Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2024

Supplementary Information for

Multi-fidelity Machine Learning for Predicting Bandgaps of Nonlinear Optical Crystals

Zhaoxi Yu,[a] Pujie Xue,[a] Bin-Bin Xie,[b] Lin Shen,[a,c*] and Wei-Hai Fang[a,d]

a) Key Laboratory of Theoretical and Computational Photochemistry of Ministry of Education, College of Chemistry, Beijing Normal University, Beijing 100875, P. R. China

b) Hangzhou Institute of Advanced Studies, Zhejiang Normal University, Hangzhou 311231, Zhejiang, P. R. China

c) Yantai-Jingshi Institute of Material Genome Engineering, Yantai 265505, Shandong, P. R. China

d) Shandong Laboratory of Yantai Advanced Materials and Green Manufacturing, Yantai 264006, Shandong, P. R. China

---

[*] Author to whom correspondence should be addressed.

Email: lshen@bnu.edu.cn

Table of Contents

Section S1. Supplementary information on multilevel descriptors

Section S2. Supplementary results of machine learning predictions

References

Section S1. Supplementary information on multilevel descriptors

Note S1. Multilevel descriptors for nonlinear optical crystals

The multilevel descriptors were proposed by our group (*J. Phys. Chem. C*, 2021, 125, 25175), which includes the atomic properties, the characters of fundamental structural groups, and the crystal structures to describe inorganic NLO crystals. The first-level features consist of the basic properties of atoms in the chemical composition of NLO crystals, for example, the Pauling electronegativity, atomic mass and the difference of d and f electrons between the atom and the noble gas in the preceding element period. The second-level features can be obtained as follows. First, based on the chemical compositions of crystals and the chemical valences of atoms, two types of structural groups, namely, acid radicals (ARs) and metallic oxides (MOs), were identified. This is inspired by the well-known anionic group theory[S1] in the research field of NLO materials. Subsequently, the geometries of isolated ARs and MOs were optimized, followed by further electronic structure calculations to obtain total dipole moments, anisotropic quadrupole moments, anisotropic dipole polarizabilities and total first hyper-polarizabilities. These properties were employed to approximate the properties of anionic and cationic groups in crystals and collected as the second-level features. The third-level features are derived from the crystal structure of materials, including space group, cell parameters, band gap and multiplicity of Wyckoff positions.

In our previous machine learning tasks[S2] towards the prediction of two NLO properties (i.e., birefringence and second-order nonlinear coefficients), machine learning classifiers that only involve the first and second-level features exhibit good performances, while the third-level features with the crystal structural information cannot further enhance the accuracy. It provides an effective filter in the initial step of high-throughput NLO materials discovery. In this study, we also applied the crystal-structure-free models with the same feature sets as that in our prior research to predict the bandgaps of NLO crystals.

Table S1. Definitions of the first-level descriptors.[S2]

ID	Description	ID	Description	ID	Description
F1[(a)]	minimal atomic mass	F14	weighted averaged Pauling electronegativity	F27	minimal atomic electron affinity
F2[(b)]	weighted averaged atomic mass	F15	maximal Pauling electronegativity	F28	weighted averaged atomic electron affinity
F3[(c)]	maximal atomic mass	F16	weighted averaged difference of *s* electrons between each atom and the noble gas atom in the preceding element period	F29	maximal atomic electron affinity
F4	minimal element period number	F17	weighted averaged difference of *p* electrons between each atom and the noble gas atom in the preceding element period	F30	minimal melting point of elements
F5	weighted averaged element period number	F18	weighted averaged difference of *d* electrons between each atom and the noble gas atom in the preceding element period	F31	weighted averaged melting point of elements
F6	maximal element period number	F19	weighted averaged difference of *f* electrons between each atom and the noble gas atom in the preceding element period	F32	maximal melting point of elements
F7	minimal element group number	F20	weighted averaged difference of unfilled *s* orbitals between each atom and the noble gas atom in the preceding element period	F33	minimal boiling point of elements
F8	weighted averaged element group number	F21	weighted averaged difference of unfilled *p* orbitals between each atom and the noble gas atom in the preceding element period	F34	weighted averaged boiling point of elements
F9	maximal element group number	F22	weighted averaged difference of unfilled *d* orbitals between each atom and the noble gas atom in the preceding element period	F35	maximal boiling point of elements
F10	minimal atomic van der Waals radius	F23	weighted averaged difference of unfilled *f* orbitals between each atom and the noble gas atom in the preceding element period	F36	minimal density of elements
F11	weighted averaged atomic van der Waals radius	F24	minimal atomic ionization energy	F37	weighted averaged density of elements
F12	maximal atomic van der Waals radius	F25	weighted averaged atomic ionization energy	F38	maximal density of elements
F13	minimal Pauling electronegativity	F26	maximal atomic ionization energy		

[(a)] In F1, F4, F7, F10, F13, F24, F27, F30, F33 and F36, "minimal" denotes the minimal value among all elements containing in the compound. Take $NaCaBe_2B_2O_6F$ as an instance. The above features are calculated as

$$\min[\mathrm{AP(Na)}, \mathrm{AP(Ca)}, \mathrm{AP(Be)}, \mathrm{AP(B)}, \mathrm{AP(O)}, \mathrm{AP(F)}]$$

where AP(X) denotes the corresponding atomic property of atom X.

[b] In F2, F5, F8, F11, F14, F16, F17, F18, F19, F20, F21, F22, F23, F25, F28, F31, F34 and F37, “weighted averaged” denotes the average with the stoichiometric ratio of each element as its pre-determined weight. Take $NaCaBe_2B_2O_6F$ as an instance. The above features are calculated as

$$\frac{\mathrm{AP(Na)} + \mathrm{AP(Ca)} + 2\mathrm{AP(Be)} + 2\mathrm{AP(B)} + 6\mathrm{AP(O)} + \mathrm{AP(F)}}{1 + 1 + 2 + 2 + 6 + 1}$$

[c] In F3, F6, F9, F12, F15, F26, F29, F32, F35 and F38, “maximal” denotes the maximal value among all elements containing in the compound. Take $NaCaBe_2B_2O_6F$ as an instance. The above features are calculated as

$$\max[\mathrm{AP(Na)}, \mathrm{AP(Ca)}, \mathrm{AP(Be)}, \mathrm{AP(B)}, \mathrm{AP(O)}, \mathrm{AP(F)}]$$

Table S2. Definitions of the second-level descriptors.[S2]

ID [(a)]	**Description**	**ID** [(a)]	**Description**
SN1[(b)]	minimal HOMO-LUMO gap of ARs	SP1[(c)]	minimal HOMO-LUMO gap of MOs
SN2[(d)]	averaged HOMO-LUMO gap of ARs	SP2[(e)]	averaged HOMO-LUMO gap of MOs
SN3[(f)]	maximal HOMO-LUMO gap of ARs	SP3[(g)]	maximal HOMO-LUMO gap of MOs
SN4[(h)]	minimal total dipole moment of ARs	SP4	minimal total dipole moment of MOs
SN5	averaged total dipole moment of ARs	SP5	averaged total dipole moment of MOs
SN6	maximal total dipole moment of ARs	SP6	maximal total dipole moment of MOs
SN7[(i)]	minimal aniso-quadrupole moment of ARs	SP7	minimal aniso-quadrupole moment of MOs
SN8	averaged aniso-quadrupole moment of ARs	SP8	averaged aniso-quadrupole moment of MOs
SN9	maximal aniso-quadrupole moment of ARs	SP9	maximal aniso-quadrupole moment of MOs
SN10[(j)]	minimal aniso-polarizability of ARs	SP10	minimal aniso-polarizability of MOs
SN11	averaged aniso-polarizability ARs	SP11	averaged aniso-polarizability MOs
SN12	maximal aniso-polarizability of ARs	SP12	maximal aniso-polarizability of MOs
SN13[(k)]	minimal total first hyper-polarizability of ARs	SP13	minimal total first hyper-polarizability of MOs
SN14	averaged total first hype-polarizability of ARs	SP14	averaged total first hype-polarizability of MOs
SN15	maximal total first hyper-polarizability of ARs	SP15	maximal total first hyper-polarizability of MOs
SN16[(l)]	total charge of ARs	SP16	total charge of MOs
SN17	total multiplicity of ARs	SP17	total multiplicity of MOs
SN18[(m)]	minimal flexibility index of ARs	SP18	minimal flexibility index of MOs
SN19	total flexibility index of ARs	SP19	total flexibility index of MOs
SN20	averaged flexibility index of ARs	SP20	averaged flexibility index of MOs
SN21	maximal flexibility index of ARs	SP21	maximal flexibility index of MOs
SN22	averaged charge of ARs		

[(a)] SP and SN denote the second-level features extracted from the properties of isolated metallic oxides (MOs) and acid radicals (ARs), respectively.

[(b)] In SN1, SN4, SN7, SN10, SN13 and SN18, "minimal" denotes the minimal value among all ARs extracted from the compound. Take $NaCaBe_2B_2O_6F$ as an instance, SN1 is calculated as $\min[\Delta(BeO_2^{2-}), \Delta(BO_3^{3-}), \Delta(F^-)]$, where $\Delta(BeO_2^{2-})$, $\Delta(BO_3^{3-})$ and $\Delta(F^-)$ denote the HOMO-LUMO gap of BeO_2^{2-}, BO_3^{3-} and F^-, respectively.

[(c)] In SP1, SP4, SP7, SP10, SP13 and SP18, "minimal" denotes the minimal value among all MOs extracted from the compound. Take $NaCaBe_2B_2O_6F$ as an instance, SP1 is calculated as $\min[\Delta(Na_2O), \Delta(CaO)]$, where $\Delta(Na_2O)$ and $\Delta(CaO)$ denote the HOMO-LUMO gap of Na_2O and CaO, respectively.

[(d)] In SN2, SN5, SN8, SN11, SN14, SN20 and SN22, "averaged" denotes the averaged value over all ARs extracted from the compound. Take $NaCaBe_2B_2O_6F$ as an instance, SN2 is $[\Delta(BeO_2^{2-}) + \Delta(BO_3^{3-}) + \Delta(F^-)]/3$.

[(e)] In SP2, SP5, SP8, SP11, SP14 and SP20, "averaged" denotes the averaged value over all MOs extracted from the compound. Take $NaCaBe_2B_2O_6F$ as an instance, SP2 is $[\Delta(Na_2O) + \Delta(CaO)]/2$.

(f) In SN3, SN6, SN9, SN12, SN15 and SN21, "maximal" denotes the minimal value among all ARs extracted from the compound. Take $NaCaBe_2B_2O_6F$ as an instance, SN3 is calculated as $\max[\Delta(\mathrm{BeO}_2^{2-}), \Delta(\mathrm{BO}_3^{3-}), \Delta(\mathrm{F}^-)]$.

(g) In SP3, SP6, SP9, SP12, SP15 and SP21, "maximal" denotes the minimal value among all ARs extracted from the compound. Take $NaCaBe_2B_2O_6F$ as an instance, SN3 is calculated as $\max[\Delta(\mathrm{Na_2O}), \Delta(\mathrm{CaO})]$.

(h) In SN4, SN5, SN6, SP4, SP5 and SP6, the total dipole moment of an isolated AR or MO is

$$\mu = \sqrt{\mu_x^2 + \mu_y^2 + \mu_z^2}$$

(i) In SN7, SN8, SN9, SP7, SP8 and SP9, the aniso-quadrupole moment of an isolated AR or MO is

$$\Delta Q = \sqrt{\frac{(Q_{xx} - Q_{yy})^2 + (Q_{yy} - Q_{zz})^2 + (Q_{zz} - Q_{xx})^2 + 6(Q_{xy}^2 + Q_{yz}^2 + Q_{xz}^2)}{2}}$$

(j) In SN10, SN11, SN12, SP10, SP11 and SP12, the aniso-polarizability of an isolated AR or MO is

$$\Delta\alpha = \sqrt{\frac{(\alpha_{xx} - \alpha_{yy})^2 + (\alpha_{yy} - \alpha_{zz})^2 + (\alpha_{zz} - \alpha_{xx})^2 + 6(\alpha_{xy}^2 + \alpha_{yz}^2 + \alpha_{xz}^2)}{2}}$$

(k) In SN13, SN14, SN15, SP13, SP14 and SP15, the total first hyper-polarizability of an isolated AR or MO is

$$\beta = \sqrt{\beta_x^2 + \beta_y^2 + \beta_z^2}$$

where

$$\beta_x = \beta_{xxx} + (\beta_{xyy} + \beta_{xzz} + \beta_{yxy} + \beta_{zxz} + \beta_{yyx} + \beta_{zzx})/3$$
$$\beta_y = \beta_{yyy} + (\beta_{yxx} + \beta_{yzz} + \beta_{xyx} + \beta_{zyz} + \beta_{xxy} + \beta_{zzy})/3$$
$$\beta_z = \beta_{zzz} + (\beta_{zyy} + \beta_{zxx} + \beta_{yzy} + \beta_{xzx} + \beta_{yyz} + \beta_{xxz})/3$$

(l) In SN16, SN17, SN19, SP16, SP17 and SP19, "total" denotes the summation over all ARs or MOs extracted from the compound. Take $NaCaBe_2B_2O_6F$ as an instance, SN16 is $q(\mathrm{BeO}_2^{2-}) + q(\mathrm{BO}_3^{3-}) + q(\mathrm{F}^-) = -6$.

(m) In SN18, SN19, SN20, SN21, SP18, SP19, SP20 and SP21, the flexibility index of an isolated AR or MO is defined as

$$F = \frac{1}{N}\sum_{\mathrm{AB}} \frac{e^{(R_{\mathrm{AB}}^0 - R_{\mathrm{AB}})/L}}{(\sqrt{C_{\mathrm{A}}} + \sqrt{C_{\mathrm{B}}})^2 / R_{\mathrm{AB}}^2} \times \frac{1}{|X_{\mathrm{A}} - X_{\mathrm{B}}|}$$

where R_{AB} is the real distance between two connected atoms A and B in the isolated AR or MO, R_{AB}^0 is the tabulated ideal bond length in the bond-valence theory, L is a parameter that is typically set as 0.37 Å, C_{A} and C_{B} are the number of valence electrons of atom A and B, respectively, X_{A} and X_{B} are the electronegativity of atom A and B, respectively, N is the number of bonds in the isolated AR or MO. Note that the flexibility index is set as 0 for monatomic molecules (e.g., F^-).

Section S2. Supplementary results of machine learning predictions

Table S3. Performance of GBRT models that only involve the first-level descriptors.

target	**fidelity**	**features**	**training set**		**test set**	
			MAE (eV)	**RMSE (eV)**	**MAE (eV)**	**RMSE (eV)**
E_g-PBE	single	multilevel[(a)]	0.243	0.319	0.423	0.565
		first-level	0.238	0.317	0.443	0.568
E_g-EXP	single	multilevel[(b)]	0.277	0.347	0.355	0.468
		first-level	0.284	0.353	0.374	0.512
E_g-EXP	multiple	multilevel[(c)]	0.261	0.331	0.293	0.391
		first-level	0.282	0.348	0.341	0.440

[(a)] Illustrated in Figure 2(a) in the main text.

[(b)] Illustrated in Figure 2(b) in the main text.

[(c)] Illustrated in Figure 3(b) in the main text.

Table S4. Performance of single-fidelity GBRTs with different random numbers for splitting.

target	split	training set		test set	
		MAE (eV)	RMSE (eV)	MAE (eV)	RMSE (eV)
E_g-PBE	1[(a)]	0.243	0.319	0.423	0.565
	2	0.221	0.294	0.411	0.608
	3	0.229	0.302	0.407	0.544
	4	0.217	0.286	0.433	0.661
	5	0.210	0.274	0.488	0.685
	average	0.224	0.295	0.432	0.613
E_g-EXP	1[(b)]	0.277	0.347	0.355	0.468
	2	0.283	0.351	0.359	0.461
	3	0.270	0.338	0.384	0.527
	4	0.273	0.344	0.385	0.513
	5	0.273	0.344	0.312	0.416
	average	0.275	0.345	0.359	0.477

[(a)] Illustrated in Figure 2(a) in the main text.

[(b)] Illustrated in Figure 2(b) in the main text.

Note S2. Details and results of single and multi-fidelity neural network models

The multi-fidelity framework based on the backpropagation[S3-S5] of neural networks (NNs) can be expressed as:

$$\frac{\partial L_1\left(y_1^{pred}(\{x\};\theta_1),y_1^{true}\right)}{\partial\theta_1}=\frac{\partial L_1}{\partial y_1^{pred}}\frac{\partial y_1^{pred}}{\partial\theta_1}\quad\text{(S1)}$$

where L_1 is the loss function at low fidelity, $\{x\}$ denote the input features of ML, and y_1^{true} and y_1^{pred} denote the reference and predicted values of low-fidelity labels, respectively. Eq. (S1) was used to determine θ_1 and construct $M_1(\theta_1)$. Similarly, $M_2(\theta_2)$ was built according to

$$\frac{\partial L_2\left(y_2^{pred}\left(\{x\},y_1^{pred}(\{x\};\theta_1);\theta_2\right),y_2^{true}\right)}{\partial\theta_2}=\frac{\partial L_2}{\partial y_2^{pred}}\frac{\partial y_2^{pred}}{\partial\theta_2}\quad\text{(S2)}$$

where L_2 is the loss function at high fidelity, and y_2^{true} and y_2^{pred} denote the reference and predicted values of high-fidelity labels, respectively. Furthermore, the low-fidelity $M_1(\theta_1)$ was involved as:

$$\frac{\partial L_2\left(y_2^{pred}\left(\{x\},y_1^{pred}(\{x\};\theta_1);\theta_2\right),y_2^{true}\right)}{\partial\theta_1}=\frac{\partial L_2}{\partial y_2^{pred}}\frac{\partial y_2^{pred}}{\partial y_1^{pred}}\frac{\partial y_1^{pred}}{\partial\theta_1}\quad\text{(S3)}$$

In practice, we retrained $M_1(\theta_1)$ using Eq. (S3) after the construction of $M_2(\theta_2)$. It is worth noting that the concurrent training of two neural networks at different fidelities has been proposed by Meng and Karniadakis,[S6] but its reliability on the small dataset of NLO crystals is still unclear.

In this work, we built multi-fidelity neural networks with the same datasets. Feature selection procedure was performed at the beginning. First, the features with a variance lower than 0.03 after normalization were removed. Second, the principal component analysis was applied, resulting in 20 input nodes for D_1 and 10 for D_2. The predicted E_g-PBE was concatenated as an additional input node for D_2. Using the grid-search method, one hidden layer was adopted with 16 and 8 nodes for D_1 and D_2, respectively. The connection between neighboring layers was activated by the sigmoid function. The Adam optimizer[S7] was applied to optimize parameters with the learning rate as 0.001. All NNs were implemented using PyTorch (version 2.0.1).[S8]

The values of E_g-PBE were first obtained with one NN (referred to as M_1), and the predicted E_g-PBE as well as nine other features was applied to another NN (referred to as M_2) to obtain E_g-EXP. When M_1 and M_2 were built with Eqs (S1) and (S2), the MAE and RMSE of E_g-EXP on the test set were 0.447 and 0.540 eV, respectively. After retraining $M_1(\theta_1)$ using Eq (S3), the MAE and RMSE decreased to 0.400 and 0.487 eV, respectively. It suggests that the backpropagation from L_2 at high fidelity to θ_1 at low fidelity is beneficial to multi-fidelity ML. However, the accuracy of NNs was below the baseline, which can be attributed to the small size of NLO dataset.

References

S1 C. Chen, Y. Wu and R. Li, *Int. Rev. Phys. Chem.*, 1989, **8**, 65–91.

S2 Z.-Y. Zhang, X. Liu, L. Shen, L. Chen and W.-H. Fang, *J. Phys. Chem. C*, 2021, **125**, 25175–25188.

S3 S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, N.J., 2nd edition, 1998.

S4 P. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science*, PhD Thesis, Harvard University, 1974.

S5 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533–536.

S6 X. Meng and G. E. Karniadakis, *J. Comput. Phys.*, 2020, **401**, 109020.

S7 D. P. Kingma and J. Ba, arXiv:1412.6980, 2017.

S8 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, arXiv:1912.01703, 2019.