Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: 2766-0400 (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Data-driven analysis and visualization of dielectric properties curated from scientific literature

Tomoki Murata, Naoto Saito, Eiji Koyama, Ton Nu Thanh Phuong, Ryusuke Misawa, Satoshi Yokomizo, Tomoya Mato, Yu Takada, Sakyo Hirose & Yukari Katsura

To cite this article: Tomoki Murata, Naoto Saito, Eiji Koyama, Ton Nu Thanh Phuong, Ryusuke Misawa, Satoshi Yokomizo, Tomoya Mato, Yu Takada, Sakyo Hirose & Yukari Katsura (2025) Data-driven analysis and visualization of dielectric properties curated from scientific literature, Science and Technology of Advanced Materials: Methods, 5:1, 2485018, DOI: 10.1080/27660400.2025.2485018

To link to this article: https://doi.org/10.1080/27660400.2025.2485018

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2025, VOL. 5, NO. 1, 2485018
https://doi.org/10.1080/27660400.2025.2485018

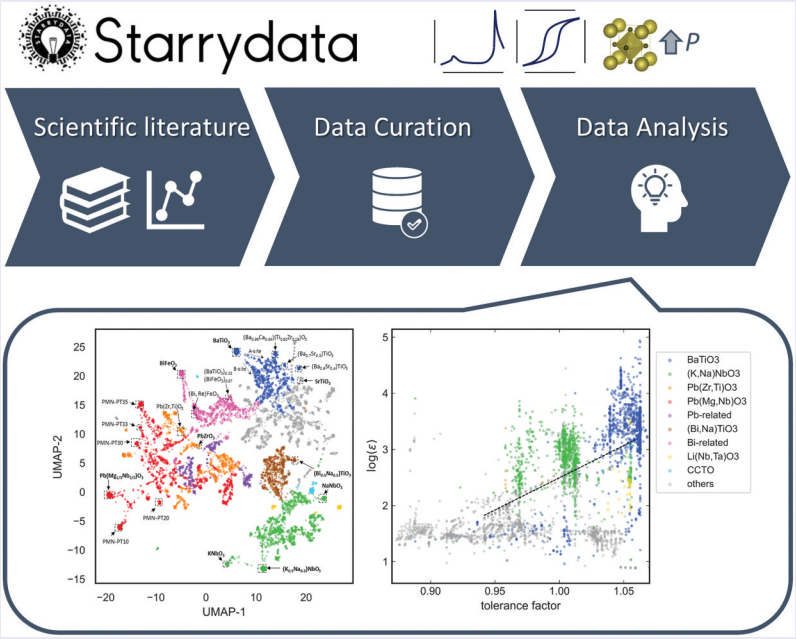
Data-driven analysis and visualization of dielectric properties curated from scientific literature

Tomoki Murata [a], Naoto Saito [b], Eiji Koyama [b], Ton Nu Thanh Phuong [b], Ryusuke Misawa [a], Satoshi Yokomizo [a], Tomoya Mato [b], Yu Takada [b], Sakyo Hirose [a] and Yukari Katsura [b,c,d]

[a]R&D Department for Frontier Technology, Murata Manufacturing Co, Ltd, Nagaokakyo, Japan; [b]Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba, Japan; [c]Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan; [d]RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

ABSTRACT

Data-driven methods are powerful tools for understanding and discovering materials. However, in certain domains of materials science, the lack of available datasets significantly restricts the research scope. To address this issue, we utilized Starrydata2 web system to compile a comprehensive dataset on dielectric materials, collecting experimental data on over 20,000 samples across a wide compositional space. This dataset enabled the development of machine learning models with high predictive performance and facilitated the identification of important descriptors through recursive feature eliminations. Since the models worked as complete black boxes and hindered intuitive understanding, we employed additional techniques such as dimensionality reduction and clustering to visualize compositional landscape and trends in dielectric properties. By combining the identified important factors with material clustering, we attempted to visualize the effect of crystal lattice on dielectric permittivity within ABO_3 systems, revealing a roughly linear relationship. Our preliminary analyses and visualizations demonstrate the potential of the Starrydata dielectric dataset collected in this study, offering an important foundation for advanced data-driven materials research.



IMPACT STATEMENT

This study presents a comprehensive dielectric database curated from the scientific literature with data analyses and visualizations demonstrating its potential, offering an important foundation for advanced data-driven materials research.

ARTICLE HISTORY

Received 27 November 2024
Revised 28 February 2025
Accepted 24 March 2025

KEYWORDS

Materials data analysis; database; data curation; dielectric materials; ferroelectric; machine learning; dimensionality reduction

CONTACT Tomoki Murata tomoki.murata258@murata.com Murata Manufacturing Co, Ltd, 10-1, Higashikotari 1-chome, Nagaokakyo, Kyoto 617-8555, Japan; Yukari Katsura KATSURA.Yukari@nims.go.jp Center for Basic Research on Materials, National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan

Supplemental data for this article can be accessed online at https://doi.org/10.1080/27660400.2025.2485018

1. Introduction

Data-driven exploration has emerged as a powerful approach for understanding and discovering materials across various domains in materials science. Advanced data analysis methods, including machine learning, dimensionality reduction, and clustering, have been widely applied to predict material properties and capture the overall trends of materials. These methods rely on well-structured datasets. As a result, data-driven techniques are often particularly compatible with high-throughput computational approaches, where large volumes of data are generated *in-silico* [1]. Excellent predictive performance has been reported in computed materials property such as band gap [2,3] and dielectric permittivity [4,5]. Additionally, sophisticated techniques like dimensionality reduction and clustering are highly effective on large computed datasets, capturing global trends within vast materials spaces [5–8].

Although computed data plays a major role and firmly supports data-driven research, there are still some fundamental limitations to computational approaches. Certain material properties are challenging or costly to calculate, especially those at finite temperatures, such as thermal conductivity, ionic conductivity, and structural phase transitions. In addition, while calculations for perfect crystals are feasible, complexities arise when dealing with solid solutions. Simulating solid solutions often requires constructing supercells, a process that is computationally intensive and costly [9,10]. For instance, computing the properties of a solid solution like (Ba, Sr)TiO_3 is far more demanding than for a single-phase material such as $BaTiO_3$ or $SrTiO_3$ [11]. Furthermore, some materials properties are quite feasible to extrinsic effects, which could not be represented in atomic scale computations. Due to these multiple challenges, computational methods alone cannot capture the full range of material properties, making it necessary to incorporate experimental data to build comprehensive datasets.

Various experimental databases are utilized for machine learning involving physical properties that are difficult to calculate. A simple but prominent example is the SuperCon database [12], an experimental database containing approximately 26,000 entries on superconducting critical temperatures. This dataset has been utilized to construct predictive models of superconducting transition temperatures [13,14] and to propose potential compositions for new superconductors [15]. Another example is the database of ionic conductivities for solid-state electrolytes, which contains around 2,000 entries [16]. This database has facilitated the visualization of conductivity trends across material compositions through mapping techniques using dimensionality reductions. Additional experimental databases include the Inorganic Crystal Structure Database (ICSD) for crystal structures [17], the Pauling File for material properties [18], the High Throughput Experimental Materials (HTEM) database for thin-film properties [19], and the International Centre for Diffraction Data's Powder Diffraction File (ICDD-PDF) database for diffraction data [20]. These databases have enabled a variety of data-driven research across materials science.

However, a comprehensive database that compiles experimental data on dielectric and ferroelectric materials has not existed until now. Experimental databases such as the Pauling File [18] and Sebastian's dataset [21,22] provide dielectric properties but are limited to single-temperature data. Temperature dependence is a key characteristic of dielectric materials, particularly in ferroelectrics, where both dielectric permittivity and loss exhibit peaks near the Curie temperature. Numerous studies have explored compositional adjustments to modify the shape of the permittivity peak, aiming to optimize dielectric properties for both fundamental understanding of ferroelectricity and practical applications for capacitors [23–27]. Despite the importance of these properties, collecting such temperature-dependent data is an extremely time-consuming task and has rarely been done in previous research. The Starrydata2 web system, a tool specifically designed for extracting experimental data from graphs in scientific literature [28,29], has been employed for temperature-dependent data collection. Starrydata2 facilitates the direct extraction of x-y data points from graphs, making it particularly well-suited for datasets with significant parameter dependencies. Previous studies have utilized Starrydata2 to compile temperature-dependent thermoelectric properties [28,29]. In this research, we construct a large experimental database of dielectric materials which includes the temperature dependence of dielectric permittivity and dielectric loss, using the Starrydata2 web system.

Another important aspect for collecting experimental data, rather than solely relying on computations, lies in the mechanism of dielectric permittivity and limitation of computational approach. The dielectric permittivity of inorganic materials consists of three distinct contributions from different physical origins, as formulated below:

$$\varepsilon = \varepsilon_{\text{electronic}} + \varepsilon_{\text{ionic}} + \varepsilon_{\text{dipole}} \quad (1)$$

where $\varepsilon_{\text{electronic}}$ is the electronic contribution, $\varepsilon_{\text{ionic}}$ is an ionic contribution, and $\varepsilon_{\text{dipole}}$ is the dipole contribution. Paraelectric material only has electronic and ionic contribution, whereas ferroelectric and antiferroelectric material has the additional term of dipole component. The dipole contribution is typically much larger than the electronic and ionic contributions and plays a critical role in the dielectric permittivity [30,31]. The contributions of electrons and ions can

be predicted using density functional perturbation theory (DFPT) [32]. Since DFPT calculations are computationally rather efficient, large-scale calculation on the entire materials database is feasible [4,33]. Previous data-driven studies on dielectric materials have primarily focused on the computed ionic and electronic components [4,5,33–37]. On the other hand, the simulation of dipole contribution remains highly challenging. As the dipole contribution originates from the rotation of dipole moments at finite temperature, this requires finite-temperature molecular dynamics (MD) simulations with an expanded supercell. Full prediction of temperature dependence using MD has been reported to date employing an effective Hamiltonian [38,39] and machine learned interatomic potentials [40], though it still lacks established methodology and is far more computationally demanding, making it unsuitable for high-throughput computations. Due to this limitation, data analysis of ferroelectric materials, which exhibit high dielectric permittivity due to their substantial dipole component, remains largely unexplored. While excellent predictive models for the temperature dependence of dielectric permittivity have been reported, they work in a limited compositional space due to the limited data availability [41]. Thus, collecting experimental data on the temperature dependence of dielectric permittivity across a wide range of material compositions is essential for further data-driven research.

In this research, we compiled a large-scale experimental database on dielectric properties. The schematic of this study is presented in Figure 1. We carefully curated experimental data on dielectric permittivity and dielectric loss from graphs in scientific literature. To extract and organize the numerical data from graphs, we utilized the Starrydata2 web system. This process involved manual curation supported by our web system and was not fully automated. We gathered data from over 20,000 samples across more than 5,000 publications. Since this is the first data analysis of this unprecedented large-scale experimental dataset, we prioritized understanding and showcasing the nature and scope of the dataset in this study. To effectively visualize the trend within the curated dataset, we applied dimensionality reduction and clustering techniques, effectively revealing distinct clusters among known ferroelectric materials. In addition, we found that the overall trend in dielectric permittivity could be explained by six descriptors, including previously recognized important factors for ferroelectrics, such as the tolerance factor and electronegativity. Our visualizations and findings align with established scientific knowledge, underscoring how data-driven analysis can corroborate and complement human insights. The data analysis presented here is a preliminary result to showcase the nature of the dataset. The Starrydata dielectric dataset curated in this study offers researchers worldwide an important foundation to perform diverse data-driven research to uncover new insights and explore new materials.

2. Methods

2.1. Data curation from literatures

We created a list of academic papers using Scopus [43]. We searched for keywords such as ‘dielectric property’ and selecting articles primarily from ceramic-focused

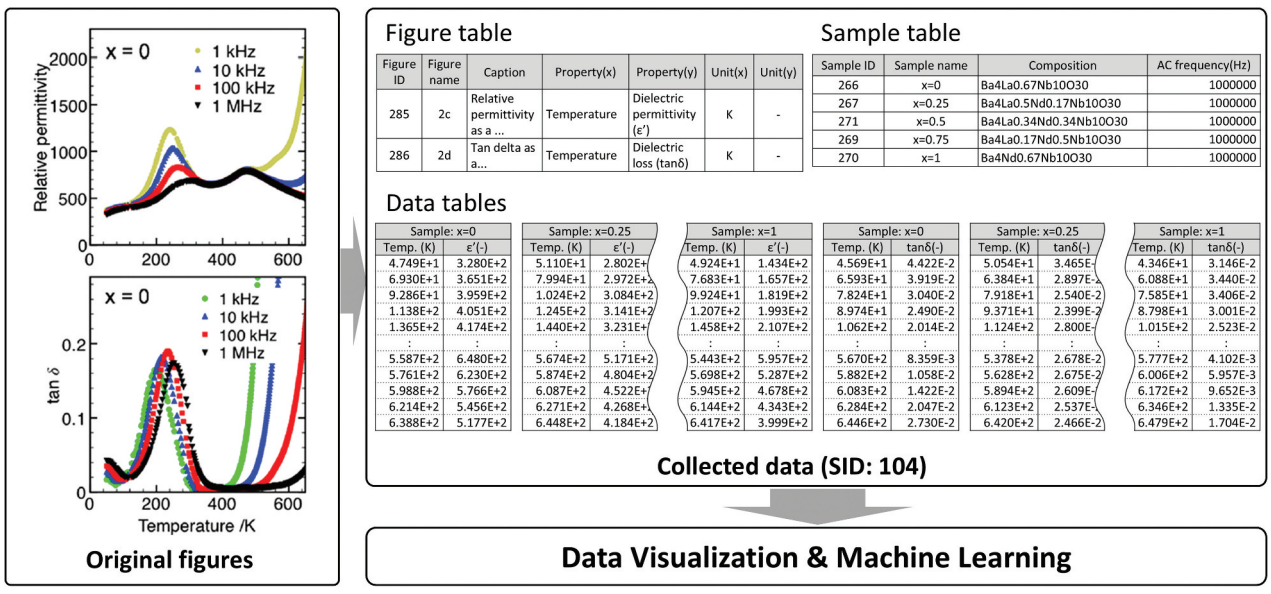


Figure 1. Schematics of data curation procedure from scientific literature. We extracted figures and tables containing dielectric properties, digitized the data, and registered them in the database. The dataset includes material composition, temperature dependence of dielectric permittivity and dielectric loss, and measurement frequency. Original figures are reprinted with permission from ref [42]. Copyright 2020 American Chemical Society.

journals, as our focus was on ceramics. Following this list, we downloaded and reviewed each paper in PDF format. If data on dielectric permittivity or dielectric loss were present, we manually collected it using softwares for x-y data collection, WebPlotDigitizer [44] and our original StarryDigitizer [45] integrated in the Starrydata2 system. Details of the Starrydata2 system are available in prior studies [28]. Approximately 50–60% of the papers on our list contained relevant data for collection.

The procedure of data curation is illustrated in Figure 1. Each paper is assigned a unique identifier referred to as an SID (Starrydata ID). We show an example figure extracted from ref [42] assigned an SID of 104. Figures within each paper are also given unique Figure IDs. Each material sample reported in the paper is assigned a unique Sample ID, which links to the material composition, temperature dependence of dielectric permittivity, temperature dependence of dielectric loss, and measurement frequency. If data for dielectric loss or measurement frequency are not provided, these fields are left blank. When only a room temperature value of permittivity or loss is available, that value is recorded along with the measurement temperature of 298 K.

Several specific guidelines were followed to ensure consistency in collecting data on dielectric materials. Dielectric permittivity was recorded as a dimensionless quantity ε', and dielectric loss was collected as either the dimensionless ε'' or tanδ. After collecting the values of ε' and ε'', ε'' was converted to tanδ formulated as $\varepsilon''/\varepsilon'$. We used tan$\delta$ as the standardized measure of dielectric loss. Since dielectric permittivity and loss are frequency-dependent, graphs in the literature often display curves for multiple frequencies, as exemplified in Figure 1 [42]. In such cases, we selected a single representative frequency curve for data collection. In a large-scale data collection project like this, adding even a single type of metadata can require significant time and resources, and therefore we had to carefully prioritize which data to collect. We specifically prioritised the highest frequency data below 1 MHz based on the following assumptions: measurements at higher frequencies are less affected by extrinsic factors such as leakage currents, providing values closer to the intrinsic properties of the materials. The frequency above 1 MHz might suffer from the relaxation of the dipole contribution ε_{dipole} and underestimate the dielectric permittivity of ferroelectrics. The frequency dependence of the intrinsic dielectric permittivity is not very significant except near the peak temperature in relaxor ferroelectrics. Even in these cases, the degree of variation due to frequency dependence is relatively small compared to the changes caused by compositional variations.

During data curation, a flexible approach was adopted to interpret graphs with missing or unclear captions and labels, inferring the authors' intentions to ensure accurate data collection. Multiple reviewers verified the data, and missing or unclear labels were supplemented when possible. A particularly common situation was the lack of frequency labels on graphs displaying multiple curves for different frequencies. In these cases, we applied the following guidelines to identify each curve: for dielectric permittivity, the curve at a higher frequency consistently appears lower than those at lower frequencies. For dielectric loss, the curve at a higher frequency shows a peak at a higher temperature.

After collecting the data, we examined the distribution of each property to ensure data accuracy. Since data in the Starrydata2 web system are manually curated, occasional human errors are inevitable. To mitigate this, we reviewed the distribution of dielectric permittivity and dielectric loss values, flagging any outliers for further inspection. When an abnormal value was detected, we returned to the original paper to verify whether an error had occurred during data collection. We also examined the material composition data, and if any composition appeared anomalous based on valency considerations, we conducted a similar verification process by referring to the source. Through these measures, we aimed to minimize human error as much as possible throughout the data curation process.

2.2. Data categorization

In this section, we describe the datasets used in the subsequent data analysis. The data collected through the Starrydata2 web system are referred to as the 'Starrydata' dataset. In addition, we incorporated an existing dataset of microwave dielectric materials, which we designate as the 'Sebastian' dataset [21,22]. We used approximately 1,300 samples with an interpretable composition data from Sebastian dataset. Furthermore, to compare material composition distributions, we extracted an oxide dataset from the ICSD [17], which we label as 'ICSD oxide' dataset.

The dielectric permittivity data are categorized based on measurement conditions. Data with temperature-dependent measurements are labelled as 'temperature dependence'. The remaining data are those measured at a single temperature, typically room temperature. Of these, the data measured at frequencies above 100 MHz are categorized as 'High frequency', while the rest are designated as 'other RT data'. The distribution of each labelled dataset is illustrated in Figure 2 and will be discussed later.

2.3. Material descriptors

To perform data analysis on materials data, information about material composition and crystal structure

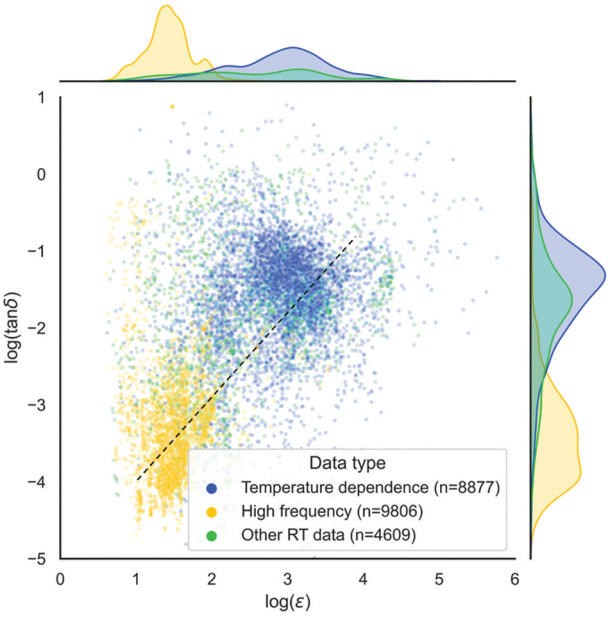


Figure 2. Dielectric loss tanδ versus dielectric permittivity ε at room temperature (297 K) for all available data. Each sample is represented by a single point, with color indicating the data category. The dashed line is just a guide for eye. The marginal distributions along the top and right axes illustrates the distribution of dielectric permittivity and dielectric loss, respectively.

are converted into numerical values, referred to as material descriptors. In this study, as crystal structure data were not available, we created composition-based descriptors implemented in XenonPy module [46]. This descriptor calculates various statistical values such as mean, variance, maximum, and minimum of physical properties for each element in the composition, converting each chemical formula into a set of 290 numerical values. The procedures for handling material compositions are also firmly supported by the pymatgen library [47].

As XenonPy descriptors are challenging to interpret intuitively, analyses using these descriptors mostly function as a black box. In an attempt to enhance interpretability, we introduced an additional set of composition descriptors, which we refer to as the 'ABO_3 descriptor', specifically designed for ABO_3-type compounds. First, we extracted compositions that could potentially form the ABO_3 structure, specifically selecting samples with an anion-to-cation ratio close to 2/3 (= 0.666). This process will be discussed later with Figure 3 and described in detail in Supplemental Material (S1). Then, cations were arranged in order of ionic radius, assigning the larger half to the A-site and the remainder to the B-site. Compositions that could not be clearly divided into A- and B-sites were excluded from the ABO_3-type dataset. Ionic radii were determined by first assigning stable valence states under standard conditions for each element and then using an extended Shannon ionic radii table, where missing values were

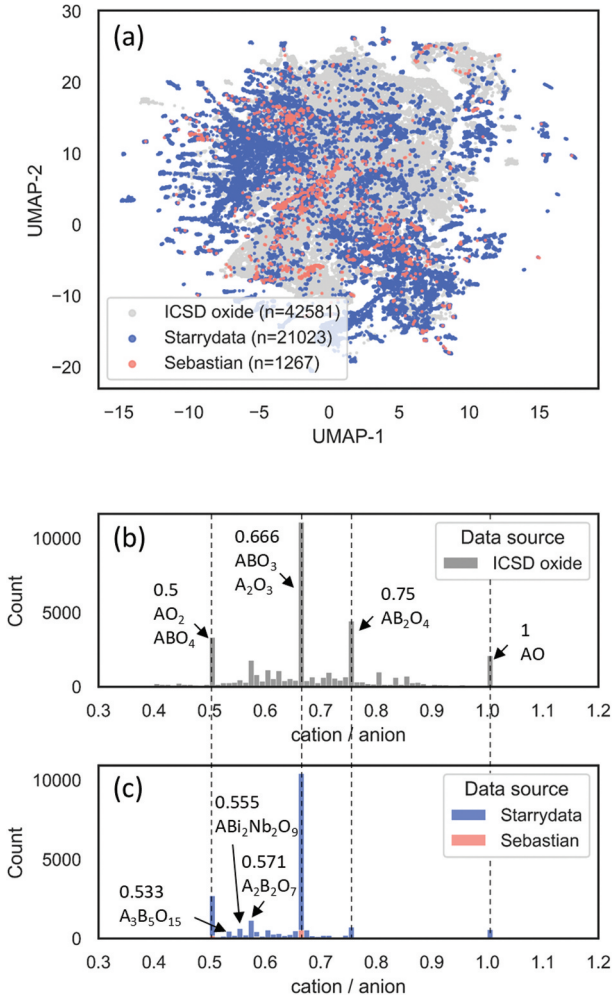


Figure 3. (a) Compositional mapping of oxide compounds recorded in ICSD oxide, Starrydata, and Sebastian dataset. Each sample is represented as a single dot. The vertical and horizontal axes are obtained by dimensionality reduction of the material composition. (b), (c) distribution of material composition against cation-to-anion molar ratio shown as histograms for (b) ICSD oxide and (c) Starrydata and Sebastian dataset.

supplemented through machine learning [48,49]. After assigning A- and B-sites, we calculated several features for A- and B-sites, including the ionic radii and electronegativity of each site, as well as the tolerance factor, yielding a total of 44 descriptors. This descriptor set is slightly modified from the reported set of descriptors in previous research [41,50]. Details and a complete list of descriptors are provided in Supplemental Material (Table S1).

2.4. Machine learning and dimensionality reduction

We constructed regression models to predict the room-temperature dielectric permittivity, dielectric loss, and the peak temperature of dielectric permittivity. The models were trained on either the full dataset or the ABO_3 dataset, with features based on either XenonPy descriptors or the ABO_3 descriptors. We

used random forest regression model [51] as implemented in scikit-learn [52], and optimized the hyperparameters to improve model performance, setting the number of trees to 120 and maximum depth to 15. Model performance was evaluated using a holdout method with an 85:15 train-test split. To identify the important features, recursive feature elimination [53] was used to reduce the number of descriptors.

Uniform manifold approximation and projection (UMAP) [54] was used for dimensionality reduction. Descriptors were alternately used from either the XenonPy set or the ABO_3 descriptors, depending on the analysis focus. For ABO_3-type compounds, we performed clustering based on Euclidian distance for ABO_3 descriptors using the k-means++ [55] algorithm implemented in scikit-learn [52]. The number of clusters was set to 30, and the results were visually inspected for further analysis. Various kinds of data visualizations, including outlier analysis, trend analysis, verification of machine learning results, and mapping through dimensionality reduction, was performed using Matplotlib [56] and Seaborn [57].

3. Results & discussions

We begin by describing the distribution of the collected data across categories. Figure 2 illustrates the relationship between dielectric permittivity and dielectric loss, using values at room-temperature (298 K) for each sample, with one data point representing each sample. To better visualize the data distribution, a common logarithm transformation has been applied to both dielectric permittivity and dielectric loss. Without this transformation, the data distribution is highly skewed, making analysis more challenging.

A positive correlation is observed between dielectric permittivity and dielectric loss. The data labelled as 'Temperature dependence' tend to exhibit high dielectric permittivity values mainly above 100, while 'High frequency' data are more commonly associated with lower permittivity values typically in the range of 10–100. Samples with temperature-dependent measurements generally show high permittivity, suggesting that many of these data points correspond to ferroelectric or antiferroelectric materials. These materials often display a permittivity peak near the Curie temperature, making temperature-dependent measurements necessary for capturing this characteristic. On the other hand, data measured at high frequencies typically have lower permittivity. These samples are likely to be paraelectric materials, which were developed for applications in microwave resonators, where high permittivity is less important than low loss at higher frequencies.

Figure 3 analyses the compositional distribution of the data. To compare the compositional distributions and coverage of each dataset, we aimed to visualize them. As compositional data span complex, multi-dimensional spaces, direct comparisons are challenging and require careful interpretation. Here, we applied a dimensionality reduction technique as one approach to project the compositional space of each database into a more interpretable format. Figure 3(a) compares the material composition distributions of the ICSD oxide dataset, the Starrydata dataset, and the previously reported Sebastian dielectric dataset [21,22]. The vertical and horizontal axes represent the XenonPy descriptors of each sample, reduced to two dimensions using UMAP, providing a rough overview of the composition distribution. The distribution of the Starrydata samples (shown in blue) covers a wide range of compositions, comparable to that of the ICSD oxide samples (shown in gray). Starrydata also has a broader range of compositions compared to Sebastian dataset, highlighting its diversity. Notably, the composition distribution of around 20,000 samples in Starrydata does not correspond one-to-one with the distribution of approximately 45,000 samples in ICSD oxide. This difference suggests a partial discrepancy between materials with crystal structure data (in ICSD oxide) and those with dielectric property data (in Starrydata). This observation aligns with our experience reviewing the literature, where studies presenting dielectric permittivity data do not necessarily include crystal structure analysis.

The Starrydata dataset, unlike ICSD oxide, contains compositional data and lacks structural information, making it challenging to categorize samples by crystal structure. To gain insights from a different perspective, we focused on the cation-to-anion ratio of the materials. Figure 3(b) and (c) display histograms of data distributions by cation-to-anion ratio for ICSD oxide (Figure 3(b)), and Starrydata and Seabstian dataset (Figure 3(c)). The data show clustering around several characteristic cation-to-anion ratios including AO_2 (0.5), or A_2O_3 or ABO_3 (0.667), AB_2O_4 (0.75), and AO (1), each corresponding to distinct types of oxide compounds. While the overall distribution is similar between ICSD oxide and Starrydata, there is a higher concentration of samples around a ratio of 0.667 in Starrydata. This suggests an abundance of ABO_3-type compounds including perovskites in Starrydata. Additionally, other characteristic concentration points appear in Starrydata such as 0.533, which corresponds to tetragonal tungsten bronze $A_3B_5O_{15}$, 0.555 to Aurivillius-type $ABi_2Nb_2O_9$, and 0.571 to pyrochlore $A_2B_2O_7$. These findings indicate that the cation-to-anion ratio is effective for roughly stratifying groups of dielectric materials. Accordingly, we decided to perform data analysis not only on the full dataset but also specifically on the subset of ABO_3-type compounds.

We constructed machine learning models for both the full dataset and the ABO_3 dataset. As shown in Figure 4(a)(b), both dielectric permittivity ε and dielectric loss tanδ exhibit distributions without skewness when a common logarithm transformation is applied. Without the logarithmic transformation, the data show significant skew, making it challenging to train the machine learning models effectively. In contrast, the peak temperature of the dielectric permittivity T_{peak} shown in Figure 4(c) did not exhibit significant skewness, so no logarithmic transformation was applied to this variable. After excluding outliers, each dataset was split into training and test sets. As described in the Methods section, random forest regression models were trained on the training set and scores were then calculated based on the predictive performance of the models on the test data. Table 1 shows the coefficient of determination (R^2) and mean absolute error (MAE) scores of the machine learning models, with MAE scores indicated in parentheses. The R^2 score reflects the proportion of variance in the target variable with a value of 1 indicating a perfect fit. MAE represents the average magnitude of errors in predictions, indicating how far predictions are from actual values on average. Our models achieved favourable scores across datasets and properties. It is important to note that the MAE scores for both permittivity and loss are calculated on the common logarithm-transformed values. For example, an MAE score of 0.16 for dielectric permittivity indicates an average deviation of approximately 1.45 times the actual permittivity value. In constructing the models, we also considered other models including linear regressions and support vector regression. However, these models did not yield good scores compared to decision-tree-based models (Supplemental Materials S2). The fact that linear regressions did not yield good scores suggests that the dependence of dielectric permittivity on descriptors is nonlinear and intricately intertwined.

Table 1. R^2 scores (upper) and MAE scores (lower, in parentheses) of machine-learning models trained on each property and dataset. Columns represent properties, while rows denote datasets and descriptor sets. MAE scores for the bottom row are not available because RFE was conducted based on R^2 scores.

dataset	Permittivity ε	Loss tanδ	T_{peak}
All data XenonPy descriptors	0.929 (0.13)	0.820 (0.30)	0.832 (40.3)
ABO_3 dataset XenonPy descriptors	0.879 (0.16)	0.748 (0.29)	0.887 (29.3)
ABO_3 dataset ABO_3 desciptors	0.866 (0.16)	0.749 (0.29)	0.874 (32.1)
ABO_3 dataset reduced desciptors	0.854	0.707	0.861

Table 1 provides prediction scores on the different datasets and descriptor sets. The R^2 scores for dielectric loss were relatively low compared to the dielectric permittivity and peak temperature. There could be various potential reasons including: (i) insufficient representational power of the compositional descriptors, (ii) the rapid variation of dielectric loss with respect to composition, (iii) differences in dataset sizes, and (iv) high levels of noise in the data. In this case, the factors (iii) and (iv) are the most plausible contributors. The numbers of data points used in the model evaluations were 17,587 (9,275) for dielectric permittivity, 9,488 (5,604) for dielectric loss, and 6,491 (5,090) for peak temperature the entire dataset (in the ABO_3 dataset). Dielectric loss has fewer data points than that of dielectric permittivity because literatures

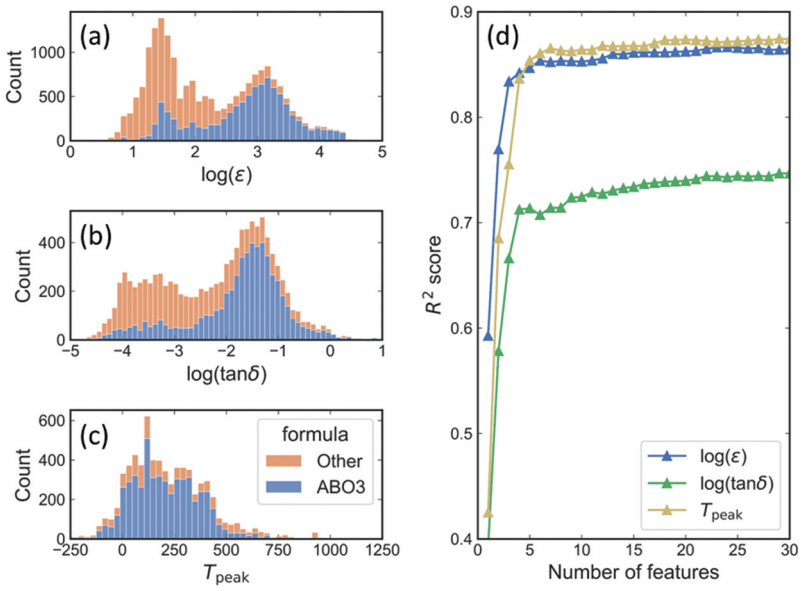


Figure 4. (a)–(c) distribution of target property for machine-learnings. (a) dielectric permittivity, (b) dielectric loss and (c) peak temperature in dielectric permittivity. (d) Machine-learning R^2 scores against number of descriptors obtained through recursive feature elimination.

often provide only dielectric permittivity without the value of the loss. Additionally, due to the nature of experimental datasets collected in this study, dielectric loss is more susceptible to variations stemming from sample quality or measurement precision compared to dielectric permittivity. Factors such as ceramic sintering quality, insulation properties, or calibration accuracy of measurement instruments can strongly influence tanδ values. This could lead to deviations from the intrinsic material properties, and thus reducing the prediction accuracy of the machine learning models to some extent.

The first and second rows of Table 1 compare the scores when the dataset is changed from the full dataset to the ABO_3 dataset. In this case, the R^2 and MAE scores for permittivity and loss decrease, while the scores for peak temperature improve. This suggests that including the full dataset enhances the score for permittivity and loss, whereas focusing specifically on the ABO_3 dataset improves the score for peak temperature. The improvement in peak temperature scores when using the ABO_3 dataset indicates that compositional descriptors alone may not sufficiently distinguish ABO_3-type compounds out of the full dataset. This result implies that while composition descriptors are quite useful, they are not universally applicable, and stratifying the data into relevant groups beforehand can enhance model performance [13].

For the ABO_3 dataset, we compared two sets of descriptors. The second and third rows of Table 1 show the scores based on XenonPy and ABO_3 descriptor sets, respectively. There are several compositional descriptors available to date, including matminer [58], Magpie [59], JARVIS-Tools [60], and XenonPy. Among these, XenonPy offers a large number of descriptors derived from chemical properties with ease of use. While having more descriptors does not necessarily mean better comprehensiveness, we selected XenonPy as sufficiently extensive set of descriptors. While XenonPy has 290 descriptors, we compared its performance with a smaller descriptor set of 44 descriptors specifically designed for ABO_3 compounds. The score changes between XenonPy and the ABO_3 descriptors are minimal for all properties. This indicates that the representational power of the descriptor set, meaning their ability to capture key information needed to predict target variables accurately and represent underlying data patterns effectively, has not diminished. Although the ABO_3 descriptors are limited to 44 variables for interpretability, they retain similar representational power compared to the 290 XenonPy descriptors.

We applied recursive feature elimination (RFE) to reduce the number of descriptors while retaining their representational power [53]. RFE was performed on the 44 descriptors of ABO_3 descriptor set. In RFE, descriptors with the lowest importance are sequentially removed, and the model is reconstructed with the remaining descriptors at each step. Figure 4(d) shows how R^2 scores change with the number of descriptors, where descriptors were removed one at a time, starting from the right side. The scores achieved with the reduced set of six descriptors are shown in the fourth row of Table 1. Notably, for both dielectric permittivity and peak temperature, there is minimal decline in performance, indicating that the main variations in these target properties can be captured effectively with as few as six descriptors.

Table 2 lists the top six descriptors identified by RFE for each target parameter: dielectric permittivity, dielectric loss, and the peak temperature of dielectric permittivity. Based on the scores, the sets of the six descriptors demonstrate comparable representational power to the full set of 290 XenonPy descriptors or the 44 ABO_3 descriptors. These six descriptors are likely strongly correlated with the target property. For example, tolerance factor and electronegativity frequently appear among the selected descriptors, both of which have been previously discussed in the literature as influential factors in dielectric properties [61–65]. The high ranking of these parameters in RFE aligns well with past academic insights. In decision tree-based models, it is difficult to isolate the individual effects of explanatory variables. The relationship between dielectric permittivity and its explanatory variables is non-linear and intertwined, as reflected in the unsuccessful results in linear regression models (Supplemental Materials S2). This makes it difficult to understand the intuitive meanings of the six descriptors. However, our finding illustrates that even with a dataset of this scale, the dielectric permittivity could be explained with six key factors, yielding consistent results with established academic knowledge. The tolerance factor will be revisited later for further visualization and discussion.

Although the predictive model works well and achieve high scores, it functions as a black box, preventing the intuitive understanding of the data and underlying trends. To effectively visualize the compositional trend within the curated data, we applied dimensionality reduction to the ABO_3 dataset. In Figure 3(a), we visualized the overall data distribution for each database, but here we focus specifically on ABO_3-type. Figure 5(a) shows the compositional mapping of the ABO_3 dataset using UMAP dimensionality reduction applied to the ABO_3 descriptors. To distinguish material compositions more effectively, we applied k-means++ clustering to the ABO_3 dataset with the ABO_3 descriptor set, assigning labels to each cluster. The number of clusters was set to 30 in the k-means++ model. Each cluster was reviewed manually and re-labeled into 10 broader categories representing distinct material groups. As a result, ABO_3-

Table 2. Ranked descriptors for each property extracted through recursive feature elimination. In each cell of the table, we indicated whether it is a descriptor for the A-site or B-site. 'Ave' stands for average, and 'std' stands for standard deviation. Further details of each descriptor are provided in supplemental material.

Rank	Permittivity ε	Loss tanδ	T_{peak}
1	A & B-sites tolerance factor	A-site, ave, electron affinity	A-site, ave, electron affinity
2	A & B-sites, ave, electronegativity	B-site, ave, dipole polarizability	B-site, std, nominal charge
3	B-site, ave, dipole polarizability	A-site, ave, nuclear effective charge	A & B-sites tolerance factor
4	A-site, ave, covalent radius	B-site, std, nominal charge	B-site, std, atomic weight
5	A & B-sites, difference electronegativity	A & B-site tolerance factor	A & B-sites, ave, electron affinity
6	B-site, std, nominal charge	B-site, std, electron affinity	A-site, std, covalent radius

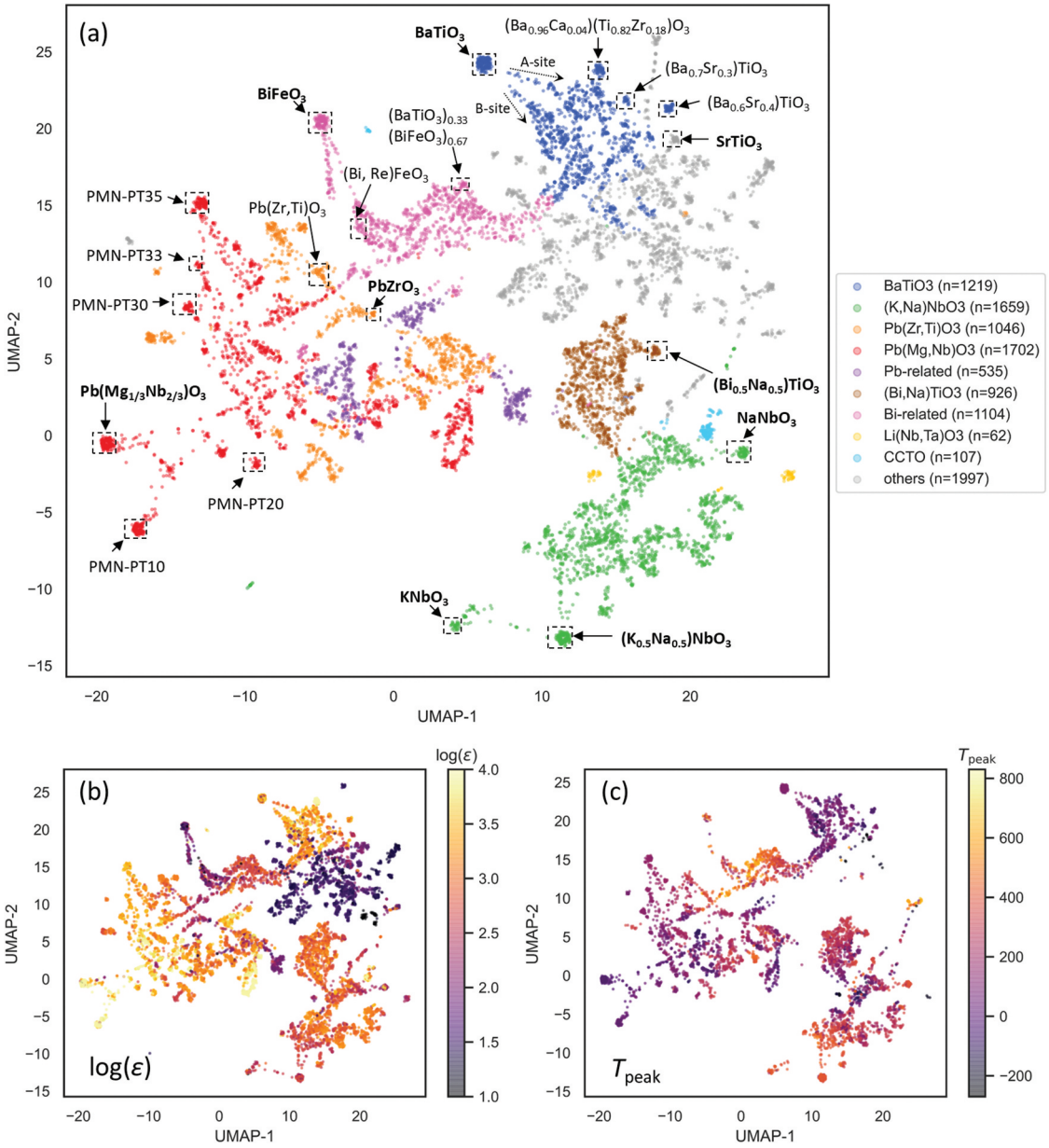


Figure 5. (a) Compositional mapping of ABO_3 materials. Each sample is represented as a single dot with color showing compositional clusters. 9 prominent clusters of dielectric materials are shown, while those not classified into these clusters are shown as others. (b) Dielectric permittivity at room temperature (297 K) and (c) peak temperature in dielectric permittivity plotted against the same axes as in (a).

type materials were divided into 9 primary clusters, with miscellaneous samples grouped as 'others'.

In Figure 5(a), which shows the ABO_3-type data mapped with UMAP, the 9 clusters are not always distinctly separated. For example, the clusters for Pb(Zr, Ti)O_3 and CCTO contain some isolated islands, which is a characteristic of UMAP. K-means++ clustering groups data points based on Euclidean distances within the high-dimensional descriptor space, assigning nearby points to the same cluster. UMAP, on the other hand, first constructs a graph connecting nearby points in the original descriptor space and then projects this graph into two dimensions while preserving local structure. Because UMAP does not retain precise pairwise distances, isolated islands can appear when projecting k-means++ clusters onto the UMAP-reduced space. Nevertheless, UMAP is advantageous for in preserving local structures, and is often more effective than methods like principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE) in visualizing broad ranges of material composition data when interpreting compositional distributions.

Based on Figure 5, we examine the characteristics and distribution of our ABO_3 dataset. Among the nine clusters, seven represent major ferroelectric families with the perovskite structure: $BaTiO_3$, (K, Na)NbO_3, Pb(Zr, Ti)O_3, Pb(Mg, Nb)O_3, Pb-related, (Bi, Na)TiO_3, and Bi-related. Starting with the $BaTiO_3$ cluster (shown in blue) in the upper right, we see a large island representing pure (or nearly pure) $BaTiO_3$. This blue cluster extends downward to the right, containing compositions where elements are substituted in $BaTiO_3$. Branches extend for substitutions at the A-site and B-site, showing distributions for elemental substitutions such as Ca, Sr, and Zr. The fully Sr-substituted $SrTiO_3$ composition is categorized in the 'others' cluster adjacent to the $BaTiO_3$ cluster. The seven ferroelectric families are positioned close to each other, with boundary regions often containing solid solution formed by mixing the adjacent clusters. For instance, the Bi-related cluster (in pink) includes an island of $BiFeO_3$, with branches containing elemental substitutions to $BiFeO_3$. Near the boundary with the $BaTiO_3$ cluster, we observe compositions such as $(BaTiO_3)_{0.33}(BiFeO_3)_{0.67}$. This indicates that the dataset not only includes doped compositions within each of the seven primary families but also compositions formed by mixing different ferroelectric families. This observation is consistent with our experience, as solid solutions of ferroelectrics are frequently investigated for tuning dielectric properties.

Among the remaining two clusters, Li(Nb, Ta)O_3 represents piezoelectric materials with the $LiNbO_3$-type structure, and its data appear as a localized island. The CCTO cluster includes compositions with $CaCu_3Ti_4O_{12}$, a material group known for its large dielectric permittivity, though its origin is attributed to electronic effects rather than dielectric polarization. Although neither of these two material groups has a perovskite structure, they are classified as ABO_3-type under the current compositional categorization method. Both appear as isolated islands without branches extending to other material systems, indicating a limited number of reported solid solutions with other material groups.

Figure 5(b) and (c) display maps using the same axes as Figure 5(a), with room-temperature dielectric permittivity and peak temperature of dielectric permittivity shown on color scales, respectively. The property values exhibit clear gradients across the map, which appear to be well captured by the previously mentioned random forest regression models. In Figure 5(b), room-temperature dielectric permittivity values are available for nearly all data points. On the other hand, in Figure 5(c), some data points lack peak temperature values, reflecting the nature of the Starrydata dataset. Peak temperature data may be absent when only a room-temperature value was recorded or when the temperature dependence of dielectric permittivity is monotonically increasing or decreasing. The peak temperature of dielectric permittivity generally corresponds to the ferroelectric transition temperature (Curie temperature), so data points lacking this information are likely paraelectric rather than ferroelectric. Observing the distribution from this perspective, we find that most points in the 'others' category lack peak temperature data, indicating that these materials are primarily paraelectric. This is not surprising, as the number of perovskite materials exhibiting ferroelectricity is indeed limited [64]. Although exceptions exist, such as $CdTiO_3$ [66], which is a ferroelectric outside the main seven families, most of the reported data on perovskite ferroelectrics are confined to these seven families.

In the analyses presented above, we successfully reviewed the compositional distribution and characteristics within the Starrydata dataset curated in this study. This could not be achieved solely by predictive models which function as black boxes. It is crucial to combine the visualization technique. As a final step, we visualize the key descriptor identified through RFE analysis of the machine learning models using the clusters identified through the visualization. We selected the tolerance factor from the important features listed in Table 2. The tolerance factor [61] is a parameter in perovskite materials that reflects the geometric fit between the ions in the crystal structure, formulated as follows:

$$t = \frac{r_A + r_O}{\sqrt{2}(r_B + r_O)} \tag{2}$$

where r_A, r_B, and r_O denote the ionic radii at each site [48,49]. It is a well-known parameter that has long

been recognized as influential in perovskite stability and dielectric properties [62–65,67]. A tolerance factor close to 1 indicates a stable perovskite structure, while values significantly above or below 1 can lead to structural distortions or phase transitions. In this study, the tolerance factor is again recognized through the purely statistical approach. Though the intuitive understanding of the descriptor is not straightforward in general, we attempted to reveals a visible correlation between dielectric properties and the tolerance factor, as discussed below.

Figure 6(a) shows a plot of dielectric permittivity on the vertical axis against the tolerance factor on the horizontal axis, with data points colored according to the clusters in Figure 5(a). In Figure 6(a), clusters with high dielectric permittivity can be broadly categorized into three groups. The first group, centered around a tolerance factor of approximately 1.06, primarily consists of the blue $BaTiO_3$ family. The second group, around a tolerance factor of 1, is composed mainly of families with Pb or Bi on the A-site. The third group, shown in light blue, is the CCTO family, where the high permittivity arises from electronic effects rather than dielectric polarization and can thus be disregarded in the context of dielectric properties.

The lone pairs on Pb and Bi atoms at the A-site are known to significantly influence the dielectric properties of perovskites [26,64,68]. For example, $SrTiO_3$ with a tolerance factor of 1 is nearly cubic and quantum paraelectric, while $PbTiO_3$ with a tolerance factor of 1.019 is ferroelectric with a Curie temperature of 490°C [69]. Chemical modulation of $PbTiO_3$ produces $PbZr_{0.7}Ti_{0.3}O_3$ with a tolerance factor of 1, still maintaining a high Curie temperature of approximately 440°C [68,70]. This difference arises because Pb^{2+} readily shifts off-center due to its lone pair electrons, creating spontaneous polarization. Similarly, Bi^{3+} also has lone pair electrons. Since Pb^{2+} and Bi^{3+} tend to shift off-center regardless of lattice size, they likely induce dielectric polarization independent of tolerance factor. The second group is located around a tolerance factor of 1 mainly due to the ionic radii of Bi^{3+} (1.38 Å) and Pb^{2+} (1.49 Å) and their respective combinations with possible B-site elements.

To better isolate the effects of lattice factors on dielectric permittivity, we excluded other contributions. Figure 6(b) displays the data from Figure 6(a) with all samples containing Pb or Bi at the A-site or belonging to the CCTO cluster removed. By excluding lone-pair elements and electronic effects, we focused on the impact of the crystal lattice. In Figure 6(b), excluding paraelectric materials in the 'others' category, points belonging to ferroelectric families coarsely scale with tolerance factor as indicated by the dashed line. Although the contributions from the descriptors seemed non-linear and intertwined as discussed, this trend can be observed across the blue $BaTiO_3$ family and the green (K, Na)NbO_3 family. Numerous factors affect the dielectric properties of ceramics, making it challenging to clarify a single cause. One possible explanation is that larger tolerance factors stretch the perovskite lattice, expanding the BO_6 octahedra and facilitating second-order Jahn-Teller distortions, leading to enhanced ferroelectricity and large dielectric permittivity. This phenomenon is analogous to the dielectric property changes observed in $BaTiO_3$-$SrTiO_3$ solid solutions. $SrTiO_3$ with a tolerance factor of 1 is a quantum paraelectric, but adding Ba expands the lattice, increasing Ti distortion and enhancing ferroelectricity [23,68]. The trend observed in the large Starrydata dataset may follow a similar analogy. Through combining feature elimination and clustering techniques, we visualized effects of the crystal lattice on dielectric permittivity by eliminating other factors and finding a roughly linear relationship. Since this observation is quite coarse,

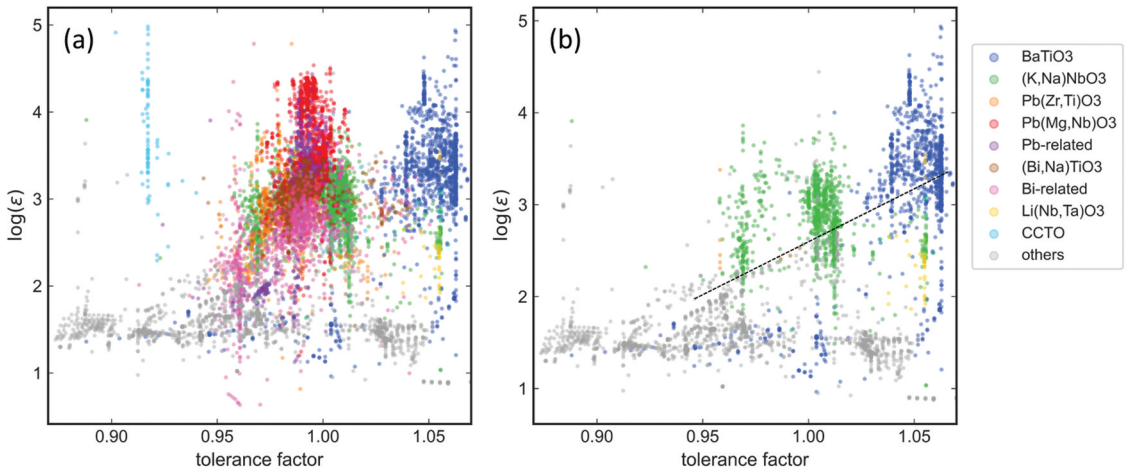


Figure 6. (a) Dielectric permittivity at room temperature (297 K) against tolerance factor in ABO_3 materials. Each sample is represented as a single dot with color indicating the same compositional clusters as in Figure 6. (b) The same plot as (a) but excluding samples containing A-site atoms with lone pairs (Pb and Bi) and the CCTO cluster. A rough correlation between dielectric permittivity and tolerance factor is observed, as indicated by the dotted line.

further detailed data analysis will be necessary to gain a more comprehensive understanding.

The data analysis presented in this study serves as preliminary results to demonstrate the potential of this unprecedented dataset. It is expected that various types of data-driven research can be conducted using this dataset in the future. Even by applying the same machine learning and visualization methods used in this study, changing the descriptors or focusing on different compositional systems could lead to entirely new insights. Additionally, more advanced approaches leveraging data science techniques could be explored as follows: Building a neural network model capable of directly predicting temperature spectra would enable more comprehensive property predictions compared to the predictive models for single properties presented in this research. Incorporating crystal structure data from sources such as the Materials Project or ICSD could help identify structural factors that enhance dielectric properties. Another possibility is employing optimization algorithms, such as genetic algorithms or particle swarm optimization, based on the constructed models. This approach could directly facilitate the discovery of new materials.

When employing such approaches, it will be crucial to pay close attention to the applicable range of the dataset [71,72]. Material compositions and crystal structure data possess high-dimensional degrees of freedom, but within this vast high-dimensional space, the actual regions where relevant data exist are extremely limited. Without careful consideration of both the scope of available data and the applicability range of models, it may prove difficult to derive practical guidelines for materials design. For instance, there is a discrepancy between material compositions covered in Starrydata and those in ICSD. As discussed above, they do not correspond one-to-one. To effectively integrate and utilize these databases together would require specifically tailored transfer learning schemes. Furthermore, when conducting materials exploration with machine learning-based methods, meticulous attention must also be paid to ensuring that models are applied within their valid range. Nevertheless, having access to robust datasets enables such data-driven research possibilities. The Starrydata dataset curated through this study provides an essential foundation for advancing future research aimed at uncovering fundamental principles governing materials and discovering novel materials with improved properties.

4. Conclusions

To advance data-driven materials research, we constructed a comprehensive database of dielectric property for inorganic materials by curating data from scientific literature. Utilizing the Starrydata2 web system, we collected dielectric permittivity and dielectric loss data alongside material compositions. Since dielectric permittivity exhibits characteristic peaks near the Curie temperature, capturing the temperature dependence is critical for assessing material property. Our data collection process involved digitizing numerical data from graphs and figures in original publications, allowing us to capture these key features. We assembled a dataset of over 20,000 material samples extracted from more than 5,000 papers, covering a broad compositional space comparable to all oxide materials recorded in ICSD. The collected data revealed a significant concentration on ABO_3 materials which we further analysed in detail.

We developed machine learning models to predict room-temperature dielectric permittivity, dielectric loss, and the peak temperature of permittivity based on material compositions. We obtained preferrable predictive performance and identified key factors influencing the dielectric properties of ABO_3-type materials through recursive feature elimination (RFE). This analysis yielded six significant descriptors including tolerance factor and electronegativity, which align well with previous academic findings. However, the contributions of these descriptors are nonlinear and entangled, making them difficult to intuitively understand. The machine learning model works as a black box. To address this gap and understand the nature of the collected data, we employed UMAP for compositional mapping and utilized k-means ++ clustering. The ABO_3 dataset was successfully categorized into nine major material groups, seven of which were identified as prominent ferroelectric families. These seven families include atomic substitutions and solid solutions with other ferroelectric families, indicating their frequent investigation in previous research. Combining the identified key descriptors and clusters, we attempted to visualize the impact of the crystal lattice on dielectric property, by isolating the effects from lone pairs and electronic contributions. We observed a positive correlation between the dielectric permittivity and the tolerance factor. A plausible explanation is that larger tolerance factors expand the perovskite lattice, stretch the BO_6 octahedra, and facilitate second-order Jahn-Teller distortions. These structural changes enhance ferroelectricity and increase dielectric permittivity, aligning well with established knowledge of ferroelectricity.

The data analysis presented in this study serves as a preliminary example to showcase the nature of this unprecedented experimental dataset of dielectric materials, and further research would be required to gain a comprehensive understanding of overall trends in dielectric materials. Beyond the analysis and visualization demonstrated in this research, this dataset can support diverse types of data analyses, such as predicting the full temperature spectrum, exploring correlations between structural and dielectric properties, and navigating the discovery of new materials through optimization methods based on the predictive models. The

Starrydata dielectric dataset established in this study offers an important foundation for data-driven materials science, opening new pathways for innovative analyses in materials research.

Acknowledgements

We are grateful to Masaya Kumagai, Atsumi Tanaka and other data curators for their assistance in data curation and management of the Starrydata project. We also thank Daisuke Hirai, Keitaro Fuji and Tomoya Gake for fruitful discussions on machine learning procedures.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was partially supported by the Japan Science and Technology Agency (JST) CREST grant number JPMJCR19J1.

Data availability statement

The link for the dataset used in this study will be shared in Starrydata Dataset page (URL: https://github.com/starrydata/starrydata_datasets) in April 2026.

ORCID

Tomoki Murata http://orcid.org/0000-0001-8698-9953
Tomoya Mato http://orcid.org/0000-0002-0918-6468
Yu Takada http://orcid.org/0009-0002-0709-1817
Sakyo Hirose http://orcid.org/0000-0003-4090-7806
Yukari Katsura http://orcid.org/0000-0002-8905-2995

References

[1] Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. 2013;1(1):011002. doi: 10.1063/1.4812323
[2] Pilania G, Wang C, Jiang X, et al. Accelerating materials property predictions using machine learning. Sci Rep. 2013;3(1):2810. doi: 10.1038/srep02810
[3] Zhuo Y, Mansouri Tehrani A, Brgoch J. Predicting the band gaps of inorganic solids by machine learning. J Phys Chem Lett. 2018;9(7):1668–1673.
[4] Takahashi A, Kumagai Y, Miyamoto J, et al. Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations. Phys Rev Mater. 2020;4(10):103801. doi: 10.1103/PhysRevMaterials.4.103801
[5] Shimano Y, Kutana A, Asahi R. Machine learning and atomistic origin of high dielectric permittivity in oxides. Sci Rep. 2023;13(1):22236. doi : 10.1038/s41598-023-49603-2
[6] Suzuki Y, Taniai T, Saito K, et al. Self-supervised learning of materials concepts from crystal structures via deep neural networks. Mach Learn. 2022;3(4):045034.
[7] Park H, Onwuli A, Butler KT, et al. Mapping inorganic crystal chemical space. Faraday Discuss. 2025;256(0):601–613. doi: 10.1039/D4FD00063C
[8] Sato N, Takahashi A, Kiyohara S, et al. Target material Property-Dependent cluster analysis of inorganic compounds. Adv Intell Syst. 2024;6(12):2400253. doi: 10.1002/aisy.202400253
[9] Dickson RC, Manning TD, Raj ES, et al. Predicting spinel solid solutions using a random atom substitution method. Phys Chem Chem Phys. 2022;24(26):16374–16387. doi: 10.1039/D2CP02180C
[10] Kavanagh SR, Squires AG, Nicolson A, et al. Doped: python toolkit for robust and repeatable charged defect supercell calculations. JOSS. 2024;9(96):6433.
[11] Dimou A, Biswas A, Grünebohm A. Ab Initio-Based study on atomic ordering in (ba, Sr) TiO $_3$. Physica Rapid Res Ltrs. 2024;18(4):2300380. doi: 10.1002/pssr.202300380
[12] Materials Database Group. MDR SuperCon Datasheet. National Institute for Materials Science; 2022. Available from: https://mdr.nims.go.jp/datasets/5d8000f3-a8cd-4ad5-bcdb-2447a5166839
[13] Stanev V, Oses C, Kusne AG, et al. Machine learning modeling of superconducting critical temperature. Npj Comput Mater. 2018;4(1):29.
[14] Konno T, Kurokawa H, Nabeshima F, et al. Deep learning model for finding new superconductors. Phys Rev B. 2021;103(1):014509. doi: 10.1103/PhysRevB.103.014509
[15] Fujii A, Shimizu K, Watanabe S. Efficient exploration of high-tc superconductors by a gradient-based composition design [Internet]. arXiv. [cited 2024 Mar 28]. Available from: http://arxiv.org/abs/2403.13627
[16] Hargreaves CJ, Gaultois MW, Daniels LM, et al. A database of experimentally measured lithium solid electrolyte conductivities evaluated with machine learning. Npj Comput Mater. 2023;9(1):9. doi: 10.1038/s41524-022-00951-z
[17] Zagorac D, Müller H, Ruehl S, et al. Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. J Appl Crystallogr. 2019;52(5):918–925. doi: 10.1107/S160057671900997X
[18] Villars P, Cenzual K, Gladyshevskii R, et al. Pauling File: toward a holistic view. Materials informatics [Internet]. John Wiley & Sons, Ltd; 2019. p. 55–106. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527802265.ch3
[19] Zakutayev A, Wunder N, Schwarting M, et al. An open experimental database for exploring inorganic materials. Sci Data. 2018;5(1):180053. doi: 10.1038/sdata.2018.53
[20] Kabekkodu SN, Dosen A, Blanton TN. PDF-5+: a comprehensive powder diffraction File™ for materials characterization. Powder Diffr. 2024;39(2):47–59. doi: 10.1017/S0885715624000150
[21] APPENDIX 2. In: Sebastian MT, editor. Dielectric materials for wireless communication [internet]. Amsterdam: Elsevier; 2008. p. 541–652. Available from: https://www.sciencedirect.com/science/article/pii/B9780080453309000182
[22] Sebastian MT, Ubic R, Jantunen H. Low-loss dielectric ceramic materials and their properties. Int Mater Rev. 2015;60(7):392–412.
[23] Lemanov VV. Barium titanate-based solid solutions. Ferroelectrics. 2007;354(1):69–76.

[24] Pan M-J, Randall CA. A brief introduction to ceramic capacitors. IEEE Electr Insul Mag. 2010;26(3):44–50. doi: 10.1109/MEI.2010.5482787
[25] Tan X, Ma C, Frederick J, et al. The antiferroelectric ↔ ferroelectric phase transition in Lead-Containing and Lead-Free perovskite ceramics. J Am Ceram Soc. 2011;94(12):4091–4107.
[26] Setter N. What is a ferroelectric–a materials designer perspective. Ferroelectrics. 2016;500(1):164–182. doi: 10.1080/00150193.2016.1232104
[27] Buscaglia V, Randall CA. Size and scaling effects in barium titanate. An overview. J Eur Ceramic Soc. 2020;40(11):3744–3758.
[28] Katsura Y, Kumagai M, Kodani T, et al. Data-driven analysis of electron relaxation times in PbTe-type thermoelectric materials. Sci Technol Adv Mater. 2019;20(1):511–520.
[29] Kumagai M, Ando Y, Tanaka A, et al. Effects of data bias on machine-learning–based material discovery using experimental property data. Sci Technol Adv Mater. 2022;2(1):302–309. doi: 10.1080/27660400.2022.2109447
[30] Tsurumi T, Li J, Hoshina T, et al. Ultrawide range dielectric spectroscopy of $BaTiO_3$-based perovskite dielectrics. Appl Phys Lett. 2007;91(18):182905. doi: 10.1063/1.2804570
[31] Teranishi T, Sogabe T, Hayashi H, et al. Ferroelectric domain contribution to the tunability of $Ba_{0.8}Sr_{0.2}TiO_3$ ceramics. Jpn J Appl Phys. 2013;52(9S1):09KF06.
[32] Gonze X. Perturbation expansion of variational principles at arbitrary order. Phys Rev A. 1995;52(2):1086–1095. doi: 10.1103/PhysRevA.52.1086
[33] Lee M, Youn Y, Yim K, et al. High-throughput ab initio calculations on dielectric constant and band gap of non-oxide dielectrics. Sci Rep. 2018;8(1):14794. doi: 10.1038/s41598-018-33095-6
[34] Umeda Y, Hayashi H, Moriwake H, et al. Prediction of dielectric constants using a combination of first principles calculations and machine learning. Jpn J Appl Phys. 2019;58(SL):SLLC01. doi: 10.7567/1347-4065/ab34d6
[35] Qin J, Liu Z, Ma M, et al. Machine learning approaches for permittivity prediction and rational design of microwave dielectric ceramics. J Materiomics. 2021;7(6):1284–1293.
[36] Noda Y, Otake M, Nakayama M. Descriptors for dielectric constants of perovskite-type oxides by materials informatics with first-principles density functional theory. Sci Technol Adv Mater. 2020;21(1):92–99.
[37] Morita K, Davies DW, Butler KT, et al. Modeling the dielectric constants of crystals using machine learning. J Chem Phys. 2020;153(2):024503. doi: 10.1063/5.0013136
[38] Paul J, Nishimatsu T, Kawazoe Y, et al. Polarization rotation, switching, and electric-field–temperature phase diagrams of ferroelectric $BaTiO_3$: a molecular dynamics study. Phys Rev B. 2009;80(2):024107.
[39] Qi Y, Liu S, Grinberg I, et al. Atomistic description for temperature-driven phase transitions in $BaTiO_3$. Phys Rev B. 2016;94(13):134308. doi: 10.1103/PhysRevB.94.134308
[40] Gigli L, Veit M, Kotiuga M, et al. Thermodynamics and dielectric response of $BaTiO_3$ by data-driven modeling. Npj Comput Mater. 2022;8(1):209. doi: 10.1038/s41524-022-00845-0
[41] He J, Wang C, Li J, et al. Machine learning assisted prediction of dielectric temperature spectrum of ferroelectrics. J Adv Ceram. 2023;12(9):1793–1804.
[42] McNulty JA, Pesquera D, Gardner J, et al. Local structure and order–disorder transitions in "empty" ferroelectric tetragonal tungsten bronzes. Chem Mater. 2020;32(19):8492–8501. doi: 10.1021/acs.chemmater.0c02639
[43] Elsevier Science Publishers. Scopus quick reference guide. CRC Press; 2015. p. 14.
[44] Rohatgi A. Web plot digitizer. Available from: http://arohatgi.info/WebPlotDigitizer
[45] Tomoya M. Starry Digitizer. Available from: https://digitizer.starrydata.org/
[46] Yamada H, Liu C, Wu S, et al. Predicting materials properties with little data using shotgun transfer learning. ACS Cent Sci. 2019;5(10):1717–1730.
[47] Ong SP, Richards WD, Jain A, et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci. 2013;68:314–319. doi: 10.1016/j.commatsci.2012.10.028
[48] Shannon RD. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Crystallogr Sect A. 1976;32(5):751–767.
[49] Baloch AAB, Alqahtani SM, Mumtaz F, et al. Extending Shannon's ionic radii database using machine learning. Phys Rev Mater. 2021;5(4):043804. doi: 10.1103/PhysRevMaterials.5.043804
[50] He J, Su X, Wang C, et al. Machine learning assisted predictions of multi-component phase diagrams and fine boundary information. Acta Materialia. 2022;240:118341. doi: 10.1016/j.actamat.2022.118341
[51] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
[52] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12(null):2825–2830.
[53] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1/3):389–422. doi: 10.1023/A:1012487302797
[54] McInnes L, Healy J, Saul N, et al. UMAP: uniform manifold approximation and projection. JOSS. 2018;3(29):861.
[55] Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms; USA. Society for Industrial and Applied Mathematics; 2007. p. 1027–1035.
[56] Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9(3):90–95. doi: 10.1109/MCSE.2007.55
[57] Waskom M. Seaborn: statistical data visualization. JOSS. 2021;6(60):3021. doi: 10.21105/joss.03021
[58] Ward L, Dunn A, Faghaninia A, et al. Matminer: an open source toolkit for materials data mining. Comput Mater Sci. 2018;152:60–69. doi: 10.1016/j.commatsci.2018.05.018
[59] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials. Npj Comput Mater. 2016;2(1):16028.
[60] Choudhary K, DeCost B, Tavazza F. Machine learning with force-field-inspired descriptors for materials: fast screening and mapping energy landscape. Phys

Rev Mater. 2018;2(8):083801. doi: 10.1103/PhysRevMaterials.2.083801
[61] Goldschmidt VM. Die Gesetze der Krystallochemie. Naturwissenschaften. 1926;14(21):477–485.
[62] Cohen RE. Origin of ferroelectricity in perovskite oxides. Nature. 1992;358(6382):136–138.
[63] Reaney IM, Enrico L Colla C, Setter NSN. Dielectric and structural characteristics of Ba- and Sr-based complex perovskites as a function of tolerance factor. Jpn J Appl Phys. 1994;33(7R):3984. doi: 10.1143/JJAP.33.3984
[64] Benedek NA, Fennie CJ. Why are there so few perovskite ferroelectrics? J Phys Chem C. 2013;117(26):13339–13349.
[65] Shimizu H, Guo H, Reyes-Lillo SE, et al. Lead-free antiferroelectric: $xCaZrO_3$ -$(1-x)NaNbO_3$ system ($0 \leq x \leq 0.10$). Dalton Trans. 2015;44(23):10763–10772. doi: 10.1039/C4DT03919J
[66] Jin Shan Hi Hajime Mori Y, Tezuka K, Itoh M. Ferroelectric phase transition in $CdTiO_3$ single crystal. Ferroelectrics. 2003;284(1):107–112.
[67] Fu D, Itoh M. Ferroelectricity in silver perovskite oxides. In: Lallart M, editor. Ferroelectrics - material aspects [internet]. InTech. 2011 [cited 2024 Oct 27]. Available from: http://www.intechopen.com/books/ferroelectrics-material-aspects/ferroelectricity-in-silver-perovskite-oxides
[68] Jaffe B, Cook WR Jr., Jaffe H. Piezoelectric ceramics. London and (NY): Academic Press; 1971.
[69] Shirane G, Hoshino S. On the phase transition in lead titanate. J Phys Soc Jpn. 1951;6(4):265–270.
[70] Sawaguchi E. Ferroelectricity versus Antiferroelectricity in the solid solutions of $PbZrO_3$ and $PbTiO_3$. J Phys Soc Jpn. 1953;8(5):615–629.
[71] Tetko IV, Sushko I, Pandey AK, et al. Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. J Chem Inf Model. 2008;48(9):1733–1746. doi: 10.1021/ci800151m
[72] Kaneko H, Arakawa M, Funatsu K. Applicability domains and accuracy of prediction of soft sensor models. AlChE J. 2011;57(6):1506–1513. doi: 10.1002/aic.12351