Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Benchmark datasets incorporating diverse tasks, sample sizes, material systems, and data heterogeneity for materials informatics[☆]

Ashley N. Henderson, Steven K. Kauwe, Taylor D. Sparks[*]

Materials Science & Engineering Department, University of Utah, Utah 84112, USA

ARTICLE INFO

Article history:
Received 1 May 2021
Revised 10 June 2021
Accepted 11 June 2021
Available online 13 July 2021

Keywords:
Machine learning
Materials discovery
Benchmark data
Materials datasets
Materials informatics

ABSTRACT

Materials discovery via machine learning has become an increasingly popular method due to its ability to rapidly predict materials properties in a time-efficient and low-cost manner. However, one limitation in this field is the lack of benchmark datasets, particularly those that encompass the size, tasks, material systems, and data modalities present in the materials informatics literature. This makes it difficult to identify optimal machine learning model choices including algorithm, model architecture, data splitting, and data featurization for a given task. Here, we attempt to address this lack of benchmark datasets by assembling a unique repository of 50 different datasets for materials properties. The data contains both experimental and computational data, data suited for regression as well as classification, sizes ranging from 12 to 6354 samples, and materials systems spanning the diversity of materials research. Data were extracted from 16 publications. In addition to cleaning the data where necessary, each dataset was split into train, validation, and test splits. For datasets with more than 100 values, train-val-test splits were created, either with a 5-fold or 10-fold cross-validation method, depending on what each respective paper did in their studies.

[☆] Given his role as Section Editor of this journal, Taylor D. Sparks had no involvement in the peer-review of this article and has no access to information regarding its peer-review.

DOI of original article: 10.1016/j.ijantimicag.2020.105949

[*] Corresponding author.

E-mail address: sparks@eng.utah.edu (T.D. Sparks).

https://doi.org/10.1016/j.dib.2021.107262

Datasets with less than 100 values had train-test splits created using the Leave-One-Out cross-validation method. These benchmark data can serve as a basis for a more diverse benchmark dataset in the future to further improve their effectiveness in the comparison of machine learning models.

Specifications Table

Subject	Computational Materials Science
Specific subject area	Machine learning models for materials informatics
Type of data	Table
How data were acquired	Gathered data from past literature
Data format	Analyzed, Filtered
Parameters for data collection	Data were gathered from papers that had easily accessible data, had been published relatively recently (since 2013), and had used either a regression or classification machine learning model on any kind of material property. Any kind of material system, data type (experimental vs calculated), data size, and organic/inorganic materials were selected if said data fit the above parameters.
Description of data collection	The data collected were taken from past material science machine learning model papers whose data were either publicly available or were provided when one contacted the corresponding author. Each dataset was downloaded and analyzed using Python's seaborn.distplot and subsequently cleaned if needed. The specific papers that were used for this dataset are described in the Data Source Location section below.
Data source location	Primary data sources: See Table 1
Data accessibility	Repository name: GitHub http://doi.org/10.5281/zenodo.4903958 Direct URL to data: https://github.com/anhender/mse_ML_datasets/tree/v1.0 Instructions for accessing these data: It is recommended that one accesses the GitHub repository using the direct URL provided, where further instructions for accessing these data are provided.

Value of the Data

- Because these data are from past literature, they include a variety of materials properties, as well as both experimental and calculated values. Therefore, this collection of data acts as a unique benchmark dataset that can be used to accurately and efficiently compare different machine learning models for materials informatics, which can aid in improving current practices in the field of Computational Materials Science.
- Beyond the merits of the data themselves, as described in their respective publications, the aggregation of these benchmark data brings researchers closer to having a single unified collection of materials data for machine learning and statistical methods. As such, these benchmark data allow researchers to effectively compare machine learning models to one another, which will ultimately aid the process of finding the most efficient method for materials discovery.
- These benchmark data can be used as a basis for creating a more diverse benchmark dataset in the future. Larger, more diverse datasets will allow researchers to explore the generalizability of machine learning models. The MatBench project [1] could also benefit by incorporating some of these data to improve the diversity and types of learning problems on which they test various machine learning approaches.
- For 2 of the 16 papers, the data that was collected and presented here was not previously available publicly. Therefore, this dataset constitutes the first public repository which can be easily accessed for subsequent learning.

1. Data Description

The current dataset is a compilation of 50 datasets that were collected from 16 previous machine learning materials informatics papers, which were published between the years 2013 to 2019. Table 1 lists the papers used for this collection of benchmark data.

Fig. 1 provides a general overview of the kinds of datasets that were collected for this benchmark data. This figure gives information about the kind of material systems, data sizes, organic nature of material, data types, and task types that exist in this new benchmark data. Each category and respective specifications will be briefly described below.

The Material System category describes the kind of materials that each dataset studied. From Fig. 1, it can be seen that the majority of these datasets studied either polymers or superlattice materials. The Misc. subcategory is comprised of composite materials, component solids, semiconductors, metal alloys, glass, and MXenes. While some of these are considered inorganic solids, they were placed in the Misc. subcategory because the respective papers specifically described the kind of inorganic solid that was studied.

Table 1
A list of the 16 primary sources that were used to create this collection of benchmark data. The sources are listed in alphabetical order.

Balachandran, Prasanna V., et al. “Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning.” Nature communications 9.1 (2018): 1–9. https://doi.org/10.1038/s41467–018–03821–9
Carrete, Jesús, et al. “Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling.” Physical Review X 4.1 (2014): 011,019. https://doi.org/10.1103/PhysRevX.4.011019
Lee, Joohwi, et al. “Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques.” Physical Review B 93.11 (2016): 115,104. https://doi.org/10.1103/PhysRevB.93.115104
Li, Wei, Ryan Jacobs, and Dane Morgan. “Predicting the thermodynamic stability of perovskite oxides using machine learning models.” Computational Materials Science 150 (2018): 454–463. https://doi.org/10.1016/j.commatsci.2018.04.033
Liu, Yue, et al. “The onset temperature (Tg) of AsxSe1-x glasses transition prediction: A comparison of topological and regression analysis methods.” Computational Materials Science 140 (2017): 315–321. https://doi.org/10.1016/j.commatsci.2017.09.008
Mannodi-Kanakkithodi, Arun, et al. “Machine learning strategy for accelerated design of polymer dielectrics.” Scientific reports 6 (2016): 20,952. https://doi.org/10.1038/srep20952
Pilania, Ghanshyam, et al. “Accelerating materials property predictions using machine learning.” Scientific reports 3.1 (2013): 1–6. https://doi.org/10.1038/srep02810
Pilania, Ghanshyam, et al. “Machine learning bandgaps of double perovskites.” Scientific reports 6 (2016): 19,375. https://doi.org/10.1038/srep19375
Pilania, Ghanshyam, and X-Y. Liu. “Machine learning properties of binary wurtzite superlattices.” Journal of materials science 53.9 (2018): 6652–6664. https://doi.org/10.1007/s10853–018–1987-z
Rajan, Arunkumar Chitteth, et al. “Machine-learning-assisted accurate band gap predictions of functionalized MXene.” Chemistry of Materials 30.12 (2018): 4031–4038. https://doi.org/10.1021/acs.chemmater.8b00686
Seko, Atsuto, et al. “Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids.” Physical Review B 89.5 (2014): 054,303. https://doi.org/10.1103/PhysRevB.89.054303
Wei, Han, et al. “Predicting the effective thermal conductivities of composite materials and porous media by machine learning methods.” International Journal of Heat and Mass Transfer 127 (2018): 908–916. https://doi.org/10.1016/j.ijheatmasstransfer.2018.08.082
Wu, K., et al. “Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials.” Journal of Polymer Science Part B: Polymer Physics 54.20 (2016): 2082–2091. https://doi.org/10.1002/polb.24117
Xue, Dezhen, et al. “Accelerated search for materials with targeted properties by adaptive design.” Nature communications 7.1 (2016): 1–9. https://doi.org/10.1038/ncomms11241
Zeng, Shuming, et al. “Machine learning-aided design of materials with target elastic properties.” The Journal of Physical Chemistry C 123.8 (2019): 5042–5047. https://doi.org/10.1021/acs.jpcc.9b01045
Zhuo, Ya, Aria Mansouri Tehrani, and Jakoah Brgoch. “Predicting the band gaps of inorganic solids by machine learning.” The journal of physical chemistry letters 9.7 (2018): 1668–1673. https://doi.org/10.1021/acs.jpclett.8b00124

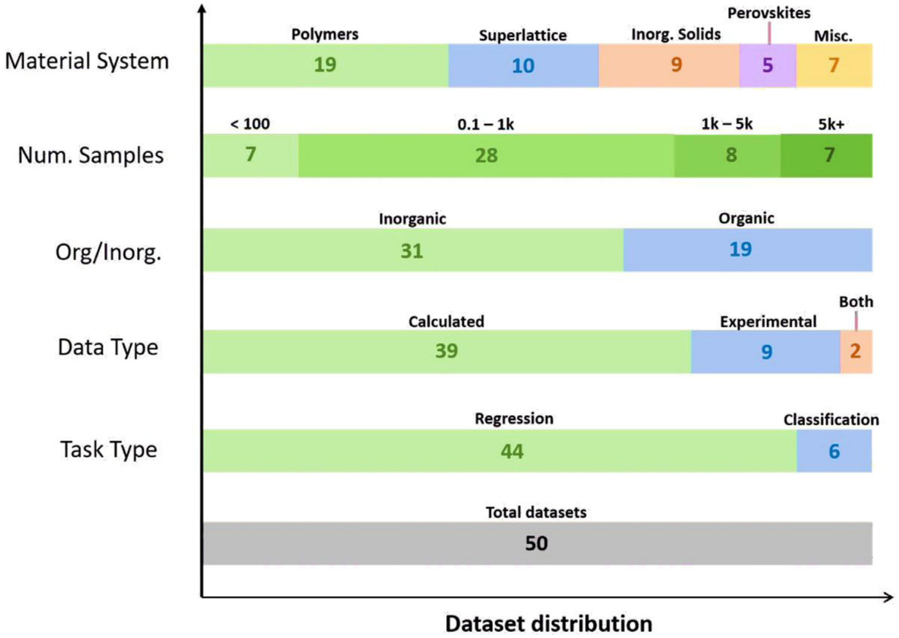


Fig. 1. A categorial dataset distribution of the 50 datasets compiled from 16 previous machine learning materials informatics papers. The categorization methods used are listed on the left and specific descriptors are listed above each colored bar of the graph. The number in each bar describes the number of datasets that fit that specification (e.g., 44 of the 50 datasets are regression tasks).

The Num. Samples category describes the sizes of the datasets. The size range is from below 100 values to over 5000, with the smallest dataset having a size of 12 and the largest dataset with a size of 6354. The majority of datasets are between 100 and 1000 samples.

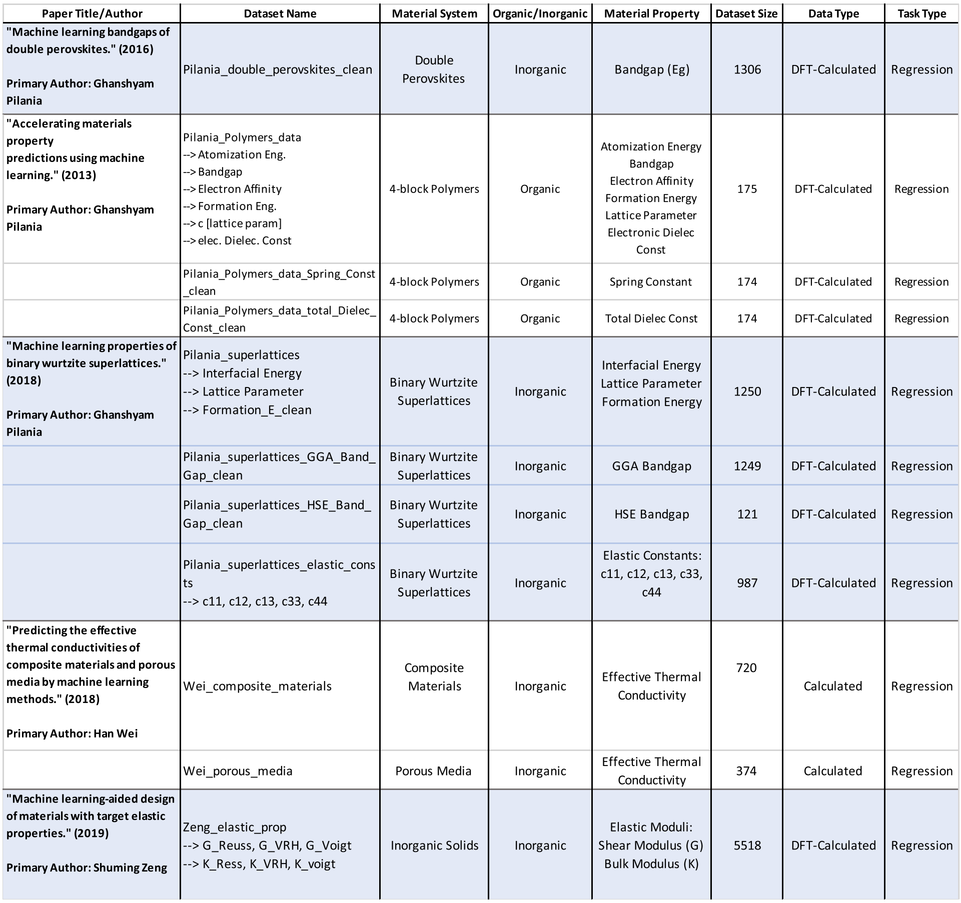
The Org/Inorg. category describes whether a dataset studied organic or inorganic materials. 31 of the 50 total datasets studied inorganic materials while the other 19 studied organic materials.

The Data Type category describes if a dataset used calculated or experimental data for its tasks. The majority of datasets used calculated data (39 of 50). It should be noted that two datasets (Zhuo_classification_data and Liu_Tg_AsSe_glass) used calculated and experimental data together.

Finally, the Task Type category describes the machine learning task type of a dataset. Forty-four of the datasets are regression tasks and the remaining six are classification tasks.

After all of the 50 datasets were gathered and cleaned (when necessary), they were split into train-val-test splits or train-test splits, depending on their size. Three different methods were used: 5-Fold cross-validation, 10-Fold cross-validation, or Leave-One-Out cross-validation. Fig. 2 gives specific information of the datasets that had train-val-test splits created using the 5-Fold method, while Fig. 3 gives information about the datasets whose train-val-test splits were created with the 10-Fold method and Fig. 4 describes the datasets that had train-test splits created using the Leave-One-Out method. For all three of these figures, the information given for each dataset includes: its respective paper, its name as given in the repository, the specific material system studied, the organic nature of said material, the kind of material property, the size of the dataset, the type of data, and the task type.

The method of how the Dataset Name column is set up will be briefly described here. If multiple datasets originated from the same paper, their names are organized in two different ways, depending on whether or not a dataset was cleaned.



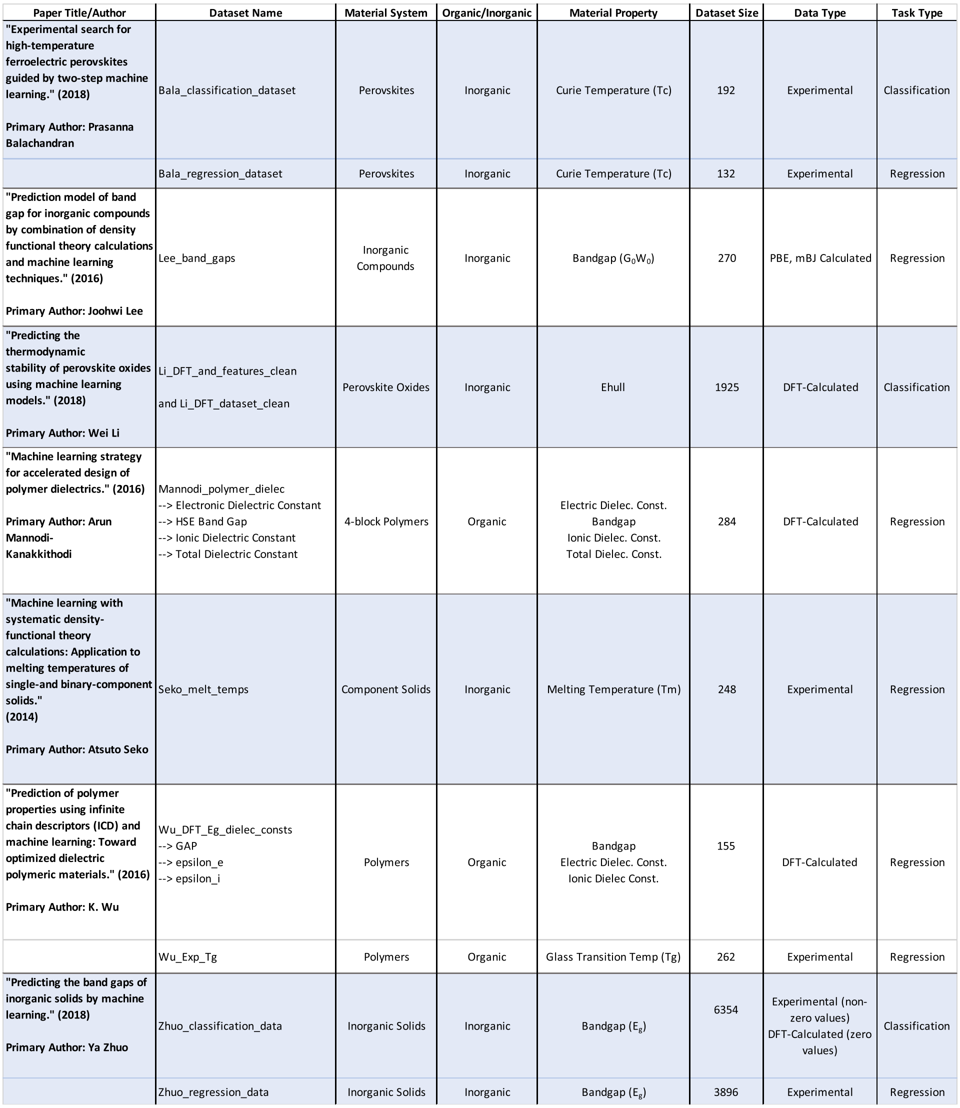
Paper Title/Author	Dataset Name	Material System	Organic/Inorganic	Material Property	Dataset Size	Data Type	Task Type
"Machine learning bandgaps of double perovskites." (2016) Primary Author: Ghanshyam Pilania	Pilania_double_perovskites_clean	Double Perovskites	Inorganic	Bandgap (Eg)	1306	DFT-Calculated	Regression
"Accelerating materials property predictions using machine learning." (2013) Primary Author: Ghanshyam Pilania	Pilania_Polymers_data -->Atomization Eng. -->Bandgap -->Electron Affinity -->Formation Eng. -->c [lattice param] -->elec. Dielec. Const	4-block Polymers	Organic	Atomization Energy Bandgap Electron Affinity Formation Energy Lattice Parameter Electronic Dielec Const	175	DFT-Calculated	Regression
	Pilania_Polymers_data_Spring_Const_clean	4-block Polymers	Organic	Spring Constant	174	DFT-Calculated	Regression
	Pilania_Polymers_data_total_Dielec_Const_clean	4-block Polymers	Organic	Total Dielec Const	174	DFT-Calculated	Regression
"Machine learning properties of binary wurtzite superlattices." (2018) Primary Author: Ghanshyam Pilania	Pilania_superlattices --> Interfacial Energy --> Lattice Parameter --> Formation_E_clean	Binary Wurtzite Superlattices	Inorganic	Interfacial Energy Lattice Parameter Formation Energy	1250	DFT-Calculated	Regression
	Pilania_superlattices_GGA_Band_Gap_clean	Binary Wurtzite Superlattices	Inorganic	GGA Bandgap	1249	DFT-Calculated	Regression
	Pilania_superlattices_HSE_Band_Gap_clean	Binary Wurtzite Superlattices	Inorganic	HSE Bandgap	121	DFT-Calculated	Regression
	Pilania_superlattices_elastic_consts --> c11, c12, c13, c33, c44	Binary Wurtzite Superlattices	Inorganic	Elastic Constants: c11, c12, c13, c33, c44	987	DFT-Calculated	Regression
"Predicting the effective thermal conductivities of composite materials and porous media by machine learning methods." (2018) Primary Author: Han Wei	Wei_composite_materials	Composite Materials	Inorganic	Effective Thermal Conductivity	720	Calculated	Regression
	Wei_porous_media	Porous Media	Inorganic	Effective Thermal Conductivity	374	Calculated	Regression
"Machine learning-aided design of materials with target elastic properties." (2019) Primary Author: Shuming Zeng	Zeng_elastic_prop --> G_Reuss, G_VRH, G_Voigt --> K_Ress, K_VRH, K_voigt	Inorganic Solids	Inorganic	Elastic Moduli: Shear Modulus (G) Bulk Modulus (K)	5518	DFT-Calculated	Regression

Fig. 2. Information about the datasets that had train-val-test splits created using the 5-Fold cross-validation method. Each dataset is described by its name, material system, organic nature, material property, dataset size, data type, and task type. The paper of each respective dataset is provided as well in the left-most column.

(1) If a dataset did not need cleaning, it is listed under its parent name within a single cell. For example, reference the second paper from the top of Fig. 2 ("Accelerating materials property predictions using machine learning"). In the first cell of the Dataset Name column, there are multiple datasets listed, each designated with an arrow (→) below the parent name, 'Pilania_Polymers_data'. This notation means that there are six separate datasets (Atomization Eng., Bandgap, Electron Affinity, Formation Eng., c [lattice param], and elec. Dielec. Const) of the same size that came from the same paper.

(2) If a dataset needed cleaning, it is given its own separate row and its full name is written out, including the parent name. The second paper of Fig. 2 will be used again as an example. The second cell of the Dataset Name column shows how this is written: 'Pilania_Polymers_data_Spring_Const_clean'. All in all, it can be seen that this paper contributed eight datasets in total to this benchmark data.

All datasets can be accessed at https://github.com/anhender/mse_ML_datasets/tree/v1.0 [2]. The data is provided in both its raw format (before the creation of train-val-test splits) and its processed format (in which train-val-test splits were created). In general, there are two ways

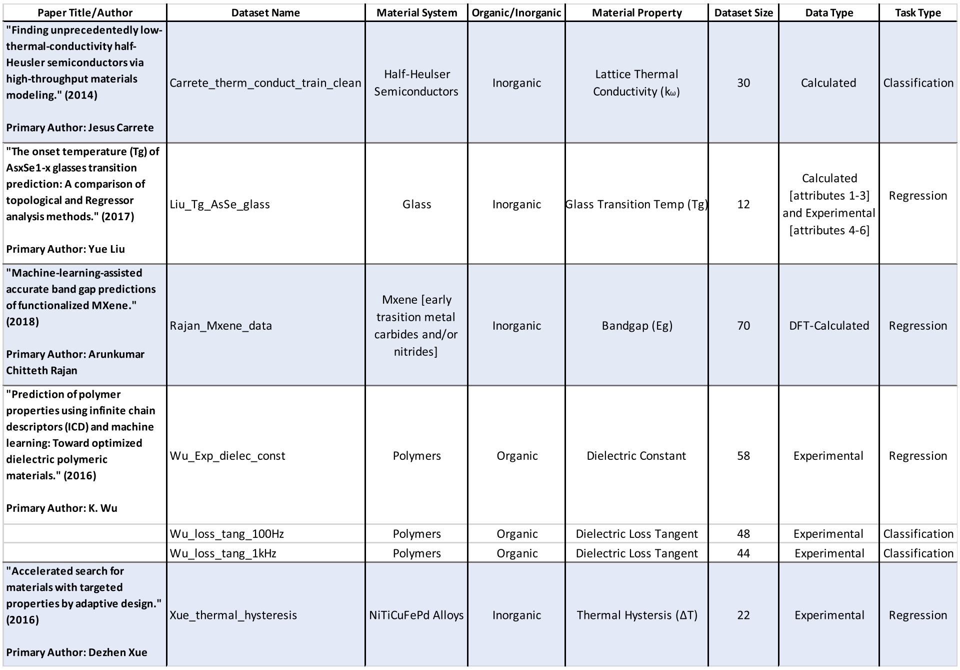


Paper Title/Author	Dataset Name	Material System	Organic/Inorganic	Material Property	Dataset Size	Data Type	Task Type
"Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning." (2018) **Primary Author: Prasanna Balachandran**	Bala_classification_dataset	Perovskites	Inorganic	Curie Temperature (Tc)	192	Experimental	Classification
	Bala_regression_dataset	Perovskites	Inorganic	Curie Temperature (Tc)	132	Experimental	Regression
"Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques." (2016) **Primary Author: Joohwi Lee**	Lee_band_gaps	Inorganic Compounds	Inorganic	Bandgap (G_0W_0)	270	PBE, mBJ Calculated	Regression
"Predicting the thermodynamic stability of perovskite oxides using machine learning models." (2018) **Primary Author: Wei Li**	Li_DFT_and_features_clean and Li_DFT_dataset_clean	Perovskite Oxides	Inorganic	Ehull	1925	DFT-Calculated	Classification
"Machine learning strategy for accelerated design of polymer dielectrics." (2016) **Primary Author: Arun Mannodi-Kanakkithodi**	Mannodi_polymer_dielec --> Electronic Dielectric Constant --> HSE Band Gap --> Ionic Dielectric Constant --> Total Dielectric Constant	4-block Polymers	Organic	Electric Dielec. Const. Bandgap Ionic Dielec. Const. Total Dielec. Const.	284	DFT-Calculated	Regression
"Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids." (2014) **Primary Author: Atsuto Seko**	Seko_melt_temps	Component Solids	Inorganic	Melting Temperature (Tm)	248	Experimental	Regression
"Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials." (2016) **Primary Author: K. Wu**	Wu_DFT_Eg_dielec_consts --> GAP --> epsilon_e --> epsilon_i	Polymers	Organic	Bandgap Electric Dielec. Const. Ionic Dielec Const.	155	DFT-Calculated	Regression
	Wu_Exp_Tg	Polymers	Organic	Glass Transition Temp (Tg)	262	Experimental	Regression
"Predicting the band gaps of inorganic solids by machine learning." (2018) **Primary Author: Ya Zhuo**	Zhuo_classification_data	Inorganic Solids	Inorganic	Bandgap (E_g)	6354	Experimental (non-zero values) DFT-Calculated (zero values)	Classification
	Zhuo_regression_data	Inorganic Solids	Inorganic	Bandgap (E_g)	3896	Experimental	Regression

Fig. 3. Information about the datasets that had train-val-test splits created using the 10-Fold cross-validation method. Each dataset is described by its name, material system, organic nature, material property, dataset size, data type, and task type. The paper of each respective dataset is provided as well in the left-most column.

data is provided, either with features or without. Table 2 lists the datasets that do not contain any features, while Table 3 lists the datasets that contain features.

Datasets with no features (see Table 2) only have two or three columns, which describe material compositions, material property values, and (when a third column exists) the Materials Project ID for the corresponding compositions. The first ten lines of the Zhuo_classification_data dataset are shown in Fig. 5 as an example of what a typical Table 2 dataset looks like. It can be seen that only two columns exist, where each composition given in column one has a corresponding band gap value in column two.



Paper Title/Author	Dataset Name	Material System	Organic/Inorganic	Material Property	Dataset Size	Data Type	Task Type
"Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling." (2014) Primary Author: Jesus Carrete	Carrete_therm_conduct_train_clean	Half-Heulser Semiconductors	Inorganic	Lattice Thermal Conductivity (k_{ω})	30	Calculated	Classification
"The onset temperature (Tg) of AsxSe1-x glasses transition prediction: A comparison of topological and Regressor analysis methods." (2017) Primary Author: Yue Liu	Liu_Tg_AsSe_glass	Glass	Inorganic	Glass Transition Temp (Tg)	12	Calculated [attributes 1-3] and Experimental [attributes 4-6]	Regression
"Machine-learning-assisted accurate band gap predictions of functionalized MXene." (2018) Primary Author: Arunkumar Chitteth Rajan	Rajan_Mxene_data	Mxene [early trasition metal carbides and/or nitrides]	Inorganic	Bandgap (Eg)	70	DFT-Calculated	Regression
"Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials." (2016) Primary Author: K. Wu	Wu_Exp_dielec_const	Polymers	Organic	Dielectric Constant	58	Experimental	Regression
	Wu_loss_tang_100Hz	Polymers	Organic	Dielectric Loss Tangent	48	Experimental	Classification
	Wu_loss_tang_1kHz	Polymers	Organic	Dielectric Loss Tangent	44	Experimental	Classification
"Accelerated search for materials with targeted properties by adaptive design." (2016) Primary Author: Dezhen Xue	Xue_thermal_hysteresis	NiTiCuFePd Alloys	Inorganic	Thermal Hystersis (ΔT)	22	Experimental	Regression

Fig. 4. Information about the datasets that had train-test splits created using the Leave-One-Out cross-validation method. Each dataset is described by its name, material system, organic nature, material property, dataset size, data type, and task type. The paper of each respective dataset is provided as well in the left-most column.

Datasets that contain features (see Table 3) range in their number of columns from 4 to 966, though all but one has 16 columns or less. This wide range is due to the fact that each dataset considered different material systems and properties, so some features were important to consider while others were not. The dataset that provides the most robust number of features is Li_DFT_and_features_clean, which contains 966 columns in total. A slimmed down version of this dataset is provided as well, Li_DFT_dataset_clean, which has 12 columns. Along with the extra information the features provide, all of these datasets still contain the material composition and material property columns as with the Table 2 datasets.

2. Experimental Design, Materials and Methods

The datasets were manually gathered from 16 previous materials science machine learning (ML) model papers. The parameters for this process were briefly explained in the Specifications Table above, but will be restated here. A dataset was chosen if: 1) its paper was relatively recent (published in 2013 or later), 2) the data was publicly available or easily attainable by contacting the corresponding author, and 3) the study used some kind of regression and/or classification ML model for any kind of material property. Datasets were gathered independent of data type, data size, and material system as long as the above three parameters were met and the data itself matched what its respective paper described it to be.

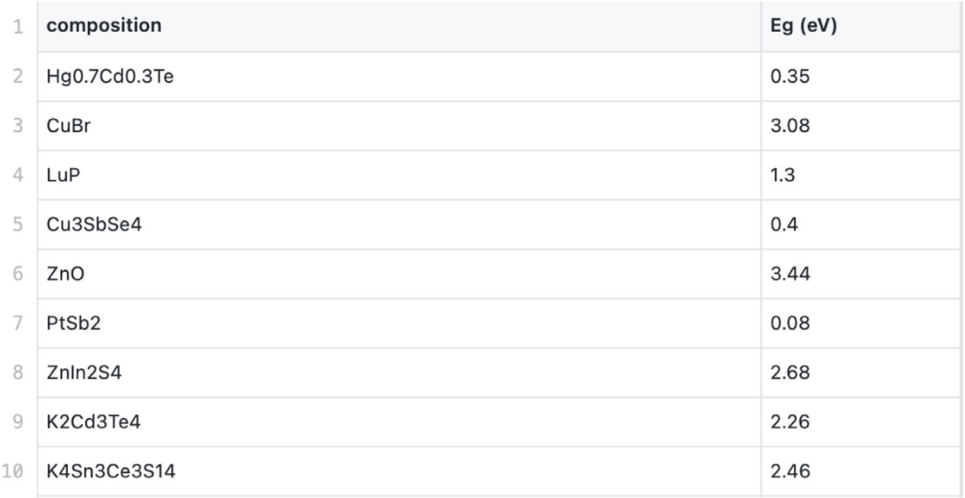
Once all data had been gathered, the initially collected datasets were split such that data were separated either by model type (classification or regression) or by data type (experimental or calculated) for each respective paper, if applicable, which led to the creation of 25 datasets. This was done manually, without any coding, as many of the initial datasets came in PDF format and had to be converted to CSV format. Then, if applicable, these data were split further such

Table 2
Datasets that contain no features, only information regarding the material property. These datasets contain only two or three columns, as described in the text.

Carrete_therm_conduct_train_clean	Pilania_superlattices_HSE_Band_Gap_clean
Mannodi_polymer_dielec/Electronic Dielectric Constant	Pilania_superlattices/Interfacial Energy (eV-angstrom^2)
Mannodi_polymer_dielec/HSE Band Gap (eV)	Pilania_superlattices/Lattice Parameter (angstrom)
Mannodi_polymer_dielec/Ionic Dielectric Constant	Seko_melt_temps
Mannodi_polymer_dielec/Total Dielectric Constant	Wu_DFT_Eg_dielec_consts/epsilon_e
Pilania_Polymers_data_Spring_Const_clean	Wu_DFT_Eg_dielec_consts/epsilon_i
Pilania_Polymers_data_total_Dielec_Const_clean	Wu_DFT_Eg_dielec_consts/GAP
Pilania_Polymers_data/Atomization Eng. (eV)	Wu_Exp_dielec_const
Pilania_Polymers_data/Bandgap (eV)	Wu_Exp_Tg
Pilania_Polymers_data/c [lattice param] (angstrom)	Wu_loss_tang_100Hz
Pilania_Polymers_data/elec. Dielec. Const	Wu_loss_tang_1kHz
Pilania_Polymers_data/Electron Affinity (eV)	Zeng_elastic_prop/G_Reuss
Pilania_Polymers_data/Formation Eng. (eV)	Zeng_elastic_prop/G_Voigt
Pilania_superlattices_elastic_consts/c11 (GPa)	Zeng_elastic_prop/G_VRH
Pilania_superlattices_elastic_consts/c12 (GPa)	Zeng_elastic_prop/K_Ress
Pilania_superlattices_elastic_consts/c13 (GPa)	Zeng_elastic_prop/K_voigt
Pilania_superlattices_elastic_consts/c33 (GPa)	Zeng_elastic_prop/K_VRH
Pilania_superlattices_elastic_consts/c44 (GPa)	Zhuo_classification_data
Pilania_superlattices_Formation_E_clean	Zhuo_regression_data
Pilania_superlattices_GGA_Band_Gap_clean	

Table 3
Datasets that contain extra features besides only the material property and material composition. The features used in each dataset vary due to the different material system and material properties that were studied in each respective dataset.

Bala_classification_dataset	Pilania_double_perovskites_clean
Bala_regression_dataset	Rajan_MXene_data
Lee_band_gaps	Wei_composite_materials
Li_DFT_and_features_clean	Wei_porous_media
Li_DFT_dataset_clean	Xue_thermal_hysteresis
Liu_Tg_AsSe_glass	



1	composition	Eg (eV)
2	Hg0.7Cd0.3Te	0.35
3	CuBr	3.08
4	LuP	1.3
5	Cu3SbSe4	0.4
6	ZnO	3.44
7	PtSb2	0.08
8	ZnIn2S4	2.68
9	K2Cd3Te4	2.26
10	K4Sn3Ce3S14	2.46

Fig. 5. The first ten lines of the Zhuo_classification_data dataset. The left column describes compositions of inorganic solids while the right column gives the corresponding band gap values. This is an example of a dataset without features.

that, per paper, each material property had its own dataset. This led to a total of 50 datasets. This process was done in order to make the data as accessible as possible for others, since many of the papers chosen for this benchmark dataset studied multiple material properties at once.

Data visualization was then done on each dataset using seaborn.distplot to determine if any outliers existed. If an outlier was found, it was removed from its respective dataset. From the 50 total datasets, only nine had to be cleaned. The cleaning process entailed the removal of duplicates, NaN values, and outlier points. Generally, outliers were deemed to be outliers if a single point excessed several standard deviations away from the mean in clear contradiction from other datapoints in the dataset.

After all of the necessary datasets were effectively cleaned, train-val-test or train-test splits were created, depending on the size of each dataset. For datasets with more than 100 values, train-val-test splits were created via the scikit-learn K-Fold cross-validation method. The number of folds, 5 versus 10, was determined by following what each respective paper did in their studies. For datasets with less than 100 values, test-train sets were created via the scikit-learn Leave-One-Out cross-validation method. The code for this last step is available in the repository, as well as the raw data that was used to create the final splits for each dataset.

Ethics Statement

The work did not involve the use of human subjects, animal experiments, nor data collected from social media platforms.

CRediT Author Statement

Ashley N. Henderson: Data curation, Formal analysis, Writing – original draft; **Steven K. Kauwe:** Supervision, Writing – review & editing; **Taylor D. Sparks:** Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the University of Utah's Undergraduate Research Opportunities Program (UROP) for the Fall 2020 Semester. TDS and SKK gratefully acknowledge support from the National Science Foundation CAREER Award, Grant No. 1651668.

References

[1] A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm, npj Comput. Mater. 6 (2020) 138, doi:10.1038/s41524-020-00406-3.

[2] A. Henderson, S. Kauwe, T. Sparks, Benchmark Datasets Incorporating Diverse Tasks, Sample Sizes, Material Systems, and Data Heterogeneity For Materials Informatics, Zenodo, 2021 v1.0, doi:10.5281/zenodo.4903958.