SCIENTIFIC REPORTS
nature research

OPEN

Data-driven prediction of diamond-like infrared nonlinear optical crystals with targeting performances

Rui Wang[1], Fei Liang[1,2] & Zheshuai Lin[1,2*]

Combining high-throughput screening and machine learning models is a rapidly developed direction for the exploration of novel optoelectronic functional materials. Here, we employ random forests regression (RFR) model to investigate the second harmonic generation (SHG) coefficients of nonlinear optical crystals with distinct diamond-like (DL) structures. 61 DL structures in Inorganic Crystallographic Structure Database (ICSD) are selected, and four distinctive descriptors, including band gap, electronegativity, group volume and bond flexibility, are used to model and predict second-order nonlinearity. It is demonstrated that the RFR model has reached the first-principles calculation accuracy, and gives validated predictions for a variety of representative DL crystals. Additionally, this model shows promising applications to explore new crystal materials of quaternary DL system with superior mid-IR NLO performances. Two new potential NLO crystals, Li_2CuPS_4 with ultrawide bandgap and $Cu_2CdSnTe_4$ with giant SHG response, are identified by this model.

As a class of optoelectronic functional materials, nonlinear optical (NLO) crystals enable many important applied communities in laser frequency conversion, quantum information, optical communications and other fields[1–4]. Especially, in the mid-IR spectral range of 3–25 μm, as the fingerprint region of organic and inorganic molecules, the searching of good NLO crystals is in urgent demand to obtain coherent generation. To date, all commercial mid-IR NLO crystals are semiconductor chalcopyrites, such as $AgGaS_2$, $AgGaSe_2$ and $ZnGeP_2$[5]. However, they suffer from some intrinsic shortcomings, e.g., low laser damage threshold (LDT)[6], strong two-photon absorption[7], and difficult/dangerous crystal growth[8], which hinder their wide applications in high-power laser industry. Therefore, it is still a challenging task to promote and develop new mid-IR crystals with superior NLO performances.

Generally, a practical mid-IR NLO crystal should meet the following requirements[9]: (i) good IR transparency in the important mid-IR atmospheric window (3–5 μm and 8–12 μm); (ii) large SHG coefficient d_{ij} (at least larger than 10 × KDP ($d_{36} \approx 0.39$ pm/V) and at best larger than $AgGaS_2$ ($d_{36} \approx 13.4$ pm/V)); (iii) high LDT. For a good mid-IR NLO crystal, the energy band gap E_g should be more than 3.0 eV; (iv) moderate birefringence Δn ($\sim 0.03 - 0.10$); (v) easy crystal growth and chemical stability. It should be noted that a critical problem for the development of mid-IR NLO crystals is to fulfill the suitable balance between the large SHG response (d_{ij}) and enough band gap (E_g) in the light of their inverse dependence. Though many researchers aimed at mid-IR crystals and involved structure-property relationship in recent years, the traditional trial-and-error experiments and first-principle simulations are still time-consuming and laborious. Therefore, first-principles prediction combining high-throughput screening and machine learning is a burning issue for rapid development of mid-IR NLO crystals.

With the introduction of "Material Genome Project"[10], it has become a new research hotspot in the field of material science for combining High-Throughput Computing (HTC) with Machine Learning (ML) models to clarify the inherent structure-property relationship of materials and to accelerate the research and development of new materials, especially of functional materials, such as thermoelectric materials, low dimensional materials, solar cell, superconductors and superhard materials[11–18]. For example, in 2018, Brgoch *et al*. identified a new promising phosphor ($NaBaB_9O_{15}$) for solid state lighting via ML[19]. The further experiments verified this new

[1]University of Chinese Academy of Sciences, Beijing, 100190, China. [2]Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Beijing, 100190, China. *email: zslin@mail.ipc.ac.cn

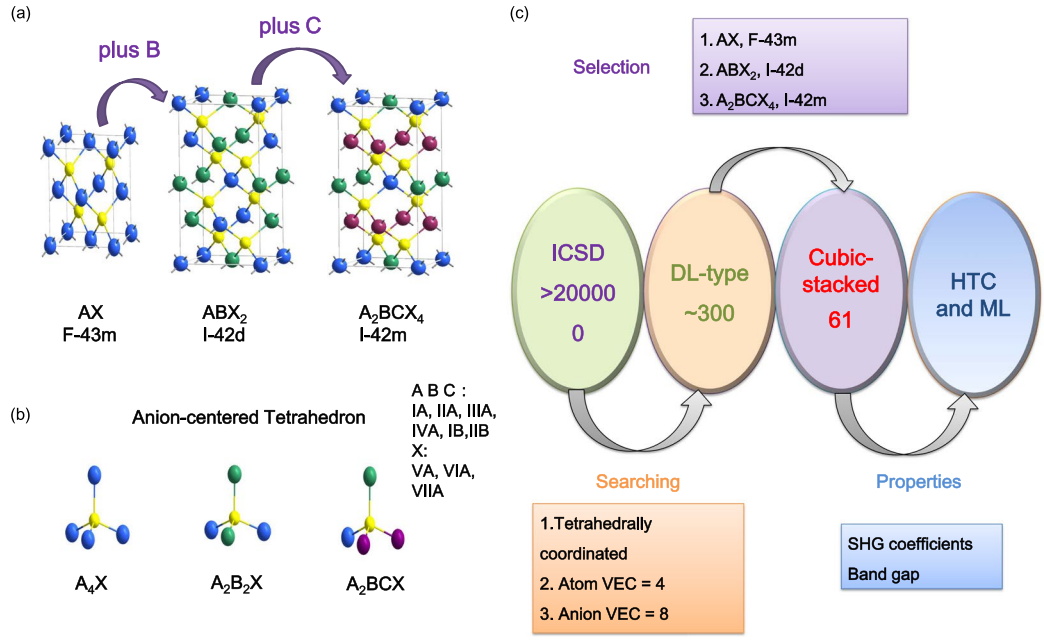

Figure 1. (**a**) Crystal structures of cubic-stacked DL compounds, AX, ABX_2 and A_2BCX_4. (**b**) The anion-centered tetrahedron as the basic unit in DL structures. (**c**) The screening workflow for DL-type crystals stacked cubically.

phosphor with a high quantum yield (95%) and excellent thermal stability. In 2019, Braatz *et al.* accurately applied machine-learning tools to predict and classify lithium-ion battery cycle life before capacity degradation with test errors less than 10%, providing a promising route for prognostics and diagnostics of lithium-ion batteries[20]. This work also highlighted the prospects of data-driven modeling to predict the behavior of complex systems.

Herein, for the first time, we employ ML models into the area of NLO crystals, aiming to provide more insights for exploring new mid-IR NLO crystals fulfilling the good balance between d_{ij} and E_g. In particular, the cubic close- packed diamond-like (DL) structures are selected to be the candidate dataset because of their promising NLO properties. The model's predictions on the SHG coefficients of commercial NLO DL crystals are in good agreement with first-principles simulations. Remarkably, two superior NLO crystals, Li_2CuPS_4 with a wide forbidden gap and $Cu_2CdSnTe_4$ with giant nonlinearity, are predicted and wait for further experimental verifications.

Dataset

In past few years, metal chalcogenides with DL structures have received increasing attention as they provide an attractive performance tuning dataset for a range of important applications, including lithium-ion conductors[21], band gap adjustable semiconductors[22], solar cells[23], thermoelectric materials[24] and nonlinear optics[25]. Several commercially NLO crystals belong to the DL structure. $AgGaS_2$ exhibits a large SHG effect (~13 pm/V) and possesses a wide transparent window (from 0.74 to 13 μm); it also acts as a standard crystal to evaluate other crystals' performance[26,27]. $ZnGeP_2$ (ZGP) is the most important and widely used NLO crystal in 3–5 μm due to its multiple advantages[28]. $CdGeAs_2$ owns the currently largest SHG coefficient ($d_{36} = 236$ pm/V) and has been successfully applied to the difference-frequency generation (DFG) and optical parametric generation (OPG)[6,29]. In 2017, Liang *et al.*[30] systematically summarized and analyzed the structure and property relationships in mid-IR NLO metal chalcogenides, and proposed that polar aligned DL structure would be the most favorable system due to its large band gap, sufficient SHG effect, moderate birefringence, and good crystal growth habit and chemical stability. After then, quite a few experiments demonstrated this prediction, such as for Hg_2GeSe_4[31], $Li_4HgGe_2S_7$[32], and Ga_2Se_3[33]. Therefore, the similar stacking type and abundant composition in the DL structure definitely enables the exploration for superior mid-IR crystals in the context of ML and HTC.

In this paper, we mainly considered three classes of DL crystals, AX, ABX_2 and A_2BCX_4 (A, B, C = IA, IIA, IIIA, IVA, IB, IIB cations; X = VA, VIA, VIIA anions) which belong to the space groups F-43m,I-42d and I-42m, respectively. As shown in Fig. 1a,b, the splitting of the cation site in AX leads to the ternary compound ABX_2 and quaternary compound $ABCX_4$, in which all of their basic structural units are tetrahedral. In these three kinds of crystal structures, all anion-centered tetrahedrons are arranged along the [111] direction, which is very beneficial for the superposition of microscopic second-order susceptibility of units. After a screening (Fig. 1c) in the Inorganic Crystallographic Structure Database (ICSD)[34], totally 61 DL structures (26 binary compounds, 29 ternary compounds and 6 quaternary compounds) were collected. It should note that all these compounds, except the quaternary compound Li_2CuPS_4, was experimentally synthesized and their crystal structure data has been determined.

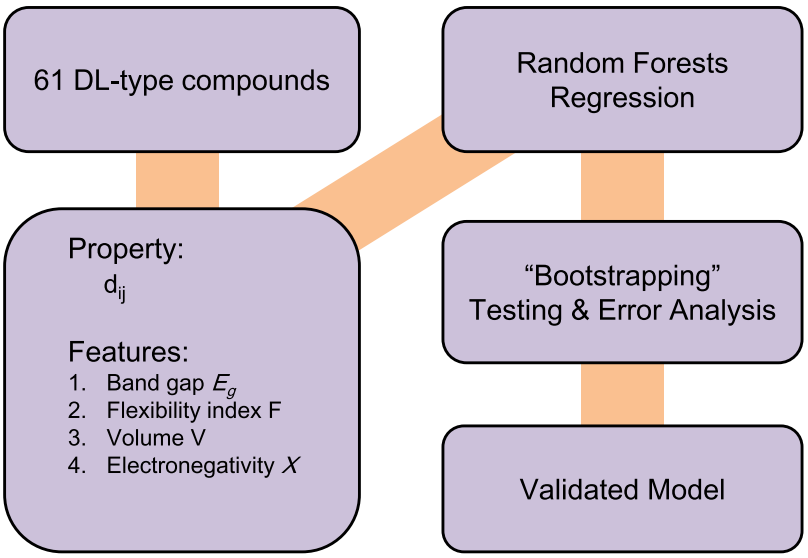

Figure 2. Overall workflow for the machine learning model. Four atomic or structural features are generated for the Random Forests Regression, in which "bootstrapping"[55] and performance evaluation are used to validate the model, leading to a predictive model for the SHG coefficient of NLO crystals.

First-Principles Methods

To provide the learning data for machine learning, the first-principles simulations were fulfilled by the plane-wave pseudopotential method[35] based on density functional theory (DFT)[36] implemented in the CASTEP module[37]. The cell parameters and atomic positions in the unit cells of all crystals were firstly optimized using the BFGS method[38] with the convergence criterion of 5×10^{-6} eV/atom, 0.01 eV/Å, 0.02 GPa, and 5.0×10^{-4} Å for energy change, maximum force, maximum stress, and maximum displacement, respectively, between two consecutive processes. The exchange-correlation functionals were described by the local density approximation (LDA)[39]. A kinetic energy cutoff 880 eV and Monkhorst-Pack k-point meshes[40] spanning less than 0.07 $Å^{-3}$ in the Brillouin zone were chosen. We used the screened-exchange local density approximation (sx-LDA)[41] in the calculation of electronic structure in order to obtain the accurate band gap E_g. The static SHG coefficients d_{ij} are calculated using an expression originally proposed by Rashkeev *et al.*[42]. It has been revealed from previous researches that our first-calculations provided a good estimate of NLO properties in IR sulfide crystals, as demonstrated in $LiGaS_2$[43], $AgGaS_2$[44], $BaGa_4S_7$[9], and $LiZnPS_4$[9]. The calculated values of band gaps E_g and SHG coefficients d_{ij} for the common DL NLO materials are listed in the supplementary information file.

Machine Learning Methods

Random forest regression. We propose to use Random Forests Regression (RFR) to predict the target variable, SHG coefficients in the selected NLO crystals (Fig. 2). For the RFR, the Scikit-learn package in Python was used[45].

RF is an integrated algorithm that combines multiple decision trees, with the advantages of good generalization performance, insensitivity to data outliers and fewer hyper parameters. RFR's training procedure was first proposed by Breiman[46]. (1) Acquire a random bootstrap sample from the data set. (2) For each bootstrap sample, nurture a tree with the following rule: at each node, find the best split point by a specific feature. The split criterion is to maximize the Information Gain (IG), which can be defined (for a binary split) as

$$IG(D_p, x_i) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right}) \quad (1)$$

x_i is the feature, N_p, N_{left} and N_{right} are the number of sample at the parent node and two child nodes, respective, I_p, I_{left} and N_{right} represent the impurity function for the parent and child nodes, respectively. The impurity index $I(t)$ at the node t can be calculated as

$$I(D_t) = MSE(t) = \frac{1}{N_t} \sum_{i \in D_t} (y^{(i)} - \hat{y}_t)^2 \quad (2)$$

N_t is the number of sample at t node, $y^{(i)}$ is the true target value and $\hat{y}_t$ is the average target value of the sample.

The RFR model uses the mean squared error (MSE) criteria to nurture every decision tree and the average value of the decision trees predicts the target variable.

Feature selection. The representations of a crystal, called "descriptors" or "features", play an indispensable role in applying ML model to predict its physical properties. In materials science, descriptors can be divided into elemental or structural representations[47]. The selection of descriptor candidates is an essential part in

constructing a ML model, which need not only satisfy the necessary specific requirements (e.g. dimensional invariance for chemical compositions), but also reflect physical meanings related to the target variable. Based on the existing knowledge, we chose four features to represent the DL crystals, *i.e.*, band gap, flexibility index, ionic group's volume and the electronegativity of anions.

(i) Energy band gap Eg. The band gap is a physical quantity closely related to the SHG coefficients. Generally, for the similar structure and composition, they have a negative correlation[48]. The band gap of a crystal can be acquired by first-principles simulations or experiments (e.g. Photoluminescence). Here, we choose the experimental value of the band gap as one of features.

(ii) Flexibility index F. In 2014, Jiang et al.[49] proposed a flexible dipole model based on the concept of bond-valence. The results showed that the magnitude of NLO effects was determined by the compliance of the dipole moment in response to external disturbances. "Flexibility index" between two connected atoms is defined as

$$F = \frac{exp[(R_0 - R_l)/C]}{(\sqrt{C_A} + \sqrt{C_B})^2/R_l^2} \times \frac{1}{(X_A - X_B)} \tag{3}$$

where R_l is the distance of two atoms, R_0 is the tabulated ideal bond length in the bond-valence theory, C_A (or C_B) is the number of valence electrons of atom A (or B), X_A (or X_B) is the electronegativity of atom A (or B) and C is an empirical constant, typically 0.37 Å. For a tetrahedron, the F was calculated by averaging four values of the coordinated atoms.

(iii) Volume V. The optical properties of a crystal have closely relations with the volume and density of groups[50]. Here we only consider the volume of the ionic groups (tetrahedra) because all crystal structures are in closely packed configurations.

(iv) Pauling Electronegativity PE. As the scale of the ability of atoms to attract electrons, the Pauling electronegativity is related to the strength of covalent bonds, which highly influence the SHG coefficients. For example, the substitution of S (PE = 2.58) to Se (PE = 2.55) to Te (PE = 2.10) enhances the SHG coefficients (from AGS (d_{36} = 13.4 pm/V), AGSe (d_{36} = 33.0 pm/V) to AGTe (d_{36} = 99.5 pm/V)). Therefore, we choose the electronegativity of the anion X as the forth feature.

Performance evaluation. The performance of the RFR model is evaluated through two common quantitative measures, root mean squared error (*RMSE*) and coefficient of determination (R^2)

$$RMSE = \sqrt{\sum_{i=1}^{m} \frac{1}{m}(f(x_i) - y_i)^2} \tag{4.1}$$

$$R^2 = 1 - \frac{MSE}{Var(y)} \tag{4.2}$$

$f(x_i)$ is the predicted value of the model, y_i is the target variable and $Var(y)$ is the variance of the sample data. A model with smaller *RMSE* and R^2 closer to 1 will have a higher level of prediction ability.

Results and Discussion

We selected 6 typical NLO crystals (CdSe, GaAs, $AgGaS_2$, $AgGaSe_2$, $ZnGeP_2$, $CdGeAs_2$) and 3 new quaternary compounds (Cu_2CdSnS_4, Li_2SrGeS_4, Li_2SrSnS_4) from the screened 61 DL structures as the test set. These crystals have been fully investigated so that the accurate SHG coefficients or credible powder SHG data were obtained. Thus, the RFR model's results with the first-principles and experimental values can be well compared to get a persuasive evaluation of the model. Meanwhile, another 4 compounds (Cu_3SbS_4, $Cu_2CdSnSe_4$, $Cu_2CdSnTe_4$, Li_2CuPS_4, also in the 61 DL structures), which don't have the experimental SHG values but have been investigated computationally, were added. So, the final test set includes 13 NLO crystals (listed in Table 1). The rest 48 crystals in the dataset were randomly split using "bootstrapping" to produce the training and validation set. After repeating 100 random "bootstrapping" and training, we found that the RFR model with the number of estimators 50 yielded a small *RMSE* as well as a high R^2 (indicated in the black circle in Fig. 3a). Then, the test set was input into the trained RFR model with 50 estimators. Figure 3b shows the RFR predicted results and their comparison with first-principles or experimental values.

The results in Fig. 3 and Table 1 demonstrate that the RFR model performs well on the studied DL NLO crystals. The predicted SHG coefficients are in good agreement with the first-principles values, proving that the RFR model has reached within the accuracy of first-principles simulations.

Notably, the RFR model does not need the scissors operator which is vital in the first-principles simulations. When calculating the optical properties of crystals with DFT, the scissors operator is usually used to shift upward all the conduction bands to agree with the experimental band gap[9]. But for those crystals with a small band gap E_g (<1 eV), this type of scissors sometimes make the first-principles simulation fail to give the accurate optical properties[51]. Another useful method to determine the scissors operator is by aligning the first peak of the calculated conduction bands with the corresponding experimental peak. In Table 1 the first-principles SHG coefficient values using these two kinds of scissors operators and the RFR results for three NLO crystals with small E_g ($CdGeAs_2$, $Cu_2CdSnSe_4$ and $Cu_2CdSnTe_4$) are listed. $CdGeAs_2$ is an important ternary NLO crystal because the experiments demonstrated that it has extremely high SHG coefficient (d_{36} = 236 pm/V)[29,51]. The first type

Formula	Space Group	E_g (eV)	SHG d_{ij}(pm/V)		
			P. v.[§]	C.v.[§]	E.v.[§]
CdSe	F-43m	1.74	39.18	42.16	40[56]
GaAs	F-43m	1.42	94.73	126.46	119[6]
$AgGaS_2$	I-42d	2.64	11.23	16.64	13.4[44]
$AgGaSe_2$	I-42d	1.80	73.83	67.99	41.4[44]
$ZnGeP_2$	I-42d	2.05	74.03	78.57	68.9[28]
$CdGeAs_2$	I-42d	0.57	254.32	194.04/(904.08*)	236[51]
Cu_2CdSnS_4	I-42m	1.80	34.46	25.42	31[57]
Li_2SrGeS_4	I-42m	3.75	5.48	4.75	0.5*AGS[58]
Li_2SrSnS_4	I-42m	3.10	5.76	6.64	0.8*AGS[58]
Cu_3SbS_4	I-42m	0.88	56.47	66.06	
$Cu_2CdSnSe_4$	I-42m	0.98	61.68	71.80/(223*)	
$Cu_2CdSnTe_4$	I-42m	0.80	239.05	209.05/(528*)	
Li_2CuPS_4	I-4	3.30	7.66	6.40	

Table 1. Space groups, band gaps and SHG coefficients of DL-type crystals in the test set. *First-principles value when upshifting the bands to agree the experimental band gap. §P.v., C.v. and E.v. refer to RFR predicted value, first-principles value and experimental value, respectively.

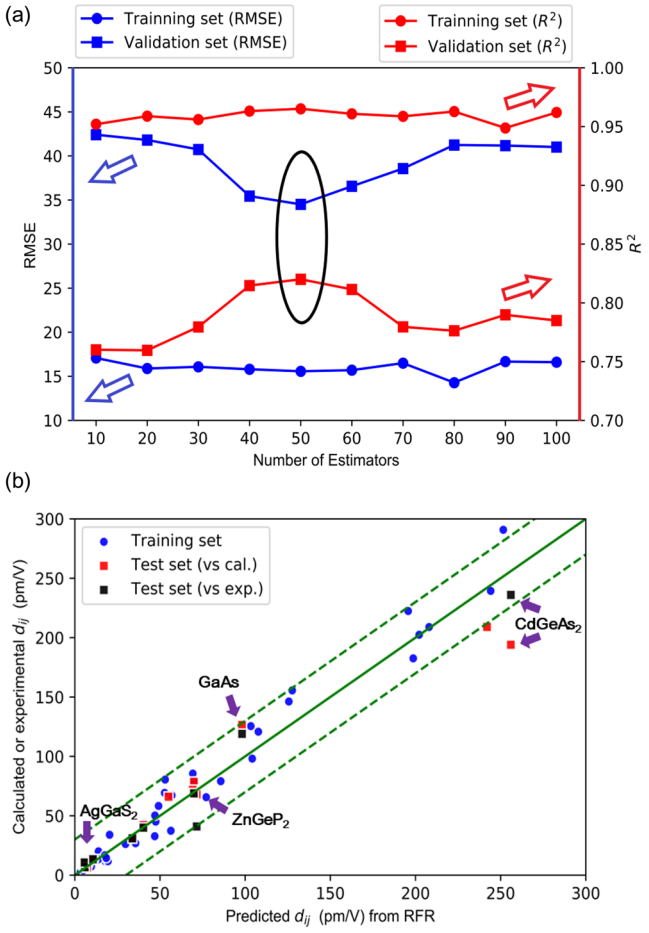

Figure 3. Performance evaluation and model prediction. (**a**) The *RMSE* and R^2 of training and validation set with the change of number of estimators. (**b**) Comparison of DFT training data or experimental data with RFR model predictions for d_{ij}. Blue circles represent the training and validation data; red squares represent the test data and the y axis is the calculated value; black squares represent the test data and the y axis is the experimental value. The error is less as the point approaching the green line (y = x).

of scissors operator gives $d_{36} = 904.08$ pm/V, while the second gives $d_{36} = 194.04$ pm/V. The remaining difference between the calculation and the experiment may come from the fact that the first-principle parameters are selected to balance the cost of time and accuracy. But clearly, the latter first-principles result agrees better with the experimental value, which is also correctly predicted by the RFR model (254.32 pm/V). The similar situations occur in the cases of $Cu_2CdSnSe_4$ and $Cu_2CdSnTe_4$; the comparison of the first-principles results by adopting the first and second types of scissors operators for these two compounds are 223 pm/V *vs*. 71.80 pm/V ($Cu_2CdSnSe_4$) and 528 pm/V *vs*. 209.05 pm/V ($Cu_2CdSnTe_4$), respectively. Considering the fact that the experimental value of Cu_2CdSnS_4 is $d_{36} = 31$ pm/V, the choice of the second type of scissors operator is more reasonable. In comparison, the RFR predicted SHG coefficients for $Cu_2CdSnSe_4$ and $Cu_2CdSnTe_4$ are $d_{36} = 61.68$ pm/V and 239.05 pm/V, respectively, which are also very reasonable and independent of the choice of scissors operator. Therefore, the RFR model bypasses the scissors operator problem presented in the first-principles calculations and can obtain the accurate predictions, since it performs on the basic chemical and physical information in crystals.

Moreover, it should be emphasized that our RFR model is successful for the prediction of SHG coefficients in considerable variation of chemical constituents from binary, ternary to quaternary crystals. Thus, this method has the good capability to explore new NLO crystals in the DL system. In particular, two new NLO crystals, Li_2CuPS_4 and $Cu_2CdSnTe_4$ with superior mid-IR NLO performances were identified by this model.

$Cu_2CdSnTe_4$ crystallizes in the stannite structure type with the space group I-42m. Dong *et al*.[52] synthesized this compound by direct reaction of the corresponding elements and discussed its temperature dependent transport properties. In addition, the structural, optoelectronic and thermoelectric properties of $Cu_2CdSnTe_4$ were theoretically studied, which showed that this compound is a potential candidate in the fields of solar cell and thermoelectric[53]. However, the NLO properties of $Cu_2CdSnTe_4$ have not been paid attention. We predict that $Cu_2CdSnTe_4$ is a promising NLO crystal with a band gap 0.8 eV and the SHG coefficient value $d_{36} = 239.05$ pm/V. As a comparison, $CdGeAs_2$ owns the currently largest SHG coefficient value ($d_{36} = 236$ pm/V) in known inorganic materials, but its band gap is only 0.57 eV. Therefore, $Cu_2CdSnTe_4$ is expected to have a higher LDT than $CdGeAs_2$ in the case of comparable SHG effects. We encourage further experiments to verify our predictions.

Li_2CuPS_4 was predicted by Zhu *et al*.[54] as a sulfide-based super ionic conductor with the kesterite structurel. The higher ionic conductivity arising from the smaller Li ion binding and the reduced electronegativity difference between the anion element and non-lithium cation elements enables Li_2CuPS_4 as a promising solid-state electrolyte material. Though this crystal hasn't been synthesized experimentally, the first-principle calculations have provided the reliable structure information and electronic properties. Its calculated band gap is 3.3 eV using the HSE06 hybrid functional. After inputting the related features of Li_2CuPS_4, our model predicts that it has the SHG coefficient 7.66 pm/V (~0.6 × AGS), which can be a potential NLO crystal once experimentally synthesized. Compared to AGS ($E_g = 2.64$ eV), Li_2CuPS_4 is likely to have the smaller SHG response but the higher LDT, which accords with the inverse relationship between the band gap and the SHG response.

Conclusion

An HTC and ML workflow has been designed and performed to screen for DL crystals with good balance on the band gap and the SHG coefficient. By selecting four distinctive descriptors, i.e., band gap, electronegativity, group volume and bond flexibility, the predicted results using RFR model are in good agreement with the first-principle calculations, especially on some representative crystals like $AgGaS_2$, $ZnGeP_2$ and $CdGeAs_2$. More interestingly, this fast workflow is independent of the selection of scissors operators, making it more practical. Additionally, this model can be used to facilitate the research of new DL systems. Two unexplored quaternary crystals with good NLO properties, Li_2CuPS_4 and $Cu_2CdSnTe_4$, are predicted by this model and wait for the experimental investigations. In summary, this new method opens opportunities for the fast design of NLO crystals with targeting properties.

Received: 10 September 2019; Accepted: 11 February 2020;
Published online: 26 February 2020

References

1. Beck, M. *et al*. Continuous wave operation of a mid-infrared semiconductor laser at room temperature. *Science* **295**, 301–305 (2002).
2. Pushkarsky, M. B. *et al*. High-sensitivity detection of TNT. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 19630–19634 (2006).
3. Boskey, A. & Camacho, N. P. FT-IR imaging of native and tissue-engineered bone and cartilage. *Biomaterials* **28**, 2465–2478 (2007).
4. Petrov, V., Rempel, C., Stolberg, K. P. & Schade, W. Widely Tunable Continuous-Wave Mid-Infrared Laser Source Based on Difference-Frequency Generation in $AgGaS_2$. *Appl Opt* **37**, 4925–4928 (1998).
5. Ohmer, M. C. & Ravindra, P. Emergence of Chalcopyrites as Nonlinear Optical Materials. *Mrs Bulletin* **23**, 16–22 (1998).
6. Nikogosyan, D. N. *Nonlinear Optical Crystals: A Complete Survey*. (Springer, New York, NY, 2005).
7. Schunemann, P. G. Crystal Growth and Properties of Nonlinear Optical Materials. *AIP Conference Proceedings* **916**, 541 (2007).
8. Verozubova, G. A., Gribenyukov, A. I., Ohmer, M. C., Fernelius, N. C. & Goldstein, J. T. Growth and characterization of epitaxial films of $ZnGeP_2$. *Mrs Proceedings* **744**, M8.46.41–46.47 (2002).
9. Kang, L. *et al*. Metal Thiophosphates with Good Mid-infrared Nonlinear Optical Performances: A First-Principles Prediction and Analysis. *Journal of the American Chemical Society* **137**, 13049–13059 (2015).
10. Jain, A., Ong, S. P., Hautier, G., Wei, C. & Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *Apl Materials* **1**, 1049 (2013).
11. Zhu, H. *et al*. Computational and experimental investigation of $TmAgTe_2$ and XYZ_2 compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening. *Journal of Materials Chemistry C* **3**, 10554–10565 (2015).
12. Oliynyk, A. O. *et al*. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chemistry of Materials* **28**, 7324–7331 (2016).
13. Pilania, G. *et al*. Machine learning bandgaps of double perovskites. *Sci Rep* **6**, 19375 (2016).
14. van Roekeghem, A., Carrete, J., Oses, C., Curtarolo, S. & Mingo, N. High-Throughput Computation of Thermal Conductivity of High-Temperature Solid Phases: The Case of Oxide and Fluoride Perovskites. *Physical Review X* **6**, 041061 (2016).

15. Jalem, R. *et al*. A general representation scheme for crystalline solids based on Voronoi-tessellation real feature values and atomic property data. *Sci. Technol. Adv. Mater.* **19**, 231–242 (2018).
16. Legrain, F., Carrete, J., van Roekeghem, A., Madsen, G. K. H. & Mingo, N. Materials Screening for the Discovery of New Half-Heuslers: Machine Learning versus ab Initio Methods. *Journal of Physical Chemistry B* **122**, 625–632 (2018).
17. Stanev, V. *et al*. Machine Learning Modeling of Superconducting Critical Temperature. *Npj Computational Materials* **4**, 29 (2018).
18. Zhang, T. *et al*. Catalogue of Topological Electronic Materials. *Nature* **566**, 475-+ (2019).
19. Zhuo, Y., Tehrani, A. M., Oliynyk, A. O., Duke, A. C. & Brgoch, J. Identifying an efficient, thermally robust inorganic phosphor host via machine learning. *Nature Communications* **9**, 4377 (2018).
20. Severson, K. A. *et al*. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4**, 383–391 (2019).
21. Brant, J. A. *et al*. A new class of lithium ion conductors with tunable structures and compositions: Quaternary diamond-like thiogermanates. *Solid State Ionics* **278**, 268–274 (2015).
22. Ford, G. M., Guo, Q. J., Agrawal, R. & Hillhouse, H. W. Earth Abundant Element $Cu_2Zn(Sn_{1-x}Ge_x)S_4$ Nanocrystals for Tunable Band Gap Solar Cells: 6.8% Efficient Device Fabrication. *Chemistry of Materials* **23**, 2626–2629 (2011).
23. Guo, Q. *et al*. Fabrication of 7.2% Efficient CZTSSe Solar Cells Using CZTS Nanocrystals. *Journal of the American Chemical Society* **132**, 17384–17386 (2010).
24. Li, R. *et al*. High-Throughput Screening for Advanced Thermoelectric Materials: Diamond-Like ABX_2 Compounds. *ACS applied materials & interfaces* **11**, 24859–24866 (2019).
25. Liang, F., Kang, L., Lin, Z. S., Wu, Y. C. & Chen, C. T. Analysis and Prediction of Mid-IR Nonlinear Optical Metal Sulfides with Diamond-like Structures. *Coord. Chem. Rev.* **333**, 57–70 (2017).
26. Roberts, D. A. Dispersion equations for nonlinear optical crystals: KDP, $AgGaSe_2$, and $AgGaS_2$. *Applied Optics* **35**, 4677–4688, https://doi.org/10.1364/ao.35.004677 (1996).
27. Reshak, A. H. Linear, nonlinear optical properties and birefringence of $AgGaX_2$ (X = S, Se, Te) compounds. *Physica B* **369**, 243–253 (2005).
28. Kato, K. Second-harmonic and sum-frequency generation in $ZnGeP_2$. *Applied Optics* **36**, 2506–2510 (1997).
29. Schunemann, P. G. & Pollak, T. M. Single crystal growth of large, crack-free $CdCeAs_2$. *Journal of Crystal Growth* **174**, 272–277 (1997).
30. Liang, F., Kang, L., Lin, Z. S. & Wu, Y. C. Mid-Infrared Nonlinear Optical Materials Based on Metal Chalcogenides: Structure-Property Relationship. *Cryst. Growth Des.* **17**, 2254–2289 (2017).
31. Guo, Y. W. *et al*. Nonbonding Electrons Driven Strong SHG Effect in Hg_2GeSe_4: Experimental and Theoretical Investigations. *Inorganic Chemistry* **57**, 6795–6798 (2018).
32. Wu, K., Yang, Z. H. & Pan, S. The first quaternary diamond-like semiconductor with 10-membered LiS_4 rings exhibiting excellent nonlinear optical performances. *Chemical Communications* **53**, 3010–3013 (2017).
33. Guo, S. P. *et al*. Large Second Harmonic Generation (SHG) Effect and High Laser-Induced Damage Threshold (LIDT) Observed Coexisting in Gallium Selenide. *Angew. Chem.-Int. Edit.* **58**, 8087–8091 (2019).
34. Alec, B., Mariette, H., Vicky Lynn, K. & Peter, L. New developments in the Inorganic Crystal Structure. *Database (ICSD): accessibility in support of materials research and design. Acta Crystallographica* **58**, 364–369 (2010).
35. Payne, M. C., Arias, T. A. & Joannopoulos, J. D. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Reviews of Modern Physics; (United States)* **64**(4), 1045–1097 (1992).
36. Kohn & W. Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics* **71**, 1253–1266 (1999).
37. Milman, V., Refson, K., Clark, S. J., Pickard, C. J. & Segall, M. D. Electron and vibrational spectroscopies using DFT, plane waves and pseudopotentials: CASTEP implementation. *Journal of Molecular Structure Theochem* **954**, 22–35 (2010).
38. Pfrommer, B. G., Coté, M., Louie, S. G. & Cohen, M. L. Relaxation of Crystals with the Quasi-Newton Method. *Journal of Computational Physics* **131**, 233–240 (1997).
39. Rappe, A. M., Rabe, K. M., Kaxiras, E. & Joannopoulos, J. D. Optimized pseudopotentials. *Physical Review B Condensed Matter* **41**, 1227–1230 (1990).
40. Pack, J. D. & Monkhorst, H. J. "Special points for Brillouin-zone integrations"— a reply. *Physical Review B Condensed Matter* **16**, 1748–1749 (1976).
41. Asahi, R., Mannstadt, W. & Freeman, A. Optical properties and electronic structures of semiconductors with screened-exchange LDA. *Applied Physics Letters* **21**, 165–176 (1999).
42. Rashkeev, S. N., Lambrecht, W. R. L. & Segall, B. Efficient ab-initio method for the calculation of frequency dependent non-linear optical response in semiconductors: application to second harmonic generation. *Physics* **46**, 3848–3859 (1997).
43. Bai, L., Lin, Z. S., Wang, Z. Z. & Chen, C. T. Mechanism of Linear and Nonlinear Optical Effects of Chalcopyrites $LiGaX_2$ (X = S, Se, and Te) Crystals. *J. Appl. Phys.* **103**, 083111 (2008).
44. Bai, L., Lin, Z. S., Wang, Z. Z., Chen, C. T. & Lee, M. H. Mechanism of Linear and Nonlinear Optical Effects of Chalcopyrite $AgGaX_2$ (X = S, Se, and Te) Crystals. *Journal of Chemical Physics* **120**, 8772–8778 (2004).
45. Swami, A. & Jain, R. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2013).
46. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
47. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 11 (2017).
48. Jackson, A. G., Ohmer, M. C. & LeClair, S. R. Relationship of the second order nonlinear optical coefficient to energy gap in inorganic non-centrosymmetric crystals. *Infrared Phys. Technol.* **38**, 233–244 (1997).
49. Jiang, X. *et al*. The Role of Dipole Moment in Determining the Nonlinear Optical Behavior of Materials: Ab-initio Studies on Quaternary Molybdenum Tellurite. *Crystals. J. Mater. Chem. C* **2**, 530–537 (2014).
50. Kang, L. *et al*. Ab initio studies on the optical effects in the deep ultraviolet nonlinear optical crystals of the $KBe_2BO_3F_2$ family. *J Phys Condens Matter* **24**, 335503 (2012).
51. Yu, Y. *et al*. Ab Initio Study of the Linear and Nonlinear Optical Properties of Chalcopyrite $CdGeAs_2$. *Journal of Solid State Chemistry* **185**, 264–270 (2012).
52. Dong, Y. *et al*. Synthesis, transport properties, and electronic structure of Cu2CdSnTe4. *Applied Physics Letters* **104**, 252107 (2014).
53. Hussain, S. *et al*. First principles study of structural, optoelectronic and thermoelectric properties of Cu2CdSnX4 (X = S, Se, Te) chalcogenides. *Materials Research Bulletin* **79**, 73–83 (2016).
54. Xu, Z., Chen, R. & Zhu, H. Li_2CuPS_4 Superionic Conductor: A New Sulfide-Based Solid-State Electrolyte. *J. Mater. Chem. A* **7**, 12645–12653 (2019).
55. Efron, B. & Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science* **1**, 54–75 (1986).
56. Boyd, G. D., Buehler, E. & Storz, F. G. Linear and Nonlinear Optical Properties of $ZnGeP_2$ and CdSe. *Applied Physics Letters* **18**, 301-& (1971).
57. Rosmus, K. A. *et al*. Optical Nonlinearity in Cu_2CdSnS_4 and alpha/beta-Cu_2ZnSiS_4: Diamond-like Semiconductors with High Laser-Damage Thresholds. *Inorganic Chemistry* **53**, 7809–7811 (2014).
58. Wu, K., Chu, Y., Yang, Z. & Pan, S. $A_2SrM^{IV}S_4$ (A = Li, Na; M^{IV} = Ge, Sn) Concurrently Exhibiting Wide Bandgaps and Good Nonlinear Optical Responses as New Potential Infrared Nonlinear Optical Materials. *Chemical Science* (2019).

Acknowledgements
This work was supported by National Natural Science Foundation of China (51872297, 51702330, 51890864 and 51802321), and Fujian Institute of Innovation (FJCXY18010201) in CAS. Z.S. Lin acknowledges the support from the Youth Innovation Promotion Association, CAS.

Author contributions
Z.L. and F.L. conceived the research, R.W. wrote the code and most of the manuscript. R.W. and F.L. built the data sets and designed the study. All author analyzed the results, reviewed and edited the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-60410-x.

Correspondence and requests for materials should be addressed to Z.L.

Reprints and permissions information is available at www.nature.com/reprints.

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