4-10-2022

MaterialsAtlas.org: A Materials Informatics Web App Platform for Materials Discovery and Survey of State-Of-The-Art

Jianjun Hu
University of South Carolina - Columbia, jianjunh@cec.sc.edu

Stanislav Stefanov

Yuqi Song

Sadman Sadeed Omee

Steph-Yves Louis

See next page for additional authors

Publication Info

Published in *npj Computational Materials*, Volume 8, 2022, pages 65-.

npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

MaterialsAtlas.org: a materials informatics web app platform for materials discovery and survey of state-of-the-art

Jianjun Hu[1]✉, Stanislav Stefanov[1], Yuqi Song[1], Sadman Sadeed Omee[1], Steph-Yves Louis[1], Edirisuriya M. D. Siriwardane[1,2], Yong Zhao[1] and Lai Wei[1]

The availability and easy access of large-scale experimental and computational materials data have enabled the emergence of accelerated development of algorithms and models for materials property prediction, structure prediction, and generative design of materials. However, the lack of user-friendly materials informatics web servers has severely constrained the wide adoption of such tools in the daily practice of materials screening, tinkering, and design space exploration by materials scientists. Herein we first survey current materials informatics web apps and then propose and develop MaterialsAtlas.org, a web-based materials informatics toolbox for materials discovery, which includes a variety of routinely needed tools for exploratory materials discovery, including material's composition and structure validity check (e.g. charge neutrality, electronegativity balance, dynamic stability, Pauling rules), materials property prediction (e.g. band gap, elastic moduli, hardness, and thermal conductivity), search for hypothetical materials, and utility tools. These user-friendly tools can be freely accessed at http://www.materialsatlas.org. We argue that such materials informatics apps should be widely developed by the community to speed up materials discovery processes.

npj Computational Materials (2022)8:65 ; https://doi.org/10.1038/s41524-022-00750-6

INTRODUCTION

Machine learning (ML) models and algorithms are increasingly applied in materials science for a wide variety of tasks ranging from materials characterization, property prediction, and to structure/composition generation design as reviewed in[1–11]. These data-driven algorithms have dramatically sped up the exploration in the vast chemical design space and have helped to discover many novel functional materials[12]. However, compared to the mature bioinformatics field with thousands of web servers (>9000)[13,14], the ecosystem of materials informatics is still in the embryo stage with <100 web servers, most of them being data infrastructures[15]. This can also be seen in our survey in Table 1 which focuses on inorganic crystal materials. We also find that the ecosystem of chemical informatics web apps is also in the primitive stage as reviewed in[16]. In contrast, the bioinformatics field even has a search engine named bio.tools which indexes and tracks biological scientific web servers throughout their lifetime.

Here we argue that despite the increased sharing of data, programs or source code in the materials informatics community, the missing web apps for these tools have significantly impeded the progress of our field as most experimental teams do not have the expertise to implement, train and deploy these tools locally and many of the proposed materials informatics algorithms are under-used. Furthermore, compared to bioinformatics, materials informatics web tools are much fewer in terms of quantity, diversity, and quality. Developing and providing web servers can make complex algorithms accessible to a broad research and user community. In addition to providing user-friendly services to materials researchers, a recent study has found that there exists a positive association between the number of citations and the probability of a web server being reachable[14].

Currently, the most widely used web services in materials include Materials Project(MP)[17], Aflow-lib[18], and OQMD[19], which are all mainly used as data sources. Even though these major databases come with several related analysis tools, there are many missing web apps that are strongly needed in exploratory materials discovery research. This process can be generally divided into four major stages each needing specific convenient web apps: characterization, property prediction, synthesis, theory discovery, and materials design[20].

Starting from the composition exploration, one would need tools and models that can check the charge neutrality and electronegativity balance and estimate its formation energy. Composition-based prediction of crystal symmetry or lattice constants or even crystal structures is also highly desirable. When structures can be predicted or obtained via element substitution, tools such as structural relaxation, formation energy calculation, e-above-hull energy calculation, Pauling rule check, phonon calculation, and synthesizability are all useful to evaluate the feasibility of the candidate materials. The second major category of tools needed is property prediction web apps as provided by several existing servers[18,21]. However, many of these property prediction web apps do not support screening multiple inputs, which limits their usage in high-throughput screening for new materials. Nowadays, the modern deep generative materials design models can easily generate millions of candidate compositions[22] and structures[23]. Also, many of these tools do not support a convenient download of the prediction results. In addition, it is desirable that databases of hypothetical new materials can be made available for users to find novel functional materials.

In this paper, we first survey current state-of-the-art (SOTA) web services in the inorganic materials community and identify the requirements of a sufficient materials web app and the limitations of current web apps. We then introduce MaterialsAtlas.org, our materials informatics web app platform for supporting the whole life cycle of materials discovery. It includes multiple candidate materials composition and structure validations/checks, materials

[1]Department of Computer Science and Engineering, University of South Carolina, 550 Assembly Street, Columbia, SC 29201, USA. [2]Department of Physics, University of Colombo, PO Box 1490, Colombo 00300, Sri Lanka. ✉email: jianjunh@cse.sc.edu

Table 1. Survey of current web apps for materials discovery.

App name	URL	Institute	App functions	Comment
MaterialsAtlas	www.materialsatlas.org	UofSC	Composition/structure validation, property prediction, screening of materials, ML, composition enumeration, and more	This work. Easy to use.
Materials project[17]	materialsproject.org	Lawrence Berkeley National Lab	Crystal toolkit, structure predictor, phase diagram, Pourbaix Diagram, reaction calculator, interface reaction, nanoporous materials analysis, and synthesis description search	Major public repository. Good web apps.
Aflowlib[18]	aflowlib.org	Duke	Elastic, thermal, prototype, chull, aflow-ML for superconductor Tc, free energy and entropy, metal/insulator classification, band gap energy, bulk/shear moduli, Debye temperature, and heat capacities	Outdated descriptor methods
OQMD[19]	oqmd.org/analysis	Northwestern Univ.	Phase diagram, structure visualizer, and ground state analysis	Limited analysis web apps
JARVIS[21]	jarvis.nist.gov	NIST	Web ML tools for diverse property predictions (regression/classifications)	Account needs approval. CFID descriptors.
Crystal.AI[31]	crystals.ai	UCSD	Prediction models of formation energy, bandgap, elastic constants, perovskite/garnet stability, and coordination from X-ray absorption spectroscopy	Characterization and property prediction
Matgenie[38]	matgenie.materialsvirtuallab.org	UCSD	Materials analysis web app. Structure file format conversion;symmetry analysis; structure similarity comparison; XRD calculation; and surface generation	Utility tools
Materials Cloud[39]	materialscloud.org/work/tools	EPFL	QE input generator, chemical shift, molecular polarizability, phonon visualizer, synthesis condiction finder, predicting oxidation states, atomic environment finder, electron transport, and simulation in cloud (AiiDA)	Mainly utility tools
Bilbao crystallographic server[111]	www.cryst.ehu.es	Univ. of Basque Country	Show Wyckoff positions, symmetry, and structure utility	Utility tools
Thermoelectric[33]	thermoelectrics.citrination.com	Citrine	Predict thermoelectric materials properties	Commercial solution
NIMS[34]	mits.nims.go.jp/en/	Japan Nat. inst. of Mat. Sci.	Various databases and Composite Design & Property Prediction System	Rich databases & Data Conversion Tools
SUNCAT[41]	catalysis-hub.org	Stanford Univ.	Database and tools for interface science and catalysis design	Diagrams, ML models, and diverse tools
Polymer design[40]	reccr.chem.rpi.edu/polymerdesign	RPI	ML for polymer design	Materials design tool
Matlearn[36]	matlearn.org	Univ.of Houston	Predict Formation energy and create composition diagrams using ML to guide synthetic chemistry	Inorganic materials design tool
USPEX[42]	uspex-team.org/en	Skoltech	Crystal structure prediction	Binary program
CALYPSO[43]	calypso.cn/cdg	Jilin Univ. China	Crystal structure prediction	Binary program
JAMIP[112]	www.jamip-code.com/	Jilin Univ. China	Platform for feature engineering, data preprocessing, ML model building, property calculation, hpc computing management	Not web server. Tool to run DFT jobs ML

property prediction modules, hypothetical materials databases, and utility tools. Our web apps are developed with high-throughput materials discovery processes in mind with a user-friendly web interface and an easy download of results.

RESULTS

Survey of existing web apps for materials discovery

While there are many known AI or ML studies applied to the materials discovery process[10,24], many of them do not offer or share their code, programs, web apps, or even datasets, which significantly lower their potentials in materials research. Compared to thousands of bioinformatics web apps, the number of materials informatics web apps is much fewer and are developed in an ad hoc way without considering the high-throughput screening requirement from the materials discovery process. Table 1 shows a list of web apps and tools that support the materials discovery process.

Materials characterization is a key step in experimental analysis which is especially true with the progress of high-throughput materials characterization that generates huge amounts of data. There are an increasing number of algorithmic studies on phase mapping of X-ray diffraction data[25,26], symmetry determination in electron diffraction[27], predicting crystallographic dimensionality and space group from a limited number of thin-film XRD patterns[28], predicting accurate scale factor, lattice parameter and crystallite size maps for all phases[29], and tuning of parameters

in the Rietveld method[30]. However, most of these studies provide user-friendly web services. In our survey, only USCD team provides a web tool for coordination environment prediction from X-ray absorption spectroscopy[31].

The second major category of web tools is for materials property prediction. This includes aflow-ML[18], Javis-ML[21], Crystal. AI[32], thermoelectric predictor[33], NIMS tools[34], SUNCAT catalysis property predictor[35], and matlearn[36]. These web apps cover a variety of material's properties. For example, JARVIS-ML from NIST can predict formation energies, exfoliation energies, bandgaps, magnetic moments, refractive index, dielectric, thermoelectric, and maximum piezoelectric and infrared modes. However, many of these web apps are developed in an ad hoc way; they usually only accept one composition or structure at a time and cannot be used for screening. They usually do not come up with a performance measure to indicate the prediction confidence. More importantly, many of the algorithms or descriptors are outdated. For example, a recent benchmark study[37] showed that the best algorithms for formation energy and bandgap prediction are based on Graph Neural Networks (GNN), which are all much better than other structural descriptor-based methods as used in[18] and[21].

The third category of web apps is diverse utility tools for structure and composition analysis including crystal toolkit, phase diagram, and others from Materials projects[17], prototype finder from aflowlib[18], phase diagram tool from OQMD[19], analysis tools from JARVIS[21], Matgenie from USCD[38], phonon visualizer from MaterialsCloud[39], and crystal symmetry tool from Bilbao crystallographic server.

The fourth category of web tools is the materials design tools including polymer designer[40], Matlearn composition explorer[36], SUNCAT catalysis designer[41], and heterostructure designer in JARVIS[21].

There are several offline tools that are very useful for materials discovery including the crystal structure prediction softwares such as USPEX[42] and CALYPSO[43]. There are also platform tools such as JAMIP which includes property ML models and first-principle calculation job managements.

MaterialsAtlas: platform of materials discovery tools

The MaterialsAtlas platform includes four types of web apps for supporting exploratory materials discovery including: composition and structure check and validation, materials property prediction, screening of hypothetical materials, and utility tools.

TOOLS FOR COMPOSITION AND STRUCTURE VALIDATION

Chemical validity check

Given a predicted or generated material composition or structure, there are several steps to verify their physical feasibility. The first quick check of the chemical validity is the charge neutrality and electronegativity balance check (Fig. 1). These two check algorithms are based on the SMACT package[44] with improvements to speed up the enumeration and search process. For both checks, only composition information is needed. Another chemical validation check is the Pauling rules check. Here we only check the input structure against the first three Pauling rules[45].

Formation energy and e-above-hull energy check

Another structure validation step is to check the thermodynamical stability in terms of formation energy calculation. This step is usually done by DFT relaxation and then the calculation of their total energy and formation energy. However, this computation is expensive for a large amount of structures. Here, we can first

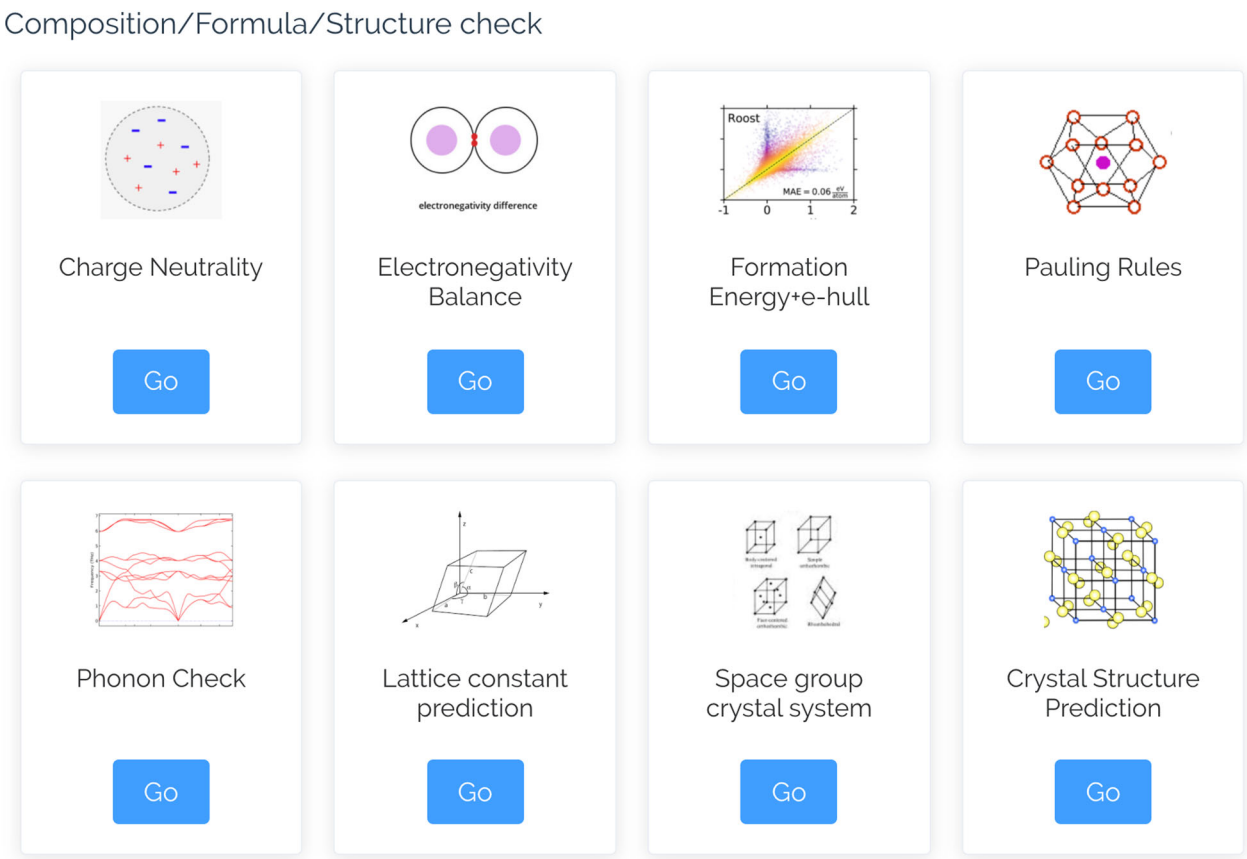

Fig. 1 Tools for composition and structure validation and check.

optimize input materials using Bayesian optimization with symmetry relaxation as introduced by Zuo et al.[46]. Here, we implemented two ML models for formation energy prediction, one is based on the Roost algorithm[47] with only the composition as input. This model has demonstrated exceptionally good performance for compound stability prediction among composition-only ML models[48]. The other structure-based energy prediction model is based on our deep global attention graph neural networks (DeeperGATGNN)[49] due to its exceptional performance based on our systematic benchmark studies. The e-above-hull energy prediction module has been implemented based on Pymatgen APIs: given an input materials composition and its total energy, it will report the e-above-hull energy.

Prediction of crystal symmetry (space group and crystal systems) and lattice parameters

Given a materials composition, predicting its structure is very valuable as its many macro-properties such as ion conductivity, thermal conductivity, band gap, and formation energy can be calculated using first-principle calculations. However, currently crystal structure prediction is an unsolved problem. In this case, predicting the crystal symmetry such as crystal systems or space groups can be very useful to estimate some of its properties. Here we implement neural network models for space group and crystal system prediction[50] which have achieved SOTA performance. Another important structure information of crystals is the unit cell parameters, whose precise estimation can greatly help the crystal structure prediction step. Here we implemented a deep neural network model for lattice parameter estimation, which has demonstrated exceptionally good performance for cubic systems and reasonably good results for other crystal systems[51].

Template-based crystal structure prediction

We have developed and implemented a template-based crystal structure prediction algorithm TCSP for fast structure determination[52]. By exploiting the vast known crystal structures, our algorithm has demonstrated good performance in CSP as benchmarked on the Materials Project dataset. The only input to this app is a material formula with an optional space group parameter, it will then generate multiple hypothetical crystal structures along with the template information used.

MATERIALS PROPERTY PREDICTION WITH COMPOSITION OR STRUCTURES

Depending on the types of features used to train the algorithm, we can categorize the ML properties predictive models as either composition-based or structure-based. Composition-based prediction algorithms have been demonstrated to be reliable, accurate, and even preferred at times[53]. The composition-based category includes models that primarily use chemical composition-induced descriptors such as elemental representation or chemical composition features[54,55]. Algorithms used in these composition-based ML models range from basic ML techniques such as decision trees[56] to more complex deep learning algorithms such as Convolutional Neural Networks[57] or Graph Neural Networks[47].

Composition-based ML models for property prediction come with both advantages and disadvantages. Because these models only use chemical composition descriptors as inputs, their predictive performance heavily relies on the quality of these features and the dataset. Therefore, the application of these models requires careful curative steps[53]. As the composition ML models omit the structural information of the materials, these models generally offer results with inferior predictive performance compared to structure-based ML models, especially when the size of the dataset is sufficiently large[37,58]. However, thanks to this omission of structural information, composition-based models are more computationally efficient than structure-based ones and can be used to screen much larger chemical space as material compositions are much easier to acquire than crystal structure data[47]. This omission can be very beneficial in some scenarios since structural-feature extraction is generally very complex and need to be symmetrically invariant[53]. With just composition descriptors, composition-based ML models can adapt any simple ML algorithms such as decision trees and support vector machines and still obtain accurate results[53]. Composition-based models can also adapt more robust ML algorithms from Deep Learning as shown in several deep learning models for property prediction including ElemNet (17 fully-connected layers)[58], Roost (GNN)[47], and Periodic-table based Convolutional Neural Network[59]. We note that composition-based predictors have one inherent limitation due to the polymorphic structures that may correspond to a given composition, which may bring bias to these models.

Another category of ML models for materials property prediction is structure-based ML models. As almost all materials properties are heavily dependent on their structures, the structure-based ML models for materials property prediction usually achieve greater accuracy than composition-based ML models[60,61]. Structure-based models use structure-based descriptors or features learned from raw structure information[60,62,63]. Structure Graph, Voxel Grids[64], Coulomb Matrix[65], and Voronoi Tessellation[12] are some of the most popular techniques to represent materials based on knowledge of their structure. Although models of this category accomplish better prediction results, they can only predict properties of materials whose structures are already known from repositories like Inorganic Crystal Structure Database (≈165,000 materials)[66] or Materials Project Database (≈125,000 materials)[17] (whereas the cardinality of chemical materials is infinite) and hypothetical materials generated using generative models[22,67].

Recent studies have shown that when structural descriptors are learned by deep neural network models, they can predict materials properties with much better accuracy than methods that use descriptors based on physicochemical information[37,68]. For doing this, GNN models have been intensively used as they have shown great success in this task[60,63,69]. GNN models have been found to achieve SOTA performance for various materials property prediction tasks. CGCNN[60], MEGNet[63], GATGNN[68], SchNet[69], and MPNN[70] are some of the well-known GNN models for materials property prediction that use graph representation learning. One of the problems of these existing GNN models is that they cannot go deep, i.e., their performance decreases with increasing number of graph convolution layers as the representation of all the node vectors becomes indistinguishable. This problem is known as the over-smoothing problem[71–74], and almost all the GNN models are victims of it. But recently, we designed a deeper and much improved version of the GATGNN model (DeeperGATGNN[49]) using Differentiable Group Normalization (DGN)[75] and skip-connections[76,77] which allows our DeeperGATGNN to use a large number of graph convolution layers to predict materials property with better accuracy than all the above mentioned GNN models for the five datasets used in a recent large-scale benchmark study[37] and the Band Gap dataset from Materials Project Database. In our current system, the structure-based formation energy predictor is based on CGCNN and the structure-based predictors for band gap, elastic moduli, hardness, thermal conductivity are based on our DeeperGATGNN trained with samples from Materials Project. The details of the datasets used to train our DeeperGATGNN models are presented in Table 2.

Table 2. Datasets used for training structure-based DeeperGATGNN models.

Dataset	Source	# of elements	# of samples
Bulk materials band gap	MaterialsProject	87	36,837
Hardness	MaterialsProject	85	12,854
Bulk modulus	MaterialsProject	89	13,176
Shear modulus	MaterialsProject	89	13,176
Young's modulus	MaterialsProject	85	12,854
Thermal conductivity	MaterialsProject	38	2701
Poisson ratio	MaterialsProject	85	12,858

MATERIALS PROPERTY PREDICTION TOOLS

Predicting 2D materials from composition

We train a Random Forest classification model to predict whether a given composition forms a 2D or layered structure[78]. As for the training data, 6351 2D materials (positive samples) are collected from the 2DMatPedia dataset[79]; 15,959 negative samples are gathered from The Materials Project by removing 2D materials. After training, our model achieves a classification accuracy of 88.98%. For a given input formula, our model outputs a predicted label (True or False) with corresponding probability in the downloaded results file. Inputs of multiple formulas are also supported either as a CSV file or by typing them into the input box separated by a comma or space. Clicking the 'Check now' button will show the found 2D materials; clicking the 'Download results' link, the detailed results will be downloaded.

Predicting noncentrosymmetric materials from composition

A Random Forest classification model is trained to predict whether a material is noncentrosymmetric[80]. For training this model, a total of 82,506 samples are collected from the Materials Project by removing those compositions belonging to multiple space groups with conflicting centrosymmetric tendencies; here, 60,687 of them are positive samples and 21,919 are negative samples. The predicted accuracy reaches 84.8%. The input format and output form are the same as the above method.

Predicting band gap from composition or structure

The band gap prediction models are trained with the dataset downloaded from the Materials Project. There are a total of 36,837 samples downloaded. The composition ML model is based on the CrabNet[81], which uses a transformer self-attention mechanism[82] in the compositionally restricted attention-based network for materials property prediction. Evaluations of over 28 datasets have shown good performance compared to other models. The structure-based band gap predictor is based on the dataset downloaded from the Materials Project and trained using the DeeperGATGNN graph attention network model[49]. For a given input formula, this model outputs the predicted band gap values.

Predicting elastic moduli from composition or structure

We trained two types of prediction models for elastic moduli prediction: composition-based prediction models and structure-based ones. The former type are Roost neural network models[47] trained with only materials compositions. Our structure-based elastic Moduli prediction models are based on our recent work of DeeperGATGNN algorithm[49], which is a global attention-based scalable deep graph neural network model with the state-of-the-art performance for structure-based materials property prediction. Both types of models are trained using the known materials with elastic information in the MaterialsProject database. For each category, we train four models to predict bulk modulus, shear modulus, Young's modulus, and poisson ratio based on the composition or structure information.

Predicting hardness from composition or structure

The most recent study uses deep learning for hardness prediction which has shown good performance[83]. Another study[84] uses 1062 experimentally measured load-dependent Vickers hardness data extracted from the literature to train the XGBoost ML algorithm using composition-only descriptors with boosting with excellent accuracy ($R2 = 0.97$). In a related study[85], XGBoost has been applied to build a temperature-dependent Vickers hardness prediction model with $R2 = 0.91$ performance using only 593 labeled samples. Here we trained a Roost ML model for composition-based hardness prediction and trained a graph neural network model for structure-based hardness prediction using our DeeperGATGNN algorithm[49].

Predicting thermal conductivity from composition or structure

The most recent study on thermal conductivity prediction is from[86] in which GNNs (CGCNN) and random forest approaches are combined to build the prediction model. The prediction model is trained with 2668 ordered and stoichiometric inorganic structures from the ICSD. Here we build a graph neural network model Roost[47] model for a composition-based prediction model and a CGCNN graph neural network model[60] for structure-based predictions. The dataset is downloaded from[87], which contains thermal conductivity values for 2701 crystal structures contained in the ICSD database. Due to the limited data size, the prediction performance is only for experimental purposes.

Predicting superconductor transition temperature from composition

We also train a random forest model and a CrabNet model to predict the superconductor transition temperature. The dataset is collected from the superCon database[88].

In our current implementation of materials predictors, all models only generate a single-point prediction without uncertainty estimation as shown in almost all materials prediction algorithms[37]. However, in practice, it is desirable to obtain robust predictions with accurate uncertainty estimation[89], which can be achieved via methods such as ensemble[90], Bayesian[91], or evidential deep learning regression models[92]. While such methods have been rarely used in materials property predictions, we expect their wider adoption in the future and will be added to our models in future upgrades.

GENERATIVE DESIGN AND SCREENING FOR MATERIALS DISCOVERY

Deep generative design of materials compositions/formulas

Generative models, such as variational autoencoder (VAE)[93] and Wasserstein generative adversarial network(WGAN)[94], play an important part in computer vision, audio processing, natural language processing, and molecular science. However, limited works have focused on using generative models to generate virtual inorganic materials (e.g., compositions and crystal structures). There are mainly two directions that researchers use generative models in material science. The first is we use generative models to generate compositions[22,95]. Dan et al. propose[22] to use WGAN models to generate hypothetical materials compositions that are trained using the ICSD dataset. Their models not only can rediscover most compositions from existing materials databases but also generate many novel

Table 3. Summary of materials property prediction tools.

Property prediction	Model	Training dataset	Performance	Output
2D materials	Random Forest	2DMatPedia Material Project	88.98% (Acc)	Label Probability
Noncentro symmetry	Random Forest	Material Project	84.8% (Acc)	Label Probability
Band gap	Roost DeeperGATGNN	Material Project	0.465 (MAE)	Band gap (eV)
Elastic moduli	CrabNet DeeperGATGNN	12858 samples from MP	15.7 (MAE, Bulk) 18 (MAE, Shear) 76.8 (MAE, Young's) 8.7 (MAE, Poisson)	Bulk mod (GPa) Shear mod (GPa) Young's mod (psi) Poisson ratio
Hardness	Roost DeeperGATGNN	12854 samples from MP	2.42 (MAE)	Hardness (N/mm^2)
Thermal conductivity	CrabNet DeeperGATGNN	2688 samples from ICSD	5.03 (MAE)	Thermal conductivity (W/(mK))
Ionic conductivity	under development	N/A	N/A	Ionic conductivity
Superconductivity	Random Forest CrabNet	25378 samples from supercon	4.76 (MAE)	Transition temperature (K)

compositions that are chemically valid. Here we provide the screening tools for such hypothetical materials.

Deep generative design of cubic crystal materials

Compared to generating virtual materials compositions, generating virtual crystal structures is more helpful for practitioners to find novel materials since many materials' properties can only be calculated with structural information. Several works[96–98] based on VAE and[23,67,99,100] based on GAN have been proposed to generate material structures. CubicGAN proposed by Zhao et al.[23] is the first method that can achieve the large-scale generative design of novel cubic materials. The authors not only can rediscover most of the cubic materials in The Materials Project and ICSD but also can discover new prototypes with stable materials. In their work[23], they found 31 new prototypes for space groups of $Fm\bar{3}m$, $F\bar{4}3m$, and $Pm\bar{3}m$, of which 4 prototypes contain stable materials. A total of 506 cubic materials have been verified stable by phonon dispersion calculation. Here in our web app platform, we provide the search function for those materials (Table 3 and Fig. 2).

Tools for hypothetical materials screening

One of the major goals for the materials informatics community is to expand the existing materials repositories in terms of materials compositions, structures, and properties, which can help accelerate materials with novel functions. Using our recently developed materials composition generative models (MATGAN)[22], we have generated a large number of hypothetical material compositions which are deposited to the database for screening, hence the Hypothetical composition database (Fig. 3). For convenience, we also selected those lithium compound candidates and built the Hypothetical lithium materials database. Using our crystal structure generator, the CubicGAN[23], we have created a cubic materials database for screening. Hypothetical materials compositions can also be combined with element substitution based structure prediction to generate new materials database. Finally, we trained a 2D materials classifier which is used to screen the whole hypothetical compositions generated by MATGAN, which are then deposited as the hypothetical 2d materials database.

UTILITY TOOLS

Several utility tools (Fig. 4) to assist the materials discovery process have been developed and deployed on our platform. This includes chemical composition enumeration tool, feature generation and click-and-run machine learning models for users' datasets, composition and structure search, and supercell generator and structure file format converter.

Composition enumerator

Given several elements, what are the possible chemically valid formulas that can be synthesized and stable? Based on the SMACT materials informatics package[44,101], we develop this composition enumerator to generate target materials compositions given a set of elements or an existing formula with one or more dopant elements. Due to the oxidation preferences, the number of possibilities is limited and this tool can help the investigator to narrow down the search space. A case study on how to use this module for discovering new materials is reported in our work[52]. With the hypothetical compositions, one can then apply crystal structure prediction to get their crystal structure and then predict their properties using composition-based or structure-based predictors.

Feature generation

The very first step for developing materials property prediction models is to generate and select a set of good descriptors. Here we implemented a pipeline that allows users to choose feature combinations from diverse feature types such as composition features, structure features, electronic features, etc. This will greatly simplify the steps for materials scientists without a strong materials informatics background to develop ML models.

Composition-based ML models for user-specified property prediction

We have built an ML pipeline that allows the user to specify the datasets and target property values and the algorithm, the web tool, will then build composition-based ML models and report the prediction performance. The test input will be a group of materials formulas.

Structure-based ML models for user-specified property prediction

We have built a pipeline that allows the user to train a structure-based ML model for their custom-property prediction at http://materialsatlas.org/mlstructure which can greatly help the materials

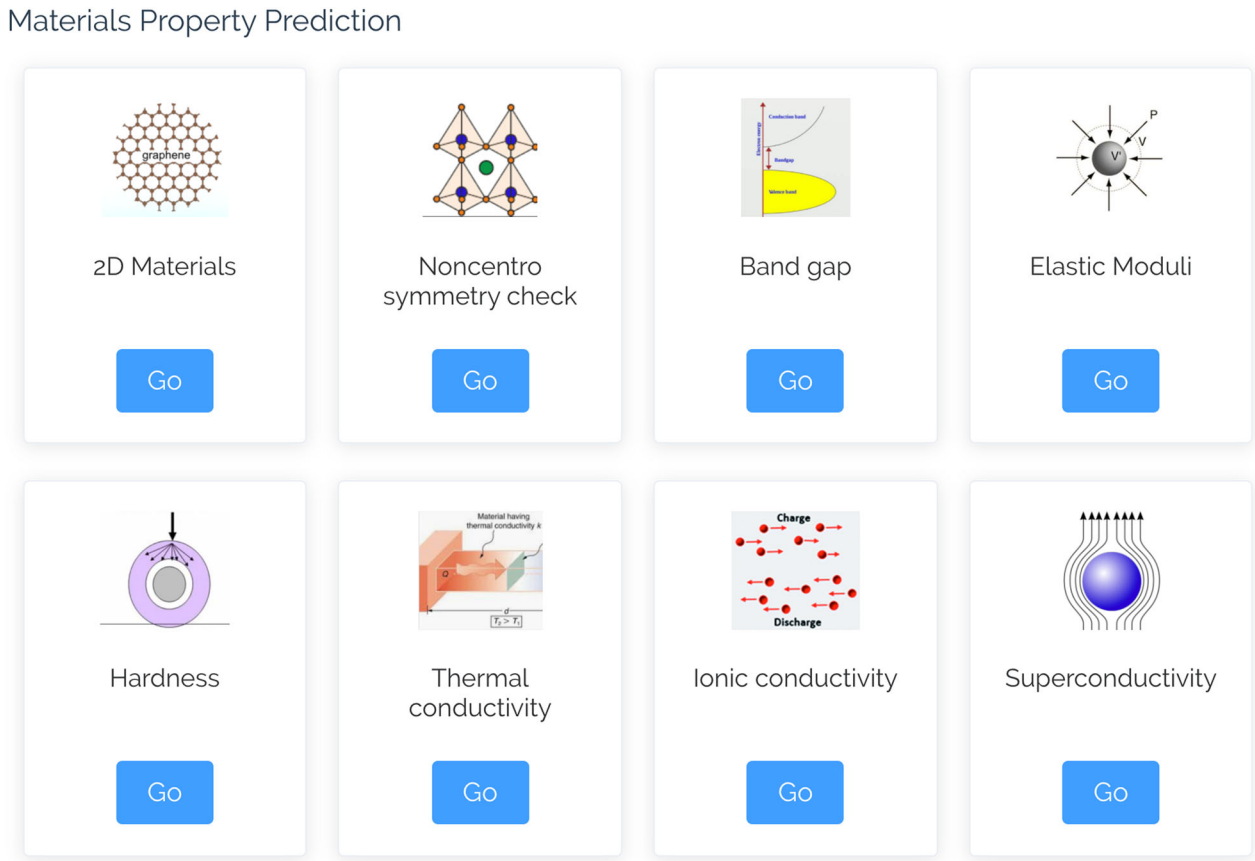

Fig. 2 Materials property prediction tools.

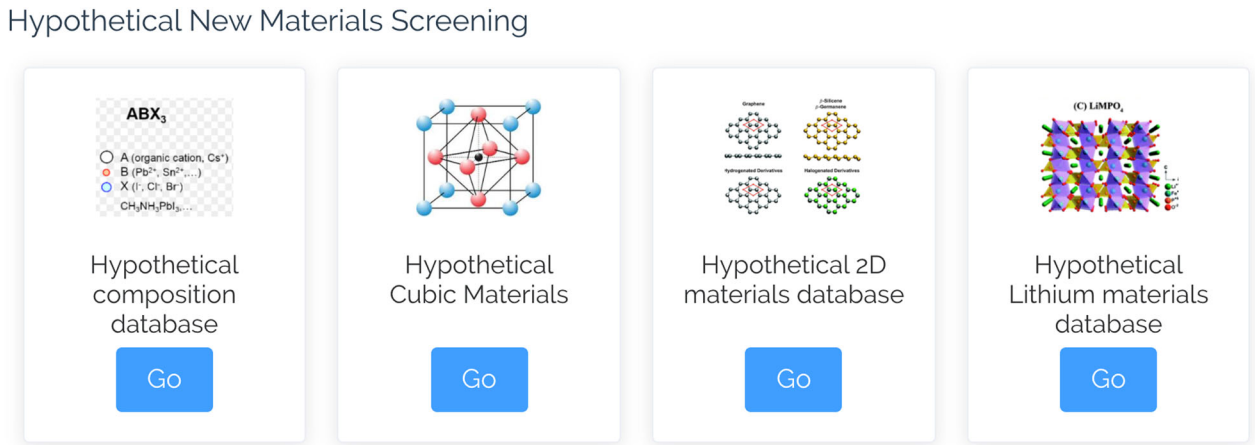

Fig. 3 Screening hypothetical materials generated by machine learning or deep learning models.

scientists to try different representations and ML algorithms to get the best performance.

Finding similar compositions and structures

In many of the tinkering and exploratory studies of the materials design space, it is very helpful to find similar materials and explore their property changes. We use the Earth Mover's Distance[102] to search top N most similar formulas from different databases. For structure similarity, we use the computed XRD features[103] to search similar structures. This search function will help with that.

For the convenience of the community, we have included other utility tools such as structure file conversion and supercell generation apps.

DISCUSSION

In addition to candidate materials composition and structure validation, materials property prediction, and screening of materials, several additional tools and services are highly desirable for exploratory materials discovery and will be added to our platform to lower the barrier for materials scientists in data-driven exploratory materials discovery.

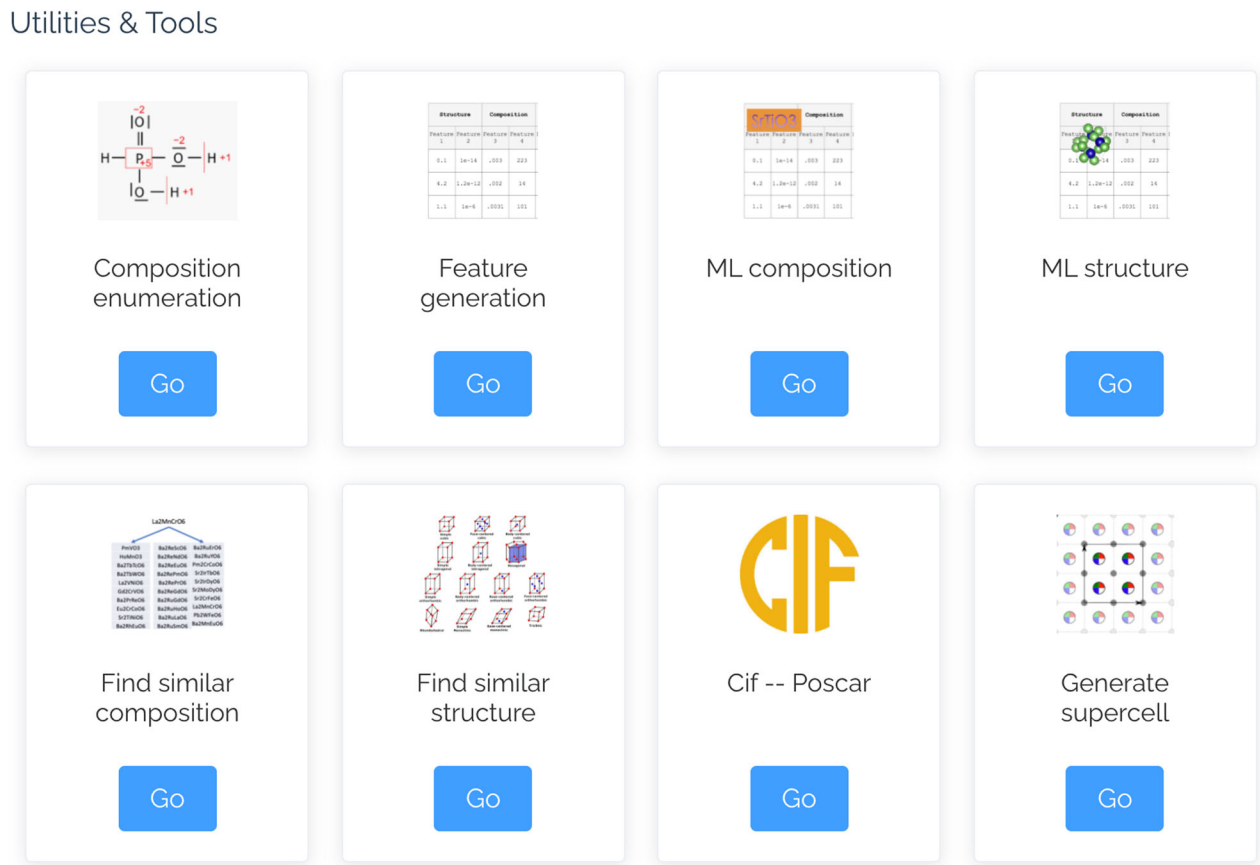

Fig. 4 Utility tool web apps.

Phonon prediction, synthesizablity prediction, additional crystal structure prediction algorithms

One important validation step for newly proposed hypothetical materials is to calculate its mechanical dynamic stability. This can be done by calculating the phonon dispersion spectrum and checking whether the material is dynamically stable at 0K temperature when there are no imaginary frequencies. The phonon dispersion relations for hypothetical materials are important to study the k-space dependence of frequencies of normal modes. However, first principle phonon dispersion calculation is computationally expensive. Based on recent work on phonon density of states prediction[104] and phonon vibration frequency prediction[105], we are developing a graph neural network model for phonon dispersion prediction aiming to use for checking the dynamic stability of structures. Another module under development is the material synthesizability prediction model, which has been shown to be able to achieve good performance for inorganic materials using semi-supervised ML models[106,107]. In addition, we find that crystal structure prediction plays an important role in exploratory materials discovery and current DFT-based global optimization-based algorithms are applicable only to small systems due to the inherent challenges in crystal structure prediction. In addition to the template-based crystal structure prediction service[52], we are planning to develop deep learning-based crystal structure algorithms by exploiting the databases of known crystal structures.

Predicting ion conductivity from composition or structure

Due to the extremely limited datasets, prediction of ion conductivity has been very challenging with moderate success by using a set of hand-crafted structural descriptors[108,109]. This module is under development and will be added in future to our platform.

Extensible servers and API services

To expand the coverage of functionalities, our MaterialsAtlas web server is open to include third-party web apps for materials research. We welcome any investigator to collaborate with us and deploy their applications on our platform. Only executable code or python code in a Linux environment is needed. Another useful feature is the REST API services so that other web services can call our APIs to do some query or calculation, which has shown great success in Materials Project's Pymatgen APIs.

Visualization and interactive exploration of design space

Interactive exploration in the materials design space has big potential to help researchers. We will add modules that support the visualization of materials property distribution among materials in the structural or composition space as shown in Fig. 5. In this figure, we map the structures into a 2D space using t-sne[110] and XRD representation of the structures. We then annotate those red dots as the samples with annotated thermal conductivity with the dot size representing the magnitude of the thermal conductivity. Such interactive maps will greatly facilitate the search for high performance materials.

Despite the rapid progress of ML for materials research, a lot of studies have only led to papers without sharing their software while some of them shared their source code but without creating a user-friendly web service or web apps for them. Based on the experience of the bioinformatics field, it is critical for materials informatics researchers to develop and share easy-to-use web apps that wrap their developed algorithms for maximum adoption

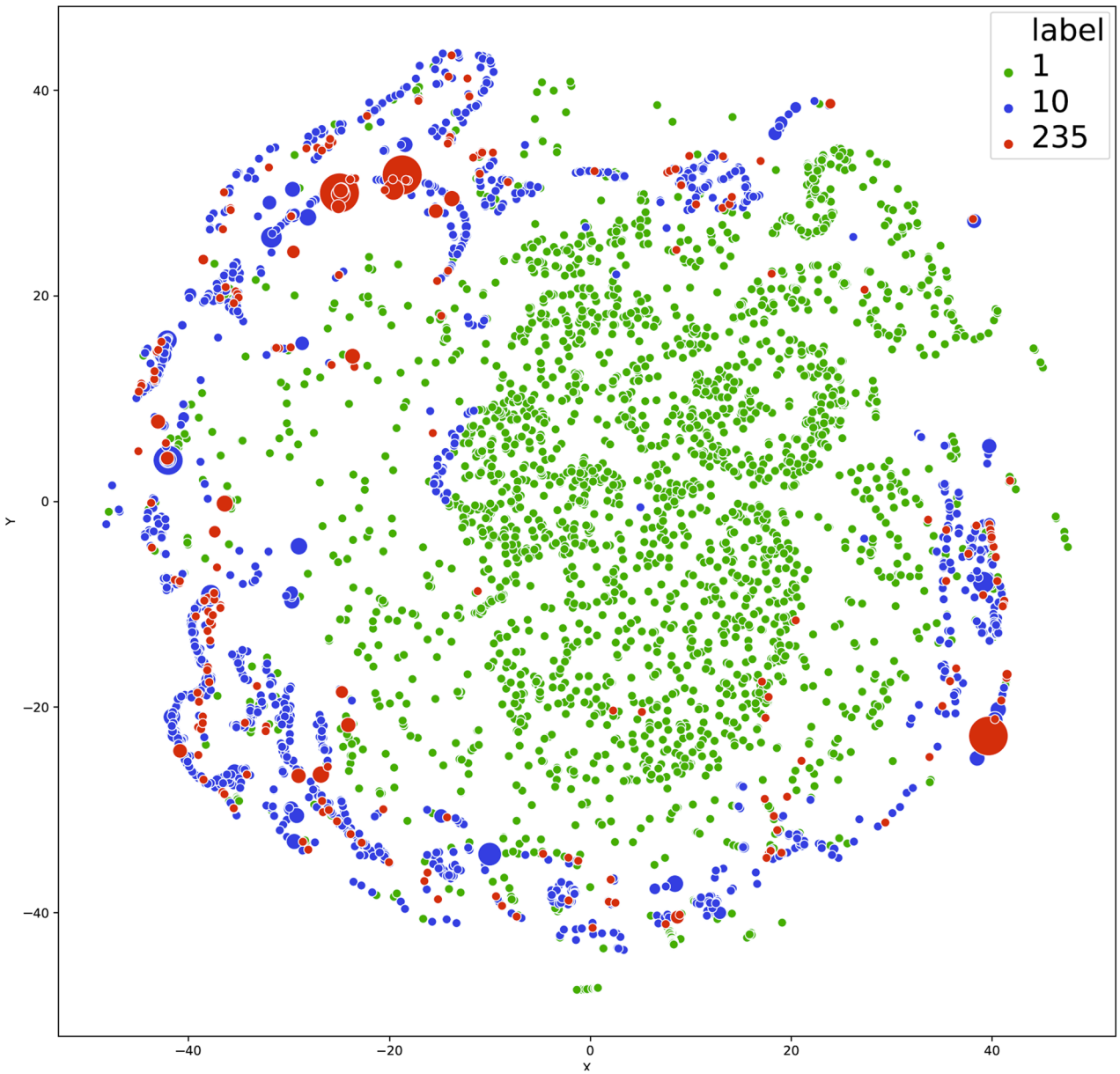

Fig. 5 Interactive exploration of thermal conductivity. We map the materials by projecting their XRD spectra to 2D space. Each red point represents a material with known thermal conductivity with the size of dots representing their thermal conductivity values. The green and blue dots represent the materials with unknown thermal conductivity.

and usage of such data-driven tools in real-life materials discovery and analysis. We have surveyed the status quo of materials informatics web apps and find that they drastically lag behind the bioinformatics community. Here we report our MaterialsAtlas.org web platform that implements and integrates a variety of user-friendly tools for aiding the materials design space exploration, generation of candidates, and validating the candidates. These tools and those planned together will greatly decrease the barrier for materials researchers without deep computing or ML backgrounds to effectively exploit such tools.

METHODS

System architecture and web app

MaterialsAtlas uses Django's built-in SQLite3 database for storing hypothetical materials found by our generative materials design models[22,23,78]. Moreover, a RESTful API framework is used to send data from the Django back-end to the Vue.js front-end and vice versa. For example, a user will input either a chemical formula or element in one of the apps which will then be interpreted through the Django REST framework. The data is then queued as a job using Redis and subsequently, a Python worker is used to input the data into the corresponding app function. Once the worker and job have finished, the result is returned to the front-end to be viewed by the user. MaterialsAtlas also uses Ajax for some of the applications to communicate to our API. On a separate note, Nginx is used as the web application's HTTP server. Additionally, MaterialsAtlas utilizes Nginx to proxy to the back-end and front-end server. For easier deployment, Docker is used to assemble each web-service as containers allowing the web application to work as a whole.

Backend models

Python is used as MaterialsAtlas' primary back-end language to compute each application result and write to the Django database.

Job submission

When integrating a web application with any ML model, latency is a large concern. Using Redis' job queue and fast in-memory data storage functionality allows a web application of this nature to run smoothly.

DATA AVAILABILITY

All data used to train the models are from the pubic database MaterialsProject or are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

Most of the codes of the prediction models are available as open source code as stated in the cited references. Other codes may be obtained from from the corresponding author upon reasonable request.

Received: 12 September 2021; Accepted: 10 March 2022;
Published online: 11 April 2022

REFERENCES

1. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
2. 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).
3. Gubernatis, J. & Lookman, T. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Phys. Rev. Mater.* **2**, 120301 (2018).
4. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
5. Wei, J. et al. Machine learning in materials science. *InfoMat* **1**, 338–358 (2019).
6. Morgan, D. & Jacobs, R. Opportunities and challenges for machine learning in materials science. *Annu. Rev. Mater. Res.* **50**, 71–103 (2020).
7. Wang, A. Y.-T. et al. Machine learning for materials scientists: an introductory guide toward best practices. *Chem. Mater.* **32**, 4954–4965 (2020).
8. Chen, A., Zhang, X. & Zhou, Z. Machine learning: accelerating materials development for energy storage and conversion. *InfoMat* **2**, 553–576 (2020).
9. Moosavi, S. M., Jablonka, K. M. & Smit, B. The role of machine learning in the understanding and design of materials. *J. Am. Chem. Soc.* **142**, 20273–20287 (2020).
10. Saal, J. E., Oliynyk, A. O. & Meredig, B. Machine learning in materials discovery: confirmed predictions and their underlying approaches. *Annu. Rev. Mater. Res.* **50**, 49–69 (2020).
11. Sparks, T. D., Kauwe, S. K., Parry, M. E., Tehrani, A. M. & Brgoch, J. Machine learning for structural materials. *Annu. Rev. Mater. Res.* **50**, 27–48 (2020).
12. Chen, C. et al. A critical review of machine learning of energy materials. *Adv. Energy Mater.* **10**, 1903242 (2020).
13. Fehlmann, T. et al. Aviator: a web service for monitoring the availability of web services. *Nucleic Acids Res.* **49**, W46–W51 (2021).
14. Kern, F., Fehlmann, T. & Keller, A. On the lifetime of bioinformatics web services. *Nucleic Acids Res.* **48**, 12523–12533 (2020).
15. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
16. Medina-Franco, J. L., Sánchez-Cruz, N., López-López, E. & Díaz-Eufracio, B. I. Progress on open chemoinformatic tools for expanding and exploring the chemical space. *J. Comput. Aided Mol. Des.* **18**, 1–14 (2021).
17. Ceder, G. & Persson, K. The materials project: A materials genome approach. DOE Data Explorer, http://www.osti.gov/dataexplorer/biblio/1077798 (2010). [2016-08-28].
18. Curtarolo, S. et al. Aflowlib. org: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
19. Kirklin, S. et al. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *Npj Comput. Mater.* **1**, 1–15 (2015).
20. Li, J. et al. Ai applications through the whole life cycle of material discovery. *Matter* **3**, 393–432 (2020).
21. Choudhary, K. et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *Npj Comput. Mater.* **6**, 1–13 (2020).
22. Dan, Y. et al. Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials. *Npj Comput. Mater.* **6**, 1–7 (2020).
23. Zhao, Y. et al. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Adv. Sci.* **8**, 2100566 (2021).
24. Lu, Z. Computational discovery of energy materials in the era of big data and machine learning: a critical review. *Mater. Rep. Energy* **1**, MRE100047 (2021).
25. Stanev, V. et al. Unsupervised phase mapping of x-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. *Npj Comput. Mater.* **4**, 1–10 (2018).
26. Xiong, Z., He, Y., Hattrick-Simpers, J. R. & Hu, J. Automated phase segmentation for large-scale x-ray diffraction data using a graph-based phase segmentation (gphase) algorithm. *ACS Comb. Sci.* **19**, 137–144 (2017).
27. Kaufmann, K. et al. Crystal symmetry determination in electron diffraction using machine learning. *Science* **367**, 564–568 (2020).
28. Oviedo, F. et al. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *Npj Comput. Mater.* **5**, 1–9 (2019).
29. Dong, H. et al. A deep convolutional neural network for real-time full profile analysis of big powder diffraction data. *Npj Comput. Mater.* **7**, 1–9 (2021).
30. Ozaki, Y. et al. Automated crystal structure analysis based on blackbox optimisation. *Npj Comput. Mater.* **6**, 1–7 (2020).
31. Zheng, C., Chen, C., Chen, Y. & Ong, S. P. Random forest models for accurate identification of coordination environments from x-ray absorption near-edge structure. *Patterns* **1**, 100013 (2020).
32. Crystals.AI. crystals.ai. University of California. Accessed: 3-September-2021.
33. Gaultois, M. W. et al. Perspective: web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Materials* **4**, 053213 (2016).
34. Tanifuji, M., Matsuda, A. & Yoshikawa, H. *Materials Data Platform-a Fair System For Data-driven Materials Science*, p. 1021–1022 (IEEE, 2019).
35. SUNCAT. catalysis-hub.org. Accessed: 3-September-2021.
36. Peterson, G. & Brgoch, J. Materials discovery through machine learning formation energy. *J. Phys. Energy* **3**, 022002 (2021).
37. Fung, V., Zhang, J., Juarez, E. & Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *Npj Comput. Mater.* **7**, 1–8 (2021).
38. Matgenie materials analysis web app. http://matgenie.materialsvirtuallab.org/. Accessed: 3-September-2021.
39. Talirz, L. et al. Materials cloud, a platform for open computational science. *Sci. Data* **7**, 1–12 (2020).
40. Polymer design. reccr.chem.rpi.edu/polymerdesign. Accessed: 3-September-2021.
41. Winther, K. T. et al. Catalysis-hub. org, an open electronic structure database for surface reactions. *Sci. Data* **6**, 1–10 (2019).
42. Glass, C. W., Oganov, A. R. & Hansen, N. Uspex—evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175**, 713–720 (2006).
43. Wang, Y. et al. Materials discovery via calypso methodology. *J. Phys. Condens. Matter* **27**, 203203 (2015).
44. Davies, D. W. et al. Smact: semiconducting materials by analogy and chemical theory. *J. Open Source Softw.* **4**, 1361 (2019).
45. George, J. et al. The limited predictive power of the pauling rules. *Angew. Chem.* **132**, 7639–7645 (2020).
46. Zuo, Y. et al. Accelerating materials discovery with bayesian optimization and graph deep learning. *Mater. Today* **51**, 126–135 (2021).
47. Goodall, R. E. & Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* **11**, 1–9 (2020).
48. Bartel, C. J. et al. A critical examination of compound stability predictions from machine-learned formation energies. *Npj Comput. Mater.* **6**, 1–11 (2020).
49. Omee, S. S. et al. Scalable deeper graph neural networks for high-performance materials property prediction. *Preprint at* https://arxiv.org/abs/2109.12283 (2021).
50. Li, Y., Dong, R., Yang, W. & Hu, J. Composition based crystal materials symmetry prediction using machine learning with enhanced descriptors. *Comput. Mater. Sci.* **198**, 110686 (2021).
51. Li, Y., Yang, W., Dong, R. & Hu, J. Mlatticeabc: generic lattice constant prediction of crystal materials using machine learning. *ACS Omega* **6**, 11585–11594 (2021).
52. Wei, L. et al. Tcsp: a template based crystal structure prediction algorithm and web server for materials discovery. *Preprint at* https://arxiv.org/abs/2111.14049 (2021).
53. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**, 1–36 (2019).
54. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 144110 (2017).
55. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2**, 1–7 (2016).

56. Safavian, S. R. & Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **21**, 660–674 (1991).
57. O'Shea, K. & Nash, R. An introduction to convolutional neural networks. *Preprint at* https://arxiv.org/abs/1511.08458 (2015).
58. Jha, D. et al. Elemnet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 1–13 (2018).
59. Zheng, X., Zheng, P. & Zhang, R.-Z. Machine learning material properties from the periodic table using convolutional neural networks. *Chem. Sci.* **9**, 8426–8432 (2018).
60. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
61. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *Npj Comput. Mater.* **6**, 1–10 (2020).
62. Kajita, S., Ohba, N., Jinnouchi, R. & Asahi, R. A universal 3d voxel descriptor for solid-state material informatics with deep convolutional neural networks. *Sci. Rep.* **7**, 1–9 (2017).
63. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
64. Zhao, Y. et al. Predicting elastic properties of materials from electronic charge density using 3d deep convolutional neural networks. *J. Phys. Chem. C* **124**, 17262–17273 (2020).
65. Rupp, M., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
66. Rühl, S. The inorganic crystal structure database (icsd): a tool for materials sciences. *Materials Informatics: Methods, Tools and Applications* 41–54 (John Wiley & Sons, Inc., 2019).
67. Nouira, A., Sokolovska, N. & Crivello, J.-C. CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering* (2019).
68. Louis, S.-Y. et al. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* **22**, 18141–18148 (2020).
69. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
70. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. *Neural Message Passing For Quantum Chemistry*, 1263–1272 (PMLR, 2017).
71. Li, Q., Han, Z. & Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. in *32nd AAAI Conference on Artificial Intelligence* (AAAI, 2018).
72. Chen, D. et al. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. Vol. 34, 3438–3445 (AAAI, 2020).
73. Oono, K. & Suzuki, T. Graph Neural Networks Exponentially Lose Expressive Power For Node Classification https://openreview.net/forum?id=S1ldO2EFPr (2020).
74. Louis, S.-Y., Nasiri, A., Rolland, F. J., Mitro, C. & Hu, J. Node-select: a graph neural network based on a selective propagation technique. *Preprint at* https://arxiv.org/abs/2102.08588 (2021).
75. Zhou, K. et al. Towards deeper graph neural networks with differentiable group normalization. *Adv. Neural Inf. Process. Syst.* **33**, 4917–4928 (2020).
76. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning For Image Recognition*. 770-778 (IEEE, 2016).
77. Jha, D. et al. Enabling deeper learning on big data for materials informatics applications. *Sci. Rep.* **11**, 1–12 (2021).
78. Song, Y., Siriwardane, E. M. D., Zhao, Y. & Hu, J. Computational discovery of new 2d materials using deep learning generative models. *ACS Appl. Mater. Interfaces* **13**, 53303–53313 (2021).
79. Zhou, J. et al. 2dmatpedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci. Data* **6**, 1–10 (2019).
80. Song, Y. et al. Machine learning based prediction of noncentrosymmetric crystal materials. *Comput. Mater. Sci.* **183**, 109792 (2020).
81. Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Comput. Mater.* **7**, 1–10 (2021).
82. Vaswani, A. et al. Attention is all you need. *31st Conference on Neural Information Processing Systems* p. 5998–6008 (2017).
83. Mazhnik, E. & Oganov, A. R. Application of machine learning methods for predicting new superhard materials. *J. Appl. Phys.* **128**, 075102 (2020).
84. Zhang, Z., Mansouri Tehrani, A., Oliynyk, A. O., Day, B. & Brgoch, J. Finding the next superhard material through ensemble learning. *Adv. Mater.* **33**, 2005112 (2021).
85. Zhang, Z. & Brgoch, J. Determining temperature-dependent vickers hardness with machine learning. *J. Phys. Chem. Lett* **12**, 6760–6766 (2021).
86. Zhu, T. et al. Charting lattice thermal conductivity for inorganic crystals and discovering rare earth chalcogenides for thermoelectrics. *Energy Environ. Sci.* **14**, 3559–3566 (2021).
87. Gorai, P. et al. Te design lab: a virtual laboratory for thermoelectric material design. *Comput. Mater. Sci.* **112**, 368–376 (2016).
88. Lütkebohle, I. National Institute of Materials Science, Materials Information Station,SuperCon. http://supercon.nims.go.jp/index_en.html (2011). Accessed 19-July-2021.
89. Nigam, A. et al. Assigning confidence to molecular property prediction. *Expert Opin. Drug Discov.* 16, 1–15 (2021).
90. Busk, J. et al. Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks. *Mach. Learn. Sci. Technol.* **3**, 015012 (2021).
91. Olivier, A., Shields, M. D. & Graham-Brady, L. Bayesian neural networks for uncertainty quantification in data-driven materials modeling. *Comput. Methods Appl. Mech. Eng.* **386**, 114079 (2021).
92. Soleimany, A. P. et al. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).
93. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *Preprint at* https://arxiv.org/abs/1312.6114 (2013).
94. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223 (PMLR, 2017).
95. Sawada, Y., Morikawa, K. & Fujii, M. Study of deep generative models for inorganic chemical compositions. *Preprint at* https://arxiv.org/abs/1910.11499 (2019).
96. Noh, J. et al. Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
97. Court, C. J., Yildirim, B., Jain, A. & Cole, J. M. 3-d inorganic crystal structure generation and property prediction via representation learning. *J. Chem. Inf. Model.* **60**, 4518–4535 (2020).
98. Korolev, V., Mitrofanov, A., Eliseev, A. & Tkachenko, V. Machine-learning-assisted search for functional materials over extended chemical space. *Mater. Horiz.* **7**, 2710–2718 (2020).
99. Long, T. et al. CCDCGAN: Inverse design of crystal structures. Bulletin of the American Physical Society 66 (2020).
100. Kim, S., Noh, J., Gu, G. H., Aspuru-Guzik, A. & Jung, Y. Generative adversarial networks for crystal structure prediction. *ACS Cent. Sci.* **6**, 1412–1420 (2020).
101. Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chem* **1**, 617–627 (2016).
102. Hargreaves, C. J., Dyer, M. S., Gaultois, M. W., Kurlin, V. A. & Rosseinsky, M. J. The earth mover's distance as a metric for the space of inorganic compositions. *Chem. Mater.* **32**, 10610–10620 (2020).
103. de Gelder, R. *Quantifying The Similarity Of Crystal Structures*, 59 (Citeseer, 2006).
104. Chen, Z. et al. Direct prediction of phonon density of states with euclidean neural networks. *Adv. Sci.* **8**, 2004214 (2021).
105. Nguyen, N. et al. Predicting lattice phonon vibrational frequencies using deep graph neural networks. *Preprint at* https://arxiv.org/abs/arXiv:2111.05885 (2021).
106. Jang, J., Gu, G. H., Noh, J., Kim, J. & Jung, Y. Structure-based synthesizability prediction of crystals using partially supervised learning. *J. Am. Chem. Soc.* **142**, 18836–18843 (2020).
107. Gleaves, D., Siriwardane, E. M. D., Zhao, Y., Fu, N. & Hu, J. Semi-supervised teacher-student deep neural network for materials discovery. *Preprint at* https://arxiv.org/abs/2112.06142 (2021).
108. Sendek, A. D. et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306–320 (2017).
109. Sendek, A. D. et al. Machine learning-assisted discovery of solid li-ion conducting materials. *Chem. Mater.* **31**, 342–352 (2018).
110. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
111. Aroyo, M. I. et al. Bilbao crystallographic server: I. databases and crystallographic computing programs. *Z Kristallogr. Cryst. Mater.* **221**, 15–27 (2006).
112. Zhao, X.-G. et al. Jamip: an artificial-intelligence aided data-driven infrastructure for computational materials informatics. *Sci. Bull.* **66**, 1973–1985 (2021).

ACKNOWLEDGEMENTS

The research reported in this work was supported in part by National Science Foundation under the grant and 1940099, 1905775, and OIA-1655740. The views, perspectives, and content do not necessarily represent the official views of the NSF. We appreciate the help from Xerrak Agha, Daniel Varivoda, Sourin Dey for proofreading.

AUTHOR CONTRIBUTIONS

Conceptualization by J.H.; methodology by J.H., Y.S., S.Y.L., E.M.D.S., and Y.Z.; software by J.H., S.S.,Y.S., and S.S.O.; resources by J.H.; writing–original draft preparation by J.H., S.S., Y.S., S.S.O., S.Y.L., E.M.D.S., and Y.Z.; writing–review and editing by J.H.; visualization by J.H. and S.S.; supervision by J.H.; funding acquisition by J.H.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Jianjun Hu.

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.