of datasets available in the JARVIS-DFT.

Material classes	Numbers
3D-bulk	34,622
2D-bulk	2293
1D-bulk	235
0D-bulk	413
2D-monolayer	1105
2D-bilayer	102
Molecules	12
Heterostructure	3
Total DFT calculated systems	38,785

JARVIS-beyond-DFT

While quantum mechanical methods in single-particle theories such as DFT or DFT+U methods (mainly GGA) are fast and can predict accurate results for most structural parameters, even when relatively strong electron correlations are present, qualitative predictions of excited state properties may require beyond-DFT methods[70]. Beyond-DFT calculations have been applied to many materials systems, including cuprates and Fe-based high-temperature superconductors, Mott insulators, heavy Fermion systems, semiconductors, photovoltaics, and topological Mott insulators[70]. In the last few decades, both perturbative and stochastic approaches have been developed to understand these strongly correlated materials. These approaches, including Dynamical Mean Field Theory (DMFT)[71], the GW approximation, or hybrid exchange-correlation functionals are often called beyond-DFT methods since they go beyond the limit of semilocal DFT. The materials design community often requires benchmarking for particular cases, where it is necessary to use beyond-DFT methods, in order to assess accuracy of the results. In the JARVIS-Beyond-DFT database we are building a database of spectral functions and related quantities as computed using meta-GGA, GW, hybrid functionals, and LDA+DMFT for head-to-head comparison on 100+ materials.

In the JARVIS-Beyond-DFT[70] database we try to answer a few key questions regarding discoveries through a materials database for quantum materials. First, where is it necessary to use a beyond-DFT method, and which method to be use? Second, how do different "beyond-DFT" methods compare with experiments? Target materials include but are not limited to various transition metal oxides, perovskites and mixed perovskites, nickelates, transition metal dichalcogenides, and a wide range of metals starting from alkali metals to transition metals, and various Iron-based superconductors. JARVIS-Beyond-DFT will be distributed through the website: https://jarvis.nist.gov/jarvisbdft/.

JARVIS-FF

Classical force-field-/interatomic-potential-based simulations are the workhorse technique for large scale atomistic simulations.

Table 3. A brief summary of functionals used in optimizing crystal geometry in the JARVIS-DFT.

Functionals	Numbers
vdW-DF-OptB88 (OPT)	38785
vdW-DF-OptB86b (MK)	109
vdW-DF-OptPBE (OR)	111
PBE	99
LDA	92

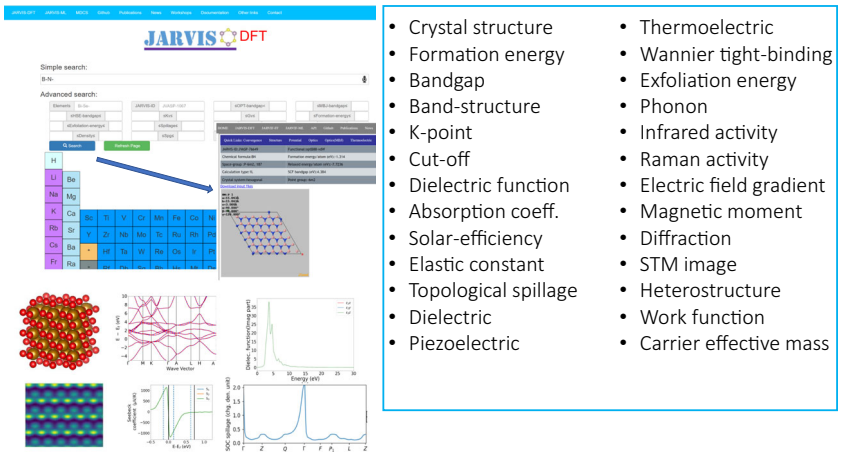

Fig. 7 A snapshot of JARVIS-DFT website and summary of its contents.

Table 4. A brief summary of material-properties available in the JARVIS-DFT.

Property	Numbers
Optimized crystal-structure (OPT)	38,785
Formation-energy (OPT)	38,785
Bandgap (OPT)	38,785
Exfoliation energy (OPT)	819
Bandgap (TBmBJ)	15,655
Bandgap (HSE06)	40
Bandgap (PBE0)	40
Bandgap (G_0W_0)	15
Bandgap (DMFT)	11
Frequency dependent dielectric tensor (OPT)	34,045
Frequency dependent dielectric tensor (TBmBJ)	15,655
Elastic-constants (OPT)	15,500
Finite-difference phonons at Γ-point (OPT)	15,500
Work-function, electron-affinity (OPT)	1105
Theoretical solar-cell efficiency (SLME) (TBmBJ)	5097
Topological spin-orbit spillage (PBE+SOC)	11,500
Wannier tight-binding Hamiltonians (PBE+SOC)	1771
Seebeck coefficient (OPT, BoltzTrap)	22,190
Power factor (OPT, BoltzTrap)	22,190
Effective mass (OPT, BoltzTrap)	22,190
Magnetic moment (OPT)	37,528
Piezoelectric constant (OPT, DFPT)	5015
Dielectric tensor (OPT, DFPT)	5015
Infrared intensity (OPT, DFPT)	5015
DFPT phonons at Γ-point (OPT)	5015
Electric field gradient (OPT)	15,187
Non-resonant Raman intensity (OPT, DFPT)	250
Scanning tunneling microscopy images (PBE+SOC)	770

The database is continuously expanding[5,6,23–25,28,29,45,49,50].

They are especially suited for temperature-dependent and defect-related phenomena. Several varieties of FFs differ based on the materials system and the underlying phenomena under investigation, e.g., whether they include bond-angle information and fixed or dynamic charges. Also, they are generally designed for particular applications and phases, making it difficult to ascertain whether they will perform well in simulations for which they were not explicitly trained. JARVIS-FF[25,53] is a collection of LAMMPS calculation-based data consisting of crystal structures, formation energies, phonon densities of states, band structures, surface energies and defect formation energies. There are ≈110 FFs in the database, for which the corresponding crystal structures are obtained from JARVIS-DFT, converted to LAMMPS format inputs, and used in a series of LAMMPS calculations to produce the aforementioned properties. These properties, when compared with corresponding DFT data, can help a user analyze the quality of a force-field for a particular application. Examples include the comparison of DFT convex hull with FF, elastic modulus, surface energy and vacancy formation energy data. Some types of FFs included are EAM, MEAM, Bond-order and Tersoff, COMB, and ReaxFF as shown in Table 5. Furthermore, we plan to include several recently developed machine learning force-fields into JARVIS-FF. A snapshot of the JARVIS-FF website is also shown in Fig. 8.

Table. 5. A summary of various types of force-fields available in the JARVIS-FF[25,53].

Force-fields	Numbers
EAM	92
Tersoff	9
ReaxFF	5
COMB	6
AIREBO	2
MEAM	1
EIM	1

JARVIS-ML

Machine learning has several applications in materials science and engineering[12,72,73], such as automating experimental data analysis, discovering functional materials, optimizing known ones by accelerating conventional methods such as DFT, automating literature searches, discovering physical equations, and efficient clustering of materials and their properties. There are several data types that can be used in ML such as scalar data (e.g., formation energies, bandgaps), vector/spectra data (e.g., density of states, dielectric function, charge density, X-ray diffraction patterns, etc.),

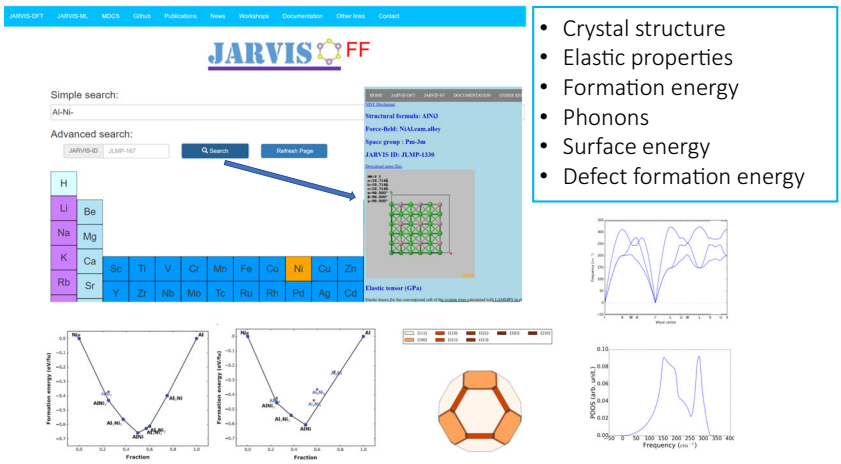

Fig. 8 A snapshot of JARVIS-FF website and summary of its contents.

Table. 6. A summary of classical force-field inspired descriptors (CFID)-descriptor datasets available in the JARVIS-ML.

CFID-Dataset	Number of materials
JARVIS-DFT 3D	39,240
JARVIS-DFT 2D	1105
AFLOW	820,082
OQMD	460,046
Materials-project	83,964
Crystallography Open Database (COD)	11,783
QM9	13,385
Total	1,429,605

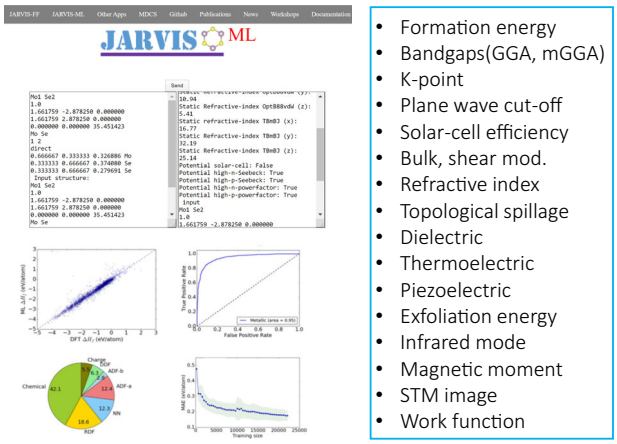

Fig. 9 A snapshot of JARVIS-ML website and summary of its contents.

image-based data (such as scanning tunneling microscopy and transmission electron microscopy images), and natural language processing-based data (such as scientific papers). In addition, ML can be applied on a variety of materials classes such as bulk crystals, molecules, proteins and free-surfaces.

Currently, there are two types of data that are machine-learned in JARVIS-ML[45,49,50,54,55]: discrete and image-based. The discrete target is obtained from the JARVIS-DFT database for 3D and 2D materials. There have been several descriptor developments as attempts to capture the complex chemical-structural information of a material[66]. We compute CFID descriptors for most crystal structures in various databases (as shown in Table 6). Many of these structures are non-unique but can still be used for pre-screening applications[45]. The CFID can also be applied to other materials classes such as molecules, proteins, point defects, free surfaces, and heterostructures, which are currently ongoing projects. These descriptor datasets, along with JARVIS-DFT and other databases, act as input and outputs for machine learning algorithms. The CFID consists of 1557 descriptors for each material: 438 average chemical, 4 simulation-box-size, 378 radial charge-distribution, 100 radial distribution, 179 angle-distribution up to first neighbor, and another 179 for the second neighbor, 179 dihedral angle up to fist neighbor and 100 nearest neighbor descriptors. More details can be found in ref. [54]. Currently, we provide CFID descriptors only, but other descriptors such as Coulomb-matrix, and sine-matrix will be provided soon. With CFID descriptors, we trained several classification and regression tasks. Once these models are trained, parameters are stored that can predict the properties of an arbitrary compound quickly. We developed a web-based application to host the trained models, as shown in Fig. 9, and a list of the trained properties are displayed there as well. We note that classical quantities such as bulk modulus, maximum infrared (IR) active mode, and formation energies can be accurately trained, especially with regression models. For other properties such as bandgaps, magnetic moments, piezoelectric coefficients, thermoelectric coefficients, high accuracy models are obtained for classification tasks only. In addition to the descriptor-based data, we develop Scanning Tunneling Microscopy (STM)[49] image classification models that can be used to accelerate the analysis of STM data. The images are converted into a black/white image to identify spots with/without atoms. The model's accuracy is compared with respect to DFT data or experiments wherever applicable.

Derived apps

The knowledge developed through the above-mentioned databases and tools can serve as static content, as well as accessed through dynamic user-defined inputs. Derived applications (apps) are designed to help a user analyze the combinatorics in the data. Based on the databases and tools discussed above, several apps are derived from JARVIS such as JARVIS-Heterostructure[29], JARVIS-Wannier TB, and JARVIS-ODF. JARVIS-Heterostructure (as shown in Fig. 10a) can be used to characterize heterojunction type and modeling interfaces for exfoliable 2D materials. We classify these heterostructures into type-I, II, and III systems according to Anderson's rule, which is based on the band-alignment with respect to the vacuum potential of non-interacting monolayers, obtained from JARVIS-DFT. The app also

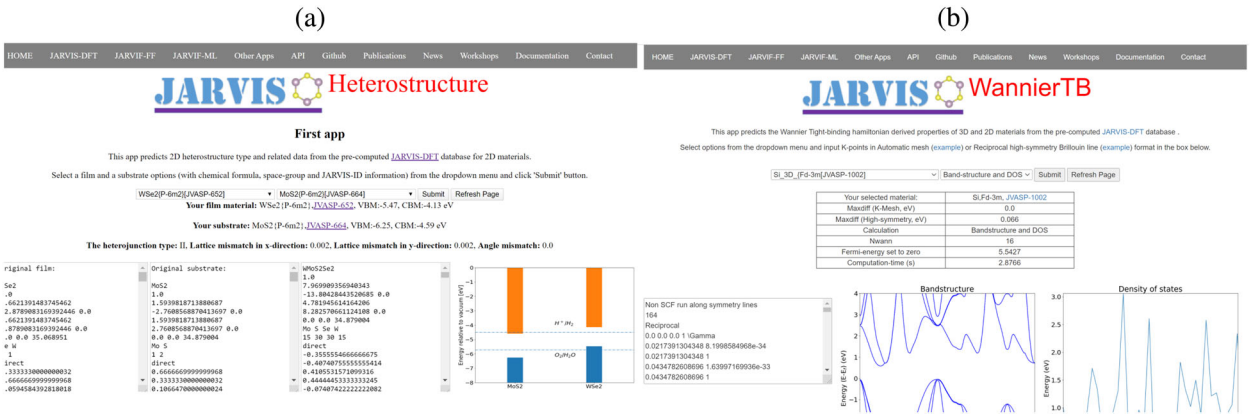

Fig. 10 Snapshots of JARVIS-DFT derived apps. a JARVIS-Heterostructure and (**b**) JARVIS-Wannier Tight Binding.

Table. 7. Mean absolute error (MAE) for JARVIS-DFT data with respect to available experimental data for various material properties.

Property	#Materials	MAE	Typical range
Formation energy (eV/atom)	1317	0.128	−4 to 2
OptB88vdW-bandgaps (eV)	54	1.33	0 to 10
TBmBJ-bandgaps (eV)	54	0.51	0 to 10
Bulk modulus (GPa)	21	5.75	0 to 250
Electronic (휀_{11}) OPT	28	3.2	0 to 60
Electronic (휀_{11}) MBJ	28	2.62	0 to 60
Solar-eff. (SLME) (%) (MBJ)	5	6.55	0 to 33.7
Max. piezoelectric strain coeff (Cm^{-2})	16	0.21	0 to 2
Dielectric constant (휀_{11}) (DFPT)	16	2.46	0 to 60
Seebeck coefficient (μV/K)	14	54.7	−600 to 600
Electric field gradient V_{zz} ($10^{21}Vm^{-2}$)	37	1.17	0 to 100
IR mode (cm^{-1})	8	8.36	0 to 4000

generates crystallographic positions for the heterostructure that could be used as input for subsequent calculations. JARVIS-WannierTB (as shown in Fig. 10b) can be used to solve Wannier Tight Binding Hamiltonians on arbitrary k-points for 3D and 2D materials. Properties such as the band structure and the density of states can be predicted on the fly from this app. In addition, many other apps are being developed, which are primarily based on the Flask python package[69].

The JARVIS-ODF (Orientation Distribution Function) library is under development, which aims to calculate volume-averaged (meso-level) material properties, including the elasto-plastic deformation behavior, using the property data available for single crystals in the JARVIS database. Once generated, the JARVIS-ODF library will be capable of obtaining such material properties for all crystalline structures.

Accuracy and precision analysis

In simulations, accuracy refers to the degree of closeness between a calculated value and a reference value, which can be from an experiment or a high-fidelity theory. Precision refers to the degree of closeness between numerical approaches to solving a certain model, including the effect of convergence and other simulation parameters.

In JARVIS-DFT, the accuracy of the DFT data is obtained by comparing it to available experimental results (see Supplementary Tables 1–9). The accuracy of JARVIS-FF and JARVIS-ML, instead, is given with respect to DFT results. Note that the numbers of high-quality experimental measurements or high-fidelity calculations for a given property are often low. Therefore, the accuracy metrics we derive in our works are obtained only for the few cases we can directly compare, not for the entire dataset. In Table 7, we provide accuracy metrics for some material properties in the JARVIS-DFT with respect to experiments. In addition to the scalar data, vector/continuous data, such as frequency dependent dielectric function and Scanning Tunneling Microscopy (STM) images, are compared to a handful of experimental data points as well. Details of individual properties can be found in refs. [6,28,45,48–50,54,55].

JARVIS-FF data accuracy is calculated with respect to the DFT data, for properties such as the convex hull, bulk modulus, phonon frequencies, vacancy formation energies and surface energies. In refs. [25,53], we showed this through several examples, including the comparison of Ni–Al and Cu–O–H systems convex hulls to DFT data. We also showed examples of comparing defect formation energies, surface energies and its effects on Wulff-shape. Although these accuracy analyses are based on 0 K DFT data, they are useful in predicting temperature-dependent and dynamical behavior because we consider several crystal prototypes of a system.

JARVIS-ML model accuracy is evaluated on the test-set (usually 10%) representing previously unseen DFT data for both regression and classifications models. Accuracy of regression and classification models are reported in terms of mean absolute error (MAE) and Receiver Operating Characteristic (ROC) Area Under Curve (AUC) metric, respectively. A brief summary of regression and classification model accuracy results is given below in Tables 8 and 9. Details of the accuracy analyses are provided in refs. [45,49,50,54,55].

Precision analysis can refer to a wide variety of optional selections of simulation set-ups. Examples of precision analysis in JARVIS-DFT are using our convergence protocols for k-points and plane-wave cutoff, and the convergence of Wannier tight-binding Hamiltonians. Using a converged k-point mesh and plane-wave cutoff[26] for each individual material is necessary to obtain high-quality data. Note that these DFT convergences are carried out for energies of the system only, and not for other properties. However, we impose tight convergence parameters for both k-points and energy cutoff (0.001 eV/cell), which typically results in other physical quantities being converged as well. In JARVIS-FF, comparison across structure-minimization methods for calculating surface and vacancy formation energy values are examples of precision analysis[25]. We find that the FF simulation setups ('refine' and 'box' methods) have minimal effect on the FF-based predictions. For classification ML models, precision is the ratio $\frac{TP}{TP+FP}$ where TP is the number of true positives and FP the number of false positives, which can be derived from the confusion. Precision analysis for classification ML model for STM Bravais-lattices are available in ref. [49]. We find high precision (more than

Table 8. Performance of regression machine learning models in JARVIS-ML with JARVIS-DFT data using OptB88vdW (OPT) and TBmBJ (MBJ) with mean absolute error (MAE).

Property	Training data	MAE	MAD
Formation energy (eV/atom)	24549	0.12	0.81
OPT bandgap (eV)	22404	0.32	1.05
MBJ bandgap (eV)	10499	0.44	1.60
Bulk mod., Kv (GPa)	10954	10.5	49.95
Shear mod., Gv (GPa)	10954	9.5	23.26
Refr. Index(x) (OPT)	12299	0.54	1.15
Refr. Index(x) (MBJ)	6628	0.45	1.03
IR mode (OPT) (cm^{-1})	3411	77.84	316.7
Max. Born eff. charge (OPT) (e)	3411	0.60	1.48
Plane-wave cutoff (OPT) (eV)	24549	85.0	370.6
K-point length (OPT) (Å)	24549	9.09	22.23
2D-Exfoliation energy(OPT) (eV/atom)	616	37.3	46.09

The mean absolute deviation (MAD) of properties are also included.

Table 9. Performance of the classification machine learning models in JARVIS-ML with JARVIS-DFT data using OptB88vdW (OPT) and TBmBJ (MBJ) with Receiver Operating Characteristic (ROC) Area Under Curve (AUC) metric.

Property	Number of datapoints	ROC AUC
Metal/non-metal (OPT)	24549	0.95
Magnetic/Non-magnetic (OPT)	24549	0.96
High/low solar-cell efficiency (TBmBJ)	5097	0.90
High/low piezoelectric coeff	3411	0.86
High/low Dielectric	3411	0.93
High/low n-Seebeck coeff	21899	0.95
High/low n-power factor	21899	0.80
High/low p-Seebeck coeff	21899	0.96
High/low p-power factor	21899	0.82

Random guessing and perfect ROC AUC are 0.5 and 1, respectively.

0.87) for all of the 2D-Bravais lattices. Precision analysis for regression tasks are still ongoing and will be available soon.

Future work

Given that the number of all possible materials[74] could be of the order of 10^{100}, and furthermore existing materials properties can be computed at increasing levels of accuracy/cost, the JARVIS databases will always be incomplete. This represents an opportunity for JARVIS to be drastically expanded in the future. Future work will be aimed at addressing some of the limitations of the existing databases, and may include additions like defect/disorder properties, magnetic ordering, non-linear optoelectronics, more beyond-DFT calculations, temperature-dependent properties, integration with experiments, and more detailed uncertainty analysis. Moreover, several ML models and methods for data-prediction and uncertainty quantification will be developed for 'explainable AI' (XAI) and transfer-learning (TL)-based research. Other derived apps such as JARVIS-ODF, JARVIS-Beyond-DFT, JARVIS-GraphConv, and JARVIS-STM are also being developed. In addition to the technical aspects, the broader impact of the infrastructure will be to provide a research platform that will allow maximum participation of worldwide researchers. NIST-JARVIS currently hosts pre-computed data and would host on-the-fly calculation resources also. To make the data-processing user-friendly, we have a few filtering options on the JARVIS-DFT website. Furthermore, advanced filtering tools will be available through ElasticSearch package soon. ElasticSearch integration will allow cross-filtering among several databases. We are also working on several visualization tool integration using Plotly, Javascript and XSLT which will be available on the web soon.

In summary, we described the Joint Automated Repository for Various Integrated Simulations (JARVIS) platform, which consists of several databases and computational tools to help accelerate materials design and enhance industrial growth. JARVIS includes three major databases: JARVIS-DFT for density functional theory calculations, JARVIS-FF for classical force-field calculations, and JARVIS-ML for ML predictions. In addition, we provide JARVIS-tools, which is used to generate the databases. The generated data is provided publicly with several example notebooks, documentation and calculation examples to illustrate different components of the infrastructure. We believe the publicly available data and resources provided here will significantly accelerate futuristic materials-design in various areas of science and technology.

METHODS

The entire study was managed, monitored, and analyzed using the modular workflow, which we have made available (Please note that commercial software is identified to specify procedures. Such identification does not imply recommendation by the National Institute of Standards and Technology) on our JARVIS-tools GitHub page (https://github.com/usnistgov/jarvis).

Density functional theory calculations

The DFT calculations are mainly carried out using the Vienna Ab-initio simulation package (VASP)[57,58]. We use the projected augmented wave method and OptB88vdW functional[51], which gives accurate lattice parameters for both van der Waals (vdW) and non-vdW solids[28]. Both the internal atomic positions and the lattice constants are allowed to relax in spin-unrestricted calculations until the maximal residual Hellmann–Feynman forces on atoms are smaller than 0.001 eV $Å^{-1}$ and energy-tolerance of 10^{-7} eV. We do not consider magnetic orderings besides ferromagnetic yet, because of a high computational cost. We note that nuclear spins are not explicitly considered during the DFT calculations. The list of pseudopotentials used in this work is given on the GitHub page. The k-point mesh and plane-wave cut-off were converged for each material using the automated procedure described in ref. [26]. The elastic constants are calculated using the finite difference method with six finite symmetrically distinct distortions. The thermoelectric coefficients such as power factor and Seebeck coefficients are obtained with the BoltzTrap code with Constant Relaxation Time approximation (CRTA)[75]. Optoelectronic properties such as dielectric function and solar-cell efficiency are calculated using linear-optics methods mainly using OptB88vdW and TBmBJ. We also compared such data with HSE06 and G_0W_0. The piezoelectric, dielectric and phonon modes at Γ-point are calculated using Density Functional Perturbation Theory (DFPT). Topological spillage for identifying topologically non-trivial materials is calculated by comparing DFT wave functions with/without SOC[5,24]. 2D exfoliation energies are calculated by comparing bulk and 2D monolayer energy per atom. The 2D heterostructure[29] behavior is predicted using Zur and Anderson methods. Wannier tight binding Hamiltonians are generated using the Wannier90 code[64]. 2D STM images are predicted using the Tersoff-Hamman method[49].

Force-field calculations

Classical force-field calculations are carried out with the LAMMPS software package[60]. In our structure minimization calculations, we used 10^{-10} eV$Å^{-1}$ for force convergence and 10,000 maximum iterations. The geometric structure is minimized by expanding and contracting the simulation box with 'fix box/relax' command and adjusting atoms until they reach the force convergence criterion. These are commonly used computational set-up parameters. After structure optimization point vacancy defects are

created using Wycoff-position data. Free surfaces for maximum miller indices up to 3 are generated. The defect structures were required to be at least 1.5 nm long in the *x*, *y*, and *z* directions to avoid spurious self-interactions with the periodic images of the simulation cell. We enforce the surfaces to be at least 2.5 nm thick and with 2.5 nm vacuum in the simulation box. The 2.5 nm vacuum is used to ensure no self-interaction between slabs, and the slab-thickness is used to mimic an experimental surface of a bulk crystal. Using the energies of perfect bulk and surface structures, surface energies for a specific plane are calculated. We should point out that only unreconstructed surfaces without any surface-segregation effects are computed, as our high-throughput approach does not allow for taking into account specific, element dependent reconstructions yet. Phonon structures are generated mainly using the Phonopy package interface[76].

Machine learning training

Machine learning models are mainly trained using Scikit-learn[61], Keras[62], and LightGBM[63] (TensorFlow backend) software. For DFT generated scalar data such as formation energies, bandgaps, exfoliation energies etc. the crystal structures are converted into a Classical Force-field Inspired Descriptors (CFID) input array and the DFT data is used as target data, which is then train-test split in a ratio of 90: 10. Preprocessing such as 'VarianceThreshold', 'StandardScalar' are used before ML training. Regression models' performance are generally reported in terms of Mean Absolute Error (MAE) or r^2, while that for classification models using the Receiver Operating Characteristic (ROC) Area Under Curve (AUC) value which lie between 0.5 and 1.0. Several other analyses such as feature importance, k-fold cross validation and learning curve are carried out after the model training. The trained model is saved in pickle and joblib formats for model persistence. All the web-apps are developed using JavaScript, Flask, and Django packages[69].

DATA AVAILABILITY

JARVIS-related data is available at the JARVIS-API (http://jarvis.nist.gov), JARVIS-DFT (https://jarvis.nist.gov/jarvisdft/), JARVIS-FF (https://jarvis.nist.gov/jarvisff/), JARVIS-ML (https://jarvis.nist.gov/jarvisml/) websites. The metadata is also available at the Figshare repository, see https://figshare.com/authors/Kamal_Choudhary/4445539.

CODE AVAILABILITY

Python-language based codes with examples are available at JARVIS-tools page: https://github.com/usnistgov/jarvis.

Received: 3 July 2020; Accepted: 12 October 2020;
Published online: 12 November 2020

REFERENCES

1. Curtarolo, S. et al. AFLOWLIB. ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
2. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *Apl. Mater.* **1**, 011002 (2013).
3. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater* **1**, 15010 (2015).
4. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater Sci.* **111**, 218–230 (2016).
5. Choudhary, K., Garrity, K. F. & Tavazza, F. High-throughput discovery of topologically non-trivial materials using spin-orbit spillage. *Sci. Rep.* **9**, 1–8 (2019).
6. Choudhary, K., Kalish, I., Beams, R. & Tavazza High-throughput identification and characterization of two-dimensional materials using density functional theory. *Sci. Rep.* **7**, 5179 (2017).
7. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys. Mats.* **2**, 036001 (2019).
8. Chung, Y. G. et al. Computation-ready, experimental metal–organic frameworks: a tool to enable high-throughput screening of nanoporous crystals. *Chem. Mater.* **26**, 6185–6192 (2014).
9. Green, M. L. et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *J. Appl. Phys. Rev.* **4**, 011105 (2017).
10. Hattrick-Simpers, J. R., Gregoire, J. M. & Kusne, A. G. Perspective: composition–structure–property mapping in high-throughput experiments: turning data into knowledge. *APL Mater.* **4**, 053211 (2016).
11. Zakutayev, A. et al. An open experimental database for exploring inorganic materials. *Sci. Data.* **5**, 180053 (2018).
12. Vasudevan, R. K. et al. Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS Commun.* **9**, 821–838 (2019).
13. Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. *APL Maters* **4**, 053208 (2016).
14. Schleder, G. R., Padilha, A. C., Acosta, C. M., Costa, M. & Fazzio, A. J. From DFT to machine learning: recent approaches to materials science–a review. *J. Phys. Mater.* **2**, 032001 (2019).
15. Ceder, G. J. Opportunities and challenges for first-principles materials design and applications to Li battery materials. *MRS Bull.* **35**, 693–701 (2010).
16. Xi, L. et al. Discovery of high-performance thermoelectric chalcogenides through reliable high-throughput material screening. *J. Am. Chem. Soc.* **140**, 10785–10793 (2018).
17. Olson, G. B. & Kuehmann, C. Materials genomics: from CALPHAD to flight. *Scr. Mater.* **70**, 25–30 (2014).
18. Aykol, M. et al. The materials research platform: defining the requirements from user stories. *Matter* **1**, 1433–1438 (2019).
19. Callister, W. D. & Rethwisch, D. G. *Materials Science and Engineering*. Vol. 5 (John Wiley & Sons, NY, 2011).
20. de Pablo, J. J. et al. The materials genome initiative, the interplay of experiment, theory and computation. *Curr. Opin. Solid State Mater. Sci.* **18**, 99–117 (2014).
21. Sholl, D. & Steckel, J. A. *Density Functional Theory: A Practical Introduction*. (John Wiley & Sons, 2011).
22. Perdew, J. P., Burke, K. & Ernzerhof, M. J. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
23. Choudhary, K. et al. Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms. *Sci. Data* **5**, 180082 (2018).
24. Choudhary, K., Garrity, K. F., Jiang, J., Pachter, R. & Tavazza, F. Computational search for magnetic and non-magnetic 2D topological materials using unified spin–orbit spillage screening. *npj Comput. Mater* **6**, 1–8 (2020).
25. Choudhary, K. et al. High-throughput assessment of vacancy formation and surface energies of materials using classical force-fields. *J. Phys.* **30**, 395901 (2018).
26. Choudhary, K. & Tavazza, F. Convergence and machine learning predictions of Monkhorst-Pack k-points and plane-wave cut-off in high-throughput DFT calculations. *Comput. Mater. Sci.* **161**, 300–308 (2019).
27. Cooper, M. et al. Development of Xe and Kr empirical potentials for CeO_2, ThO_2, UO_2 and PuO_2, combining DFT with high temperature MD. *J. Phys.* **28**, 405401 (2016).
28. Choudhary, K., Cheon, G., Reed, E. & Tavazza, F. Elastic properties of bulk and low-dimensional materials using van der Waals density functional. *Phys. Rev. B* **98**, 014107 (2018).
29. Choudhary, K., Garrity, K. F., Pilania, G. & Tavazza, F. Efficient computational design of 2D van der Waals Heterostructures: band-alignment, lattice-mismatch, web-app generation and machine-learning. *arXiv* **2004**, 03025 (2020).
30. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford university press, 2017).
31. Kattner, U. R. Phase diagrams for lead-free solder alloys. *JOM* **54**, 45–51 (2002).
32. Acar, P., Ramazani, A. & Sundararaghavan, V. Crystal plasticity modeling and experimental validation with an orientation distribution function for ti-7al alloy. *Metals* **7**, 459 (2017).
33. Castelli, I. E. et al. New light-harvesting materials using accurate and efficient bandgap calculations. *Adv. En. Mater.* **5**, 1400915 (2015).
34. Stevanović, V., Lany, S., Zhang, X. & Zunger, A. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Phys. Rev. B* **85**, 115104 (2012).
35. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Cryst. Sect. B* **58**, 364–369 (2002).
36. Talirz, L. et al. Materials Cloud, a platform for open computational science. *arXiv* **2003**, 12510 (2020).
37. Tadmor, E. B., Elliott, R. S., Sethna, J. P., Miller, R. E. & Becker, C. A. J. The potential of atomistic simulations and the knowledgebase of interatomic models. *JOM* **63**, 17 (2011).
38. Aagesen, L. et al. Prisms: an integrated, open-source framework for accelerating predictive structural materials science. *JOM* **70**, 2298–2314 (2018).

39. Wheeler, D. et al. PFHub: the phase-field community hub. *J. Open Res. Softw.* **7**, 29–36 (2019).
40. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater Sci.* **68**, 314–319 (2013).
41. Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Cond. Mat.* **29**, 273002 (2017).
42. Mathew, K. et al. MPInterfaces: a materials project based Python tool for high-throughput computational screening of interfacial systems. *Comput. Mater Sci.* **122**, 183–190 (2016).
43. Setyawan, W., Gaume, R. M., Lam, S., Feigelson, R. S. & Curtarolo, S. High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. *ACS Comb. Sci.* **13**, 382–390 (2011).
44. Gibbs, Z. M. et al. Effective mass and Fermi surface complexity factor from ab initio band structure calculation. *npj Comput. Mater* **3**, 1–7 (2017).
45. Choudhary, K. et al. Accelerated discovery of efficient solar-cell materials using quantum and machine-learning methods. *Chem. Mater.* **31**, 5900 (2019).
46. Yu, L. & Zunger, A. Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials. *Phys. Rev. Lett.* **108**, 068701 (2012).
47. Liu, J. & Vanderbilt, D. Spin-orbit spillage as a measure of band inversion in insulators. *Phys. Rev. B* **90**, 125133 (2014).
48. Jha, D. et al. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* **10**, 1–12 (2019).
49. Choudhary, K. et al. Density functional theory and deep-learning to accelerate data analytics in scanning tunneling microscopy. *arXiv* **1912**, 09027 (2019).
50. Choudhary, K., Garrity, K. & Tavazza, F. Data-driven discovery of 3D and 2D thermoelectric materials. *J. Phys.* **32**, 47 (2019).
51. Klimeš, J., Bowler, D. R. & Michaelides, A. Chemical accuracy for the van der Waals density functional. *J. Phys. Cond. Matt.* **22**, 022201 (2009).
52. Tran, F. & Blaha, P. Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential. *Phys. Rev. Lett.* **102**, 226401 (2009).
53. Choudhary, K. et al. Evaluation and comparison of classical interatomic potentials through a user-friendly interactive web-interface. *Sci. Data* **4**, 1–12 (2017).
54. Choudhary, K., DeCost, B. & Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2**, 083801 (2018).
55. Choudhary, K. et al. High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric, and dielectric responses. *npj Comput. Mater* **6**, 64 (2020).
56. Saito, T. *Computational Materials Design*. Vol. 34 (Springer Science & Business Media, 2013).
57. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
58. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comp. Mat. Sci.* **6**, 15–50 (1996).
59. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys.* **21**, 395502 (2009).
60. Plimpton, S. *Fast Parallel Algorithms for Short-range Molecular Dynamics*. (Sandia National Labs., Albuquerque, NM, 1993).
61. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
62. Gulli, A. & Pal, S. *Deep learning with Keras*. (Packt Publishing Ltd, 2017).
63. Ke, G. et al. *Advances in Neural Information Processing Systems*. In *Proceedings of the 1995 Conference*. Vol. 8. 3146–3154 (Mit Press, 1996).
64. Mostofi, A. A. et al. wannier90: a tool for obtaining maximally-localised Wannier functions. *Comp. Phys. Comm.* **178**, 685–699 (2008).
65. Wu, Q., Zhang, S., Song, H.-F., Troyer, M. & Soluyanov, A. A. WannierTools: an open-source software package for novel topological materials. *Comput. Phys. Comm.* **224**, 405–416 (2018).
66. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater Sci.* **152**, 60–69 (2018).
67. Hammer, B. & Villmann, T. *ESANN*. 79–90 (Citeseer).
68. Musciano, C. & Kennedy, B. *HTML, the Definitive Guide*. (O'Reilly & Associates, 1996).
69. Grinberg, M. *Flask Web Development: Developing Web Applications with Python*. ("O"Reilly Media, Inc.", 2018).
70. Mandal, S., Haule, K., Rabe, K. M. & Vanderbilt, D. Systematic beyond-DFT study of binary transition metal oxides. *npjComput. Mater* **5**, 1–8 (2019).
71. Kotliar, G. et al. Electronic structure calculations with dynamical mean-field theory. *Rev. Mod. Phys.* **78**, 865 (2006).
72. Schmidt, Jonathan et al. "Recent advances and applications of machine learning in solid-state materials science.". *npj Comp. Mater* **5.1**, 1–36 (2019).
73. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater* **3**, 1–13 (2017).
74. Walsh, A. The quest for new functionality. *Nat. Chem.* **7**, 274–275 (2015).
75. Madsen, G. K. & Singh, D. BoltzTraP. A code for calculating band-structure dependent quantities. *Comp. Phys. Comm.* **175**, 67–71 (2006).
76. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).

ACKNOWLEDGEMENTS

K.C., K.F.G., and F.T. thank the National Institute of Standards and Technology for funding, computational, and data-management resources. K.C. thanks the computational support from XSEDE 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. Contributions by S.M., K.H., K.R., and D.V. were supported by NSF DMREF Grant No. DMR-1629059 and No. DMR-1629346. X.Q. was supported by NSF Grant No. OAC-1835690. B.G.S. and S.V.K. acknowledge work performed at the Center for Nanophase Materials Sciences, a US Department of Energy Office of Science User Facility. A.A. acknowledges partial support by CHiMaD (NIST award # 70NANB19H005). G.P. was supported by the Los Alamos National Laboratory's Laboratory Directed Research and Development (LDRD) program's Directed Research (DR) project #20200104DR. K.C. thanks for helpful discussion with several researchers including Faical Y. Congo, Daniel Wheeler, James Warren, Carelyn Campbell, Chandler Becker, Marcus Newrock, Ursula Kattner, Kevin Brady, Lucas Hale, Eric Cockayne, Philippe Dessauw from National Institute of Standards and Technology; Karen Sauer, Igor Mazin, Nirmal Ghimire, Patrick Vora from George Mason University; Rama Vasudevan, Maxim Ziatdinov from Oak Ridge National Lab, Deyu Lu, and Matthew Carbone from Brookhaven National Lab; Marnik Bercx, Dirk Lamoen from University of Antwerp; Yifei Mo from University of Maryland; Anubhav Jain and Sinead Griffin from Lawrence Berkeley National Laboratory; Surya Kalidindi from Georgia Tech.; Tyrel McQueen and David Elbert from Johns Hopkins University; Richard Hennig from University of Florida; Giulia Galli and Ben Blaiszik from University of Chicago; Qiang Zhu from University of Nevada-Las Vegas; Dilpuneet Aidhy from University of Wyoming; Susan B. Sinnott, Tao Liang from Pennsylvania State University.

AUTHOR CONTRIBUTIONS

K.C. designed the JARVIS workflows, carried out high-throughput calculations, analysis, and developed the websites. F.T. contributed to the development of k-point and other convergence protocol, Beyond-DFT development and several other analyses. K.G. contributed to the development of topological materials discovery and Wannier-tight binding Hamiltonian projects. A.C.E.R. assisted in the deployment of the web-apps. B.D.C., A.A., and A.G.K. contributed to the machine-learning tasks. A.J. B., A.H.R., A.C., V.S., A.D. contributed to the phonon data analysis. Z.T. contributed to the development of the JARVIS-API website. J.H.S. contributed to the experimental validation of some of the screened materials. J.J. and R.P. contributed in the solar-cell and topological materials discovery tasks. G.C., E.R., X.Q., H.Z., S.V.K., B.S., G.P. contributed to the discovery and characterization of low-dimensional materials. P.A. contributed to the elastic constant analysis task. S.M., K.R., D.V., and K.H. contributed to the Beyond-DFT project. All authors contributed to writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00440-1.

Correspondence and requests for materials should be addressed to K.C.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2020