npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Identification of high-dielectric constant compounds from statistical design

Abhijith Gopakumar[1], Koushik Pal[1] and Chris Wolverton[1]✉

The discovery of high-dielectric materials is crucial to increasing the efficiency of electronic devices and batteries. Here, we report three previously unexplored materials with very high dielectric constants ($69 < \epsilon < 101$) and large band gaps ($2.9 < E_g(\mathrm{eV}) < 5.5$) obtained by screening materials databases using statistical optimization algorithms aided by artificial neural networks (ANN). Two of these new dielectrics are mixed-anion compounds ($Eu_5SiCl_6O_4$ and HoClO) and are shown to be thermodynamically stable against common semiconductors via phase diagram analysis. We also uncovered four other materials with relatively large dielectric constants ($20 < \epsilon < 40$) and band gaps ($2.3 < E_g(\mathrm{eV}) < 2.7$). While the ANN training-data are obtained from the Materials Project, the search-space consists of materials from the Open Quantum Materials Database (OQMD)—demonstrating a successful implementation of cross-database materials design. Overall, we report the dielectric properties of 17 materials calculated using ab initio calculations, that were selected in our design workflow. The dielectric materials with high-dielectric properties predicted in this work open up further experimental research opportunities.

npj Computational Materials (2022)8:146 ; https://doi.org/10.1038/s41524-022-00832-5

INTRODUCTION

Dielectric materials are among the most vital components for microelectronic device manufacturing. They are used in memory devices, capacitor-based energy storage, field-effect transistors, etc[1–3]. The dielectric constant (denoted here as ϵ), more commonly referred to as the relative permittivity, is the factor by which the electric field strength decreases inside a material compared to the vacuum when it is placed near a finite electric charge. The ϵ values of commonly used dielectric materials range between 20 and 30[1,4,5]—for example, Ta_2O_5 ($\epsilon \sim 23$–27, $E_g = 4.2$ eV)[1,2,6,7] and TiO_2 ($\epsilon = 27$, $E_g = 3.5$ eV)[1,2,8]. There is a high demand to find novel materials with high ϵ to increase the device performance and reliability. Typically, ϵ and E_g are inversely related[2,9] in a compound. As a result, although several materials are reported to have even larger ϵ values, they often have a small E_g[9–12], making the dielectric vulnerable to leakage currents under exposure to large electric fields[1,2]. Therefore, compounds with high ϵ and large band gaps are preferred while designing charge storage applications and microelectronic devices.

One of the methods to find high-ϵ compounds is to calculate the dielectric constants and band gaps of a large number of compounds that are available in large materials databases such as the Open Quantum Materials Database (OQMD)[13,14], Materials Project (MP)[15], etc using ab initio methods such as density functional theory (DFT). However, since the accurate calculation of dielectric properties using density functional perturbation theory[16] (DFPT) is computationally very expensive, it would be practically unfeasible to estimate the dielectric constants of tens of thousands of materials available in those databases using high-throughput methods. In this work, we employ an advanced screening strategy to identify compounds with better dielectric properties. Thus, the goal of this work is to find dielectric materials with large values for both ϵ and E_g by screening materials databases but at the expense of conducting as few DFPT calculations as possible. To accomplish this task, we have employed a materials design strategy comprised of statistical optimization models and DFPT calculations on a small set of compounds. While our training set consists of a small amount of data (dielectric constants) from the MP, the search-space contains a vast set of compounds available in the OQMD.

Several online data repositories exist today that are dedicated to hosting large sets of open-sourced inorganic crystal structure data generated from high-throughput (HT) DFT calculations such as the MP[15], OQMD[13,14], and AFLOWLib[17] among others[18,19]. The design and discovery of novel materials using statistical modeling has become an active research area[20–22] in recent times, largely attributed to the availability of such HT datasets. Recently, multiple studies have reported HT-generation of dielectric data and subsequent analysis[9,23,24]. For example, Morita et al. reported[25] machine learning modeling of data from MP[11,12,15] to assess the reliability of the theoretical models currently available to describe the dielectric properties of crystals.

In this work, we use the MP dataset of 1864 dielectric tensors[11,12] to train statistical models and subsequently identify dielectrics from the set of stable materials in the OQMD. Thus the MP data forms the training-data and the set of materials from OQMD forms the search-space for the materials design. This work is a successful demonstration of the scenario where the data obtained from multiple sources can be utilized to discover new compounds. The negligible difference found between the representation vectors, which are also called as feature vectors in machine learning, generated for equivalent materials in MP and OQMD made the cross-database design possible in this work. Overall, we conducted three design cycles which required us to perform dielectric calculations for just 17 materials using DFPT. We report the dielectric constant values of all the 17 materials among which three of them (HoClO, $Eu_5SiCl_6O_4$, and Tl_3PbBr_5) have very large ϵ ($69 < \epsilon < 101$) and E_g ($2.9\ \mathrm{eV} < E_g < 5.5\ \mathrm{eV}$) values making them part of the Pareto front of the known data, and four other materials (Sr_2LuBiO_6, Bi_5IO_7, Bi_3ClO_4, and Bi_3BrO_4) have moderately large ϵ ($20 < \epsilon < 40$) and E_g ($2.3\ \mathrm{eV} < E_g < 2.7\ \mathrm{eV}$) values.

[1]Department of Materials Science and Engineering, Northwestern University, 2220 Campus Drive, Evanston, IL 60208, USA. ✉email: c-wolverton@northwestern.edu

RESULTS

Materials design strategy

Our objective is to find large band gap materials with optimal dielectric constants. Since the dielectric tensor of a compound has nine components, the optimization of all nine components leads to a nine-objective optimization problem which is difficult to solve with training-data of size ~2000. Thus, we specifically optimize the largest eigenvalue of the dielectric tensor, referred to from here onward as ϵ, via statistical modeling through the materials design workflow, as depicted in Fig. 1. The workflow is similar to the strategies that have been previously reported in literature[26,27], where each design cycle consists of three steps—data processing, statistical modeling, and ab initio DFPT calculations. The largest eigenvalue of the total dielectric tensor is chosen as the property to be optimized because that is the highest possible dielectric behavior from a single crystal when it is aligned perfectly along the corresponding direction between two metallic plates. The total dielectric tensor is calculated as the sum of ionic and electronic dielectric tensors. The good agreement between dielectric tensor eigenvalues obtained from MP's DFPT HT framework and experimentally measured dielectric constant values was reported by Petousis et al.[28]. We preferred the largest eigenvalue over the average of eigenvalues because the latter value may severely underestimate the highest possible dielectric behavior from a single crystal (Supplementary Fig. 1), even though it is a popular choice to estimate the polycrystalline dielectric constant[12,28]. The new data produced from DFPT calculations at the end of each cycle is fed into the next design cycle. In the first step, we collected the relevant data from the MP database (training-data) and OQMD (search-space). All materials in the training-data have a known value for ϵ and E_g, while the materials in the search-space have known values of E_g but their ϵ values are unknown. In the second step, Modeling, we created an ensemble of artificial neural network (ANN)[29] models, fit on the training-data, which learn to predict the ϵ value of materials when their crystal structures and E_g values are known. Using this ANN ensemble, we predicted the ϵ of each material in the search-space. Since the prediction was done from an ensemble, the results were a distribution of ϵ values for each material, contrary to the usage of a single ANN model where a single prediction value is obtained. The trained ANN ensemble was used to predict the ϵ-distributions of 11,102 stable non-metallic materials in the search-space, obtained from the OQMD.

Further, the predicted distribution of ϵ was input into the Efficient Global Optimization (EGO)[26] algorithm. EGO takes into account the distribution's mean and standard deviation to rank the materials in search-space based on their potential to increase the chances of finding high-ϵ materials in this workflow within as few design cycles as possible. In this work, the optimization in dielectrics refers to the identification of dielectrics with large ϵ values. The reason for employing an EGO algorithm to explore the search-space is to account for the uncertainty in ANN model predictions when the available training-data may not have sampled the material space uniformly. The advantages of EGO-based optimization in materials design were first reported and benchmarked by Balachandran et al.[26,30,31]. In this work, we used the EGO algorithm to select the best candidates that are either predicted to have a high ϵ value or have a large uncertainty in their ANN-ensemble predictions. Materials that belong to the latter category are from the regions of materials yet to be sampled by the training-data. The DFPT characterization of such materials is expected to increase the reliability of ANN-ensemble predictions after each design cycle and eventually lead to better optimization of dielectrics during the course of this work.

The metric that is used to rank the materials is called expected improvement, or $E(I)$. More details on how the $E(I)$ is calculated, are provided in the "Methods" section. A few (5–6) materials were selected in this step with the highest values of $E(I)$ and carried onto the next step—DFPT calculations. In this final step, the dielectric tensors of the selected materials were calculated using DFPT calculations. If DFPT results show that any of the materials have a high value of E_g and ϵ, we stop the design workflow at that point. Otherwise, a new design cycle is started after transferring the newly computed ϵ values and the corresponding materials to the training-data from the search-space. With an increased size of training-data, the ANN ensemble is expected to have less uncertainty in ϵ predictions in the new design cycle. The design cycle was repeated with feedback three times in total in this work until three materials with very large values for E_g and ϵ were found.

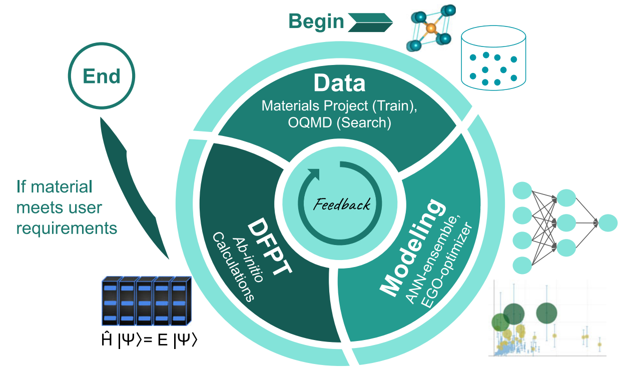

Fig. 1 Materials design workflow used in this work. The three parts of the design workflow shown here together complete a single design cycle. The newly computed DFPT results from a design cycle are fed back into the training-data for the next design cycle. Since the dielectric tensor of a crystal is of shape 3 × 3, optimizing all nine components of the dielectric tensor leads to a nine-objective optimization problem. Thus, we specifically optimize the largest value of dielectric constant (ϵ) among all crystallographic directions, also referred to as the target property hereinafter. This scalar value is quantified as the largest eigenvalue of the total dielectric tensor. The training-data in this work consist of nearly 2000 compounds from the MP for which ϵ and E_g were known and the training-data came from the OQMD, for which only E_g were known at the beginning of this work.

Data

A dataset containing information about crystal structures, chemical compositions, band gap energy values, and dielectric tensors of 1864 stable materials was obtained from the MP[11,12,15] data repository. This dataset was used to generate the training-data. The target property, ϵ, was obtained for each material in this database from its calculated dielectric tensor. Another dataset consisting of 11,102 stable, non-metallic materials containing information about crystal structures, chemical compositions, and band gap energy values was obtained from OQMD[13,14]. This OQMD dataset was used to generate the search-space in which the search to find dielectrics was conducted. The dielectric tensor data of all crystals included in the search-space were unknown at the beginning of this work.

The materials need to be represented as vectors of uniform length in order to be input into a statistical model. We generated the material representations using the Magpie[32] crystal property generator tool. Magpie generates a set of physical features (such as the mean electronegativity of constituent atoms, average coordination number inside the unit cell, etc.) from a given chemical composition and crystal structure. Within Magpie, the crystal's structure-related features are generated by building Voronoi tessellations inside the crystal and finding the nearest neighbors of each individual atom[33]. Magpie generated 271 input features that include 145 composition-based, and 126 structure-based features to represent each material. In addition to these, the

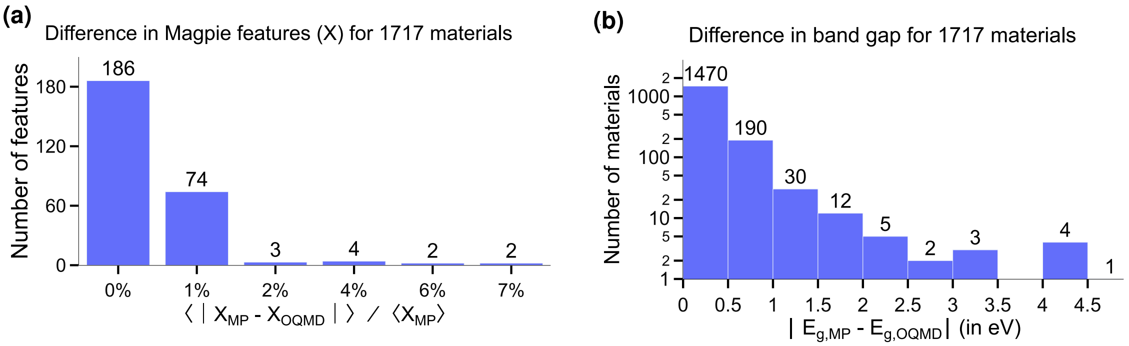

Fig. 2 Comparison of material representation vectors between the OQMD and MP structures. The difference in material representation vectors of the structures obtained from OQMD and MP for 1717 materials. **a** Mean absolute difference in the Magpie-generated representational feature vectors on structures obtained from the MP and OQMD for 1717 materials in the training-data. Crystal structures of all 1717 materials were first obtained from MP as a part of generating training-data, and further cross-referenced to find their equivalent structures in OQMD based on their ICSD Collection Codes. ICSD Collection Codes were not available for the rest of the 143 materials in the MP training-data. **b** Difference in band gap values of 1717 materials from training-data that have a corresponding structure entry in MP and OQMD, which are cross-referenced based on their ICSD Collection Codes.

material's DFT E_g value was also added as an extra feature to the representation vector since it is already known for all materials in both MP and OQMD datasets. The addition of E_g increased the size of the representation vector to 272, which was generated for each material in training-data and search-space. The input feature-vector size was further reduced to 100 using the widely-used feature reduction techniques such as principal component analysis and model-based selection, implemented in the Scikit-learn python library[34]. The set of material representation vectors of training-data and the search-space, in addition to the target values associated with the training-data, completes the first step of materials design as depicted in Fig. 1. The size of the training-dataset increases after each design cycle as a result of conducting DFPT calculations on new materials from the search-space.

Statistical modeling utilizing data from multiple computational material databases is prone to errors arising from the differences in the DFT parameters used at each database's high-throughput calculation strategy. Here, we have investigated the difference in Magpie-generated features for equivalent materials in OQMD and MP, cross-referenced based on their associated Inorganic Crystal Structure Database[35] (ICSD) Collection Codes. In total, 1717 out of 1864 materials in training-data had an ICSD Collection Code associated with them. The crystal structures from OQMD corresponding to all the 1717 ICSD materials were obtained, and their Magpie-generated features were compared against that of the structures obtained from MP as a part of the training-data. The results, as plotted in Fig. 2a, show negligible (≤2%) relative difference in 263 out of a total of 271 Magpie features, while the other eight features have low relative differences (≤7%). All 145 composition-based features are computed to be identical across the databases, as expected. The finite difference in some of the structure-based features originates because of the difference in the accuracy of crystal structural minimization across databases. Band gap, which joins the Magpie features to form the final material representation vector, was also compared between OQMD and MP for the 1717 equivalent materials, as shown in Fig. 2b. Band gap values showed a mean and median absolute deviation of 0.1 eV and 0.0 eV respectively, pointing toward a negligible difference between the calculations of band gap for materials included in the training-data across OQMD and MP. Overall, the materials representation vector considered in this design is generated in a cross-comparable manner across OQMD and MP structures with very low errors.

The ϵ values in the training-data obtained from MP are predominantly concentrated in the range of 0 to 25, making it difficult to model the data reliably for materials with large ϵ due to a possible bias toward smaller values. Less than 5% of the materials in the training-data have $\epsilon > 50$. The median of ϵ values in the MP dataset is 12.2 while the mean and standard deviation are 20.2 and 42.8 respectively. The distribution of ϵ in training-data is shown in Supplementary Fig. 2. The large spread of ϵ values is decreased upon a log-scale transformation, as shown in Fig. 3a. A smaller spread of target values helps stabilize the machine learning model during the training by reducing the probability of excessive changes in internal parameters, such as the weights in an ANN. We also analyzed the correlation between ϵ and E_g values for the materials in the training-data, and it is given in Supplementary Fig. 3.

The original dataset downloaded from MP listed BeO (MP ID: mp-1794) as having large ab initio computed values for $\epsilon (=312)$ and $E_g (=8.2\ \text{eV})$. This large value of ϵ is possibly caused by the improper relaxation of the primitive cell of BeO in MP that leads to a large volume change. Hence, the succeeding calculations on this compound such as DFPT may be incorrect. We conducted a separate DFT cell-relaxation and DFPT calculation for BeO using VASP starting with the MP's initial structure and find that the computed ϵ value for the correctly relaxed structure is 4—well in agreement with the previously reported values in literature[36]. This compound was removed from the training-data before proceeding further. We looked up other materials in training-data with very high ϵ and smaller E_g individually and confirmed that they did not have a large cell-volume change upon relaxation in MP.

Statistical modeling

The predictions from trained machine learning models, such as ANNs, are often prone to errors arising from the insufficient sampling of material space by training-data. We needed to quantify the uncertainty associated with the ϵ value predictions even though the available ANN algorithms explicitly do not provide that value from a single ANN model. So we created an ensemble of ANNs, each of which was trained on a randomly chosen subset of the training-data, and has different architectures and internal parameters. An ANN ensemble containing 2000 independent ANN models was created and trained at each design cycle. Each ANN in the ensemble predicted a single ϵ value upon inputting a material-representation vector, resulting in a distribution of 2000 predicted ϵ values for each material in the search-space. The standard deviation of each of the predicted ϵ-distribution was defined as the uncertainty of ANN modeling for the corresponding material.

Further, a statistical single-objective optimization algorithm, called EGO[26,37–40], was used in this work to evaluate the ϵ-distribution and quantify a measure of probable optimization

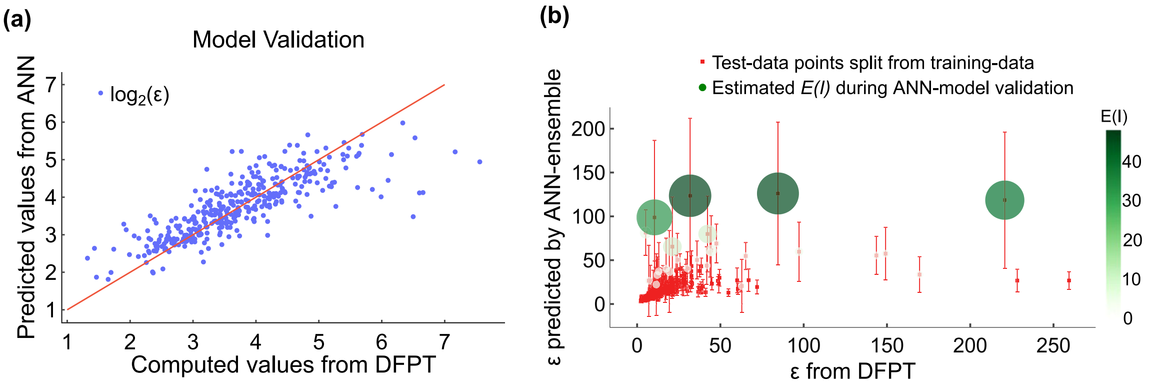

Fig. 3 Results from statistical modeling. a ANN model validation on a test set of 373 materials split from the training-data. We used an ensemble of ANNs to predict a distribution of values for each material. This particular model-fit plot is taken from a single ANN model that was part of the ensemble in design cycle 2. The 373 materials plotted here were not seen by this particular ANN model at any stage during the training. These predictions are made only for this particular ANN model to show its learning capabilities, and it is not part of the design workflow that we created. In the design workflow, each ANN model in the ensemble is exposed only to a unique subset of the full MP training-data, excluding 373 randomly chosen materials. Further, in the design workflow, this trained ANN model is used to predict the dielectric values of only the search-space materials from OQMD, not the 373 unseen materials from the MP dataset. The model was trained to predict $\log_2(\epsilon)$ because the ϵ values were highly non-uniform in the training-data with most of the values below 25, making some of the very large values outliers. A log-scale transformation of ϵ reduced the numerical difference between the largest ϵ value and the median, making the former less of an outlier in ANN modeling. The model fit shown in this plot has an R^2 score of 70%, and a Spearman's rank correlation of 85%. **b** This plot shows the predicted ϵ-distributions and corresponding *E*(*I*) values on the same test dataset consisting of 373 materials split from the training-data. The error bars represent the standard deviation in ANN-ensemble predictions which is quantified as the uncertainty of ANN modeling. For a clearer perspective, the radius and color of the circles represent the same quantity—the expected improvement, *E*(*I*), value calculated using the EGO algorithm. A point without an outer circle around it represents a material with a negligible (<10^{-3}) value for *E*(*I*). In this figure, only 25 materials have an *E*(*I*) value that is greater than 10^{-3}.

associated with each material in the search-space. EGO is not a method to model the data and predict ϵ. Instead, EGO is an algorithm to select the best candidates from a given search-space, based on their ϵ-distributions predicted by the ANN ensemble, in order to discover as many high-ϵ materials from as few design cycles as possible. Here, the desired optimization is the maximization of ϵ among all the materials in the search-space. The quantified measure of predicted optimization in EGO is called expected improvement, denoted as *E*(*I*). Conceptually, the *E*(*I*) of a material in search-space is the quantified probability with which a DFPT calculation of ϵ for that material will lead to the identification of high-ϵ material in the design workflow within as few design cycles as possible. Figure 3a shows the results from an ANN model validation as a part of model training during the second design cycle. The values of *E*(*I*) computed for the same validation data split from the training-data are shown in Fig. 3b. A simplified illustration of *E*(*I*) with the help of an example is given below.

Example illustration of *E(I)*

Suppose the predicted ϵ-distribution belonging to a material M_1 in the search-space has a large standard deviation. Then it is highly probable that the material M_1 belongs to a part of the material representation vector space which was not sampled very well in the training set. Computing the ϵ of M_1 using DFPT and feeding back that information to the training-data will lead to better ANN modeling in the subsequent design cycles. Thus, M_1 will have a large value of *E*(*I*). Now consider another material M_2 in search-space with a large mean and a small standard deviation for its predicted ϵ-distribution. The material M_2 belongs to a part of the material representation vector space that was sufficiently sampled by the training-data. So it is highly probable that M_2 will turn out to be a high-ϵ material upon DFPT calculations. Because of that, M_2 will also have a large value of *E*(*I*).

In EGO, the calculation of *E*(*I*) for a general optimization problem proceeds as follows (also shown in Fig. 4).

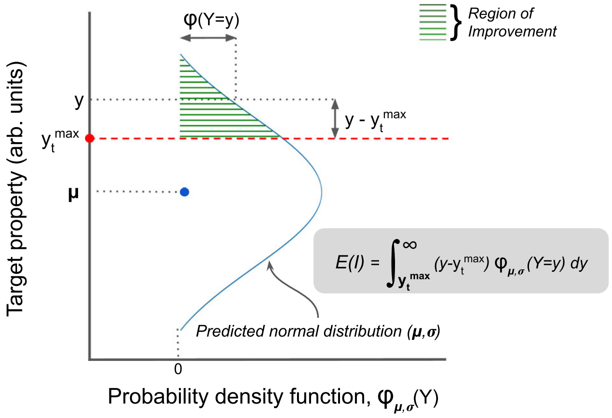
φ(Y=y)
Region of Improvement
y
y - y_t^max
y_t^max
μ
Target property (arb. units)
E(I) = ∫_{y_t^max}^∞ (y-y_t^max) φ_{μ,σ}(Y=y) dy
Predicted normal distribution (μ,σ)
0
Probability density function, φ_{μ,σ}(Y)

Fig. 4 The optimization algorithm. The value $y_t^{\max}$ represents the currently available highest value of ϵ among all materials in the training-data. μ and σ represent the mean and standard deviation of the ANN-ensemble predicted distribution of ϵ for a material (blue dot) in the search-space. Within the predicted distribution, which is assumed as a normalized Gaussian function here, the region above $y_t^{\max}$ represents the region of improvement—as shown in green. If the ab initio DFPT calculation determines that the material's ϵ value exists within the green-shaded region, it will be considered as an improvement over the current best value $y_t^{\max}$ in training-data.

Let Y be the target property to be maximized and $\varphi(Y)$ be the predicted distribution of Y for a given search-space material. The value, $\varphi(Y=y)$ is the probability when the value of Y is y. The largest value of the target property in the training-data is denoted as $y_t^{\max}$. The EGO algorithm, as formulated by Jones et al.[38], computes the expected improvement, *E*(*I*), as:

$$E(I) = \int_{y_t^{\max}}^{\infty} (y - y_t^{\max})\ \varphi(Y = y)\ dy \tag{1}$$

As mentioned in Balachandran et al.[26], if the predicted distribution is approximated as a normal (i.e., Gaussian) distribution with a mean μ and a standard deviation σ, the above equation can be re-written as:

$$E(I) = \sigma[\phi(z) + z\Phi(z)] \tag{2}$$

where, $z = \frac{\mu - y_t^{\max}}{\sigma}$, ϕ is the probability density function, and Φ is the cumulative distribution function[38] of the normal distribution, $\varphi(Y)$.

For dielectric design, Y is the dielectric constant (ϵ) of a candidate material, and $y_t^{\max}$ is the highest value of ϵ in the training-data obtained from DFPT calculations. In the MP dataset, the largest ϵ value is for TiO_2 with $\epsilon = 988$ and $E_g = 1.8$ eV. But our goal in this work is to find materials with large ϵ's, not necessarily higher than 988 as long as the E_g's are greater than 1.8 eV. Thus the $y_t^{\max}$ in this work was set at 100.0 for all design cycles, instead of setting it at 988.0, to consider the search-space materials whose ϵ values are predicted to be sufficiently high. The $\varphi(Y)$ is approximated to be a normal distribution with the same mean, μ, and standard deviation, σ, as that of the original ϵ-distribution predicted by the ANN ensemble for each search-space material.

Design cycles with feedback

The ϵ values of a few materials selected from the statistical modeling are computed from DFPT calculations, as shown in the final segment of a design cycle in Fig. 1. The results from the DFPT calculations are used to determine whether to conduct any further design cycles. In this work, we conducted the design cycles until at least one high-ϵ dielectric with a large E_g is identified. When no such materials are found during a design cycle, all the selected materials along with their newly DFPT-estimated ϵ values are transferred from search-space to training-data, resulting in a feedback of information prior to the beginning of the next design cycle. The feedback is one of the most crucial parts of our material design workflow because it results in a better sampling of material representation vector space by training-data and thus, more reliable ANN model predictions during the next design cycle. The advantage of the feedback mechanism is prominent during the quantification of uncertainty which is used directly by the EGO algorithm to identify the best candidates for the next set of DFPT calculations. After the end of a design cycle, the uncertainty on predicting the ϵ values is decreased for the set of materials which are similar to the materials whose ϵ values were calculated using DFPT in the given cycle.

In addition to the feedback mechanism, another factor that influenced the candidate selection in the design workflow is the minimum cutoff imposed on the band gap values of materials when they are included in the search-space. The reason for implementing a cutoff is to externally introduce a character of multi-objective optimization in this work. Without explicitly setting a minimum band gap limit, the candidate selection process that is dictated by the EGO algorithm tries to optimize only a single objective, which is the ϵ value. We conducted three design cycles sequentially with feedback of the newly calculated data into training-data after each cycle. In the first design cycle, we set no band gap minimum cutoffs to allow the full exploration of the search-space that consists of 11,102 non-metals from OQMD. In the second design cycle, a minimum cutoff of 2.25 eV was set, leaving 6191 materials in the search-space. In the final cycle, the minimum cutoff was increased to 5 eV to limit the candidate selection only to the materials with very high E_g. Hence, the search-space size in the final cycle was reduced to 1046 materials. The workflow that we adopted in this work deviates from the ideal situation where a dedicated multi-objective optimization statistical algorithm will be used to find a material with high ϵ and large E_g values. Since the band gap values are already available for all materials in the search-space, the best approach here was to implement a statistical optimization algorithm to quickly find high-ϵ materials while the preference for large band gap values is achieved by manually setting a minimum cutoff. This work stands as an example for the modifications required to practically implement the statistical algorithms that are often benchmarked on idealistic scenarios.

Table 1. The Pareto front of dielectric materials dataset from Materials Project.

MP ID	Material	E_g (eV)	ϵ
mp-1138	LiF	8.7161	9.3107
mp-13948	Cs_2HfF_6	7.2288	9.3281
mp-13947	Rb_2HfF_6	7.1298	9.3626
mp-7104	$CsCaF_3$	6.8955	9.7272
mp-5347	$KAlF_4$	6.7863	10.63
mp-10250	$BaLiF_3$	6.5643	14.7705
mp-3654	$RbCaF_3$	6.3974	18.8679
mp-8455	CsF	5.9329	20.3025
mp-28243	$RbLiCl_2$	5.1482	54.4788
mp-5606	$AlTlF_4$	4.2492	96.91
mp-23092	Ba_2TaBiO_6	2.5855	99.8664
mp-27832	Tl_2SnCl_6	2.4814	100.8
mp-3614	$KTaO_3$	2.0983	639.8836
mp-2657	TiO_2	1.781	988.0478

MP ID corresponds to the unique ID of material in the repository, E_g is the band gap energy, and ϵ corresponds to the largest eigenvalue in the dielectric constant tensor.

New dielectric materials

The materials that are part of the Pareto front of MP data are listed in Table 1, while the Pareto front of training-data at each design cycle is plotted in Fig. 5. Since the maximization of ϵ and E_g values are considered as optimal in this study, each material in the Pareto front has a higher value of either ϵ or E_g than any other material in the corresponding training-data. Therefore, the modification of the training-data's Pareto front by any of the newly calculated dielectric constants after each design cycle may indicate the identification of suitable, high-dielectric materials.

During the first design cycle, the EGO algorithm picked out the five most promising candidates with the largest $E(I)$ values in the search-space. The ϵ values of these five selected materials were calculated using DFPT. Two materials among them turned out to have very high ϵ values (~370) but very low E_g (~0.5 eV). The low E_g values are not unexpected since the EGO algorithm implemented in this work aims to maximize only the ϵ values. None of the materials selected in this cycle modified the Pareto front of the MP dataset, as shown in Fig. 5a. The ϵ values of these five materials were appended to the training-data prior to starting the next design cycle.

Five materials were selected in the second cycle and their dielectric constants were calculated. Our calculations predict a large dielectric constant for one of the five new materials—tetragonal Tl_3PbBr_5 ($\epsilon = 101$, $E_g = 2.9$ eV). Tl_3PbBr_5 joined the Pareto front, as shown in Fig. 5b. Three other new materials—Bi_5IO_7 ($\epsilon = 36$, $E_g = 2.7$ eV), Bi_3ClO_4 ($\epsilon = 39$, $E_g = 2.3$ eV), and Bi_3BrO_4 ($\epsilon = 39$, $E_g = 2.3$ eV), have moderately large ϵ values, even though they did not improve the existing Pareto front. All the five new materials were appended into the training-data before proceeding to begin the third design cycle.

During the third and final design cycle consisting of only materials with very large E_g in search-space, seven new candidate materials were selected to do DFPT calculations. Two among them —$Eu_5SiCl_6O_4$ ($\epsilon = 69$, $E_g = 5.5$ eV) and HoClO ($\epsilon = 75$, $E_g = 5.2$ eV) joined the Pareto front due to their large ϵ and E_g values, as shown

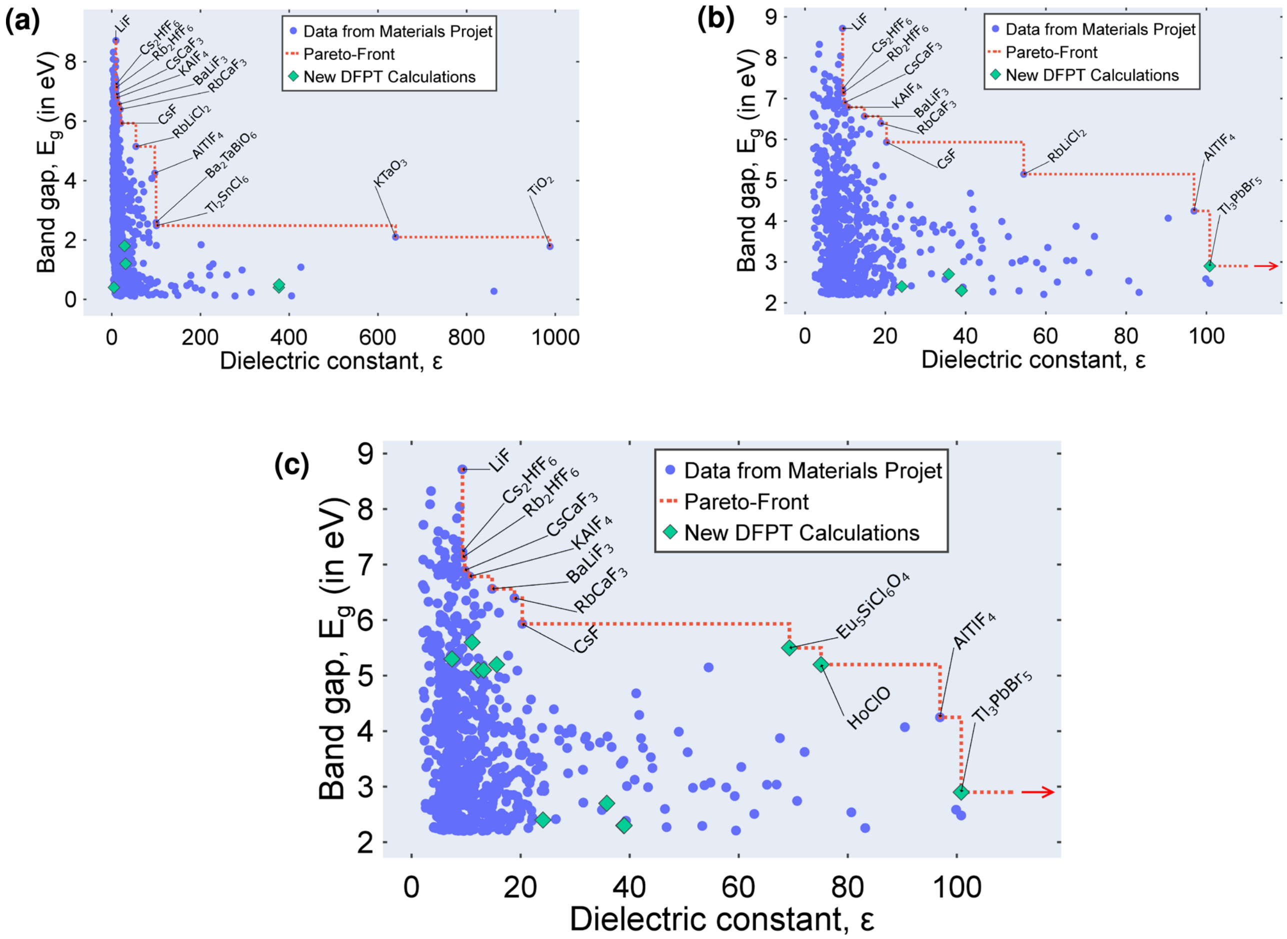

Fig. 5 Evolution of the Pareto front with design cycles. The ϵ and E_g values of the training-data and newly characterized dielectric materials are plotted for **a** design cycle 1, **b** design cycle 2, and **c** design cycle 3. All the data shown in these plots originated from DFPT calculations. Plot **a** shows the original Pareto front of the dataset from Materials Project (MP) because none of the materials measured in cycle 1 became part of the Pareto front—predominantly owing to their low band gap values. Assigning no restrictions on the band gap of search-space materials during the first cycle directed the design algorithm to pick two materials without any preference for large band gaps. The numerical values of the materials in the Pareto front of the MP dataset are given in Table 1. In both **b** and **c**, only the materials with E_g values greater than 2.0 eV are plotted to highlight the area where some of the newly discovered dielectrics in their corresponding cycles joined the Pareto front. Due to this cropping, two materials from the MP dataset which are actually in the Pareto front in plots **b** and **c** with very high ϵ values—tetragonal TiO_2 ($\epsilon = 988$, $E_g = 1.8$ eV) and cubic $KTaO_3$ ($\epsilon = 640$, $E_g = 2.1$ eV), are not shown here.

in Fig. 5c. In total, three new dielectric materials in the Pareto front were discovered after three design cycles and 17 new DFPT calculations were performed in the entire workflow. No further design cycles were conducted since we have already identified multiple compounds with high ϵ and E_g, which remained unexplored experimentally.

The ϵ values of all 17 materials which were obtained in this work are given in Table 2. The ϵ and E_g of all materials belonging to the Pareto front of the MP dataset is listed in Table 1 for comparison. Among all the newly discovered dielectrics with large ϵ values, tetragonal HoClO and monoclinic $Eu_5SiCl_6O_4$ stand out because of their very large DFT-calculated band gap energies (5.2 eV and 5.5 eV respectively). These two rare earth oxychlorides are reported to have been experimentally synthesized[41–44] but their dielectric properties remained unstudied to the extent of our knowledge. Both of these compounds are mixed-anionic inorganic compounds—a class of emerging functional materials[45]. Interestingly, the monoclinic $Eu_5SiCl_6O_4$ has 32 atoms in its primitive unit cell which often exceeds the maximum cutoff on the number of atomic sites in HT studies involving computationally expensive material properties[11,19].

Thermodynamic stability of a dielectric when in contact with Si or other semiconductors is an important requirement for it to be used in electronic applications. Several of the high-ϵ dielectrics identified in the published literature were shown to be unstable while forming an interface with Si in subsequent experimental studies conducted at or above the room temperature. The formation of SiO_x and other undesired metal oxides were reported at the interface between Si and the popular high-ϵ dielectrics such as Ta_2O_3[46–48], TiO_2[49,50], $BaTiO_3$[51], and $SrTiO_3$[52,53]. The thermodynamic stability between two compounds can be assessed from the phase diagram involving those compounds. In this work, the phase diagram is constructed by computing the convex hull[54] of formation energies of all the materials that belong to a given phase space spanned by their constituent elements. Each of the compounds that form the convex hull not only has the lowest formation energy at its composition but also has lower energy than any linear combination of other materials in that phase space. The difference between the formation energy of a compound and energy at the convex hull for the same composition is called as the hull distance (E_{hd}). By definition, each material that is on the convex hull has a hull distance of zero (i.e., $E_{hd} = 0$) and is considered to be stable. On the other hand, every material that falls above the convex hull is considered as metastable ($0 < E_{hd} \leq 50$ meV per atom) or unstable ($E_{hd} > 50$ meV per atom) depending on the magnitude of E_{hd} according to the

Table 2. Dielectric constants of 17 materials calculated using DFT in this work.

OQMD ID	Material	E_g (eV)	ϵ_x	ϵ_y	ϵ_z	Cycle
681780	$CaVO_3$	0.4	4.7	4.5	4.5	1
14476	Sr_2VN_3	1.8	28.8	16.5	16.0	1
13450	$BaZrN_2$	1.2	31.2	31.2	21.7	1
1104204	HoN	0.4	376.9	373.0	372.7	1
649584	Bi_2SeO_2	0.5	377.3	371.8	118.2	1
19571	**Sr_2LuBiO_6**	2.4	24.1	19.4	18.7	2
5958	**Bi_5IO_7**	2.7	35.8	28.2	23.1	2
24994	**Bi_3ClO_4**	2.3	38.9	24.2	25.7	2
22697	**Bi_3BrO_4**	2.3	39.0	23.7	22.1	2
118234	**Tl_3PbBr_5**	2.9	100.8	36.4	36.4	2
11916	Eu_4Cl_6O	5.3	7.4	7.3	5.5	3
18953	EuClF	5.6	11.1	11.1	10.4	3
646321	Rb_2PrCl_5	5.1	12.2	11.0	8.9	3
15191	$Cs_2NaCeCl_6$	5.1	13.2	13.2	13.2	3
4063	$EuCl_2$	5.2	15.6	12.9	11.8	3
24611	**$Eu_5SiCl_6O_4$**	5.5	69.3	15.1	12.9	3
13689	**HoClO**	5.2	75.1	37.9	15.2	3

OQMD ID refers to the materials' unique entry ID in the OQMD database, E_g refers to the band gap energy in eV, $\epsilon_{x,y,z}$ refers to the three eigenvalues (xx, yy, zz) of the of dielectric constant tensor, and the Cycle mentions the design cycle when the material was selected for the calculations of dielectric constant using DFPT. The values $\epsilon_{x,y,z}$ are ordered in such a way that $\epsilon_x > \epsilon_y > \epsilon_z$. The best seven materials found in this work are highlighted in bold letters.

heuristic conventions adopted in literature[31,55–58]. The presence of a tie-line between two compounds in a convex hull phase diagram indicates that they are thermodynamically stable phases when in contact with each other. Our thermodynamic stability analysis on Ta_2O_3, TiO_2, $BaTiO_3$, and $SrTiO_3$ in OQMD using the qmpy API[14] showed no tie-lines connecting any of them to Si, indicating they are unstable when in contact with Si. This is consistent with the published results[46–53]. We also analyzed Gd_2O_3, a high ϵ (~20[59]) that is proven to be stable against Si[60], and found that a tie-line does exist between Si and Gd_2O_3. These phase diagram plots are provided in Supplementary Fig. 6. In Fig. 6, we report a phase diagram to assess the stability of newly discovered high-ϵ dielectrics—HoClO and $Eu_5SiCl_6O_4$. The phase diagram shows that both these materials are thermodynamically stable with the semiconductors such as Si, Ge, GaAs, GaN, and SiC at 0K, a requirement for them to be used in microelectronic devices where an interface with one of the common semiconductors is often necessary[61]. The next most promising candidate, tetragonal Tl_3PbBr_5, has a very large ϵ (101) but possesses a relatively smaller band gap (2.9 eV) and is computed to be thermodynamically metastable at 0K ($E_{hd} = 16$ meV per atom) according to the data obtained from the OQMD. Tl_3PbBr_5 is also reported in the literature to have been experimentally synthesized[62–64], without any mention of its dielectric properties.

DISCUSSION

We report the identification of three dielectric materials that contain a combination of high-dielectric constant and large band gap—HoClO($\epsilon = 75$, $E_g = 5.2$ eV), $Eu_5SiCl_6O_4$($\epsilon = 69$, $E_g = 5.5$ eV), and Tl_3PbBr_5($\epsilon = 101$, $E_g = 2.9$ eV). These compounds modify the Pareto front of previously known high-throughput dielectric constants data available from the MP database. Our screening strategy also uncovers four other dielectric materials with large E_g and moderately large ϵ—Sr_2LuBiO_6($\epsilon = 24$, $E_g = 2.4$ eV), Bi_5IO_7($\epsilon = 36$, $E_g = 2.7$ eV), Bi_3ClO_4($\epsilon = 39$, $E_g = 2.3$ eV), and Bi_3BrO_4($\epsilon = 39$, $E_g = 2.3$ eV)—at the cost of conducting only 17 DFPT calculations overall. We utilize the data available in the open-source databases (OQMD, MP) to build a statistical optimization model and use it to select the best candidates after searching among 11,102 stable non-metals that are available in the OQMD. Among the newly discovered dielectrics, two mixed-anionic materials—HoClO and $Eu_5SiCl_6O_4$ are shown to have tie-lines with multiple, commonly used semiconductors on their phase diagrams, that indicate their thermodynamic equilibrium.

The presence of rare earth elements such as Ho and Eu in dielectrics can be a challenge for their use in practical applications. However, the ongoing efforts toward increasing their availability such as efficient recycling of rare earth materials[65,66] can result in a sufficient supply of elements for mass production of small electronic components. In particular, Ho is an underutilized element in the industry[67] even though it is more abundant in the earth's crust than other widely mined elements such as Mo, Bi, and precious metals[68]. Eu is more abundant on earth's crust than Ho and some of the heavily mined elements such as W and As[68]. Hence, an active exploration of cheaper and easier extraction methods for rare earth elements may make it feasible to include them in mass-produced electronics in the near future. The presence of toxic elements such as Pb and Tl can stand as a barrier against including Tl_3PbBr_5 in consumer electronics. Since mixed-anionic materials are an emerging class of functional materials, our identification of promising dielectric materials in this family opens up further research opportunities on rational design of high-performance dielectrics and their experimental characterizations.

We also assessed the thermodynamic stability of the new dielectrics by creating a large convex hull diagram containing the best two new dielectrics (HoClO and $Eu_5SiCl_6O_4$) and several commonly used materials in electronics. The relevance of this analysis is also provided in detail along with examples of previously reported high-ϵ dielectrics[46–53] that were later found out to be unstable when in contact with common electronic component materials such as SiO_2. Our convex hull analysis indicates that both HoClO and $Eu_5SiCl_6O_4$ are stable against the common electronic materials that we considered.

To understand what features of HoClO, $Eu_5SiCl_6O_4$, and Tl_3PbBr_5 make them the best dielectric candidates in this study, we have calculated their electronic structures and partial density of states (Supplementary Fig. 5). Our analysis shows that the top of the valence bands and bottom of the conduction bands in these compounds consists of primarily the contributions from the anions (Cl, Br) and cations (Ho, Eu, Tl), respectively. This analysis indicates that having lighter anions (such as Cl, Br) is advantageous as their valence orbitals making up the valence band edge in those compounds will have lower energies, hence, a relatively larger band gap that is desired in high-ϵ materials.

In addition to the identification of high-dielectrics, we successfully demonstrated an implementation of a cross-database statistical design for computational materials selection. Datasets from the MP and OQMD repositories are used in this work as training-data and search-space, respectively. The successful identification of new materials from such a workflow is another motivation for actively moving toward the interoperability of materials databases, which is one of the four pillars of FAIR data principles[69] in scientific data management. Therefore, better interoperability across databases amplifies the flexibility in utilizing materials data while solving a complex materials problem.

Lastly, this work also stands as an example of the practical implementation of a computational design strategy for property optimization via data-informed material selection. A multi-

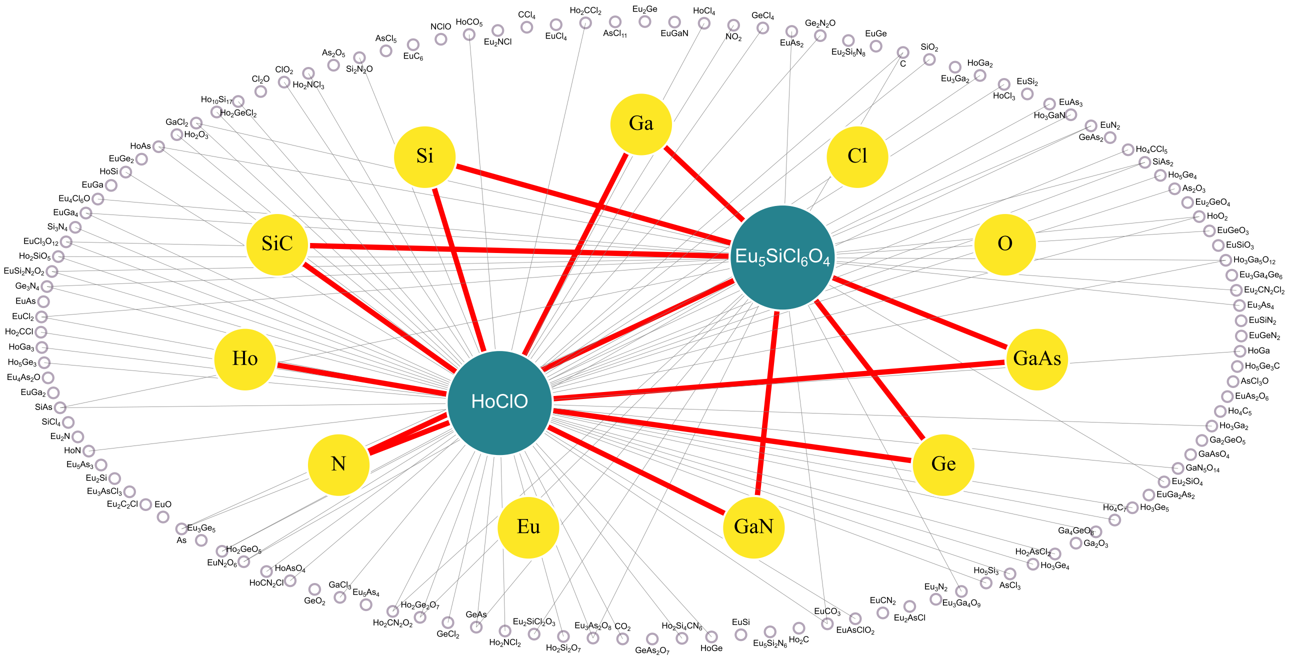

Fig. 6 Phase diagram of all stable compounds in Ho-Cl-O-Eu-Si-Ge-Ga-As-C-N phase space from OQMD (as of January 2022). The two new most promising dielectrics, HoClO and $Eu_5SiCl_6O_4$ are plotted in large green circles in the center. The elements (Ho, Eu, Si, Cl, Ge, Ga, As, C, N, and O) and semiconductors of interest (Si, Ge, GaAs, SiC, and GaN) are plotted in the middle layer in medium-sized yellow circles. All other stable compounds in the phase diagram are plotted in small dark circles in the outermost layer. Tie-lines between the new dielectrics and the semiconductors or elements are shown as thick red lines. Other tie-lines from the dielectrics to the rest of the stable materials in the outer layer are drawn as narrow gray lines. Another 2326 tie-lines exist in this phase diagram that do not include either of the dielectrics, and are not shown in this network-plot for better visibility of relevant information. The elements and compounds without any visible tie-lines in the outermost layer do not have tie-lines with HoClO or $Eu_5SiCl_6O_4$, but they have tie-lines with some of the other materials in the outer layer—making them part of this phase diagram. There exist a tie-line from each dielectric material to each semiconductor that is considered here for comparison. This indicates that both HoClO and $Eu_5SiCl_6O_4$ are in thermodynamic equilibrium with Si, Ge, GaAs, GaN, and SiC at 0K. Thermodynamic stability is a requirement for dielectrics that needs to form a stable interface with semiconductors in electronic applications[61].

objective optimization problem (maximizing ϵ and E_g) is converted into a single objective optimization using statistical methods (maximizing ϵ) combined with explicit constraining of band gap values (higher E_g) among materials since E_g is already available for all materials in the search-space. The deviation from the ideal, statistically benchmarked multi-objective optimization workflows[27] enabled the efficient utilization of resources and resulted in the identification of three high-ϵ dielectrics at the cost of just 17 new DFPT calculations.

METHODS

ANN modeling

The individual models in the ANN ensemble consisted of a single hidden layer with the number of neurons in the range of 10^2. The exact number of neurons varied randomly within a small range (10–30) to avoid any bias that may arise from model architecture since the subset of training-data for each ANN was randomly sampled. Each ANN ensemble consisted of 2000 independent ANNs. Thus, the ϵ-distribution for each material consisted of 2000 independent ϵ predictions. A new ANN ensemble was created and trained for each new design cycle to learn the incremented training-data. The Nadam optimizer is used for network optimization during the training. Both L2 layer regularization and early-stopping callback as implemented in Keras[70], are implemented for each ANN in the ensemble to prevent overfitting. On average, it took between 300 to 400 epochs to reach the local minimum of the loss function. Each epoch is a full iteration of fitting the training-data to update the internal weights of an ANN. Validation details of one of the randomly chosen ANN models from the ensemble are plotted in Fig. 3a for reference. Feature dimensional reduction prior to the training of ANNs was done using the principal component analysis algorithm implemented in scikit-learn[34]. Model validation during the training of one of the 2000 ANN models in the second design cycle is plotted in Fig. 3a.

DFPT calculations

We performed all DFT calculations using the Vienna Ab initio Simulation Package (VASP)[71,72] with potentials derived using the projector-augmented wave[73,74] method. We calculated the total dielectric constant (sum of electronic and ionic components) values for selected materials using DFPT as implemented in VASP. All the compounds were fully relaxed before the dielectric calculations. We used an energy cutoff of 520 eV, k-mesh of 6000 k-points per reciprocal atom, and an energy-threshold of 10^{-8} eV during the self-consistent calculations. The forces on the atoms after structural relaxations were less than 10^{-3} eV Å^{-1}. We used the generalized gradient approximation[75] to approximate the exchange-correlation energies of the electrons. A detailed discussion on DFPT calculations is provided in the Supplementary Methods section included within the Supplementary Material. We did DFPT calculations on a set of well-known dielectrics and a few rare earth compounds, and benchmarked the results against previously reported results in the literature. These results indicate the reliability of our calculated ϵ values, which are provided in Supplementary Table 2. Specifically, two rare earth oxides (EuO and Ho_2O_3) and one rare earth halide (EuF_2) were benchmarked to test the accuracy of the standard DFPT calculations in modeling these compounds. Furthermore, our calculations reveal that no imaginary phonon modes appear in HoClO, $Eu_5SiCl_6O_4$, and Tl_3PbBr_5, the best high-ϵ materials identified in this work. More details are provided in Supplementary Table 1 and Supplementary Fig. 4.

DATA AVAILABILITY

The data used in building statistical design models are open-sourced and available via OQMD and Materials Project databases. Other data that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

The raw, unformatted codes used in this project for statistical materials design are available via Github at https://github.com/tachyontraveler/diel-design-scripts/tree/

v0.1.0-alpha. The latest versions of the scripts upon release will be available in the future at https://doi.org/10.5281/zenodo.6515841.

Received: 10 February 2022; Accepted: 13 June 2022;
Published online: 07 July 2022

REFERENCES

1. Ortiz, R. P., Facchetti, A. & Marks, T. J. High-k organic, inorganic, and hybrid dielectrics for low-voltage organic field-effect transistors. *Chem. Rev.* **110**, 205–239 (2009).
2. Wang, B. et al. High-k gate dielectrics for emerging flexible and stretchable electronics. *Chem. Rev.* **118**, 5690–5754 (2018).
3. Kingon, A. I., Maria, J.-P. & Streiffer, S. Alternative dielectrics to silicon dioxide for memory and logic devices. *Nature* **406**, 1032 (2000).
4. Shevlin, S. A., Curioni, A. & Andreoni, W. Ab initio design of high-k dielectrics: $La_xY_{1-x}AlO_3$. *Phys. Rev. Lett.* **94**, 146401 (2005).
5. Delugas, P., Fiorentini, V., Filippetti, A. & Pourtois, G. Cation charge anomalies and high-κ dielectric behavior in $DyScO_3$: ab initio density-functional and self-interaction-corrected calculations. *Phys. Rev. B* **75**, 115126 (2007).
6. Iino, Y. et al. Organic thin-film transistors on a plastic substrate with anodically oxidized high-dielectric-constant insulators. *Jpn. J. Appl. Phys.* **42**, 299 (2003).
7. Kukli, K. et al. Properties of tantalum oxide thin films grown by atomic layer deposition. *Thin Solid Films* **260**, 135–142 (1995).
8. Ramajothi, J., Ochiai, S., Kojima, K. & Mizutani, T. Performance of organic field-effect transistor based on poly (3-hexylthiophene) as a semiconductor and titanium dioxide gate dielectrics by the solution process. *Jpn. J. Appl. Phys.* **47**, 8279 (2008).
9. Lee, M., Youn, Y., Yim, K. & Han, S. High-throughput ab initio calculations on dielectric constant and band gap of non-oxide dielectrics. *Sci. Rep.* **8**, 14794 (2018).
10. Wilk, G. D., Wallace, R. M. & Anthony, J. High-κ gate dielectrics: current status and materials properties considerations. *J. Appl. Phys.* **89**, 5243–5275 (2001).
11. Petretto, G. et al. High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 180065 (2018).
12. Petousis, I. et al. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Sci. Data* **4**, 160134 (2017).
13. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
14. Kirklin, S. et al. The open quantum materials database (OQMD): assessing the accuracy of dft formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
15. Jain, A. et al. The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
16. Giannozzi, P. & Baroni, S. *Density-Functional Perturbation Theory*, 195–214 (Springer, 2005).
17. Curtarolo, S. et al. Aflowlib. org: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
18. Draxl, C. & Scheffler, M. The nomad laboratory: from data sharing to artificial intelligence. *J. Phys.: Mater.* **2**, 036001 (2019).
19. Choudhary, K. et al. High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric, and dielectric responses. *npj Comput. Mater.* **6**, 1–13 (2020).
20. Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: the use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
21. Saal, J. E., Oliynyk, A. O. & Meredig, B. Machine learning in materials discovery: confirmed predictions and their underlying approaches. *Annu. Rev. Mater. Res.* **50**, 49–69 (2020).
22. Park, C. W. & Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **4**, 063801 (2020).
23. Umeda, Y., Hayashi, H., Moriwake, H. & Tanaka, I. Prediction of dielectric constants using a combination of first principles calculations and machine learning. *Jpn. J. Appl. Phys.* **58**, SLLC01 (2019).
24. Qu, J., Zagaceta, D., Zhang, W. & Zhu, Q. High dielectric ternary oxides from crystal structure prediction and high-throughput screening. *Sci. Data* **7**, 1–10 (2020).
25. 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).
26. Balachandran, P. V., Xue, D., Theiler, J., Hogden, J. & Lookman, T. Adaptive strategies for materials design using uncertainties. *Sci. Rep.* **6**, 19660 (2016).
27. Gopakumar, A. M., Balachandran, P. V., Xue, D., Gubernatis, J. E. & Lookman, T. Multi-objective optimization for materials discovery via adaptive design. *Sci. Rep.* **8**, 3738 (2018).
28. Petousis, I. et al. Benchmarking density functional perturbation theory to enable high-throughput screening of materials for dielectric constant and refractive index. *Phys. Rev. B* **93**, 115151 (2016).
29. Jain, A. K., Mao, J. & Mohiuddin, K. M. Artificial neural networks: a tutorial. *Computer* **29**, 31–44 (1996).
30. Balachandran, P. V., Young, J., Lookman, T. & Rondinelli, J. M. Learning from data to design functional materials without inversion symmetry. *Nat. Commun.* **8**, 14282 (2017).
31. Balachandran, P. V. et al. Predictions of new ABO_3 perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.* **2**, 043802 (2018).
32. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
33. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi Tessellations. *Phys. Rev. B* **96**, 024104 (2017).
34. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
35. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr., Sect. B: Struct. Sci.* **58**, 364–369 (2002).
36. Groh, D. et al. First-principles study of the optical properties of BeO in its ambient and high-pressure phases. *J. Phys. Chem. Solids* **70**, 789–795 (2009).
37. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
38. Jones, D. R., Schonlau, M. & Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **13**, 455–492 (1998).
39. Solomou, A. et al. Multi-objective Bayesian materials discovery: application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling. *Mater. Des.* **160**, 810–827 (2018).
40. Talapatra, A. et al. Autonomous efficient experiment design for materials discovery with bayesian model averaging. *Phys. Rev. Mater.* **2**, 113803 (2018).
41. Templeton, D. & Dauben, C. H. Crystal structures of rare earth oxychlorides. *J. Am. Chem. Soc.* **75**, 6069–6070 (1953).
42. Hölsä, J., Lahtinen, M., Lastusaari, M., Valkonen, J. & Viljanen, J. Stability of rare-earth oxychloride phases: bond valence study. *J. Solid State Chem.* **165**, 48–55 (2002).
43. Basiev, T. et al. Hydration of strontium chloride and rare-earth element oxychlorides. *Russ. J. Appl. Chem.* **78**, 1035–1037 (2005).
44. Jacobsen, H., Meyer, G., Schipper, W. & Blasse, G. Synthesis, structures and luminescence of two new Europium (II) Silicate-Chlorides, $Eu_2SiO_3Cl_2$ and $Eu_5SiO_4Cl_6$. *Z. Anorg. Allg. Chem.* **620**, 451–456 (1994).
45. Kageyama, H. et al. Expanding frontiers in materials chemistry and physics with multiple anions. *Nat. Commun.* **9**, 1–15 (2018).
46. Atanassova, E. & Spassov, D. X-ray photoelectron spectroscopy of thermal thin Ta_2O_5 films on Si. *Appl. Surf. Sci.* **135**, 71–82 (1998).
47. Schlom, D. G. & Haeni, J. H. A thermodynamic approach to selecting alternative gate dielectrics. *MRS Bull.* **27**, 198–204 (2002).
48. Alers, G. et al. Intermixing at the tantalum oxide/silicon interface in gate dielectric structures. *Appl. Phys. Lett.* **73**, 1517–1519 (1998).
49. Perego, M., Seguini, G., Scarel, G., Fanciulli, M. & Wallrapp, F. Energy band alignment at TiO_2/Si interface with various interlayers. *J. Appl. Phys.* **103**, 043509 (2008).
50. McCurdy, P. R., Sturgess, L. J., Kohli, S. & Fisher, E. R. Investigation of the PECVD TiO_2–Si (1 0 0) interface. *Appl. Surf. Sci.* **233**, 69–79 (2004).
51. George, J. P. et al. Preferentially oriented $BaTiO_3$ thin films deposited on silicon with thin intermediate buffer layers. *Nanoscale Res. Lett.* **8**, 1–7 (2013).
52. Hu, X. et al. The interface of epitaxial $SrTiO_3$ on silicon: in situ and ex situ studies. *Appl. Phys. Lett.* **82**, 203–205 (2003).
53. Goncharova, L. et al. Interface structure and thermal stability of epitaxial $SrTiO_3$ thin films on Si (001). *J. Appl. Phys.* **100**, 014912 (2006).
54. Barber, C. B., Dobkin, D. P. & Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* **22**, 469–483 (1996).
55. Sun, W. et al. The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2**, e1600225 (2016).
56. Wu, Y., Lazic, P., Hautier, G., Persson, K. & Ceder, G. First principles high throughput screening of oxynitrides for water-splitting photocatalysts. *Energy Environ. Sci.* **6**, 157–168 (2013).
57. Zakutayev, A. et al. Theoretical prediction and experimental realization of new stable inorganic materials using the inverse design approach. *J. Am. Chem. Soc.* **135**, 10048–10054 (2013).

58. Pal, K. et al. Accelerated discovery of a large family of quaternary chalcogenides with very low lattice thermal conductivity. *npj Comput. Mater.* **7**, 1–13 (2021).
59. Zhou, J.-P. et al. Properties of high k gate dielectric gadolinium oxide deposited on Si (1 0 0) by dual ion beam deposition (DIBD). *J. Cryst. Growth* **270**, 21–29 (2004).
60. Kwo, J. et al. Properties of high κ gate dielectrics Gd_2O_3 and Y_2O_3 for Si. *J. Appl. Phys.* **89**, 3920–3927 (2001).
61. Robertson, J. High dielectric constant gate oxides for metal oxide Si transistors. *Rep. Prog. Phys.* **69**, 327 (2005).
62. Keller, H.-L. Darstellung und kristallstruktur von hoch-Tl_3PbBr_5. *J. Less-Common Met.* **78**, 281–286 (1981).
63. Denysyuk, N. et al. Electronic structure of the high-temperature tetragonal Tl_3PbBr_5 phase. *J. Alloy. Compd.* **576**, 271–278 (2013).
64. Ferrier, A., Velázquez, M., Portier, X., Doualan, J.-L. & Moncorgé, R. Tl_3PbBr_5: a possible crystal candidate for middle infrared nonlinear optics. *J. Cryst. Growth* **289**, 357–365 (2006).
65. Qiu, Y. & Suh, S. Economic feasibility of recycling rare earth oxides from end-of-life lighting technologies. *Resour. Conserv. Recycl.* **150**, 104432 (2019).
66. Amato, A. et al. Sustainability analysis of innovative technologies for the rare earth elements recovery. *Renew. Sustain. Energy Rev.* **106**, 41–53 (2019).
67. Thornton, B. F. & Burdette, S. C. Homely holmium. *Nat. Chem.* **7**, 532–532 (2015).
68. Yaroshevsky, A. Abundances of chemical elements in the earth's crust. *Geochem. Int.* **44**, 48–55 (2006).
69. Wilkinson, M. D. et al. The fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9 (2016).
70. Chollet, F. et al. Keras. https://keras.io (2015).
71. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
72. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
73. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
74. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953 (1994).
75. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).

ACKNOWLEDGEMENTS

This work was funded by the SAMSUNG Global Research Outreach Program, and the U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD) award 70NANB14H012. We acknowledge the computing resources provided by (1) the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231, (2) Quest high-performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology, and (3) the Extreme Science and Engineering Discovery Environment (National Science Foundation Contract ACI-1548562).

AUTHOR CONTRIBUTIONS

A.G. devised computational strategies, wrote the manuscript, and conducted the calculations. K.P. provided important hands-on guidance in calculations and theoretical understanding. A.G. and C.W. modeled the project and analyzed the results. All authors have reviewed the manuscript.

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/s41524-022-00832-5.

Correspondence and requests for materials should be addressed to Chris Wolverton.

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

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

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

© The Author(s) 2022