PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d2cp04393a

Establishment and validation of an electron inelastic mean free path database for narrow bandgap inorganic compounds with a machine learning approach†

Xun Liu,[ab] Dabao Lu,[ab] Zhufeng Hou,[c] Kenji Nagata,[b] Bo Da,*[b] Hideki Yoshikawa,[b] Shigeo Tanuma,[d] Yang Sun[e] and Zejun Ding *[a]

Received 21st September 2022,
Accepted 29th May 2023

DOI: 10.1039/d2cp04393a

rsc.li/pccp

Narrow bandgap inorganic compounds are extremely important in many areas of physics. However, their basic parameter database for surface analysis is incomplete. Electron inelastic mean free paths (IMFPs) are important parameters in surface analysis methods, such as electron spectroscopy and electron microscopy. Our previous research has presented a machine learning (ML) method to describe and predict IMFPs from calculated IMFPs for 41 elemental solids. This paper extends the use of the same machine learning method to 42 inorganic compounds based on the experience in predicting elemental electron IMFPs. The in-depth discussion extends to including material dependence discussion and parameter value selections. After robust validation of the ML method, we have produced an extensive IMFP database for 12 039 narrow bandgap inorganic compounds. Our findings suggest that ML is very efficient and powerful for IMFP description and database completion for various materials and has many advantages, including stability and convenience, over traditional methods.

1. Introduction

Narrow bandgap inorganic compounds, especially semiconductors are frequently used nowadays in nano electronic devices.[1–3] In studying these materials, the inelastic mean free path (IMFP)[4] is one of the fundamental descriptors for surface analysis techniques based on electron scattering. The IMFP describes the mean distance that an electron travels through a solid before losing energy and plays a critical role in experimental observation facilities, such as reflection electron energy-loss spectroscopy (REELS),[5–13] X-ray photoelectron spectroscopy (XPS),[14–16] and Auger electron spectroscopy (AES).[14,17–19] Monte Carlo simulations[20–26] have shown how incident electrons scatter in materials, together with other important parameters used in surface science, such as the backscattering factor,[27,28] mean escape depth,[29] and surface excitation parameters.[30–32]

Their importance in applications makes it essential to calculate electron IMFPs at electron energies above 50 eV. Some accepted methods, such as the full-Penn algorithm (FPA),[33] Mermin algorithm,[34] and the super extended Mermin algorithm (SE-MA),[35–37] all depend on the dielectric theory. For instance, the pioneering FPA was used to establish a significant part of the IMFP database[38–43] because the FPA is currently the most systematic algorithm in applications.

From accumulated experience during applications, FPA is now very comprehensive despite some accuracy problems in specific energy regions. Specifically, the Lindhard dielectric function is used in the FPA calculation for describing the probability of electron inelastic scattering, although the finite lifetime broadening of the plasmon is neglected here.[33] In addition, the exchange–correlation potential is uncertain,[39] and the transverse differential cross section is neglected in the inelastic scattering process.[42] However, these limitations are often inconsequential to the broad application of the FPA because these uncertainties or approximations only cause FPA-calculated IMFPs to be unreliable for energies less than 50 eV and above 200 keV. Considering that practical applications of IMFPs, especially calculated IMFPs, are mainly for AES and XPS and that these applications are not related to the very low (<50 eV) and very high (>200 keV) energy ranges, the FPA

[a] Department of Physics, University of Science and Technology of China, Hefei, Anhui, 230026, P. R. China. E-mail: zjding@ustc.edu.cn
[b] Center for Basic Research on Materials, National Institute for Materials Science, Tsukuba, Ibaraki 305-0044, Japan. E-mail: DA.Bo@nims.go.jp
[c] State Key Laboratory of Structural Chemistry, Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Fuzhou 350002, China
[d] Research Network and Facility Services Division, National Institute for Materials Science, Tsukuba, Ibaraki 305-0044, Japan
[e] Department of Physics, Xiamen University, Xiamen, Fujian 361-005, China

† Electronic supplementary information (ESI) available: The complete list of the 12 039 narrow bandgap inorganic compounds (E_g < 3 eV) with their Material IDs and the predicted IMFP database. See DOI: https://doi.org/10.1039/d2cp04393a

generally meets the requirements and is very convenient for use. Moreover, the calculational availability or accuracy of FPA directly connected to the existence or reliability of experimental optical constants to derive the energy loss function (ELF).[44] Some of the ELFs for different materials[12,13,45] were already extracted from the REELS experiments with the most accurate solution to date.[11,23,24] However, some materials still lack experimental data of optical constants.

For example, the surfaces of some metals are frequently contaminated during material preparation, but reliable optical constant determination requires a very clean material surface. Often, the IMFPs cannot be calculated from uncertain ELFs because of unsatisfactory experimental capabilities or facilities. In addition, the combination of elements in inorganic compounds is tremendous, making it impossible to obtain data experimentally for all inorganic compounds. Therefore, developing predictive theoretical methods through elemental material-dependent parameters without reliance on optical constants is very necessary.

Previous researches have made some prediction attempts for IMFPs using elemental material-dependent parameters in different approaches. The most natural idea is to develop empirical formulas with the least-squares algorithm on known IMFP materials.[28–30] However, these formulas are mostly applicable only in the high-energy ranges. For example, Gries proposed the G1 formula[46] using an atomistic model:

$$\lambda = \frac{k_1 V_a E}{Z^*(\log E - k_2)}\ (\mathrm{nm}), \qquad (1)$$

where $V_a = M/\rho$ is the atomic volume (M is the atomic mass and ρ is the bulk density), and $Z^* = Z^{0.5}$ is the nominal effective number of interaction-prone electrons per atom and Z is the atomic number. The average values per atom for M and Z^* are used for inorganic compounds, and k_1 and k_2 are fitting parameters, which are the most troublesome in this formula. The original work[46] determined the fitting parameter values according to the materials' position in the periodic table. Although Tanuma *et al.*[47] have done an additional research to provide more appropriate values based on Gries' work, the use of method is inconvenient. Here we note that the G1 formula is based on the early IMFPs calculations of Tanuma *et al.* that have now been superseded by the calculations of Shinotsuka *et al.*[41,42] Moreover, although the G1 formula works well for most of materials, large deviations were found for some specific materials.

Other analytical expressions have been proposed using different approaches. For example, the expression for estimating IMFPs developed by Seah,[48] called the S1 equation, is given by:

$$\lambda = \frac{\left(4 + 0.44Z^{0.5} + 0.104E^{0.872}\right)a^{1.7}}{Z^{0.3}(1 - W)}\ (\mathrm{nm}), \qquad (2a)$$

$$a^3 = \frac{10^{21} M}{\rho N_A (g + h)}, \qquad (2b)$$

where $W = 0.02E_g$, E_g is the band gap energy ($W = 0$ for an elemental solid) and N_A is the Avogadro constant. The terms g and h in eqn (2b) represent stoichiometry coefficients for an assumed binary inorganic compound G_gH_h; therefore, $g = 1$ and $h = 0$ for an element. The S1 formula has satisfactory accuracy for elemental solids, but, it is not well adapted to inorganic compounds. In addition, the S1 formula cannot be applied to multiple inorganic compounds like $Y_3Al_5O_{12}$. Apart from its accuracy, the S1 formula is not guided by the Bethe equation[49] from its form and thus does not provide obvious physical insight.

The TPP-2M formula has been the most popular of the empirical formulas to date. The TPP-2M formula was derived from the Bethe equation,[49] and is given by

$$\begin{aligned}
&\lambda = \frac{\alpha E}{E_p^2\left[\beta_r \ln(\alpha\gamma_r E) - \frac{C_r}{E} + \frac{D_r}{E^2}\right]} && (\mathrm{nm})\\
&\alpha = \left[1 + \left(E/2m_ec^2\right)\right]\Big/\left[1 + \left(E/m_ec^2\right)\right]^2 &&\\
&\beta_r = -1.0 + 9.44\Big/\left(E_p^2 + E_g^2\right)^{0.5} + 0.69\rho^{0.1} && \left(\mathrm{eV^{-1}\,nm^{-1}}\right)\\
&\gamma_r = 0.191\rho^{-0.5} && \left(\mathrm{eV^{-1}}\right)\\
&C_r = 19.7 - 9.1U && \left(\mathrm{nm^{-1}}\right)\\
&D_r = 534 - 208U && \left(\mathrm{eV\,nm^{-1}}\right)\\
&U = \frac{N_v\rho}{M} = \left(E_p/28.816\right)^2 &&
\end{aligned}, \qquad (3)$$

where E is the electron kinetic energy (in eV), m_ec^2 = 510 998.9 eV is the electron rest energy, E_p is the free-electron plasmon energy (in eV), E_g is the band gap energy for non-conductors (in eV), ρ is the bulk density (in g cm^{-3}), and N_v is the number of valence electrons per atom or molecule.

Many adjustments have been introduced to expand the applicability of the TPP-2M formula, while the most recent adjustment is done by Shinotsuka *et al.*[42] A relativistic version of the TPP-2M formula, given by eqn (3), was produced based on the modified Bethe equation. There are four main fitting parameters in the updated formula. First, the expressions for β and γ were confirmed, and because IMFPs for energies less than 200 eV cannot be accurately described by only two fitting parameters, new fitting parameters C and D were also introduced.[40] A term $\alpha(E)$ for relativistic high energy situations was added with new fitting parameters expressed as β_r, γ_r, C_r and D_r, to extend the formula to very high energy situations. This relativistic version of the TPP-2M formula was found to enhance accuracy, making the TPP-2M formula[43] very powerful for calculating IMFPs in the electron energy range above 50 eV.

The target values used to develop the relativistic TPP-2M formula[43] were, however, FPA-calculated IMFP data. Notably, because of the uncertainty in the exchange–correlation potential[39] and because the transverse differential cross section for inelastic scattering was neglected,[42] FPA-calculated IMFPs have relatively poor accuracy at energies below 50 eV and above 200 keV. As FPA calculated-IMFP data is the base of the TPP-2M formula, similarly,

TPP-2M formula cannot be used in the very low energy (<50 eV) range. Despite its shortcomings in some special low energy (<50 eV) applications, the TPP-2M formula is no doubt another step in describing IMFPs and building relationships for material-dependent parameters. Using the TPP-2M formula, large IMFP tables can be developed, providing significant convenience in searching for IMFP values.

Artificial empirical formulas have certain shortcomings, therefore, ML algorithms have been introduced to develop new empirical formulas. Most previous IMFP equations have coefficients chosen by human judgment and thus cannot achieve the most precise description and the best predictive power when based on the modified Bethe equation. Our previous work has used the least absolute shrinkage and selection operator, LASSO, to update the empirical formula for the same IMFP database.[50] LASSO is a widely-used ML algorithm that can automatically eliminate the unimportant terms, thus pick out the principle descriptors to form a linear combination. The produced TPP-LASSO formula have dozen of items, so we refit the formula and make it more concise while without large loss of accuracy. The refitted new TPP-LASSO-S formula is:

$$\lambda = \frac{\alpha E}{E_p^2 \beta_r \ln(\alpha \gamma_r E)}\ (\text{Å}), \tag{4a}$$

$$\beta_r = -0.0012 + 0.046\left(\frac{M}{\rho N_v}\right)^{0.5} - 0.035\left(\frac{M}{\rho N_v}\right)^{0.4} + 0.0019\frac{Z}{N_v}, \tag{4b}$$

$$\gamma_r = -0.07 + 0.26\left[\rho\left(E_i + E_g\right)\right]^{-0.2} + 0.066\left(\frac{Z\rho}{M}\right)^{-0.8}, \tag{4c}$$

where E_i is the starting-point energy (in eV). For TPP-LASSO and TPP-LASSO-S formulae, more details can be found in ref. 50.

Although some ML algorithms have been introduced into the fitting process, the above-mentioned IMFP equations are still empirical formulas, and all are applicable only in the high-energy region (>50 eV). ML algorithms such as the Gaussian process regressor (GPR)[51] have demonstrated high descriptive and predictive ability in our previous work,[52] where it was demonstrated that the GPR has a strong descriptive capability based on a credible IMFP database. This conclusion was based on a thorough comparison with the robust TPP-2M formula.[52] Although the GPR cannot provide a direct formula, it has a wide application range and improved predictive accuracy. On the basis of the successful use of the GPR for IMFPs of elemental solids, the GPR is applied here to the IMFPs of inorganic compounds, and some unique characteristics of the GPR, especially related to its description and prediction of IMFPs, are also discussed.

There is no large, robust IMFP database for narrow bandgap inorganic compounds applications. Research on IMFPs for only a few narrow bandgap inorganic compounds has been reported, such as the work of Jablonski *et al.*[53] for GaP, GaSb, InP, InSb, and Si_3N_4. Shinotsuka *et al.*[54] reported first-principle-calculated energy loss functions (ELFs), which are directly related to IMFPs, for 5 elemental and 30 inorganic compound semiconductors.

In this work, we first expanded the ML algorithm to describe and predict IMFPs of inorganic compounds. Different databases are compared when used in the ML model training. Moreover, we successfully stripped out the energy dependence when predicting IMFPs. In this way, scattered experimental IMFPs can be used in the ML training data, meaning that ML has a broader application range than artificial empirical formulas. Finally, an extensive IMFP database for 12 039 narrow bandgap inorganic compounds has been established using the trained model. Robustness is demonstrated by comparing the ML data with well-known empirical formulas and analysis with another ML framework. During the ML process, we also found that the distribution behavior of E_g is different from other parameters, which may show the important effects of E_g when predicting narrow bandgap inorganic compound IMFPs.

2. Theoretical method and results

2.1 Use of GPR to describe IMFPs

When using ML, the training algorithm and database are the two main parts determining its performance.[55] For the ML algorithm, we have proven that of familiar ML algorithms, the GPR is suitable for IMFP description and prediction.[52] The GPR model is a probabilistic model belonging to a generic supervised learning method that provides a probabilistic distribution of a new output value from the descriptors based on each step's training result. In the step-by-step optimization, the joint distributions of the regressed function follow a Gaussian process. For available distributions of the function and targets, the posterior distribution, namely the distribution of regressed values for IMFPs in this work, is calculated through this Gaussian process.

We utilized the IMFPs for 42 inorganic compounds calculated by Shinotsuka *et al.*[43] They used the full-Penn algorithm (FPA)[33] in their calculations, relying on an accurate determination of the energy loss function (ELF). The pioneering FPA treats the ELF as the sum of Lindhard dielectric functions when calculating IMFPs. This database is expected to be very reliable because the FPA-calculated IMFP database calculated by Shinotsuka *et al.*[42] was also used in a previous ML study to calculate the IMFPs of elemental solids.[52]

The ML parameters used previously[52] are (average) atomic number (Z), (average) atomic mass (M), density (ρ), number of valence electrons per atom or molecule (N_v),[56] free-electron plasmon energy (E_p), band gap energy (E_g), starting-point energy (E_i), and (average) atomic radius (R). These are almost the same as those used in ML for elemental solids,[52] except that the Fermi energy (E_F) is expanded to E_i. The energy E_i, which equals the valence bandwidth plus the band gap energy, is used for inorganic compounds, while E_F is used for elemental solids. This difference originates from the different reference energies used in the original FPA calculation of the IMFP training database used in this work. Specifically, the upper limit is

($E - E_F$) for IMFPs of elemental solids,[42] while it is ($E - E_v - E_g$) for inorganic compounds.[43] For semiconductors and insulators, the location of the Fermi energy is ambiguous. In ref. 43, the x axis in plots of IMFP *versus* energy is shown as the electron energy with respect to the bottom of the conduction band. Using this criterion, we can avoid the ambiguity of the Fermi energy. Also, IMFP rises sharply at low kinetic energies. If we use the Fermi energy as the reference on the x-axis, the position of the rise is easily shifted by the band gap energy. If we use the bottom of the conduction band as the reference on the x-axis, the position of the rise is approximately similar. When discussing the shape of the IMFP in ref. 43 near the rise of different materials, it is more convenient to use the bottom of the conduction band as a reference than to use the Fermi energy. This has been validated and applied in previous work unifying the band structure description[50] and is also discussed in ref. 52.

We note that for the unavailable parameter values Z, M, and R for inorganic compounds, the average values for each atom in each molecule are used considering the unity of all materials at the atomic scale. In addition, we also attempt to introduce the (average) electron shell number (n_s) and the (average) outermost shell electron number (N_o) as parameters. We include these new parameters to capture the periodic characteristics of the elements in the material components because they are directly connected to the relationships between elements' locations in the periodic table. The parameter values used in this work can be found in Table 1. Considering that the results in previous work were acceptable, the databases from the same group are also expected to be reliable.

This work presents a new challenge because there are two different databases for elemental solids and inorganic compound materials. There is some uncertainty about whether to use the IMFP databases for elemental solids and inorganic compound materials together as a single training database or if the two material types should be trained separately. On the one hand, adding the database for elemental solids may improve the ML performance for inorganic compound materials because of the increased amount of information.

The descriptive performance of the GPR for different subsets of the Shinotsuka *et al.* datasets[42,43] is validated and compared using different subsets each consisting of 30% of the total data. The accuracy of predicted IMFPs can be measured by the root-mean-square deviation (RMSD) as:

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\log\lambda_{\mathrm{pred}}(E_k) - \log\lambda(E_k)}{\log\lambda(E_k)}\right)^2}, \qquad (5)$$

where n is the total number of data points in the testing set, E_k is the electron energy, $\lambda_{\mathrm{pred}}(E_k)$ is the predicted IMFP, and $\lambda(E_k)$ is the target IMFP FPA-calculated value. The closer the RMSD is to zero, the better the prediction.

Fig. 1 shows the accuracy of the GPR using different subsets of the IMFP data of Shinotsuka *et al.*,[42,43] demonstrating that the descriptive ability of the GPR is very similar across datasets. The poorest described IMFPs are very similar, near IMFPs of 10 angstroms. Fig. 2 compares the GPR results with other empirical formulas using the RMSD and variance. We note that the empirical formulas are based on an old version of the IMFP database, and some adjustments were made to make the comparisons persuasive, such as the relativistic modification and the exclusion of materials not within the applicable range for each formula.

Fig. 2(a) focuses on accuracy, and the RMSDs of the GPR are below 5%, while the empirical formulas cannot achieve the same accuracy for either elemental solids or inorganic compounds, even when unsuitable materials are ignored. Beyond the accuracy, Fig. 2(b) shows the RMSD variance for each result, reflecting the stability of the IMFP description using the GPR and the general ability of ML. The variances are below 0.001 for the GPR results. In contrast, there are many extremely high-RMSD materials for some empirical formulas. It can be seen that the GPR possesses a stronger descriptive ability than the empirical formulas, including our most recent TPP-LASSO formula. The GPR's advantages are its accuracy and stability, as reflected in the average RMSD and variance. In addition, different types of datasets seemingly show no significant difference in the GPR's descriptive ability.

2.2 Leave-one-out cross-validation (LOOCV)

The next step is to check the predictive ability of GPR based on different databases. A well-accepted and widely used algorithm to check the predictive ability of ML is leave-one-out cross-validation (LOOCV). LOOCV is applied for the current x materials (x = 41 for the elemental-solids-only database, x = 42 for the inorganic compounds-only database, and x = 83 for the combined databases). In LOOCV, the ML model is trained with ($x - 1$) of x materials. The predictive ability of the ML model can be seen through the predicted result of the remaining xth material. That is, the IMFP of the xth material is the test set, and the data of the ($x - 1$) materials is the training set. The process is repeated x times, calculating the predictive performance (RMSD) to assess the overall predictive ability and indicate the general ability of the ML approach.

When applying the LOOCV method here, after the training is carried out for each material, the average RMSDs represent the performance on certain kinds of materials, so this method can test the GPR's ability to predict the IMFPs of new materials. Fig. 3 shows the average RMSDs compared for different training datasets (separately train the model with elemental solids or inorganic compounds; or train the model with the combined dataset of elemental solids and inorganic compounds), different training parameter sets (only use 8 conventional parameters; or use 2 periodic information parameters together with 8 conventional parameters), and different kinds of materials using LOOCV. The predictive ability for elemental solids is increased when using a dataset of both elemental solids and inorganic compound materials instead of only elemental solids. However, the situation is reversed for inorganic compounds. According to the LOOCV results in Fig. 3, the predictive accuracy is better when using the larger dataset to predict IMFPs of elemental solids, while the

Table 1 Machine learning parameters used, including the electron shell number (n_s), outermost shell electron number (N_o), atomic number (Z), atomic mass (M), density (ρ), number of valence electrons per atom (N_v), free-electron plasmon energy (E_p), band gap energy (E_g), starting-point energy (E_i), and atomic radius (R)

	Periodic information parameters		Conventional parameters							
	n_s	N_o	Z	M	ρ (g cm^{-3})	N_v	E_p (eV)	E_g (eV)	E_i (eV)	R (pm)
Li	2.00	1.00	3.00	6.94	0.53	1.00	7.99	0.00	4.74	145.00
Be	2.00	2.00	4.00	9.01	1.85	2.00	18.44	0.00	14.30	105.00
C (graphite)	2.00	4.00	6.00	12.01	2.25	4.00	24.93	0.00	20.40	70.00
C (diamond)	2.00	4.00	6.00	12.01	3.52	4.00	31.16	5.50	20.40	70.00
C (glassy)	2.00	4.00	6.00	12.01	1.80	4.00	22.30	0.00	20.40	70.00
Na	3.00	1.00	11.00	22.99	0.97	1.00	5.92	0.00	3.24	180.00
Mg	3.00	2.00	12.00	24.31	1.74	2.00	10.89	0.00	7.10	150.00
Al	3.00	3.00	13.00	26.98	2.70	3.00	15.78	0.00	11.20	125.00
Si	3.00	4.00	14.00	28.09	2.33	4.00	16.59	1.10	12.50	110.00
K	4.00	1.00	19.00	39.10	0.86	1.00	4.28	0.00	2.12	220.00
Sc	4.00	2.00	21.00	44.96	2.99	3.00	12.86	0.00	5.80	160.00
Ti	4.00	2.00	22.00	47.87	4.51	4.00	17.68	0.00	6.00	140.00
V	4.00	2.00	23.00	50.94	6.11	5.00	22.30	0.00	6.40	135.00
Cr	4.00	1.00	24.00	52.00	7.14	6.00	26.14	0.00	7.80	140.00
Fe	4.00	2.00	26.00	55.85	7.87	8.00	30.59	0.00	8.90	140.00
Co	4.00	2.00	27.00	58.93	8.90	9.00	33.58	0.00	10.00	135.00
Ni	4.00	2.00	28.00	58.69	8.90	10.00	35.47	0.00	9.10	135.00
Cu	4.00	1.00	29.00	63.55	8.96	11.00	35.87	0.00	8.70	135.00
Ge	4.00	4.00	32.00	72.59	5.32	4.00	15.59	0.67	12.60	125.00
Y	5.00	2.00	39.00	88.91	4.47	3.00	11.18	0.00	4.40	180.00
Nb	5.00	1.00	41.00	92.91	8.57	5.00	19.56	0.00	5.30	145.00
Mo	5.00	1.00	42.00	95.94	10.28	6.00	23.09	0.00	6.50	145.00
Ru	5.00	1.00	44.00	101.07	12.41	8.00	28.54	0.00	6.90	130.00
Rh	5.00	1.00	45.00	102.91	12.41	9.00	30.00	0.00	6.90	135.00
Pd	4.00	18.00	46.00	106.42	12.02	10.00	30.61	0.00	6.20	140.00
Ag	5.00	1.00	47.00	107.87	10.50	11.00	29.80	0.00	7.20	160.00
In	5.00	3.00	49.00	114.82	7.31	3.00	12.59	0.00	4.82	155.00
Sn	5.00	4.00	50.00	118.71	7.31	4.00	14.29	0.00	5.51	145.00
Cs	6.00	1.00	55.00	132.91	1.88	1.00	3.43	0.00	1.73	260.00
Gd	6.00	2.00	64.00	157.25	8.23	9.00	19.77	0.00	3.50	180.00
Tb	6.00	2.00	65.00	158.93	8.25	9.00	19.69	0.00	4.00	175.00
Dy	6.00	2.00	66.00	162.50	8.78	9.00	20.08	0.00	3.50	175.00
Hf	6.00	2.00	72.00	178.49	13.31	4.00	15.73	0.00	7.90	155.00
Ta	6.00	2.00	73.00	180.95	16.65	5.00	19.53	0.00	8.40	145.00
W	6.00	2.00	74.00	183.85	19.30	6.00	22.86	0.00	10.10	135.00
Re	6.00	2.00	75.00	186.21	21.02	7.00	25.60	0.00	10.70	135.00
Os	6.00	2.00	76.00	190.23	22.61	8.00	28.08	0.00	11.40	130.00
Ir	6.00	2.00	77.00	192.22	22.65	9.00	29.66	0.00	11.20	135.00
Pt	6.00	1.00	78.00	195.08	21.45	10.00	30.20	0.00	10.60	135.00
Au	6.00	1.00	79.00	196.97	19.32	11.00	29.92	0.00	9.00	135.00
Bi	6.00	5.00	83.00	208.98	9.79	5.00	13.94	0.00	12.60	160.00
AgBr	4.50	4.00	41.00	93.89	6.48	9.00	22.71	2.68	6.95	137.50
AgCl	4.00	4.00	32.00	71.66	5.59	9.00	24.14	3.25	7.51	130.00
AgI	5.00	4.00	50.00	117.39	5.72	9.00	19.08	2.92	6.15	150.00
Al_2O_3	2.40	4.80	10.00	20.39	3.97	4.80	27.86	8.63	16.63	86.00
AlAs	3.50	4.00	23.00	50.95	3.73	4.00	15.59	2.16	7.43	120.00
AlN	2.50	4.00	10.00	20.49	3.26	4.00	22.99	6.00	12.03	95.00
AlSb	4.00	4.00	32.00	74.37	4.28	4.00	13.83	1.62	6.94	135.00
c-BN	2.00	4.00	6.00	12.41	3.49	4.00	30.56	7.20	15.71	75.00
h-BN	2.00	4.00	6.00	12.41	2.30	4.00	24.81	5.00	13.77	75.00
CdS	4.00	4.00	32.00	72.24	4.80	9.00	22.28	2.46	6.88	127.50
CdSe	4.50	4.00	41.00	95.69	5.66	9.00	21.03	1.70	6.23	135.00
CdTe	5.00	4.00	50.00	120.01	5.85	9.00	19.09	1.51	6.13	147.50
GaAs	4.00	4.00	32.00	72.32	5.32	4.00	15.63	1.47	8.54	122.50
GaN	3.00	4.00	19.00	41.86	6.09	4.00	21.98	3.40	10.49	97.50
GaP	3.50	4.00	23.00	50.35	4.13	4.00	16.51	2.26	8.93	115.00
GaSb	4.50	4.00	41.00	95.74	5.61	4.00	13.95	0.73	7.75	137.50
GaSe	4.00	4.50	32.50	74.34	5.07	4.50	15.96	1.98	9.71	122.50
InAs	4.50	4.00	41.00	94.87	5.67	4.00	14.09	0.36	6.48	135.00
InP	4.00	4.00	32.00	72.90	4.79	4.00	14.77	1.38	7.36	127.50
InSb	5.00	4.00	50.00	118.29	5.78	4.00	12.74	0.18	6.36	150.00
KBr	4.00	4.00	27.00	59.50	2.75	4.00	12.39	7.26	9.86	167.50
KCl	3.50	4.00	18.00	37.27	1.98	4.00	13.28	7.40	10.10	160.00
MgF_2	2.33	5.33	10.00	20.77	3.18	5.33	26.03	10.95	16.45	83.33
MgO	2.50	4.00	10.00	20.15	3.58	4.00	24.28	7.69	13.99	105.00
NaCl	3.00	4.00	14.00	29.22	2.17	4.00	15.69	9.00	13.10	140.00

Table 1 (*continued*)

	Periodic information parameters		Conventional parameters							
	n_s	N_o	Z	M	ρ (g cm^{-3})	N_v	E_p (eV)	E_g (eV)	E_i (eV)	R (pm)
$NbC_{0.712}$	3.75	2.25	26.44	59.26	7.75	4.58	22.31	0.00	7.40	113.81
$NbC_{0.844}$	3.63	2.37	24.98	55.88	7.77	4.54	22.90	0.00	7.40	110.67
$NbC_{0.93}$	3.55	2.45	24.13	53.93	7.78	4.52	23.27	0.00	7.40	108.86
PbS	4.50	5.00	49.00	119.63	7.62	5.00	16.26	0.42	5.63	140.00
PbSe	5.00	5.00	58.00	143.08	8.29	5.00	15.51	0.29	5.29	147.50
PbTe	5.50	5.00	67.00	167.40	8.27	5.00	14.32	0.32	4.86	160.00
SiC	2.50	4.00	10.00	20.05	3.22	4.00	23.10	2.31	9.26	90.00
SiO_2	2.33	5.33	10.00	20.00	2.19	5.33	22.02	9.10	19.10	85.00
SnTe	5.00	5.00	51.00	123.16	6.47	5.00	14.77	0.19	8.44	142.50
$TiC_{0.7}$	3.18	2.82	15.41	33.11	4.63	4.00	21.54	0.00	5.70	111.18
$TiC_{0.95}$	3.03	2.97	14.21	30.41	4.84	4.00	23.00	0.00	5.70	105.90
$VC_{0.76}$	3.14	2.86	15.66	34.13	5.58	4.57	24.91	0.00	7.50	106.93
$VC_{0.86}$	3.08	2.92	15.14	32.94	5.61	4.54	25.32	0.00	7.50	104.95
$Y_3Al_5O_{12}$	2.70	4.65	13.90	29.68	4.55	4.80	24.73	6.50	13.00	94.25
ZnS	3.50	4.00	23.00	48.72	4.09	9.00	25.05	3.81	9.18	117.50
ZnSe	4.00	4.00	32.00	72.17	5.26	9.00	23.34	2.68	8.10	125.00
ZnTe	4.50	4.00	41.00	96.49	5.64	9.00	20.90	2.25	7.67	137.50

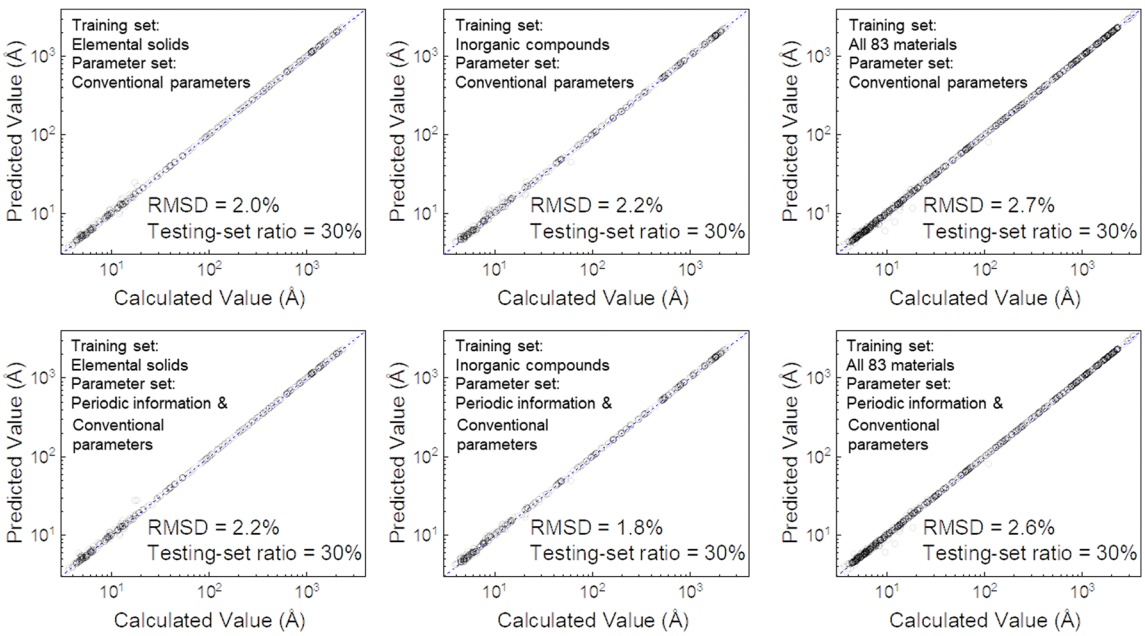

Fig. 1 IMFP learning performance for different data subsets (IMFP sets of elemental solids and/or inorganic compounds) of the IMFP data of Shinotsuka *et al.*[42,43] and parameter sets. The used training sets and parameter sets are labeled in the upper left corner of each panel. The "conventional parameters" are Z, M, ρ, N_v, E_p, E_g, E_i and R; and "periodic information parameters" are n_s and N_o. Their definitions can be found in Table 1. The x-axis is the IMFP calculated using the FPA, and the y-axis is the IMFP predicted using the GPR. The blue lines are diagonal, indicating perfect agreement between the predicted and calculated IMFPs.

smaller dataset is better when predicting IMFPs of inorganic compounds.

This conclusion is supported by the detailed RMSD distributions for different material types shown in Fig. 3(b). First, in the comparison between material types, the most poorly predicted material set is for the alkaline metals, and the predictive ability is decreased when the band gap energy increases, corresponding to the change from conductors to insulators. Considering both the distribution of elemental solids in the periodic table whose IMFPs are known and the band gap contribution for inorganic compound materials (Fig. 4), we can see that the GPR uses local information better. This is a significant conclusion that can guide the appropriate application of the GPR for IMFP predictions. The GPR may have advantages when predicting IMFPs for materials with similar physical or chemical characteristics.

A second comparison is between the different training sets used. After using the combined dataset, the average RMSD of alkaline metals sharply decreased, and the average RMSD of other elemental solids sharply decreased. The average RMSD

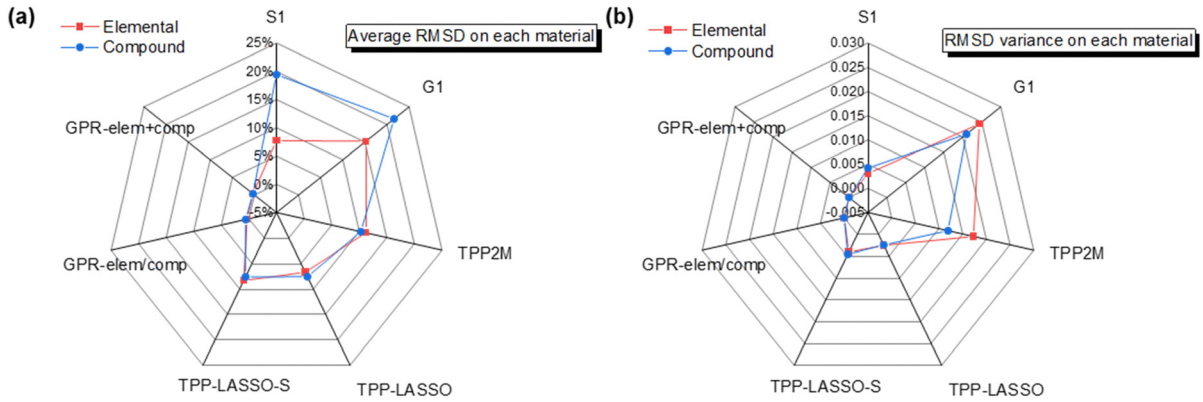

Fig. 2 Comparison of (a) average RMSDs, and (b) variance in RMSDs, for all 83 materials using the S1, G1, TPP-2M, TPP-LASSO, and TPP-LASSO-S formulas and the GPR machine learning method for different datasets. The red lines are for elemental solids, and the blue lines are for inorganic compounds. We note that the comparison is for electron energies above 200 eV for the valid energy range of the empirical formulas.

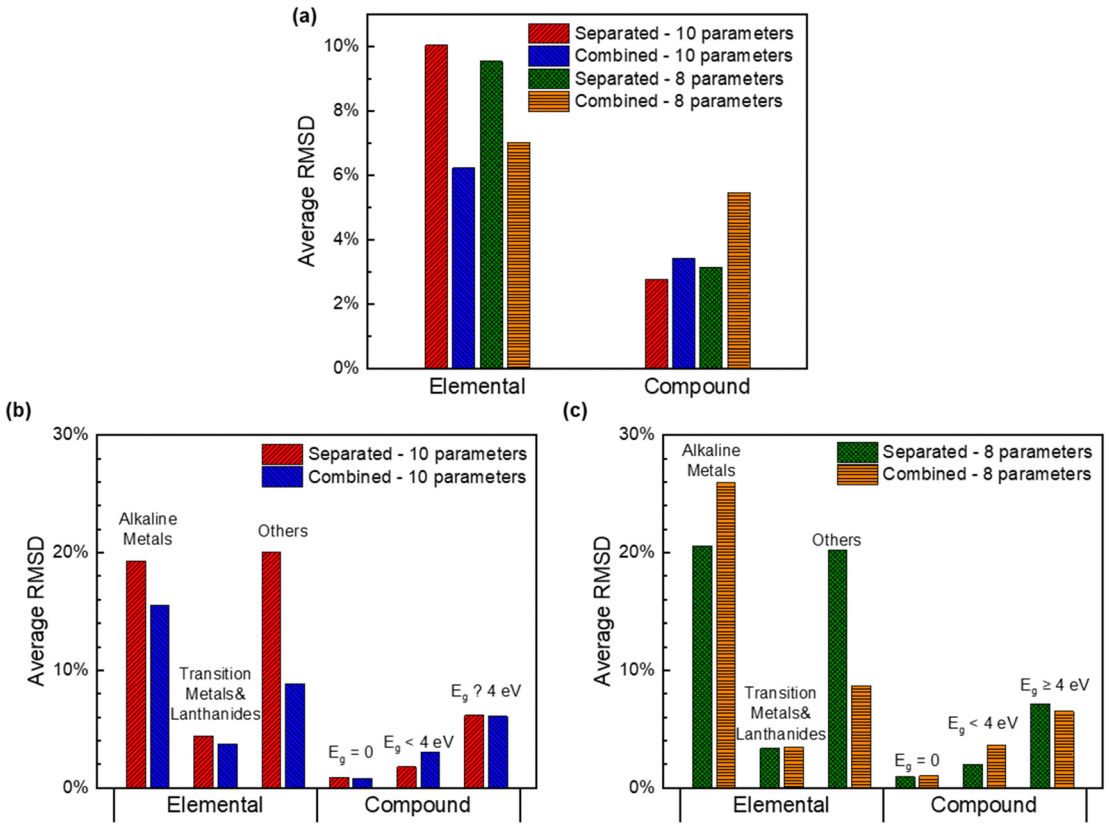

Fig. 3 (a) Average RMSDs compared for different training datasets (Separately train the model with elemental solids or inorganic compounds; or train the model with the combined dataset of elemental solids and inorganic compounds) and training parameter sets (only use 8 conventional parameters; or use 2 periodic information parameters together with 8 conventional parameters) and different material types using LOOCV; (b) detailed comparison for different material types using both periodic information and conventional parameters in (a); (c) detailed comparison for material types using only conventional parameters in (a).

of narrow bandgap inorganic compounds ($E_g < 4$ eV) slightly increased, while the average RMSD of insulators slightly decreased. This can also be explained by the conclusion above that the GPR is better at using local information.

For elemental solids, when inorganic compounds were added to the dataset, the increased local information from the inorganic compounds increased the accuracy of other elemental solids in Fig. 3, which had little information compared with the classified elemental solids. However, because the alkaline metals have a marginalized place in the periodic table, the added inorganic compounds cannot be seen as local information but as a disturbance, and the alkaline metals' accuracy is decreased.

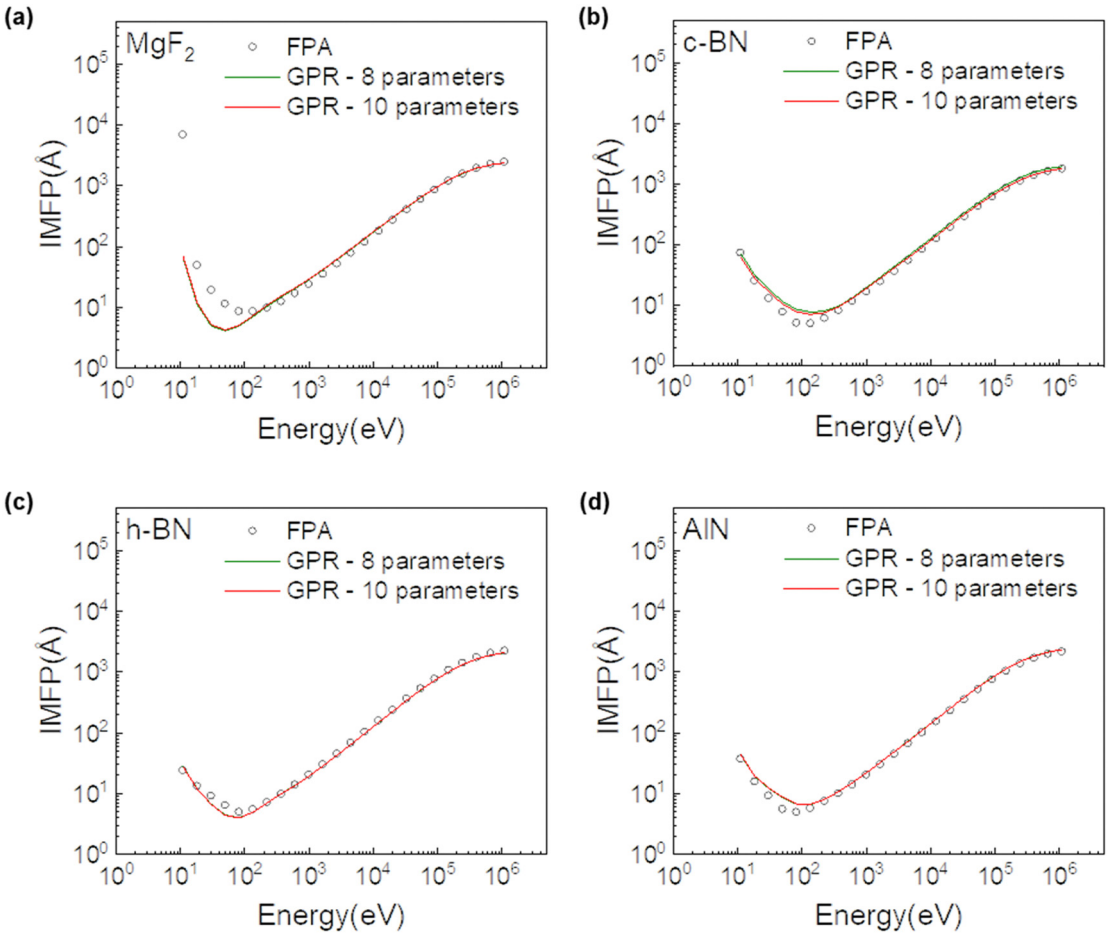

Fig. 4 Typical LOOCV examples for inorganic compounds ((a) MgF_2, (b) c-BN, (c) h-BN and (d) AlN) using only inorganic compound materials as the GPR training set. In the label, "8 parameters" means the curve for the ML model trained with only conventional parameters (Z, M, ρ, N_v, E_p, E_g, E_i, R), and "10 parameters" means the curve for the ML model trained with both periodic information and conventional parameters (n_s, N_o, Z, M, ρ, N_v, E_p, E_g, E_i, R).

Because inorganic compound materials' distribution of parameters is more focused (localized) than elemental solids, their accuracies are higher. Then, when elemental solids are added to the dataset, some dispersed information influences the inorganic compounds' IMFP accuracy. Although the gap is not very large, it is more accurate to predict the IMFPs of inorganic compound materials using a separate inorganic compound dataset than with a mixed dataset including IMFPs of elemental solids.

We also show a detailed comparison of different material types using only conventional parameters to show the influence of periodic information parameters (n_s and N_o). Fig. 3(c) shows that the predictive accuracy for alkaline metals benefits most when the periodic information parameters are applied. We note that the location of alkaline metals in the periodic table is separated from the other elements. Therefore, they cannot build strong connections with other elements if only using simple material-dependent parameters. The added parameters n_s and N_o are directly related to the period and group numbers of the elements in the periodic table, so these parameters can strengthen the relationships between different alkaline metals, leading to more stable predictions.

Fig. 4 shows typical LOOCV results, with different RMSDs for the inorganic compounds MgF_2 (19.4%), c-BN (10.3%), h-BN (6.3%), and AlN (7.1%) using only conventional parameters, and MgF_2 (18.9%), c-BN (8.1%), h-BN (6.5%), and AlN (7.5%) using both periodic information and conventional parameters. As discussed previously, periodic information parameters are not influential for materials other than alkaline metals. It is noted that MgF_2 and c-BN are the worst two materials in the LOOCV results, while $NbC_{0.93}$ is the best one. However, even the LOOCV results for MgF_2 and c-BN in Fig. 4 agree satisfactorily with the optical data in the high-energy region. Similar to the elemental solids results reported previously,[54] the ML model shows slightly larger RMSDs in the lower energy range, especially in the region below approximately several hundred electron volts. The RMSDs then gradually decrease as the energy increases, as shown for c-BN and h-BN.

The poorest prediction for MgF_2 is an outlier in the LOOCV results. For MgF_2, it is clear that the data point at 10 eV is not continuous with the points afterward. In fact, MgF_2 has the largest E_g in our training set (see Tables 1 and 2), dramatically

Table 2 RMSDs of LOOCV results for the regression only inorganic compounds

Material	AgBr	AgCl	AgI	Al_2O_3	AlAs	AlN	AlSb
RMSD-8 parameters	1.14%	2.41%	6.70%	5.10%	1.19%	7.14%	1.20%
RMSD-10 parameters	1.11%	2.44%	2.00%	2.25%	1.17%	7.45%	1.21%

Material	cubic BN	hexagonal BN	CdS	CdSe	CdTe	GaAs	GaN
RMSD-8 parameters	10.32%	6.31%	2.07%	0.54%	0.73%	1.58%	3.15%
RMSD-10 parameters	8.06%	6.52%	2.12%	0.53%	0.66%	1.19%	3.61%

Material	GaP	GaSb	GaSe	InAs	InP	InSb	KBr
RMSD-8 parameters	1.12%	0.86%	2.69%	1.04%	1.48%	1.74%	8.02%
RMSD-10 parameters	1.14%	1.08%	2.68%	1.05%	1.49%	1.17%	3.92%

Material	KCl	MgF_2	MgO	NaCl	$NbC_{0.712}$	$NbC_{0.844}$	$NbC_{0.93}$
RMSD-8 parameters	3.96%	19.44%	3.76%	3.61%	0.70%	0.48%	0.47%
RMSD-10 parameters	3.96%	18.90%	1.94%	3.71%	0.68%	0.43%	0.43%

Material	PbS	PbSe	PbTe	SiC	SiO_2	SnTe	$TiC_{0.7}$
RMSD-8 parameters	0.51%	0.69%	0.79%	3.20%	5.43%	3.46%	1.50%
RMSD-10 parameters	0.52%	0.68%	0.81%	3.29%	5.46%	3.42%	1.49%

Material	$TiC_{0.95}$	$VC_{0.76}$	$VC_{0.86}$	$Y_3Al_5O_{12}$	ZnS	ZnSe	ZnTe
RMSD-8 parameters	1.52%	1.22%	0.51%	5.42%	5.11%	2.16%	1.96%
RMSD-10 parameters	1.49%	1.14%	0.48%	5.57%	5.13%	2.16%	1.99%

affecting the IMFP behavior for the larger contribution of phonon excitation and plasmon lifetime effect, especially at lower energies as analyzed previously for MgO.[37] This IMFP outlier leads to a poor prediction when applying the GPR. Because of the data point at 10 eV for MgF_2, the entire GPR prediction in the low-energy region falls below the FPA-calculated values because the energy dependence affects the GPR results. Aside from this worst result for MgF_2, even the second-poorest result for c-BN is reasonable with an RMSD below 11%, reflecting the GPR's strong predictive power.

Fig. 5 shows the relationships between RMSDs and various parameters to illustrate the local dependence of the GPR when predicting IMFPs. The available information can be summarized similarly to Fig. 3, with the predictive accuracy (RMSD) tending to have the same distribution for materials with similar parameters. This is very apparent, especially for the band gap energy (E_g). For example, the predictive ability uniformly decreases as the band gap energy becomes larger because the material distribution of high-band-gap energy is sparse, so the GPR cannot get local information for prediction.

Above all, the GPR relies on local information when regressing IMFPs between materials, which is strongly reflected in the parameters used in the GPR. Moreover, the lines connected vertically between the red and green dots in Fig. 5 show the prediction improvement when including the periodic information parameters, with the materials containing halogenated elements (*e.g.*, MgF_2, KBr, and AgI) benefitting the most. Similar to the Fig. 3 discussion for the alkaline metals, the halogenated elements are located at the edge of the periodic table, separated from other elements. Periodic information parameters can strengthen the relationships between elements in the same group in the periodic table, thereby improving the predictive accuracy for the halides.

2.3 Material dependence check

In the GPR used above, every single IMFP data point is treated as one vector in description or prediction, meaning that the ML scheme must simultaneously consider material and energy dependencies, which is challenging. The GPR may fail to correctly balance the information from these dependencies in certain situations, so we perform LOOCV for each energy separately when predicting IMFPs of different materials. This effectively removes the electron energy from the parameters, so there is only a material dependence at each energy. When LOOCV has been performed for all the energies, the predicted data points are collected in a scatter plot presenting the IMFPs considering the material dependence only.

The average RMSD of different material types when using 10 parameters together with electron energy E (parameters with E); or using 10 parameters separately predict IMFPs for each energy (parameters without E) are shown in Fig. 6. The regression with parameters without E predicts the IMFPs of alkaline metals more accurately than parameters with E. The prediction for narrow bandgap inorganic compounds ($E_g < 4$ eV) becomes slightly better, and other materials do not change noticeably.

Fig. 7 shows the LOOCV results for Cs, Gd, MgF_2, and h-BN using regressions for parameters without E. The predictions for the low-energy region are generally not quite as smooth as for the high-energy region. Considering that the new regression only considers material dependence, separately regressing the IMFPs of the low- and high-energy regions, these energies do not influence each other. At first sight, the high-energy region

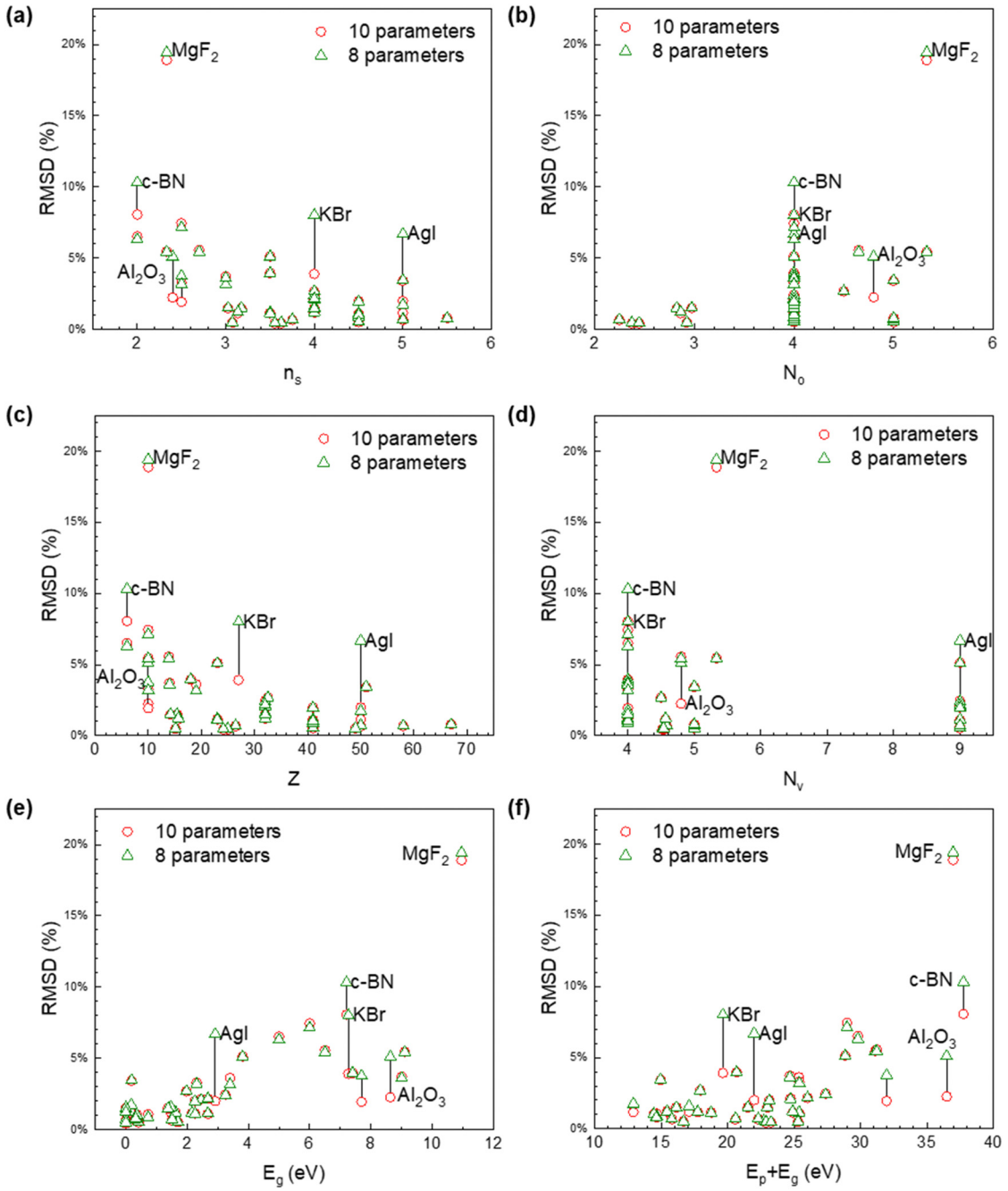

Fig. 5 Relationships between different material-dependent parameters ((a) n_s, (b) N_o, (c) Z, (d) N_v, (e) E_g and (f) E_p+E_g) and RMSDs. The dataset used here is the IMFPs of only inorganic compound materials. The red dots are for regression using both periodic information and conventional parameters (n_s, N_o, Z, M, ρ, N_v, E_p, E_g, E_i, R, See Table 2 for detail). The green dots are for regression using only conventional parameters (Z, M, ρ, N_v, E_p, E_g, E_i, R, See Table 2 for detail).

predictions are similar to the normal GPR (including the energies), except for the alkaline metals. Using Cs in Fig. 7 as an example, the IMFPs for higher energies (>1000 eV) typically show relatively larger biases with an RMSD of 59.07%, while the curve trend of the lower and intermediate energy ranges (<1000 eV) is different with a smaller RMSD of 45.18%. This example reflects how the energy dependence affects the predictions compared with the earlier GPR predictions.

When using normal GPR (including the energies), the prediction in the higher energy region (>1000 eV) influenced the prediction in the lower and intermediate energy ranges (<1000 eV) because of the energy dependence. The parameters without E regression method can increase the predictive accuracy, at least for alkaline metals: the average RMSD of alkaline metals is 15.52% using the GPR with included energies and 13.03% using the parameters without E regression method.

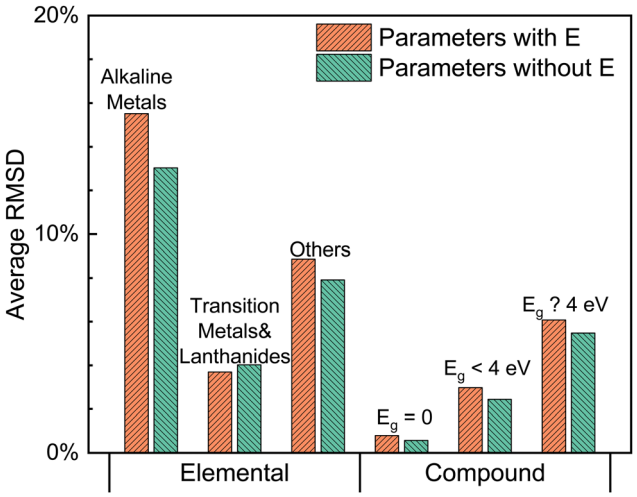

Fig. 6 Average RMSDs for different material types before and after removing energy dependence in machine learning. The dataset used here is the IMFPs for both elemental solids and inorganic compounds.

Meanwhile, there are also some elemental solids where the predictive effect is similar for the two methods. For example, the RMSDs using parameters with and without E regressions for Gd were 3.72% and 4.32%, respectively.

The LOOCVs of inorganic compounds also demonstrate similar effects. In Fig. 7, MgF_2 has separated prediction results that are not smooth in the low-energy region, similar to Cs. Meanwhile, $NbC_{0.93}$ has results consistent with the actual values, similar to Gd. However, the reasons for these behaviors may be different. As previously discussed, the FPA-calculated IMFPs constituting our training dataset have poorer accuracy in the low-energy region than in the high-energy region. We noted that the GPR prediction result with only material dependence might be more reliable than the original data point in this case. In other words, the GPR with only material dependence gives a different IMFP value than the original FPA-calculated IMFPs, and this value could be more reliable than the original training dataset because it is based on the training data of other materials. The GPR used here exhibits strong data error correction, a unique and significant capability for the data analysis problem. In the validation described in this section, the ML focuses only on material dependence by eliminating the energy dependence. As a result, the GPR model can still predict reasonable IMFPs without the energy dependence, leading to

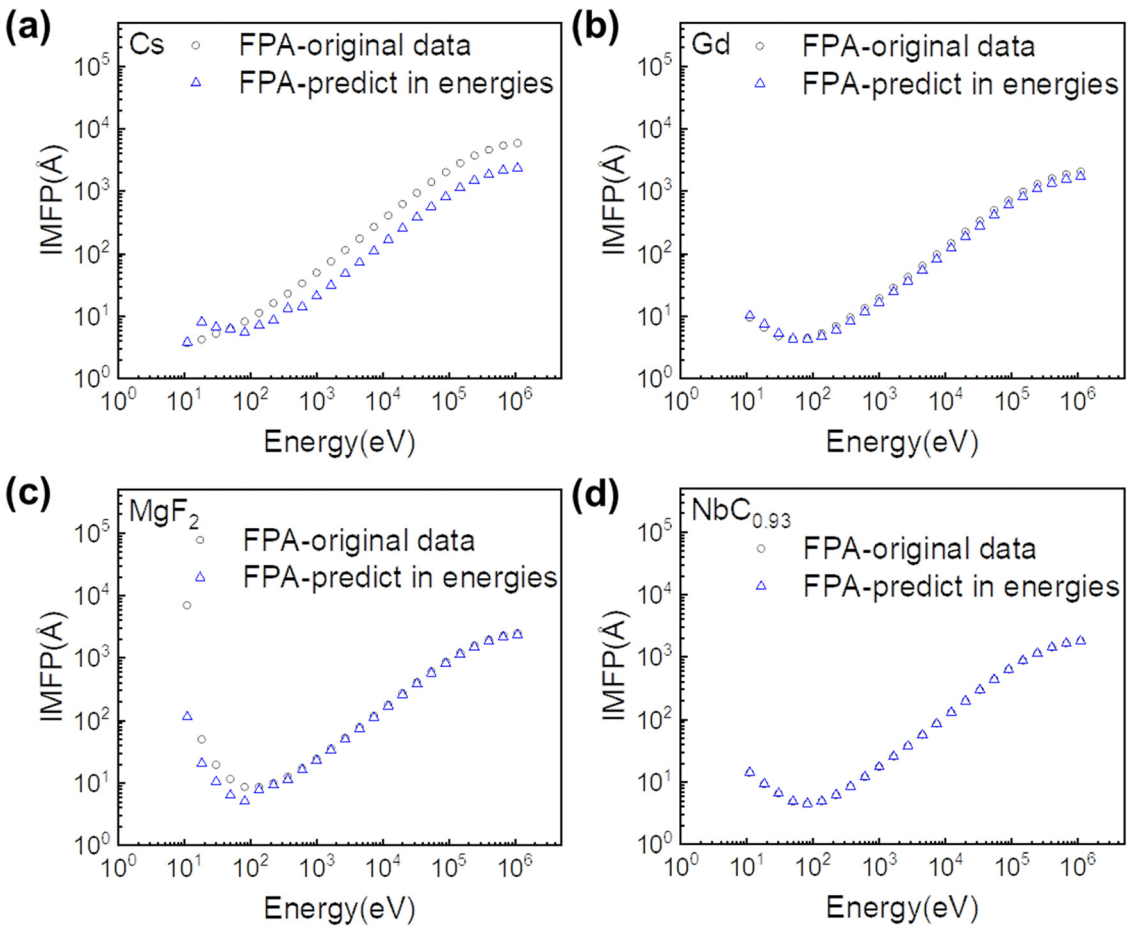

Fig. 7 Typical LOOCV examples of elemental solids ((a) Cs and (b) Gd) and inorganic compound materials ((c) MgF_2 and (d) $NbC_{0.93}$) using the GPR when considering material dependence only. The open circles are the original data, namely the testing data calculated with FPA. The blue triangles are the LOOCV results with separated energies, with other materials used as training data.

its ability to predict the IMFP database using discrete training data without large accuracy losses.

In fact, there are many experimental IMFP measurements for different materials. The experimental IMFPs are distributed in different energy regions, and data points for experimental IMFPs are often very variable from study to study. Because of this uneven distribution of electron energies, the normal GPR, which includes the energies, may be inaccurate if there are unavailable data points for a material. However, the problem can be solved more accurately if the GPR only considers material dependence with separated energies because the IMFPs are regressed separately for different energies with no missing data points. Collecting experimental IMFPs and regressing them with only material dependence is a planned topic of future study.

2.4 Predictions for new inorganic compounds

The preceding sections have shown comparisons with various formulas or methods and analyzed the RMSD distribution with different parameters, but the results are more convincing if we apply our ML approach to predict the IMFPs for the materials that are not in our training set. The absence of these materials in our training set is because suitable ELFs are not currently available for all of them. ML using only the inorganic compound training database and both periodic information and conventional parameters is used in this section because Fig. 3 shows that it is the best model for inorganic compounds.

There are, however, some applicable ELFs for some of these materials: Cu_2O, CuO, HfO_2, and ZrO_2. The ELFs in the low energy range used in the IMFP calculations were taken from Tahir and Tougaard[57] for Cu_2O and CuO (0–100 eV), Jin *et al.*[58] for HfO_2 (0–80 eV), and Tahir *et al.*[59] for ZrO_2 (0–80 eV). ELFs in the medium-energy (up to 30 keV)[60] and high-energy ranges (30 keV–10 MeV)[61] are also used. Here we calculate IMFPs of these materials base on these ELFs using FPA, as a comparison dataset to test our model.

Tanuma *et al.*[62] suggested using two sum rules to quantify the quality of ELFs; that is, the oscillator strength sum rule (f-sum rule) and the perfect screening sum rule (ps-sum rule). We use these rules to evaluate the accuracy of the ELFs of Cu_2O, CuO, HfO_2, and ZrO_2.

The f-sum rule Z_{eff} is given by,

$$Z_{\mathrm{eff}} = \frac{2}{\pi\Omega_{\mathrm{P}}^2}\int_0^{\omega_{\max}} \omega \mathrm{Im}\left\{\frac{-1}{\varepsilon(\omega)}\right\}\mathrm{d}\omega, \qquad (6)$$

where $\hbar\Omega_{\mathrm{P}} = \sqrt{4\pi n_{\mathrm{a}} e^2/m}$, $n_{\mathrm{a}} = N_{\mathrm{a}}\rho/M$ is the number density of atoms, N_{a} is the Avogadro's number, ρ is the mass density, and M is the atomic weight.

The ps-sum rule P_{eff} can be obtained from the Kramers–Kronig relation:[62]

$$P_{\mathrm{eff}} = \frac{2}{\pi}\int_0^{\omega_{\max}} \frac{1}{\omega}\mathrm{Im}\left\{\frac{-1}{\varepsilon(\omega)}\right\}\mathrm{d}\omega + \mathrm{Re}\left\{\frac{1}{\varepsilon(0)}\right\}, \qquad (7)$$

where $\mathrm{Re}\{1/\varepsilon(0)\} = 0$ for conductors. The theoretical values of Z_{eff} and P_{eff} are the atomic number and unity, respectively, in the limit $\omega_{\max} \to \infty$. In this work, $\omega_{\max}$ was 10 MeV. Table 3 lists the results from the f-sum and ps-sum rules of ELFs for Cu_2O, CuO, HfO_2, and ZrO_2.

Table 3 List of f-sum and ps-sum rule checks for four inorganic compounds

Inorganic compound	Atomic number	ps-Sum rule	Relative error (%)	f-Sum rule	Relative error (%)
Cu_2O	66	1.0905	9.05	66.067	0.102
CuO	37	1.0967	9.67	34.892	−5.70
HfO_2	88	0.8491	−15.09	75.775	−13.9
ZrO_2	56	1.0473	4.73	47.608	−15.0

Fig. 8 shows the IMFP values calculated from the ELFs above and the GPR-predicted IMFPs. The GPR-predicted IMFPs are evidently consistent with the FPA-calculated values for the high-energy region (>100 eV) but slightly shifted for HfO_2 and ZrO_2. Although the trend is correct for the lower energy region (<100 eV), there are differences between the predicted IMFPs from the GPR and the calculated IMFPs from the ELFs.

According to ref. 42, the FPA may produce less accurate IMFPs at very low energies (less than 50 eV). The Table 3 sum rule results also demonstrate that the ELFs for HfO_2 and ZrO_2 are not as accurate as for Cu_2O and CuO. The larger differences for HfO_2 and ZrO_2 at high energies may result from the uncertainty in the ELFs. Therefore, it is probable that the inaccuracy of the GPR does not cause these differences but that the FPA-calculated values disagree with the true IMFPs, for either lower or higher energies.

The band gap energy value for HfO_2 is the highest among the four materials shown in Fig. 8. Together with the previous analysis in Fig. 5, this indicates that the GPR uses local information when predicting IMFPs' detailed behavior, so the GPR naturally provides poorer predictions for high band gap energy materials. This is another powerful indicator that the GPR is superior in predicting IMFPs for narrow-bandgap materials or alloys because of its reliance on local information.

The IMFPs predicted by the TPP-2M formula for energies above 50 eV, and the TPP-LASSO-S formula for energies above 200 eV are also plotted in Fig. 8 for comparison. TPP-LASSO-S is a newer formula produced in our previous work[50] and is given in eqn (4). The TPP-LASSO-S curve is very similar to the TPP-2M curve because the TPP-LASSO-S and TPP-2M formulas are both from the modified Bethe equation,[49] although the TPP-LASSO-S formula introduced using ML is thought to improve stability. For Cu_2O, CuO, and ZrO_2, the predicted results from the GPR and the TPP-2M/TPP-LASSO-S formulas consistently agree with the FPA-calculated IMFPs.

The middle energy region (100–1000 eV) for HfO_2 is especially significant. Our GPR approach is superior to the TPP-2M/TPP-LASSO-S formulas, even though the GPR is not the most accurate when predicting large band gap materials. Unfortunately, the experimental IMFPs are still unavailable for the four materials in Fig. 8. Therefore, conclusions regarding GPR characteristics and robustness are limited to the discussion above for these materials.

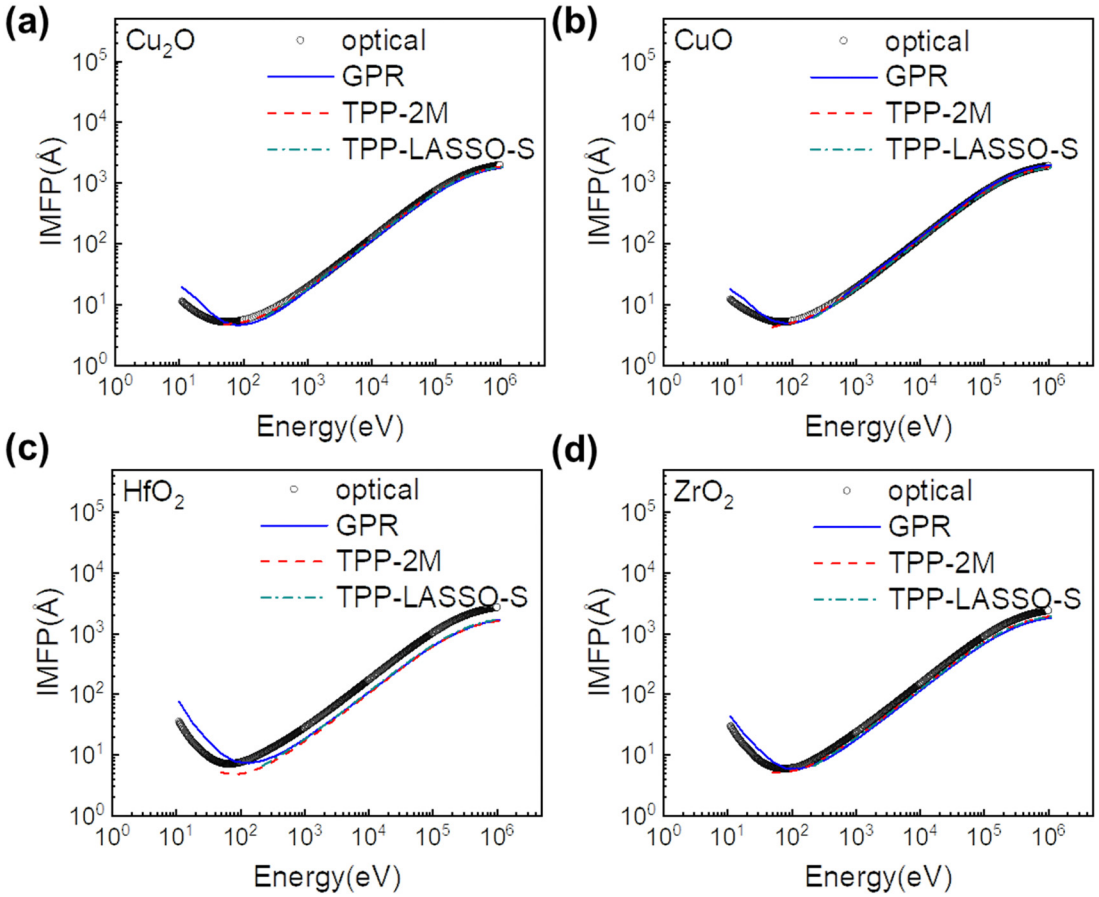

Fig. 8 Representative predictions for inorganic compounds (a) Cu_2O, (b) CuO, (c) HfO_2, and (d) ZrO_2. Blue solid curves are IMFPs predicted by the GPR; red dashed curves are IMFPs predicted by the TPP-2M formula; green dash-dotted curves are IMFPs predicted by the TPP-LASSO-S formula.

2.5 Predictions for quantitative materials using the database of materials projects

2.5.1 Use of the GPR to produce a narrow bandgap inorganic compounds (E_g < 3 eV) database. Various semiconductors are used in different electronic components in the contemporary materials industry, so an IMFP database is necessary to thoroughly explore electron transport properties inside these materials. Here we used ML model to produce an applicable database for narrow bandgap inorganic compounds (E_g < 3 eV). Fig. 3 showed that the most accurate ML prediction model for conductor and narrow bandgap inorganic compounds (E_g < 3 eV) uses only the inorganic compound IMFP training set with both periodic information and conventional parameters. With confidence in the ML model trained in this way, we thus use the trained ML model to produce an extensive narrow bandgap inorganic compounds (E_g < 3 eV) IMFP database for unknown IMFPs using parameters obtained from public material-dependent parameter databases, such as P-table[63] and the Materials Project (MP).[64]

P-Table[63] is an open access website containing periodic table elemental properties. MP[64] is an open dataset using high-throughput computing to uncover the properties of all known inorganic materials and can be accessed conveniently through an application programming interface (API) for both interactive exploration and data mining.

We note that MP is a robust and sophisticated material database, with every material having a unique, conveniently referenced Material ID, even for different calculations for the same material. Using these Material IDs as indexes, we acquired all the materials in MP as narrow bandgap inorganic compounds (E_g < 3 eV) candidates that satisfy the following conditions: (1) band gap energy (E_g) between 0.1 eV and 3 eV; (2) component elements between H (Z = 1) and Bi (Z = 83), except for the noble gas elements (He, Ne, Ar, Kr, Xe, and Rn); (3) density of states (DOS) available in MP; and (4) the material has been experimentally observed, not only theoretically existent materials. For these materials, the processing methods used for parameters are the same as for the training procedure, as discussed in the Methods section (Section 2.1).

The parameters n_s, N_o, Z, M, and R can be easily obtained from P-table. N_v is taken from,[65] where several elements are interpolated using Z values, and ρ and therefore E_p [= 28.8 $(N_v\rho/M)^{0.5}$] are obtained directly from the first principles results in MP. E_i (= E_v + E_g for inorganic compounds) and E_g are estimated from calculated DOS in MP. An example of E_i and E_g from DOS for SiC (Material ID mp-568656) is shown in Fig. 9. The DOS is the total DOS containing occupied and unoccupied states of electrons, representing conduction and valence bands for band structures. The E_g value is the energy interval between occupied and unoccupied states.

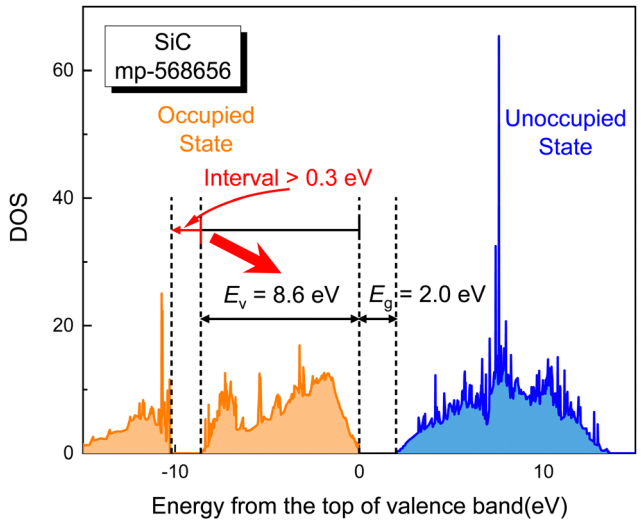

Fig. 9 Example of E_v and E_g obtained from DOS for SiC.

For the E_v value, we note that the Fermi level is set as the valence band top in the MP calculations, reflected in the relationship between the DOS and the x-axis in Fig. 9. Therefore, the E_v value is defined as the energy span of the first "section" of the occupied state in the DOS curve. Here the state "sections" are cut off such that there is a continuous energy region larger than 0.3 eV where the DOS curve has very small values (<0.001). Thus the first "section" of the occupied state is counted from the negative side of zero energy for the E_v value. This method in MP is validated with an average relative error of 10.2% compared with our training database for the narrow bandgap inorganic compounds ($E_g < 3$ eV) listed in Table 1. As a result, we collected 12 039 narrow bandgap inorganic compounds ($E_g < 3$ eV) from the Materials Project as candidates for the IMFP database.

The parameter distributions of the 12 039 narrow bandgap inorganic compounds ($E_g < 3$ eV) are shown in Fig. 10. All the parameters except for E_g naturally follow distributions similar to a peak-shaped distribution. The uniformly-distributed E_g is probably due to two factors: (1) the E_g is manually selected lower than 3 eV, which is different from other parameters; (2) E_g is undoubtedly a definitive parameter for narrow bandgap inorganic compounds ($E_g < 3$ eV), so the different distribution is probably indicative of narrow bandgap inorganic compounds' ($E_g < 3$ eV) different characteristics.

We now use these parameters for the 12 039 narrow bandgap inorganic compounds ($E_g < 3$ eV) and the trained GPR model to produce an IMFP database. The distributions of the numerous IMFP curves are shown in Fig. 11(a). For comparison, we also show the IMFPs predicted by the TPP-2M and TPP-LASSO-S formulas in Fig. 11(b) and (c), which are only valid at high energies (above 50 eV for TPP-2M and above 200 eV for

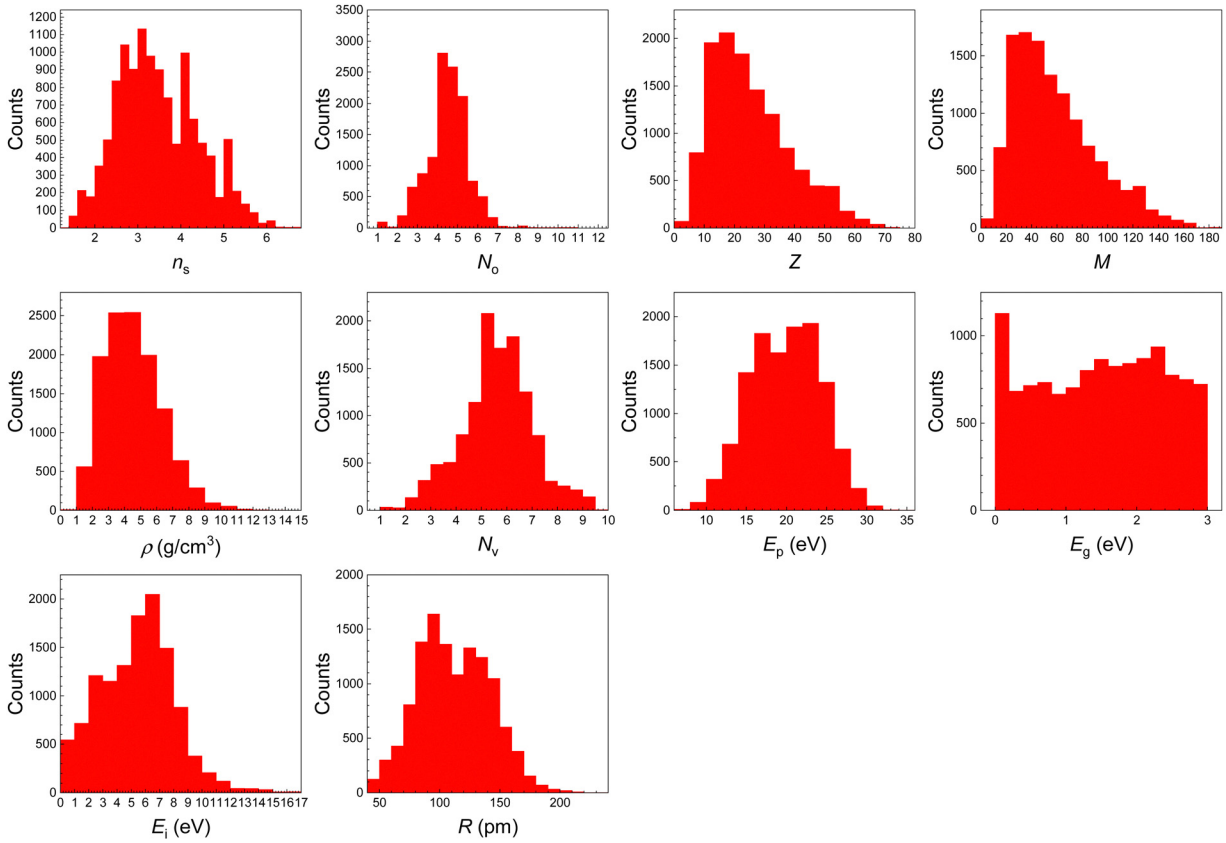

Fig. 10 Distributions of parameters for 12 039 narrow bandgap inorganic compounds ($E_g < 3$ eV).

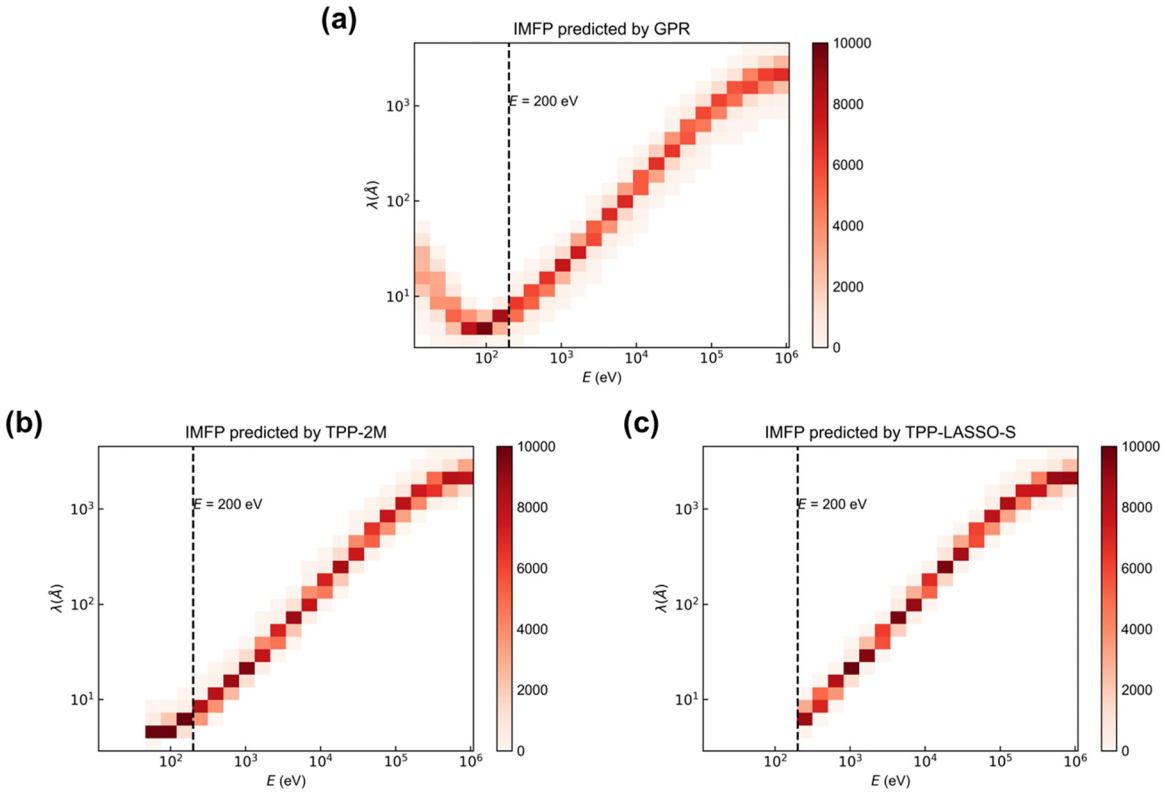

Fig. 11 Distribution of all IMFP data points predicted by: (a) GPR; (b) TPP-2M; and (c) TPP-LASSO-S. The shade of red color shows the density of IMFPs distribution.

TPP-LASSO-S). Because TPP-2M and TPP-LASSO-S empirical formulas with predictive ability, it is also appropriate to discuss the accuracy of the prediction data with experiments. We only compare the energy region above 200 eV for the three methods to ensure comparison consistency.

In the predictions by the three methods (Fig. 11), the IMFP distributions are concentrated, indicating the common surface properties of narrow bandgap inorganic compounds (E_g < 3 eV). Moreover, the prediction deviations of the three methods are very small. RMSDs between the GPR and TPP-2M/TPP-LASSO-S formulas are 3.28% and 3.12%, while the maximum RMSDs are 26.31% and 23.96%, and the minimum RMSDs are 0.13% and 0.31%. These relatively small deviations provide cross-validation that the predictions by the three methods are all trustworthy.

The distributions of the RMSDs for GPR/TPP-2M and GPR/TPP-LASSO-S and the prediction parameters for all 12 039 narrow bandgap inorganic compounds (E_g < 3 eV) are shown in Fig. 12. The distributions of RMSDs for all parameters are concentrated at <5% near the center value of each parameter in the figure, demonstrating the GPR's predictive accuracy. Moreover, the distributions in Fig. 12 are also concentrated in a certain area, except E_g is still relatively uniform comparing to the distributions of other parameters. On the basis of this special behavior of E_g in the prediction, this parameter is likely one of the principal descriptors of narrow bandgap inorganic compounds (E_g < 3 eV) IMFPs.

2.5.2 Use of the LASSO to explore the principal descriptors for narrow bandgap inorganic compounds (E_g < 3 eV). To investigate the principal descriptors, we introduce the framework in our previous work[50] based on LASSO. In ref. 50, we provided a TPP-LASSO-S formula for high energy (>200 eV) IMFPs, however, cannot predict low energy IMFPs. So it is necessary to discuss the GPR predicted IMFPs with a brand-new formula that can describe IMFPs for wider energy range. Fortunately, our proposed framework in ref. 50 is strong enough that can be widely applied on different empirical formulae. As a starting point of the framework, we note that Ziaja *et al.*[66] provided eqn (8) for a prototype formula on wider energy range:

$$\lambda = \frac{\sqrt{E}}{a(E - E_{\text{th}})^b} + \frac{E - \exp(-B/A)}{A \ln E + B}, \tag{8}$$

where E is electron energy; λ is IMFP; a, b and E_{th} are fitting parameters mainly effective on lower energies; A and B are for higher energies. In the framework, we will first use the least square fitting to transfer the regression target from the whole curve to principal fitting parameters. However, we found the complex function $\sqrt{E}\Big/\left[a(E - E_{\text{th}})^b\right]$ may cause multiple solutions for fitted a, b and E_{th}, thus lead to LASSO regression failure. After several attempts, we omitted E_{th} as follow:

$$\lambda = aE^b + \frac{E - \exp(-B/A)}{A \ln E + B}, \tag{9}$$

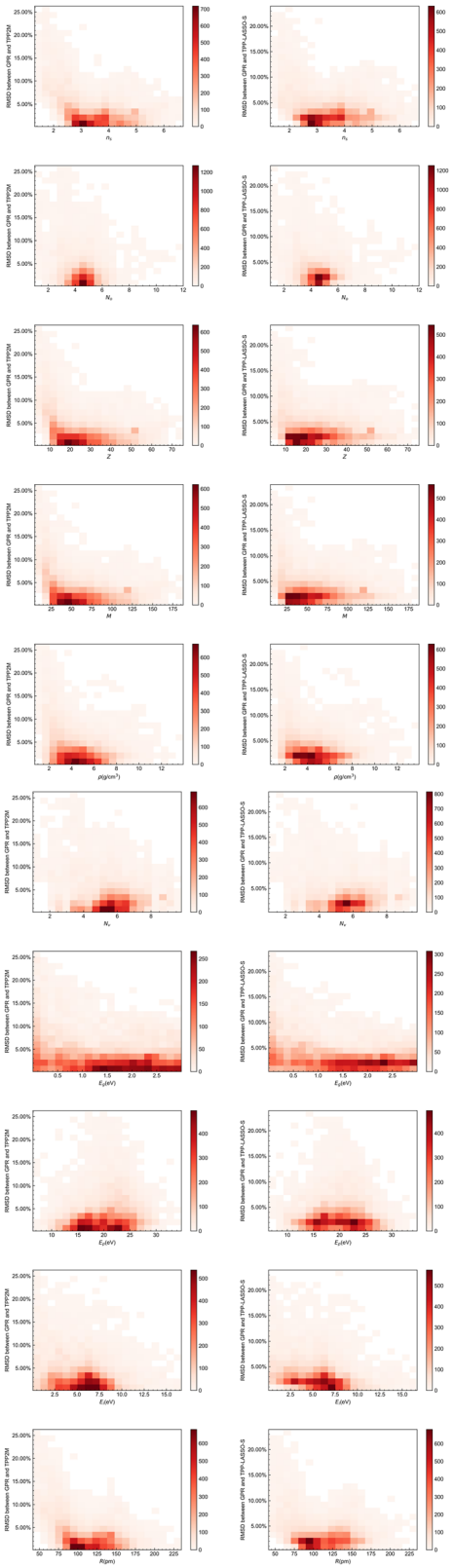

Fig. 12 Distributions between RMSDs and prediction parameters for all 12 039 narrow bandgap inorganic compounds ($E_g < 3$ eV). The left/right panel is for the RMSDs between our prediction and TPP-2M formula[42] /TPP-LASSO-S formula[50]. The shade of red color shows the density of RMSDs distribution.

Table 4 Parameter combination framework based on 10 basic parameters

ID	Description	Quantity
A1	10 basic parameters	10
B1	$(f+g)$, $(f-g)$, (f^2+g^2), (f^2-g^2); $f, g \in \{E_p, E_i, E_g\}$	159
B2	$f \cdot g$; $f, g \in \{A1\}$	
B3	f/g; $f, g \in \{A1\}$	
C1	$f \cdot g$; $f \in \{A\}$, $g \in \{B\}$	3350
C2	f/g; $f, g \in \{A, B\}$	

thus using eqn (9), least squares method is applied to fit a, b, A and B, while avoid the multiple solutions with acceptable accuracy loss. To regress the extracted a, b, A and B, a massive descriptor pool is built following Table 4.

Then LASSO is used to shrink out the most important descriptors in the descriptor pool in Table 4, with the target of fitted a, b, A and B. As results of the framework based on LASSO, we listed the most important 20 parameters in Table 5 and shown RMSD distribution in Fig. 13. First of all, thanks to the robust of our proposed framework, the LASSO regression on the GPR predicted database is accurate enough on all energies to describe narrow bandgap inorganic compounds' ($E_g < 3$ eV) IMFPs. As shown in Fig. 13, the RMSDs of most narrow bandgap inorganic compounds ($E_g < 3$ eV) are between 2% and 5%, which allow us to discuss the importance of the critical descriptors chosen by LASSO statistically.

For a, b, A and B, Table 5 list the most important 20 descriptors out of LASSO selected descriptors (20/selected) and their linear coefficients separately. We note that here the linear coefficients are for the standardized descriptors to ensure their equal weights in LASSO. In fact, for each value for each descriptor D, we used D^* in the actual LASSO procedure:

$$D^* = \frac{D-\mu}{\sigma}, \tag{10}$$

where μ is the averaged value of D, and σ is the standard variance of D for all 12 039 materials. Thus the average values for each derived descriptor D^* are 0 and the standard variance is 1. Since LASSO forms a linear combination for the principal descriptors, for example, for the LASSO regression result of a in eqn (9) (first column in Table 5), the relationship is as follows:

$$\begin{aligned} a = & -1196.90 \times \left(\left(E_g/n_s\right) \times E_g\right)^* - 1152.79 \cdot \left(\left(E_g^2 + E_i^2\right) \cdot E_g\right)^* \\ & + 1104.63 \times \left(\left(E_p \cdot E_g\right) \cdot E_g\right)^* - \cdots, \end{aligned} \tag{11}$$

where the asterisk above each descriptor indicates the standardized descriptor, according to eqn (10). Due to the fact that in eqn (11), all descriptors follow the same standardized distribution as described before, so the absolute value of the coefficient reflects the proportion of the corresponding descriptor, its relative value for the same fitting parameter (a, b, A or B) is actually the importance for each descriptor. The descriptors are ordered by the importance in Table 5, thus one can easily realize the frequency of each parameters appearance in principle descriptors' combinations. E_g appears in most of the

Table 5 The most important 20 descriptors for fitting parameters *a*, *b*, *A*, *B* in eqn (9). Formulae appeared here are not simplified in order to show how the parameter combination works as Table 4. The absolute values of coefficients reflect the importance of corresponding descriptors for same fitting parameter

a (20/346)		*b* (20/52)		*A* (20/190)		*B* (20/206)	
*D** (standardized descriptor)	Linear coefficient	*D** (standardized descriptor)	Linear coefficient	*D** (standardized descriptor)	Linear coefficient	*D** (standardized descriptor)	Linear coefficient
$((E_g/n_s)\cdot E_g)$	−1196.90	$(E_p/(E_p + E_g))$	0.23	(E_g/R)	−1.90	$((E_g^2 + E_i^2)\cdot E_g)$	8.39
$((E_g^2 + E_i^2)\cdot E_g)$	−1152.79	$(E_g/(E_p/E_g))$	0.11	$(R\cdot E_g)$	1.71	$((E_g^2 - E_i^2)\cdot E_g)$	7.25
$((E_p\cdot E_g)\cdot E_g)$	1104.63	$((M\cdot E_g)\cdot M)$	0.06	$((E_p^2 - E_g^2)\cdot E_g)$	1.46	$((M\cdot E_g)\cdot E_g)$	6.06
$((E_g^2 - E_i^2)\cdot E_g)$	−1011.64	$(R\cdot E_p)$	0.05	(n_s/R)	1.27	(n_s/R)	−5.25
(E_g/n_s)	930.90	$((N_o\cdot E_g)\cdot E_g)$	0.05	$((Z/M)\cdot n_s)$	−1.19	(E_g/R)	4.76
$(E_g/(n_s/E_g))$	−687.04	$(n_s/(\rho/E_g))$	−0.02	$(E_g/(n_s\cdot E_g))$	1.12	$((Z/M)\cdot n_s)$	3.89
$((E_g + E_i)\cdot E_g)$	444.25	$(n_s/(E_p^2 + E_i^2))$	−0.02	$((M\cdot E_g)\cdot E_g)$	−0.83	$((E_p^2 - E_g^2)\cdot E_g)$	−3.34
$((E_p\cdot E_g)\cdot n_s)$	−432.04	$((n_s\cdot E_g)/\rho)$	−0.01	(R/n_s)	−0.70	$(E_g/(N_o/E_g))$	−2.55
$((n_s\cdot E_g)\cdot E_g)$	420.62	$((E_g - E_i)\cdot E_g)$	0.01	$(E_g/(N_o/E_g))$	0.65	$((E_g/n_s)/E_p)$	2.45
$((E_p - E_g)\cdot E_g)$	−395.15	$((E_p/N_v)/n_s)$	−0.01	$(E_i/(E_g + E_i))$	0.61	$(R\cdot E_g)$	−2.34
$((E_g - E_i)\cdot E_g)$	365.54	$((E_p - E_g)\cdot E_g)$	−0.01	$((\rho\cdot E_g)\cdot\rho)$	0.51	$(E_g/(M/E_g))$	−2.29
(n_s/R)	255.35	(M/Z)	0.01	$(E_g/(M/E_g))$	0.49	$(E_g/(E_p/E_g))$	−2.25
$((E_g/n_s)/E_p)$	247.73	(R/n_s)	0.01	$((E_g/n_s)/E_p)$	−0.47	$(E_i/(E_g + E_i))$	−2.22
$((N_v\cdot E_g)\cdot E_g)$	−237.75	(E_p/R)	−0.01	$(E_g/(E_g + E_i))$	−0.47	(M/R)	−2.17
$((Z\cdot E_g)\cdot E_g)$	−224.07	$((N_v\cdot E_g)/N_o)$	−0.01	$((E_g/E_p)/n_s)$	−0.47	$((E_g/\rho)/n_s)$	2.08
$((\rho\cdot E_g)\cdot E_g)$	−208.25	$((E_g\cdot E_i)\cdot E_g)$	0.01	(M/R)	0.44	(N_o/R)	1.91
$(E_g/(N_v/E_g))$	207.98	$((Z/n_s)/n_s)$	−0.01	$((n_s/M)\cdot Z)$	−0.37	$((\rho\cdot E_g)\cdot M)$	−1.90
(E_g/R)	206.30	$((E_g/n_s)\cdot E_g)$	0.01	$((E_g/N_o)/n_s)$	−0.36	$((\rho\cdot E_g)\cdot\rho)$	−1.89
$((E_p\cdot E_g)\cdot N_o)$	−190.46	(n_s/R)	−0.01	$(E_g/(\rho/E_g))$	0.36	$(N_o/(n_s\cdot N_o))$	−1.88
$(E_g/(n_s/N_o))$	181.35	$(E_g/(\rho/E_g))$	0.01	(N_o/R)	−0.34	$((E_p^2 - E_i^2)/n_s)$	1.87

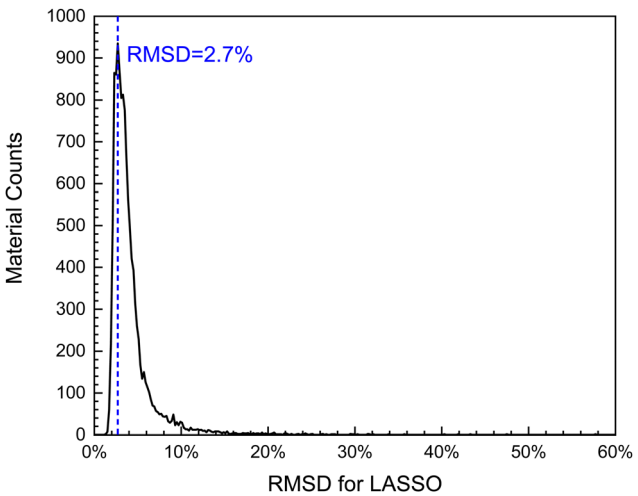

Fig. 13 RMSD distribution between LASSO results and the target GPR predicted IMFPs for 12 039 narrow bandgap inorganic compounds ($E_g < 3$ eV).

important descriptors especially for *a* and *b*, which proves that E_g is a critical parameter for describing low energy IMFPs. In fact, in our previous work,[37] we also analysis the effect of E_g, together with the plasmon lifetime or phonon-phonon interaction, through SE-MA[37] and FPA[43] calculated some narrow bandgap inorganic compounds ($E_g < 3$ eV) or insulators like Si and MgO. The most important descriptor $E_p/(E_p + E_g)$ of *b* in Table 5 reveals that the relationship between bulk plasmon energy and bandgap energy has a high probability to be the key to describe low energy IMFPs.

2.6 Prediction for organic compounds using the model trained by inorganic compounds

We tried to use our trained model based on inorganic compounds in this work to predict 14 organic compounds and liquid water IMFPs in ref. 67. The true value of organic compounds IMFP database is also taken from ref. 67. Then the prediction RMSD result is shown in Table 6. The average RMSD is larger than 10%, besides there are some materials with extra-large error, *e.g.* b-carotene and diphenyl-hexatriene.

Table 6 RMSDs of organic compounds and water when using the model trained by inorganic compounds IMFP. The true value of organic compounds IMFP database is also taken from ref. 67

Material	Molecular formula	RMSD	Material	Molecular formula	RMSD
26-*n*-Paraffin	$C_{26}H_{54}$	10.42%	Polyethylene	C_2H_4	6.09%
Adenine	$C_5H_5N_5$	7.37%	Polymethylmethacrylate	$C_5H_8O_2$	11.71%
b-Carotene	$C_{40}H_{56}$	34.89%	Polystyrene	C_8H_8	10.95%
Diphenyl-hexatriene	$C_{18}H_{16}$	33.75%	Poly(2-vinylpridine)	C_7H_7N	10.68%
Guanine	$C_5H_5N_5O$	10.73%	Thymine	$C_5N_2H_6O_2$	7.79%
Kapton	$C_{22}H_{10}N_2O_5$	15.22%	Uracil	$C_4N_2H_4O_2$	6.60%
Polyacetylene	C_2H_2	17.27%	Water	H_2O	9.43%
Poly(butene-1-sulfone)	$C_4H_8SO_2$	7.77%	**Average**		**17.18%**

Unfortunately, it is not desirable to predict the organic compounds IMFPs with only inorganic compounds.

Above all, we will try to train a model separately for organic compounds, or try to train a model for compounds together in future work, discuss various of details about machine learning application, including the broader application scenarios of ML in surface analysis.

3. Conclusions

In this work, we continue using ML to describe IMFPs. Instead of using only IMFPs for elemental solids as in ref. 52, we extend the use of ML to describe and predict IMFPs of inorganic compound materials. This work focuses on training the ML scheme with different databases, including IMFP databases for inorganic compounds only or elemental solids together with inorganic compounds. We show that the proposed GPR ML scheme is effective with both database types. However, the inorganic compounds-only database shows better results when using LOOCV to validate predictive ability.

We also consider the material and energy dependence separately when using the ML scheme. The successful prediction of each data point for different materials at the same energy reveals that the GPR can learn IMFP data without energy dependence. Therefore, this advantage can be used to apply the GPR to experimental IMFP data points to replace empirical formulas.

Finally, we produced an extensive IMFP database for 12 039 narrow bandgap inorganic compounds ($E_g < 3$ eV) based on the trained database. From the distribution between the RMSDs of the predicted IMFPs and E_g, we found that E_g may have significant descriptive power for narrow bandgap inorganic compounds ($E_g < 3$ eV) IMFPs. We also provide evidence to support this hypothesis by using our previously proposed framework[50] based on LASSO.

Author contributions

X. L. wrote the program, performed the analyzation of results, and wrote the initial manuscript. D. B. L. gave help with the code programming. Z. F. H., K. N. and Y. S. gave crucial suggestions to the ML program and the initial manuscript. B. D. and Z. J. D. supervised the research. H. Y. and S. T. gave physics picture and suggestions. All authors discussed and commented on the manuscript. All the authors developed the concepts together and participated the discussions of the work.

Data availability

All data generated and/or analyzed during this study are included in this article.

Conflicts of interest

The authors declare no competing financial or non-financial interests.

Acknowledgements

This work was supported in part by JSPS KAKENHI (No. JP21K14656), the National Institute for Materials Science under the Support system for curiosity-driven research, the Grant for Basic Science Research Projects from The Sumitomo Foundation, the Kao Foundation for Arts and Sciences, the Kurata Grants from The Hitachi Global Foundation, the Iketani Science & Technology Foundation and the Chinese Education Ministry through the "111" Project2.0 (BP0719016). The calculations in this study were performed on the Numerical Materials Simulator at NIMS. We also thank the USTC Supercomputing Center for parallel computation support.

References

1 T. Mueller, M. Kinoshita, M. Steiner, V. Perebeinos, A. A. Bol and D. B. Farmer, *et al.*, Efficient narrow-band light emission from a single carbon nanotube p–n diode, *Nat. Nanotechnol.*, 2010, **5**, 27–31.
2 H. Liu, A. T. Neal, M. Si, Y. Du and D. Y. Peide, The effect of dielectric capping on few-layer phosphorene transistors: tuning the Schottky barrier heights, *IEEE Electron. Dev. Lett.*, 2014, **35**, 795–797.
3 J. H. Lau, Recent advances and new trends in nanotechnology and 3D integration for semiconductor industry, *ECS Trans.*, 2012, **44**, 805.
4 International Organization for Standardization. ISO 18115:2010. Surface Chemical Analysis-Vocabulary-Part 1: General terms and terms used in spectroscopy. Geneva: ISO; 2010.
5 L. H. Yang, K. Tőkési, J. Tóth, B. Da, H. M. Li and Z. J. Ding, High Precision Determination of Optical Properties of Silicon and Germanium from Reflection Electron Energy Loss Spectroscopy Spectra, *Phys. Rev. B*, 2019, **100**, 245209.
6 L. H. Yang, K. Tőkési, J. Tóth, B. Da and Z. J. Ding, Revision of Optical Property of Silicon by a Reverse Monte Carlo Analysis of Reflection Electron Energy Loss Spectroscopy Spectra, *J. Phys.: Conf. Ser.*, 2020, **1412**, 202026.
7 L. Yang, B. Da, K. Tőkési and Z. J. Ding, Individual Separation of Surface, Bulk and Begrenzungs Effect Components in the Surface Electron Energy Spectra, *Sci. Rep.*, 2021, **11**, 5954.
8 L. Yang, A. Hussain, S. Mao, B. Da, K. Tőkési and Z. J. Ding, Electron Backscattering Coefficients of Molybdenum and Tungsten Based on the Monte Carlo Simulations, *J. Nucl. Mater.*, 2021, **553**, 153042.
9 D. Lu, K. Goto, B. Da, J. Liu, H. Yoshikawa, S. Tanuma and Z. J. Ding, Secondary Electron-, Auger Electron- and Reflected Electron-Spectroscopy Study on sp^2-Hybridization Carbon Materials: HOPG, Carbon Glass and Carbon Fiber, *J. Electron. Spectrosc. Relat. Phenom.*, 2021, **250**, 147086.
10 L. H. Yang, J. M. Gong, A. Sulyok, M. Menyhárd, G. Sáfrán and K. Tőkési, *et al.*, Optical Properties of Amorphous Carbon Determined by Reflection Electron Energy Loss

Spectroscopy Spectra, *Phys. Chem. Chem. Phys.*, 2021, **23**, 25335–25346.
11 H. Xu, B. Da, J. Tóth, K. Tőkési and Z. J. Ding, Absolute determination of optical constants by reflection electron energy loss spectroscopy, *Phys. Rev. B*, 2017, **95**, 195417.
12 H. Xu, L. H. Yang, J. Tóth, K. Tőkési, B. Da and Z. J. Ding, Absolute Determination of Optical Constants of Three Transition Metals Using Reflection Electron Energy Loss Spectroscopy, *J. Appl. Phys.*, 2018, **123**, 043306.
13 L. H. Yang, K. Tőkési, J. Tóth, B. Da and Z. J. Ding, Calculation of Electron Inelastic Mean Free Path of Three Transition Metals from Reflection Electron Energy Loss Spectroscopy Spectrum Measurement Data, *Eur. Phys. J. D*, 2019, **73**, 21.
14 J. D. Bourke and C. T. Chantler, Momentum-dependent lifetime broadening of electron energy loss spectra: a self-consistent coupled-plasmon model, *J. Phys. Chem. Lett.*, 2015, **6**, 314–319.
15 C. T. Chantler and J. D. Bourke, X-ray spectroscopic measurement of photoelectron inelastic mean free paths in molybdenum, *J. Phys. Chem. Lett.*, 2010, **1**, 2422–2427.
16 C. J. Powell and A. Jablonski, Surface sensitivity of X-ray photoelectron spectroscopy, *Nucl. Instrum. Methods Phys. Res., Sect. A*, 2009, **601**, 54–65.
17 W. S. M. Werner, W. Smekal, H. Stori, H. Winter, G. Stefani and A. Ruocco, *et al.*, Emission-depth-selective Auger photoelectron coincidence spectroscopy, *Phys. Rev. Lett.*, 2005, **94**, 038302.
18 Z. J. Ding, K. Salma, H. M. Li, Z. M. Zhang, K. Tokesi and D. Varga, *et al.*, Monte Carlo Simulation Study of Electron Interaction with Solids and Surfaces, *Surf. Interface Anal.*, 2006, **38**, 657–663.
19 N. Cao, B. Da, Y. Ming, S. F. Mao, K. Goto and Z. J. Ding, Monte Carlo Simulation of Full Energy Spectrum of Electrons Emitted from Silicon in Auger electron Spectroscopy, *Surf. Interface Anal.*, 2015, **47**, 113–119.
20 Z. J. Ding and R. Shimizu, A Monte Carlo modeling of electron interaction with solids including cascade secondary electron production, *Scanning*, 1996, **18**, 92–113.
21 B. Da, Z. Y. Li, H. C. Chang, S. F. Mao and Z. J. Ding, A Monte Carlo Study of Reflection Electron Energy Loss Spectroscopy Spectrum of a Carbon Contaminated Surface, *J. Appl. Phys.*, 2014, **116**, 124307.
22 B. Da, S. F. Mao, G. H. Zhang, X. P. Wang and Z. J. Ding, Monte Carlo Modeling of Surface Excitation in Reflection Electron Energy Loss Spectroscopy Spectrum for Rough Surfaces, *J. Appl. Phys.*, 2012, **112**, 034310.
23 B. Da, Y. Sun, S. F. Mao, Z. M. Zhang, H. Jin and H. Yoshikawa, *et al.*, A Reverse Monte Carlo Method for Deriving Optical Constants of Solids from REELS Spectra, *J. Appl. Phys.*, 2013, **113**, 214303.
24 B. Da, S. F. Mao and Y. Sun, A New Analytical Method in Surface Electron Spectroscopy: Reverse Monte Carlo Method, *e-J. Surf. Sci. Nanotechnol.*, 2012, **10**, 441–446.
25 B. Da, L. H. Yang, J. W. Liu, Y. G. Li, S. F. Mao and Z. J. Ding, Monte Carlo Simulation Study of Reflection Electron Energy Loss Spectroscopy of a Fe/Si Layered Nanostructure, *Surf. Interface Anal.*, 2020, **52**, 742–754.
26 Z. J. Ding, C. Li, B. Da and J. W. Liu, Charging Effect Induced by Electron Beam Irradiation: A Review, *Sci. Technol. Adv. Mater.*, 2021, **22**, 932–971.
27 Z. J. Ding, W. S. Tan and Y. G. Li, Improved Calculation of the Backscattering Factor for Quantitative Analysis by Auger Electron Spectroscopy, *J. Appl. Phys.*, 2006, **99**, 084903.
28 R. G. Zeng, Z. J. Ding, Y. G. Li and S. F. Mao, A Calculation of Backscattering Factor Database for Quantitative Analysis by Auger Electron Spectroscopy, *J. Appl. Phys.*, 2008, **104**, 114909.
29 Y. B. Zou, S. F. Mao, B. Da and Z. J. Ding, Surface sensitivity of secondary electrons emitted from amorphous solids: calculation of mean escape depth by a Monte Carlo method, *J. Appl. Phys.*, 2016, **120**, 235102.
30 B. Da, K. Salma, H. Ji, S. F. Mao, G. H. Zhang and X. P. Wang, *et al.*, Surface Excitation Parameter for Rough Surfaces, *Appl. Surf. Sci.*, 2015, **356**, 142–149.
31 B. Da, Y. Sun, S. F. Mao and Z. J. Ding, Systematic Calculation of the Surface Excitation Parameters for 22 Materials, *Surf. Interface Anal.*, 2013, **45**, 773–780.
32 Z. Zheng, B. Da, S. F. Mao and Z. J. Ding, Calculation of Surface Excitation Parameters by a Monte Carlo Method, *Chin. J. Chem. Phys.*, 2017, **30**, 83–89.
33 D. R. Penn, Electron mean-free-path calculations using a model dielectric function, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1987, **35**, 482.
34 N. D. Mermin, Lindhard dielectric function in the relaxation-time approximation, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1970, **1**, 2362.
35 B. Da, H. Shinotsuka, H. Yoshikawa, Z. J. Ding and S. Tanuma, Extended Mermin method for calculating the electron inelastic mean free path, *Phys. Rev. Lett.*, 2014, **113**, 063201.
36 B. Da, H. Shinotsuka, H. Yoshikawa and S. Tanuma, Comparison of the Mermin and Penn models for inelastic mean-free-path calculations for electrons based on a model using optical energy-loss functions, *Surf. Interface Anal.*, 2019, **51**, 627–640.
37 B. Da, X. Liu, L. H. Yang, J. M. Gong, Z. J. Ding and H. Shinotsuka, *et al.*, Evaluation of Dielectric Function Models for Calculation of Electron Inelastic Mean Free Path, *J. Appl. Phys.*, 2022, **131**, 175301.
38 S. Tanuma, C. J. Powell and D. R. Penn, Calculations of electron inelastic mean free paths for 31 materials, *Surf. Interface Anal.*, 1988, **11**, 577–589.
39 S. Tanuma, C. J. Powell and D. R. Penn, Calculations of electron inelastic mean free paths. II. Data for 27 elements over the 50–2000 eV range, *Surf. Interface Anal.*, 1991, **17**, 911–926.
40 S. Tanuma, C. J. Powell and D. R. Penn, Calculations of electron inelastic mean free paths. III. Data for 15 inorganic compounds over the 50–2000 eV range, *Surf. Interface Anal.*, 1991, **17**, 927–939.
41 S. Tanuma, C. J. Powell and D. R. Penn, Calculations of electron inelastic mean free paths. V. Data for 14 organic compounds over the 50–2000 eV range, *Surf. Interface Anal.*, 1994, **21**, 165–176.

42 H. Shinotsuka, S. Tanuma, C. J. Powell and D. R. Penn, Calculations of electron inelastic mean free paths. X. Data for 41 elemental solids over the 50 eV to 200 keV range with the relativistic full Penn algorithm, *Surf. Interface Anal.*, 2015, **47**, 871–888.

43 H. Shinotsuka, S. Tanuma, C. J. Powell and D. R. Penn, Calculations of electron inelastic mean free paths. XII. Data for 42 inorganic compounds over the 50 eV to 200 keV range with the full Penn algorithm, *Surf. Interface Anal.*, 2019, **51**, 427–457.

44 Y. Sun, H. Xu, B. Da, S. F. Mao and Z. J. Ding, Calculations of energy-loss function for 26 materials, *Chin. J. Chem. Phys.*, 2016, **29**, 663.

45 L. H. Yang, M. Menyhárd, A. Sulyok, K. Tőkési and Z. J. Ding, Optical Properties and Excitation Energies of Iridium Derived from Reflection Electron Energy Loss Spectroscopy Spectra, *Appl. Surf. Sci.*, 2018, **456**, 999–1003.

46 W. H. Gries, A universal predictive equation for the inelastic mean free pathlengths of X-ray photoelectrons and Auger electrons, *Surf. Interface Anal.*, 1996, **24**, 38–50.

47 S. Tanuma, C. J. Powell and D. R. Penn, Calculations of electron inelastic mean free paths (IMFPs) VI. analysis of the gries inelastic scattering model and predictive IMFP equation, *Surf. Interface Anal.*, 1997, **25**, 25–35.

48 M. P. Seah, An accurate and simple universal curve for the energy-dependent electron inelastic mean free path, *Surf. Interface Anal.*, 2012, **44**, 497–503.

49 H. A. Bethe, Zur Theorie des Durchgangs schneller Korpuskularstrahlen durch Materie, *Ann. Phys.*, 1930, **5**, 325–400.

50 X. Liu, Z. F. Hou, D. B. Lu, B. Da, H. Yoshikawa and S. Tanuma, *et al.*, Unveiling the principle descriptor for predicting the electron inelastic mean free path based on a machine learning framework, *Sci. Technol. Adv. Mater.*, 2019, **20**, 1090–1102.

51 C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, MIT Press, 2006.

52 X. Liu, L. H. Yang, Z. F. Hou, B. Da, K. Nagata and H. Yoshikawa, *et al.*, Machine learning approach for the prediction of electron inelastic mean free paths, *Phys. Rev. Mater.*, 2021, **5**, 033802.

53 A. Jablonski, P. Mrozek, G. Gergely, M. Menhyard and A. Sulyok, The inelastic mean free path of electrons in some semiconductor compounds and metals, *Surf. Interface Anal.*, 1984, **6**, 291–294.

54 H. Shinotsuka, H. Yoshikawa and S. Tanuma, First-principles Calculations of Optical Energy Loss Functions for 30 Compound and 5 Elemental Semiconductors, *e-J. Surf. Sci. Nanotechnol.*, 2021, **19**, 70–87.

55 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion and O. Grisel, *et al.*, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

56 S. Tanuma, C. J. Powell and D. R. Penn, Calculations of electron inelastic mean free paths (IMFPS). IV. Evaluation of calculated IMFPs and of the predictive IMFP formula TPP-2 for electron energies between 50 and 2000 eV, *Surf. Interface Anal.*, 1993, **20**, 77–89.

57 D. Tahir and S. Tougaard, Electronic and optical properties of Cu, CuO and Cu_2O studied by electron spectroscopy, *J. Phys.: Condens. Matter*, 2012, **24**, 175002.

58 H. Jin, S. K. Oh, H. J. Kang and S. Tougaard, Electronic properties of ultrathin HfO_2, Al_2O_3, and Hf–Al–O dielectric films on Si(100) studied by quantitative analysis of reflection electron energy loss spectra, *J. Appl. Phys.*, 2006, **100**, 083713.

59 D. Tahir, E. K. Lee, S. K. Oh, H. J. Kang, S. Heo and J. G. Chung, *et al.*, Dielectric and optical properties of Zr silicate thin films grown on Si(100) by atomic layer deposition, *J. Appl. Phys.*, 2009, **106**, 084108.

60 B. L. Henke, E. M. Gullikson and J. C. Davis, X-ray interactions: photoabsorption, scattering, transmission, and reflection at E = 50–30 000 eV, Z = 1–92, *At. Data Nucl. Data Tables*, 1993, **54**, 181.

61 D. E. Cullen, J. H. Hubbell and L. Kissel, *EPDL97: The evaluated data library, 1997 version*, Lawrence Livermore National Lab, CA (United States), 1997.

62 S. Tanuma, C. J. Powell and D. R. Penn, Use of sum rules on the energy-loss function for the evaluation of experimental optical data, *J. Electron Spectrosc. Relat. Phenom.*, 1993, **62**, 95–109.

63 M. Dayah, Periodic Table - Ptable [internet]; 1997 Oct 1 [updated 2022 Oct 12]. Available from: **https://ptable.com**.

64 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards and S. Dacek, *et al.*, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**, 011002.

65 S. Tanuma, C. J. Powell and D. R. Penn, Calculation of electron inelastic mean free paths (IMFPs) VII. Reliability of the TPP-2M IMFP predictive equation, *Surf. Interface Anal.*, 2003, **35**, 268–275.

66 B. Ziaja, R. A. London and J. Hajdu, Ionization by impact electrons in solids: Electron mean free path fitted over a wide energy range, *J. Appl. Phys.*, 2006, **99**, 033514.

67 H. Shinotsuka, S. Tanuma and C. J. Powell, Calculations of electron inelastic mean free paths. XIII. Data for 14 organic compounds and water over the 50 eV to 200 keV range with the relativistic full Penn algorithm, *Surf. Interface Anal.*, 2022, **54**, 534.

