npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Optimizing and extending ion dielectric polarizability database for microwave frequencies using machine learning methods

Jincheng Qin[1,2], Zhifu Liu[1,2]✉, Mingsheng Ma[1,2]✉ and Yongxiang Li[1]

Permittivity at microwave frequencies determines the practical applications of microwave dielectric ceramics. The accuracy and universality of the permittivity prediction by Clausius–Mossotti equation depends on the dielectric polarizability (α_D) database. The most influential α_D database put forward by Shannon is facing three challenges in the 5 G era: (1) Few data, (2) Simplistic relation and (3) Low frequency (kHz–MHz) oriented. Here, we optimized and extended the Shannon's database for microwave frequencies by the four-stage multiple linear regression and support vector machine model. In comparison with the conventional database, the optimized and extended databases achieved higher accuracy and expanded the amount of data from 60 to more than 900. Besides, we analyzed the relationships between α_D and ion characteristics, including ionic radius (IR), atomic number (N), valence state (V) and coordination number (CN). We found that the positive cubic law of "$\alpha_D \sim IR^3$" discussed in Shannon's work was valid for the IR changed by the N, but invalid for the change caused by the CN.

npj Computational Materials (2023)9:132; https://doi.org/10.1038/s41524-023-01093-6

INTRODUCTION

Microwave dielectric ceramics (MWDCs) are the pivotal materials for passive devices, including antennas, filters, resonators, capacitors and so forth, which are the key components for 5 G networks[1–3]. Development of novel MWDCs with good dielectric properties at microwave frequency bands (300 MHz–300 GHz) is an issue of great concern in this field. As a fundamental dielectric property, relative permittivity (ε_r) determines the practical applications of MWDCs: high ε_r is for high energy storage and miniaturization of circuits, while low ε_r is for low-latency signal propagation[4,5]. Prediction of properties guides the structure design and composition optimization of materials, reducing the cost and promoting the efficiency of experimental fabrications. For the prediction of the relative permittivity, Clausius-Mossotti (C-M) equation is the most extensively used method. The concise formula reveals that the decisive factors of the permittivity calculation are the molecular dielectric polarizability and molar volume, as expressed by[6,7]

$$\varepsilon_r = \frac{3V_m + 8\pi\alpha_D^M}{3V_m - 4\pi\alpha_D^M} \quad (1)$$

where V_m is the molar volume in the unit of Å^3, α_D^M is the total dielectric polarizability of a molecule also in Å^3. The molecular dielectric polarizability of a compound can be the summation of the dielectric polarizabilities of individual ions according to the additivity rule[8]. In this manner, the compound A_2BX_4 is taken for an example:

$$\alpha_D^M(A_2BX_4) = 2\alpha_D(A^{2+}) + \alpha_D(B^{4+}) + 4\alpha_D(X^{2-}) \quad (2)$$

The C-M equation and additivity rule were confirmed to be suitable not only for quantifying the permittivity of stoichiometric and nonstoichiometric ceramics[7,9,10], but also for the amorphous thin films and glasses[11,12]. The accuracy and universality of the two equations depend on the data of ion dielectric polarizabilities (α_D). Databases of α_D were derived by fitting the α_D^M data of different materials in the last century[8,13–18]. The most influential database, containing dielectric polarizability of 61 ions, was put forward by Shannon[8]. The α_D data was calculated by a least squares refinement technique in conjunction with the C-M equation and additivity rule, based on experimental relative permittivities of 129 oxides and 25 fluorides. However, the database is facing three challenges in the 5 G era:

(1) Few data. Only 57 elements and 61 ions (including OH^-) are covered in the Shannon's database. Only a few ions with different valence states are involved. Besides, the influence of coordination number on α_D are neglected.

(2) Simplistic relation. Shannon found the approximate linearity between α_D and the cube of the crystal radius for ions in the identical valence state. However, this law is summarized only from the ions in the states of +1, +2, +3 and +4. As a matter of fact, the variation of ion radius is not only caused by the element type and valence state, but also the coordination number, which was not taken into account in the analysis.

(3) Low frequency oriented. Most of the permittivity data used in Shannon's calculation were measured in the range of kHz ~ MHz. Frequency dispersion is an intrinsic characteristic of permittivity, because the dielectric polarization mechanism changes with the testing frequency[19,20]. Consequently, the data derived at low frequencies may not perform well at high frequencies. In the 5 G era, researchers are more concerned about the dielectric properties in GHz frequency bands.

Fortunately, a great amount of relative permittivity data of new MWDCs tested at microwave frequencies has been accumulated since Shannon tabulated the α_D database last century[21]. These studies provide plenty of data for updating the conventional database. Even though a new α_D database with more amount of data for microwave frequencies can be derived via a similar approach in Shannon's work, it can hardly cover the entire periodic table by fitting the limited data of the reported MWDCs. A complete α_D database is necessary to meet the demands of predicting and screening novel materials. The cubic law found by Shannon may be helpful for extending the

[1]CAS Key Laboratory of Inorganic Functional Materials and Devices, Shanghai Institute of Ceramics, Chinese Academy of Sciences, 201899 Shanghai, China. [2]Center of Materials Sciences and Optoelectronics Engineering, University of Chinese Academy of Sciences, 100049 Beijing, China. ✉email: liuzf@mail.sic.ac.cn; mamingsheng@mail.sic.ac.cn

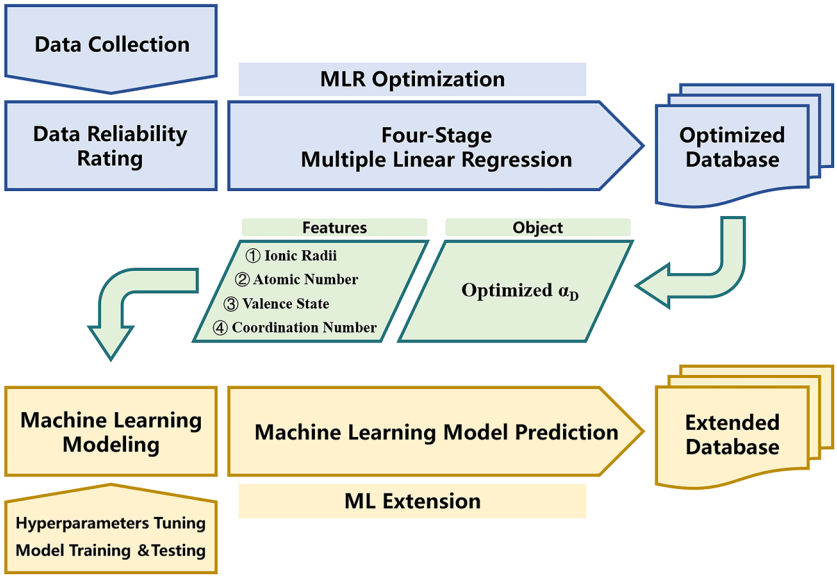

Fig. 1 The schematic workflow to illustrate the procedures of the MLR optimization and ML extension of the ion dielectric polarizability database.

database, but its reliability was not verified by a greater amount of data. The α_D value at GHz bands mainly originates from electronic and ionic polarizations, which are both related to the basic characteristics of ions[8,22]. With the assistance of machine learning methods[23–26], new relations between α_D and basic characteristics of ions could be extracted. Thus, more α_D can be generated, so that the database can be extended via the model. Recently, Baloch et al.[27] enlarged the ionic radii database from 475 types of ions to more than 900 via the Gaussian process regression model. Sufficient data of ion characteristics obtained in Baloch's study lays a foundation for us to exploit new α_D data and relations.

In this work, we optimized the Shannon's α_D database by a four-stage multiple linear regression approach, and then extended the database by the support vector machine model. Our objectives are to build up a complete and accurate α_D database for microwave frequencies, and also to better understand the overall relationships between α_D and ion characteristics.

RESULTS AND DISCUSSION

Workflow and data preparation

This work mainly consists of two parts: multiple linear regression (MLR) optimization and machine learning (ML) extension of the ion dielectric polarizability database. Figure 1 shows the main procedures: data collection, data reliability rating, four-stage multiple linear regression, machine learning modeling based on the optimized database, hyperparameters tuning, model training and testing, and model prediction to extend the database.

The data set for the MLR optimization should contain molecular dielectric polarizability (α_D^M) as the target variable, and the numbers of different ions as feature variables. The α_D^M values were calculated by the C-M equation using the collected data of single-phase microwave dielectric ceramics from published literatures. To calculate α_D^M values, data of permittivity and molar volume (V_m) was needed. All of the permittivity data was measured by the Hakki–Coleman dielectric resonator in the TE_{011} mode at around 10 GHz. The effect of porosity on the experimental permittivities were eliminated by the Penn's revision equation[28]:

$$\varepsilon_r^{exp} = \varepsilon_r\left(1 - \frac{3P(\varepsilon_r - 1)}{2\varepsilon_r + 1}\right) \quad (3)$$

where ε_r^{exp} is the experimental relative permittivity of a polycrystalline ceramic, P is the fractional porosity and ε_r is the porosity-corrected relative permittivity used for the α_D^M calculation. The V_m data of stoichiometric compounds was obtained from the Materials Project (MP) database[29]. However, the MP database has not included nonstoichiometric materials, but those materials should not be neglected, since they also have extensive applications in practice. In a different way, the V_m data of nonstoichiometric compounds was collected from literatures, in which the data obtained from Rietveld refinement results of XRD patterns was reported.

Feature variables describe the composition of ions in a compound. An ion species is determined by the element type, valence state and coordination number. The collected compounds involved 141 species of ions, so the feature variables have 141 dimensions. Finally, we built up three data sets of "α_D^M-ion number" for different purposes: (1) fitting data set for the MLR optimization, including 334 stoichiometric compounds collected from literatures based on the appendix of the book *Microwave Materials and Applications*[21], and their V_m data was from MP database; (2) verification data set for optimizing the constraints of the MLR optimization and verifying the accuracies of the new databases, including 70 new stoichiometric compounds collected from latest literatures (V_m from MP database); (3) evaluation data set for evaluating the generalization of the new databases to nonstoichiometric compounds, including 76 nonstoichiometric compounds collected from latest literatures (V_m from literatures).

Multiple linear regression optimization

Although Shannon's data may be not precise, it is extensively used and acceptable for most microwave dielectric ceramics. For the 334 compounds in the fitting data set, the calculated α_D^M by using Shannon's data show a good agreement with the real values ($R^2 > 0.9$), as shown in Fig. 2a, b. This result suggests that the majority of the Shannon's data is reliable. Besides, we cannot guarantee that all of the experimental data collected from literatures is precise. So, it is unwise to abandon the existing

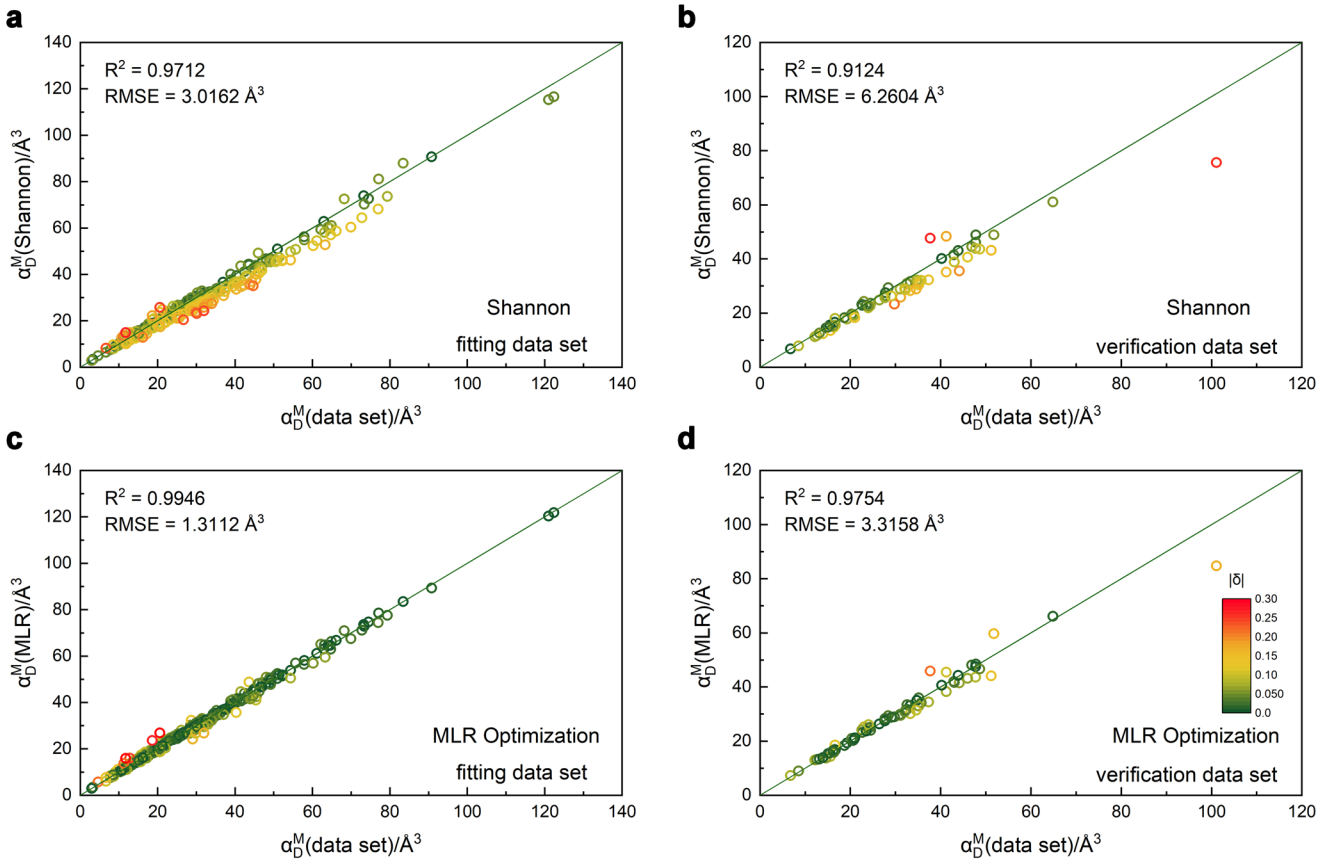

Fig. 2 The performances of the calculated molecular dielectric polarizabilities. a, b By using the Shannon's database and **c, d** by using the MLR optimized database on the fitting and verification data sets.

Shannon's database and recalculate the α_D values of all ions thoroughly. We proposed the approach of α_D database optimization, namely modifying the α_D values within limits using Shannon's data as the initial values by fitting the data set of 334 compounds. The ions with the same element type and valence state, but different coordination numbers, shared an identical Shannon's polarizability as the initial value. However, the degree of error varies with the ion species. For examples, the crystal structures with the octa-coordination Y^{3+} (CN = 8), such as $Y_3Al_5O_{12}$, YPO_4 and $BaY_2Mo_4O_{16}$ have low deviation rates ($|\delta| < 5\%$); while those involve hexa-coordination Zr^{4+} (CN = 6), including $BaZrO_3$, $Ca_2ZrSi_4O_{12}$, $CaZrO_3$, $Ca_3ZrSi_2O_9$, $SrZrO_3$ and $ZrTe_3O_8$ show large negative bias of −7.5% to −21.7% (Supplementary Table 1 and Supplementary Fig. 1). For this reason, we introduced the hierarchical optimization scheme, which means that different initial α_D values were optimized differently according to the reliability of ion. Ions were classified into four reliability levels according to the accuracy of initial α_D values and data amount (Supplementary Table 2). Higher accuracy and larger data amount bring higher level of reliability. Then, the four-stage MLR optimization was exerted. In each stage, the α_D values of ions were fitted by the least square calculation with constraints (optimized in Supplementary Fig. 2), namely the upper and lower limits. Ions at lower reliability level participated more optimization stages, and the constraints would become stricter from Stage 1 to 4 (Supplementary Table 3). Details about the reliability rating and MLR optimization are introduced in METHODS section.

Figure 2 shows improved performances of the calculated α_D^M by using the MLR optimized database comparing with the results using the Shannon's database. Evaluating by the coefficient of determination (R^2) and root means square error (RMSE), the accuracy was not only improved for the fitting data, but also for the verification data not involved in the optimization. The improved performance is ascribed to the refinement and enrichment of the α_D data. The identical α_D of the ions with the same element and valence state were split into different values due to their different coordination numbers. For the Zr^{4+} ion mention above, the Shannon's value of Zr^{4+} is 3.25, while the optimized Zr^{4+} (CN = 6) is elevated to 5.30, and Zr^{4+} (CN = 8) is lessened to 2.09. Consequently, the amount of ion dielectric polarizability data was extended from 67 to 141 (Supplementary Fig. 3).

Machine learning extension

Although the optimized α_D database contains most of the ion species involved in the currently reported MWDCs, this database does not cover the more than 900 ion species currently known[27]. A complete database is necessary to meet the demands of predicting and screening novel materials. Machine learning methods were introduced to further extend the database based on the MLR optimization results. Negative values of α_D are meaningless, but the ML models do not consider the physics, causing negative outputs possibly. To confine the outputs to the positive numbers range, the optimized α_D was transformed into the natural logarithm form ($ln\alpha_D$), serving as the target variable of the model. Seven features were selected as the input variables, including atomic number (N), valence state (V), coordination number (CN), ionic radius (IR), the number of the valence electrons (VE), atomic mass (AM) and the first ionization energy (IE1). It is worth mentioning that the discussion on ionic radius is based on the

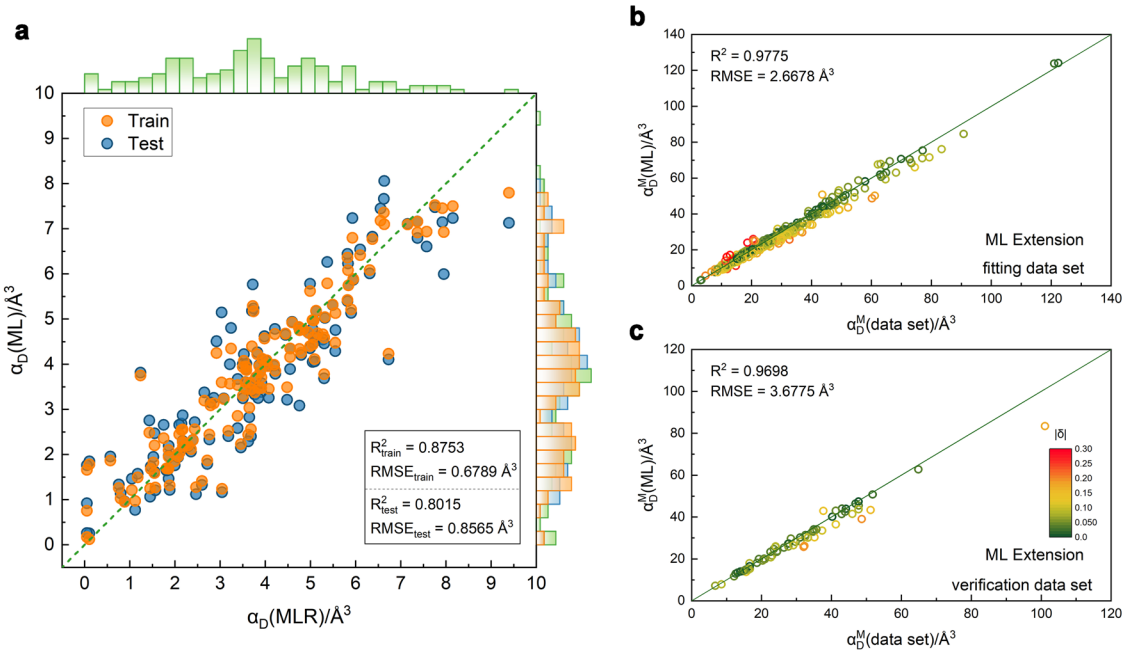

Fig. 3 Plots of predicted α_D values by the ML model. a The comparison of ML extended and MLR optimized α_D values. The performances of the calculated molecular dielectric polarizabilities on **b** the fitting data set and **c** verification data set by using ML extended α_D values.

Baloch's work, instead of the incomplete crystal radius used by Shannon. Dimension reduction was exerted to remove the redundant features according to the Pearson correlation coefficient (PCC) matrix (Supplementary Fig. 4). Among the highly correlated features (PCC > 0.99), only one was retained. Therefore, four decorrelated features, namely N, V, CN and IR were remained. The four features and MLR optimized ion dielectric polarizability of 141 ions form the data set for the subsequent ML modeling.

The prediction models of lna_D were developed by using six ML regression algorithms: linear regression (*lr*), random forest (*rf*), gradient boosting decision tree (*gbdt*), artificial neural network (*ann*), support vector regression with linear kernel (*svr.lin*) and with radial basis function kernel (*svr.rbf*). Hyperparameters of different models were optimized by grid-search method and 10-fold cross-validation (Supplementary Fig. 5 and Supplementary Table 4). The models were trained based on the training data set, which contained 70% of cases from the entire data set by random sampling. The residual 30% was adopted for model testing. To eliminate the random error, the procedures of sampling, training and testing were repeated for 1000 times with different random seeds[7,30,31]. The model prediction results of lna_D were reverted to a_D by the exponential function. The average R^2 and RMSE of the model testing results were served as the criteria for model selection. With the highest R^2 (0.8015) and lowest RMSE (0.8566), the *svr.rbf* model was selected as the best model for the a_D prediction (Supplementary Figs. 6–8). Approximate R^2 (0.87528) and RMSE (0.67895) of the model training confirmed the good generalization ability of the model. By applying the prediction model, the a_D values of 915 ions were generated. Consequently, the ML extended database covers much greater amount and wider range of data than the MLR optimized results (Supplementary Fig. 9), and also has high accuracy (Fig. 3). The Shannon's, optimized and extended α_D databases were integrated and publicly available online (https://qincas.gitee.io/idp-ml/).

The database optimization and extension were based on the stoichiometric compounds. The evaluation data set of 76 non-stoichiometric compounds was used for further evaluating the practicability and generalization of the databases. The compounds were in the structures of $Y_3Al_5O_{12}$, $CaMgSi_2O_6$, $MgZrNb_2O_8$, $Sr_3Ti_2O_7$, $RE_2Zr_3(MoO_4)_9$ (RE = rare earth) and other types (Supplementary Table 5)[32–36]. All the three databases, namely Shannon's, MLR optimized and ML extended databases, performed well, except for the α_D^M calculation of $RE_2Zr_3(MoO_4)_9$ materials[36–38]. Large deviations can be observed for the $RE_2Zr_3(MoO_4)_9$ system by using the Shannon's database ($|\overline{\delta}| = 19.5\%$), but relatively high accuracy was achieved by using MLR optimized ($|\overline{\delta}| = 10.4\%$) and ML extended ($|\overline{\delta}| = 4.7\%$) databases. The discrepancy of the observed and calculated α_D^M by using Shannon's database was ubiquitous in different studies, so measurement errors may not be the key factor[39]. Some studies attributed it to the contribution of relaxation polarizations and polyhedral distortions, but no reasonable evidence was provided[38,40]. In this study, we found that the underestimated a_D values of Mo^{6+} (CN = 4), Zr^{4+} (CN = 6) and O^{2-} (CN = 2) in Shannon's database were ascribed to the discrepancy. It is worth pointing out that the $RE_2Zr_3(MoO_4)_9$ structure materials were not in our fitting data set, but the resulting databases can generalize to them well.

Relationships analysis

The relationships between features and ion dielectric polarizability were analyzed. In accordance with Shannon's finding, we also found the positive correlations between a_D and the cube of the radii of ions with the same valence state (Fig. 4). The slopes of the "$a_D \sim IR^3$" relations disperse in a range, caused by the variations in V and CN. Larger slope is accompanied by the higher V and lower CN. In other words, a_D is sensitive to the change in IR^3 when the ion has high V and low CN. However, for the ions with the same N and V, larger CN will also result in elevating IR^3, but a_D may not positively correlate with IR^3 (Supplementary Fig. 10). When CN is increasing, a_D (Zr^{4+}) and a_D (Y^{3+}) show downtrends with respect to the IR^3, and this category takes up 61% of ions in the database. While, less than 7% show uptrends, including Hg^+ and Cs^+. Both ions with tendencies of "increase, then decrease" (for example, Bi^{3+}) and "decrease, then increase" (for example, K^+) account for about 15%, respectively (Fig. 5 and Supplementary Fig. 11). Ions demonstrating the tendency of "decrease, then increase" are in low valence states (< +3), and

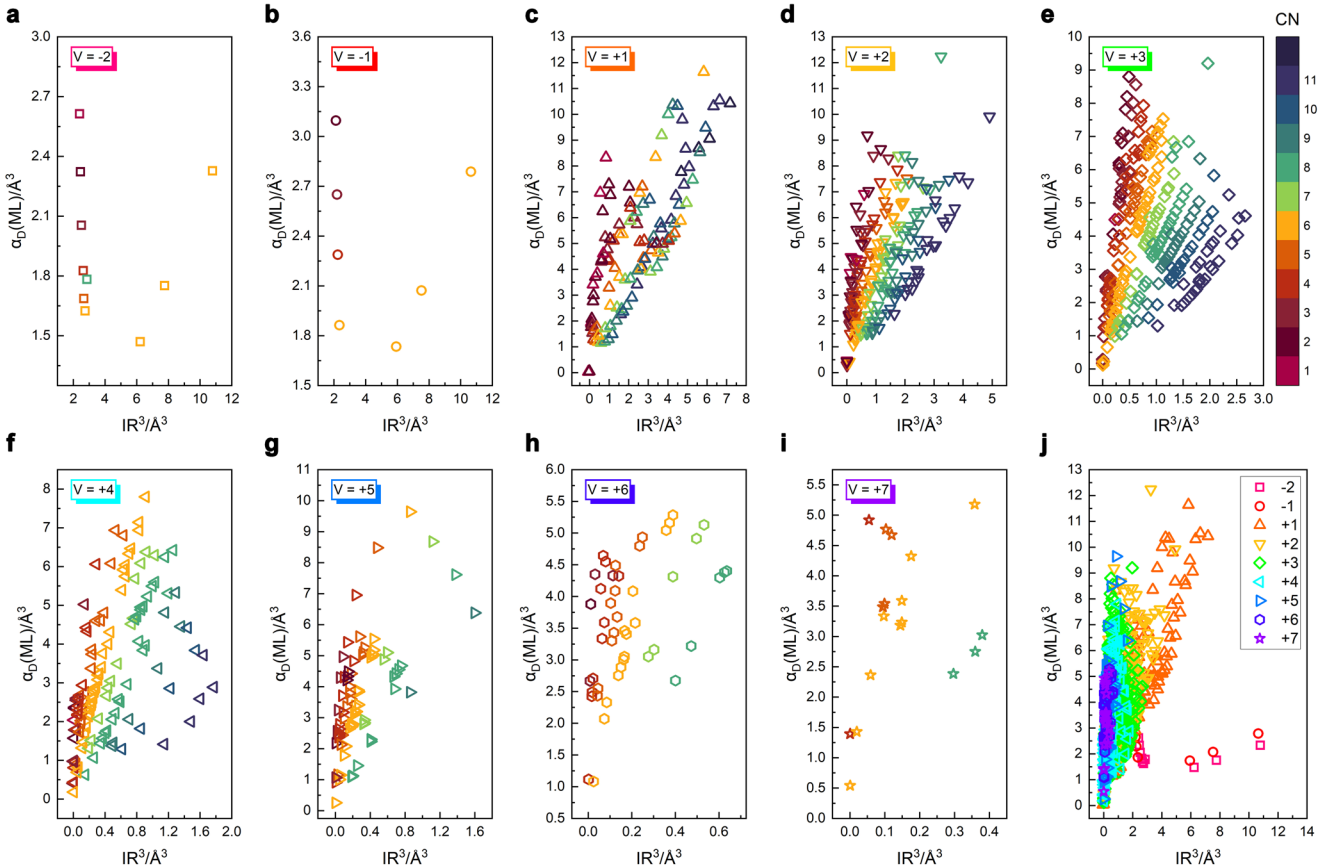

Fig. 4 ML extended dielectric polarizability vs. cube of ionic radius. a Divalent anions (-2), **b** Monovalent anions (-1), **c** Monovalent cations (+1), **d** Divalent cations (+2), **e** Trivalent cations (+3), **f** Quadrivalent cations (+4), **g** Pentavalent cations (+5), **h** Hexavalent cations (+6), **i** Heptavalent cations (+7), **j** Ions with different valence states (-2 to +7).

most ions of alkali metals and alkaline earth metals belong to this category. Ions belonging to "increase" and "increase, then decrease" categories show dispersive distributions. These findings suggest that the positive cubic law may only be valid for ionic radius changed by the atomic number, and the basic characteristics of N, V and CN play a crucial role in α_D prediction.

Further analysis of the relationships between features and α_D value was performed by Shapley additive explanations (SHAP) method. SHAP analysis provides an explainable approach to measure how the variation of a feature influences the model prediction. A SHAP program was built up based on a linear explainer to analyze the contribution of each feature to the *svr.rbf* model output. In Fig. 6, the *x*-axis is the SHAP value indicating the contribution of each feature to the predicted α_D value. The *y*-axis represents the features, and the color corresponds to the magnitude of the feature value. The features were sorted in the descending order from top to bottom based on their contribution. The results suggest that IR contributes most significantly to the model output and larger IR gives a positive contribution to α_D value. Feature CN shows the second biggest impact but a negative contribution to α_D value, which confirms the findings that the α_D of most ions demonstrate downtrends with respect to larger IR^3 caused by increasing CN (Fig. 5). Larger V and N give positive SHAP values, revealing the positive relations between the features and α_D. Those features affect α_D jointly, and we tried to explain the mechanisms in the following discussions.

In GHz bands, dielectric polarization is originated from two principal mechanisms, namely electronic and ionic polarizations. Approximately, the ion dielectric polarizability can be regarded as the sum of the electronic polarizability (α_e) and ionic polarizability (α_i). The classical model of spherical atoms indicates that the electronic polarizability is proportional to the cube of the radius of the electron cloud (r)[41]:

$$\alpha_e = 4\pi\varepsilon_0 r^3 \tag{4}$$

This formula might only fit well with the noble gases. Even though the quantitative analysis of complex compounds by this formula is unreliable, it still reflects the positive correlation between the cube of ionic radius and the electronic polarizability to some extent. The electrons far from the nucleus have low binding energy, which leads the electron cloud to a strong deformation induced by the external electric field. Thus, the contribution to polarizability is also elevated.

An ion balances at the site of the equilibrium between repulsive and Coulombic potentials. When an external field is applied, the adjacent ions offset from the equilibrium separation ($r_0 = r^+ + r^-$) with a displacement of Δr ($\Delta r \ll r_0$)[42]. In a one-dimensional chain of ions, the ionic polarizability can be derived as

$$\alpha_i = \frac{4\pi\varepsilon_0 r_0^3}{n-1} \tag{5}$$

where n is a constant related to the crystal structure and interatomic forces. By extending the formula to the three-dimensional lattice, for example, the cubic crystal structure of NaCl, the α_i can be expressed as

$$\alpha_i = \frac{12\pi\varepsilon_0 r_0^3}{M(n-1)} \tag{6}$$

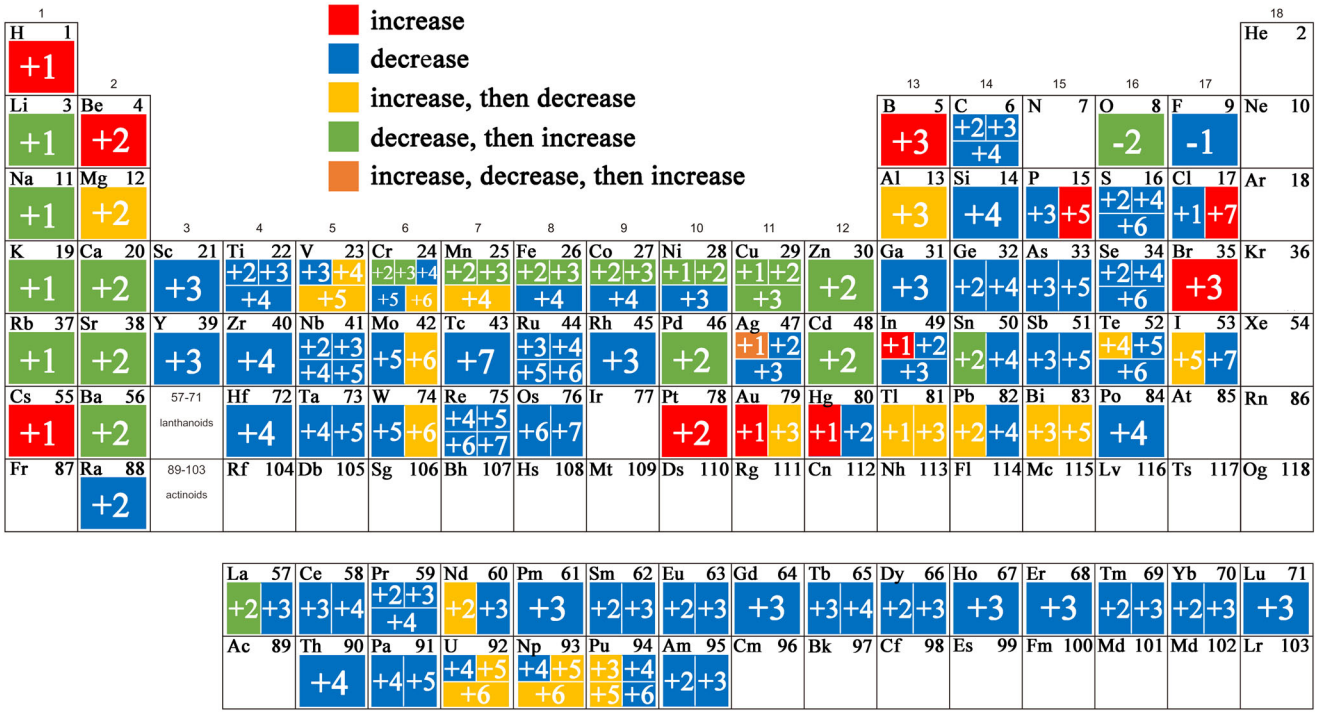

Fig. 5 The tendency of the ML extended dielectric polarizabilities with the increase of ionic radius of ions with the same N and V, but different CN.

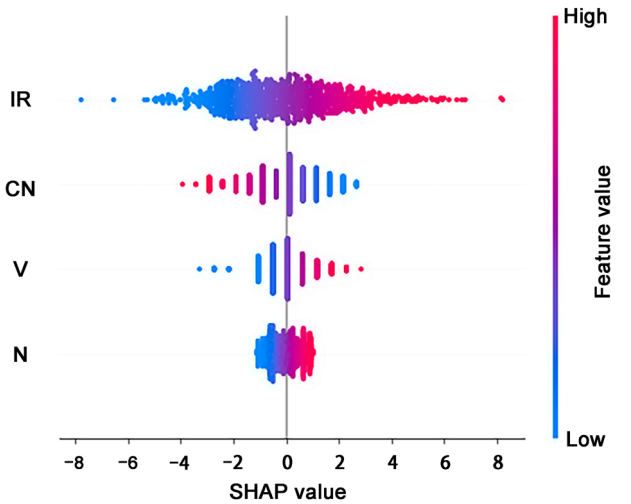

Fig. 6 Relations between feature values and their SHAP values.

where r_0 is the sum of radii of the adjacent cation and anion, so the ionic polarizability is also positively correlated to the cube of the ionic radius. The *M* value, known as the Madelung constant, is a main factor influencing the Coulomb potential energy[43]. The *M* value depends on the distances between ions, locations and charges of ions. The electrostatic interaction between adjacent ions is the strongest, so *M* value generally increases with more adjacent ions, or larger coordination number. Thus, negative contribution of CN to α_D value can be explained: the larger CN, the larger *M* value, and the smaller α_i value. However, at the same time, larger CN will also increase IR, then increase α_e value, so the "$\alpha_D \sim IR^3$" relations of different ions become complex and various (Fig. 5). Moreover, the "$\alpha_D \sim IR^3$" relations of most high-valence ions show downtrends, which might be ascribed to the fact that larger V commonly promotes *M* value. In summary, the *M* and *n* are structure-dependent constant, affected by the radius, charge and coordination number of ions.

For the complex compounds, the calculations of α_e and α_i may deviate far from the ideal models, but the latest machine learning researches on permittivity prediction of thousands of compounds also support our findings. In accordance with features of IR and V in this work, the ionic radii and charges of A and B cations of ABO_3 perovskite crystals exert significant influences on the prediction of the total permittivity[44,45]. The mean atomic mass (MAM) and oxidation state (OS) were also found as the decisive features in the prediction of the electronic permittivity[46,47]. Besides, the standard deviation of the principal quantum number (SDN) and mean neighbor distance variation (MNDV) highly contribute to the ionic permittivity[46]. The OS is the same as the feature V, while the MAM and SDN are closely related to the feature N in this work. The MNDV is determined from Voronoi tessellation, and it reveals the symmetry and bonding in a crystal, which are closely linked with the feature CN in this work.

To highlight the improvements of the new databases, the data amount and accuracy of Shannon's, MLR optimized and ML extended databases are visualized in Fig. 7. Comparing the performances of the α_D^M calculation, highest accuracy was achieved by using the MLR optimized database, which covers most of the ion species in the reported MWDCs. For the exploration of new materials systems, the ML extended database demonstrates its superiority because of the large amount of data, far exceeding the other two. Except for the radioactive and noble gas elements, the ML extended database covers the entire periodic table (Supplementary Fig. 12). Comparing with Shannon's values, the generalization was enhanced by MLR optimization, as $R^2 > 0.985$ and RMSE < 2.5 for all crystal systems. The overall accuracy of the ML extended database lies in between the other two (Supplementary Fig. 13a, b). The accuracies of the α_D^M calculation of eight representative crystal structures in MWDCs,

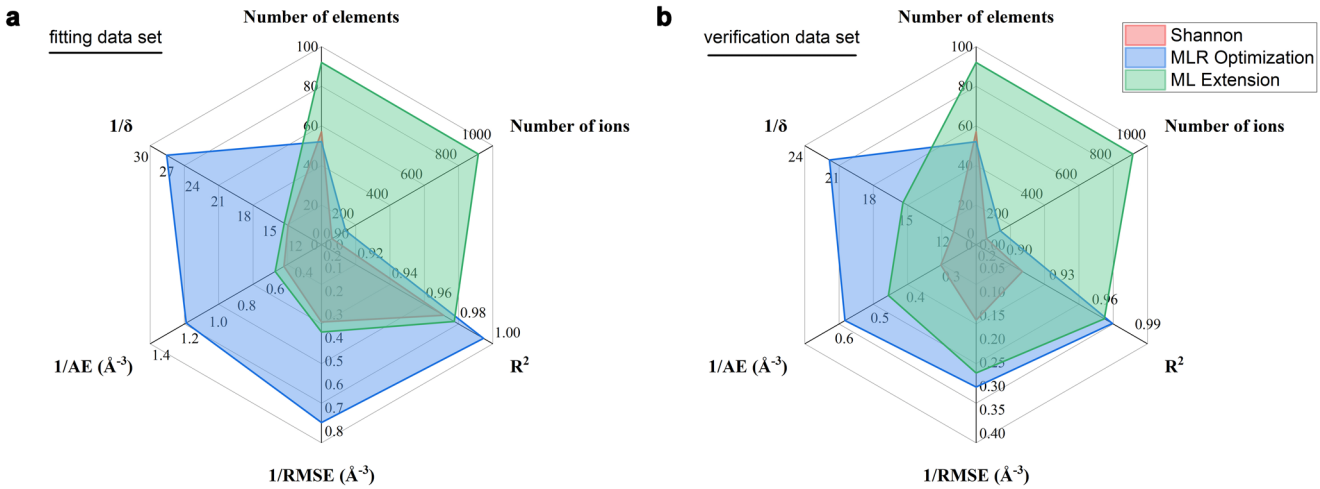

Fig. 7 The comparison of the data amounts and accuracies of databases. a The fitting data set and **b** the verification data set using the Shannon's, MLR optimized and ML extended databases.

namely ABO_3, ABO_4, $AB_2O_4 + A_2BO_4$, AB_2CO_5, $AB_2O_6 + A_2BO_6$, $A_2B_2O_7$, ABC_2O_7 and $A_3BC_2O_9$, were also compared (Supplementary Fig. 13c, d). The average R^2 values by using Shannon's, MLR optimized and ML extended databases are 0.72, 0.91 and 0.76, respectively. No obvious selectivity was observed by using MLR optimized database, while ML extended database performs well except for $AB_2O_6 + A_2BO_6$-type materials.

In summary, we optimized and extended the conventional Shannon's ion dielectric polarizability database for microwave frequencies by the four-stage multiple linear regression and support vector machine model. Both MLR optimized and ML extended databases achieved higher accuracy and possessed greater amount of data in comparison with the conventional database. Relationships between α_D and ion characteristics, including ionic radius, atomic number, valence state and coordination number, conform the dielectric theory and the findings of the latest researches. Besides, we found that the positive cubic law of "$\alpha_D \sim IR^3$" discussed in Shannon's work, was valid for the ionic radius changed by the atomic number, but invalid for the change caused by the coordination number. Although the new α_D databases exceeded the conventional database in accuracy and data amount, improvements are still in progress. On the one hand, to meet the demands of 5 G/6 G technologies, more and more novel materials are developing, and the dielectric testing frequency is rising even to the THz bands. More data tested at different frequencies would be involved in future researches to revolutionize the α_D database. On the other hand, the calculation of molecular dielectric polarizability in data set rely on the C-M equation. Although the equation is useful and commonly used for developing new microwave dielectric ceramics, it is derived based on the assumption of Lorentz local field, leading to errors when calculating the materials in low-symmetry structures. A more precise approach to calculate molecular dielectric polarizability is needed in further studies. Moreover, anisotropy is not taken into account since it is not much concerned in the ceramics field. Prediction on the tensor of polarizability would be more meaningful for the fundamental physics. It is encouraged to collect tensor data of polarizability/permittivity of single-crystal materials by experiments or first-principles calculations, then establish a model solving the problem on anisotropy.

METHODS

Data preparation

The fitting and verification data sets. The target variable, molecular dielectric polarizability (α_D^M), was calculated by the C-M equation based on the permittivity and molar volume. The permittivity was corrected by Penn's revision equation (Eq. (3)) using the experimental permittivity and fractional porosity. The fractional porosity (P) can be calculated by $P = 1 - \rho_r$, where ρ_r is the relative density. The relative densities and permittivities were collected from published literatures. If the relative density (or bulk density) was not provided, we consider the relative density as the default value of 95%. The molar volumes (V_m) were exported from the Materials Project (MP) database (https://materialsproject.org/) by pymatgen (Python Materials Genomics) library[29,48]. By using the pymatgen library, we obtained the V_m data programmatically from the MP database via the Materials Project's REpresentational State Transfer (REST) application programming interface (API)[49]. The feature variables have 141 dimensions, representing the numbers of 141 ion species. The valence state discussed in this work is the nominal valence state of ion, and the method of valence state determination is provided in Supplementary Information. The numbers and coordination numbers of different ions in a compound were counted in the VESTA software (version 3.4.7)[50]. The verification data set includes 70 stoichiometric compounds, whose data was obtained in the same way as above. The compounds in fitting and verification data sets are shown in Supplementary Table 6 and 7, respectively.

The evaluation data set. The evaluation data set includes 76 nonstoichiometric compounds. Both permittivity and V_m data was obtained from literatures, since the MP database has not included the data of nonstoichiometric compounds until now. Some of the data was visualized as charts in the literatures, and we captured them with the help of the Digitizer tool in the Origin Pro 2021 software suite.

Data set for ML extension. The ion dielectric polarizabilities as the target variables were the results of MLR optimization. The data of features N, V, CN and IR were exported from the online database (https://cmd-ml.github.io/)[27]; while features PE, AM and IE1 were from the periodic table website of Royal Society of Chemistry (https://www.rsc.org/periodic-table).

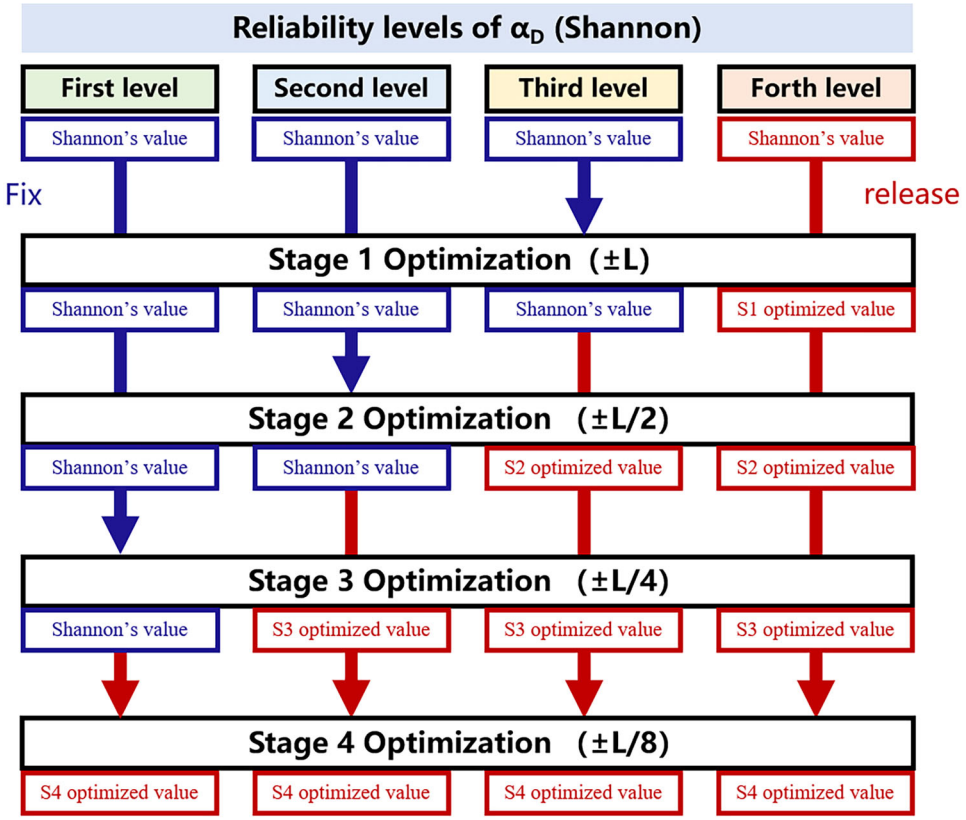

Fig. 8 The flowchart for the four-stage multiple linear regression with constraints.

MLR optimization and ML extension

Programming language and packages. The R Statistical Programming Language (version 3.6.1) with the integrated developer environment (IDE) of the RStudio software suite (version 1.1.456)[51], was used for the MLR optimization and ML extension in this work. The "lsei" (least squares with equality and inequality constraints) function in the package "limSolve" (version 1.5.6) was employed for the four stage MLR optimization with constraints. Packages "randomForest" (version 4.6.14), "gbm" (version 2.1.8), "nnet" (version 7.3.12) and "e1071" (version 1.7.2) were used for random forest (*rf*), gradient boosting decision tree (*gbdt*), artificial neural network (*ann*) and support vector regression (*svr*), respectively. The Python language (version 3.10.9) with IDE of the Spyder software suite (version 5.3.1) was used for the Shapley additive explanations (SHAP) analysis. The library of "sklearn" was used to build up a linear model for SHAP explainer, and the library of "shap" was used for SHAP analysis.

Reliability rating of the initial α_D data. The reliability rating was judged by two criteria, namely the total count (*n*) and the relative frequency in the small error set (*f*). The total count is the number of occurrences of an ion in the whole data set of 334 compounds. Among the whole data set, the α_D^M calculation of 145 compounds have low deviation rates ($|\delta| < 5\%$), forming the small error set. The *f* value of an ion is defined as the ratio between the number of occurrences in the small error set and the total count. Larger *f* value and *n* value reveal the higher accuracy and universality, respectively. The results are shown in Supplementary Table 2.

Four-stage MLR optimization. The MLR optimization in each stage was the least square calculation with constraints, expressed as

$$\min(\|Ax - b\|)^2 \tag{7}$$

subject to

$$u \geq x \geq l$$

$$A = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1m} \\ n_{21} & n_{22} & \cdots & n_{2j} & \cdots & n_{2m} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} & \cdots & n_{im} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{n1} & n_{n2} & \cdots & n_{nj} & \cdots & n_{nm} \end{pmatrix}, x = \begin{pmatrix} \alpha_{D1} \\ \alpha_{D2} \\ \vdots \\ \alpha_{Dj} \\ \vdots \\ \alpha_{Dm} \end{pmatrix},$$

$$b = \begin{pmatrix} \alpha_{D1}^M \\ \alpha_{D2}^M \\ \vdots \\ \alpha_{Di}^M \\ \vdots \\ \alpha_{Dn}^M \end{pmatrix}, l = \begin{pmatrix} \alpha_{D1}^l \\ \alpha_{D2}^l \\ \vdots \\ \alpha_{Dj}^l \\ \vdots \\ \alpha_{Dm}^l \end{pmatrix}, u = \begin{pmatrix} \alpha_{D1}^u \\ \alpha_{D2}^u \\ \vdots \\ \alpha_{Dj}^u \\ \vdots \\ \alpha_{Dm}^u \end{pmatrix}$$

where n_{ij} is the number of the j^{th} ion in the i^{th} compound, α_{Dj} is the dielectric polarizability of the j^{th} ion, α_{Di}^M is the molecular dielectric polarizability of the i^{th} compound, α_{Dj}^l and α_{Dj}^u are the lower and upper limits of the dielectric polarizability of the j^{th} ion, respectively.

In different optimization stages, the ions to be optimized and the constraints were different. Ions at lower reliability level were allowed having larger optimization range, and they also participated in more optimization stages. For examples, in Stage 1 optimization, only the α_D values of ions at the fourth level can be changed, and fixed the others as the initial values; in Stage 2

optimization, the α_D values of the ions at the fourth and third levels can be changed, and fixed the initial values of the second and first level ions. The initial values of ion dielectric polarizabilities were Shannon's values, but the database did not cover all the ions involved in the fitting data set. Therefore, the a_D values of 7 new ions were extended according to the cubic law. The seven new ions include Ag^{+} ($a_D = 2.79$), Hf^{4+} ($a_D = 3.84$), Mn^{4+} ($a_D = 1.86$), Mo^{6+} ($a_D = 3.28$), Sb^{5+} ($a_D = 2.88$), Te^{6+} ($a_D = 4.59$) and W^{6+} ($a_D = 3.2$).

In a certain stage, for the ions not involved, the initial values were fixed, i.e., $a_{Dj} = a_{Dj}^{Shannon} = a_{Dj}^{u} = a_{Dj}^{l}$; While for the ions participated in this stage of optimization, the lower limit was $a_{Dj}^{l} = a_{Dj}^{0} - \Delta a_D$ and the upper limit was $a_{Dj}^{u} = a_{Dj}^{0} + \Delta a_D$, where a_{Dj}^{0} is the optimized value of the previous optimization and the intervals Δa_D are L, L/2, L/4 and L/8 in stage 1 to 4, respectively (Fig. 8). Overlarge intervals may lead to the problem of overfitting. Different values of L were tried in fitting, and the results showed that L = 2 was the optimal option approved by the verification data set of 70 stoichiometric compounds (Supplementary Fig. 2). In summary, the fitting was limited in a certain range based on the Shannon's value or the fitting result of the previous stage fitting. Details of Ions participating in different stages of optimization are shown in Supplementary Table 3.

Model tuning in machine learning extension. The grid-search method was used to generate different hyperparameters combinations. For the models with different hyperparameters combinations, the mean R^2 was calculated via 10-fold cross-validation. The objective function of the optimization is maximizing the average cross-validation R^2 score of 30 trials with different random seeds (Supplementary Fig. 5). The optimized hyperparameters are shown in Supplementary Table 4. Details about hyperparameters optimization can be found in Supplementary Information.

DATA AVAILABILITY

The data used in and resulting from this work is available in the manuscript, Supplementary Information, Supplementary Files and the online repository (https://qinjcsiccas.github.io/idp-ml/).

CODE AVAILABILITY

The codes used in this work are available in Supplementary Files and https://gitee.com/qincas/idp-ml.

Received: 11 October 2022; Accepted: 23 July 2023;
Published online: 28 July 2023

REFERENCES

1. Kamutzki, F., Schneider, S., Barowski, J., Gurlo, A. & Hanaor, D. A. H. Silicate dielectric ceramics for millimetre wave applications. *J. Eur. Ceram. Soc.* **41**, 3879–3894 (2021).
2. Yang, H. et al. The latest process and challenges of microwave dielectric ceramics based on pseudo phase diagrams. *J. Adv. Ceram.* **10**, 885–932 (2021).
3. Sebastian, M. T., Ubic, R. & Jantunen, H. Low-loss dielectric ceramic materials and their properties. *Int. Mater. Rev.* **60**, 392–412 (2015).
4. Raveendran, A., Sebastian, M. T. & Raman, S. Applications of microwave materials: a review. *J. Electron. Mater.* **48**, 2601–2634 (2019).
5. Ohsato, H., Tsunooka, T., Sugiyama, T., Kakimoto, K.-I. & Ogawa, H. Forsterite ceramics for millimeterwave dielectrics. *J. Electroceram.* **17**, 445–450 (2006).
6. Kirkwood, J. G. On the theory of dielectric polarization. *J. Chem. Phys.* **4**, 592–601 (1936).
7. Qin, J., Liu, Z., Ma, M. & Li, Y. Machine learning approaches for permittivity prediction and rational design of microwave dielectric ceramics. *J. Materiomics* **7**, 1284–1293 (2021).
8. Shannon, R. D. Dielectric polarizabilities of ions in oxides and fluorides. *J. Appl. Phys.* **73**, 348–366 (1993).
9. Viegas, J. I. et al. Vibrational spectroscopy and intrinsic dielectric properties of $Sr_2RE_8(SiO_4)_6O_2$ (RE = rare earth) ceramics. *Mater. Res. Bull.* **146**, 111616 (2022).
10. Xiang, H. et al. Microwave dielectric high-entropy ceramic $Li(Gd_{0.2}Ho_{0.2}Er_{0.2}Y_{0.2}Lu_{0.2})GeO_4$ with stable temperature coefficient for low-temperature cofired ceramic technologies. *J. Mater. Sci. Technol.* **93**, 28–32 (2021).
11. Naoi, T. A. & van Dover, R. B. Dielectric properties of amorphous Ta-Ge-O and Ta-Si-O thin films. *J. Appl. Phys.* **123**, 244103 (2018).
12. Lanagan, M. T. et al. Dielectric polarizability of alkali and alkaline-earth modified silicate glasses at microwave frequency. *Appl. Phys. Lett.* **116**, 222902 (2020).
13. Roberts, S. Dielectric constants and polarizabilities of ions in simple crystals and barium titanate. *Phys. Rev.* **76**, 1215–1220 (1949).
14. Lasaga, A. & Cygan, R. T. Electronic and ionic polarizabilities of silicate minerals. *Am. Mineral.* **67**, 328–334 (1982).
15. Wilson, J. N. & Curtis, R. M. Dipole polarizabilities of ions in alkali halide crystals. *J. Phys. Chem.* **74**, 187–196 (1970).
16. Roberts, S. Polarizabilities of ions in perovskite-type crystals. *Phys. Rev.* **81**, 865–868 (1951).
17. Batana, A., Bruno, J. & Munn, R. W. Anion polarizability functions in alkali halide crystals. *Mol. Phys.* **20**, 1029–1034 (1997).
18. Fowler, P. W. & Madden, P. A. In-crystal polarizabilities of alkali and halide ions. *Phys. Rev. B* **29**, 1035–1042 (1984).
19. Teranishi, T. Broadband spectroscopy of dielectrics and oxygen-ion conductors. *J. Ceram. Soc. Jpn.* **125**, 547–551 (2017).
20. Jonscher, A. K. The 'universal' dielectric response. *Nature* **267**, 673–679 (1977).
21. Sebastian, M. T., Jantunen, H. & Ubic, R. *Microwave Materials and Applications* (Wiley, 2017).
22. Dick, B. G. & Overhauser, A. W. Theory of the dielectric constants of alkali halide crystals. *Phys. Rev.* **112**, 90–103 (1958).
23. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
24. Tao, Q., Xu, P., Li, M. & Lu, W. Machine learning for perovskite materials design and discovery. *npj Comput. Mater.* **7**, 23 (2021).
25. Chen, L. et al. Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Comput. Mater.* **6**, 61 (2020).
26. Gong, J., Chu, S., Mehta, R. K. & McGaughey, A. J. H. XGBoost model for electrocaloric temperature change prediction in ceramics. *npj Comput. Mater.* **8**, 140 (2022).
27. Baloch, A. A. B. et al. Extending Shannon's ionic radii database using machine learning. *Phys. Rev. Mater.* **5**, 043804 (2021).
28. Penn, S. J. et al. Effect of porosity and grain size on the microwave dielectric properties of sintered alumina. *J. Am. Ceram. Soc.* **80**, 1885–1888 (1997).
29. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
30. Qin, J., Liu, Z., Ma, M. & Li, Y. Machine learning-assisted materials design and discovery of low-melting-point inorganic oxides for low-temperature cofired ceramic applications. *ACS Sustainable Chem. Eng.* **10**, 1554–1564 (2022).
31. Yuan, R. et al. The search for $BaTiO_3$-based piezoelectrics with large piezoelectric coefficient using machine learning. *IEEE T. Ultrason. Ferr.* **66**, 394–401 (2019).
32. Wu, G. et al. Crystal structure and microwave dielectric properties of Mg^{2+}-Si^{4+} co-modified yttrium aluminum garnet ceramics. *J. Mater. Sci. Mater. Electron.* **33**, 4712–4720 (2022).
33. Yang, M. et al. Microwave dielectric properties of $Ca_{1-x}Ba_xMgSi_2O_6$ ceramics. *Ceram. Int.* **48**, 9407–9412 (2022).
34. Yang, H. et al. Improved microwave dielectric properties of wolframite $MgZrNb_2O_8$ ceramics by $(Ti_{1/2}W_{1/2})^{5+}$ ionic co-substitution. *J. Mater. Sci. Mater. Electron.* **33**, 20846–20854 (2022).
35. He, X., Ma, W., Hong, J., Ba, R. & Li, J. Microwave dielectric properties of $Sr_3Ti_2O_7$ ceramics with composite element doping of Nd & Al. *Mater. Chem. Phys.* **282**, 125961 (2022).
36. Ma, X. et al. Influence of Sn^{4+} substitution for Zr^{4+} in $Nd_2Zr_3(MoO_4)_9$ and the impact on the crystal structure and microwave dielectric properties. *J. Alloys Compd.* **902**, 162526 (2022).
37. Tian, H. et al. Structure characteristics and microwave dielectric properties of $Pr_2(Zr_{1-x}Ti_x)_3(MoO_4)_9$ solid solution ceramic with a stable temperature coefficient. *J. Mater. Sci. Technol.* **116**, 121–129 (2022).
38. Feng, Z. et al. Effects of $(Cr_{1/2}Nb_{1/2})^{4+}$-substitution on the chemical bond characteristics, and microwave dielectric properties of cerium zirconium molybdate ceramics. *Mater. Chem. Phys.* **287**, 126261 (2022).
39. Zheng, J. et al. Structure, infrared reflectivity spectra and microwave dielectric properties of a low-firing microwave dielectric ceramic $Pr_2Zr_3(MoO_4)_9$. *J. Alloys Compd.* **826**, 153893 (2020).
40. Feng, C., Zhou, X., Tao, B., Wu, H. & Huang, S. Crystal structure and enhanced microwave dielectric properties of the $Ce_2[Zr_{1-x}(Al_{1/2}Ta_{1/2})_x]_3(MoO_4)_9$ ceramics at microwave frequency. *J. Adv. Ceram.* **11**, 392–402 (2022).

41. Von Hippel, A. *Dielectric and Waves* (Chapman and Hall, 1995).
42. Ashcroft, N. W. & Mermin, N. D. *Solid State Physics* (Harcourt College Publishers, 1976).
43. Mamode, M. Computation of the Madelung constant for hypercubic crystal structures in any dimension. *J. Math. Chem.* **55**, 734–751 (2016).
44. Kim, E., Kim, J. & Min, K. Prediction of dielectric constants of ABO_3-type perovskites using machine learning and first-principles calculations. *Phys. Chem. Chem. Phys.* **24**, 7050–7059 (2022).
45. 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.* **21**, 92–99 (2020).
46. Takahashi, A., Kumagai, Y., Miyamoto, J., Mochizuki, Y. & Oba, F. Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations. *Phys. Rev. Mater.* **4**, 103801 (2020).
47. Morita, K., Davies, D. W., Butler, K. T. & Walsh, A. Modeling the dielectric constants of crystals using machine learning. *J. Chem. Phys.* **153**, 024503 (2020).
48. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
49. Ong, S. P. et al. The Materials Application Programming Interface (API): a simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Comput. Mater. Sci.* **97**, 209–215 (2015).
50. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **44**, 1272–1276 (2011).
51. Racine, J. S. RStudio: a platform-independent IDE for R and sweave. *J. Appl. Econ.* **27**, 167–172 (2012).

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support from the National Natural Science Foundation of China (61871369). M.M. acknowledges the Youth Innovation Promotion Association of CAS and Shanghai Rising-Star Program (20QA1410200).

AUTHOR CONTRIBUTIONS

J.Q. initiated the concept, designed and conducted the project, wrote the original draft and built up the online repository. Z.L. took the lead in supervising the project and reviewing the draft. M.M. supervised the project, reviewed and edited the draft. Z.L. and Y.L. gave scientific and technical advices throughout the project and helped with the analysis. All authors provided feedback and contributed to the discussion of the project.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-023-01093-6.

Correspondence and requests for materials should be addressed to Zhifu Liu or Mingsheng Ma.

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

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