Communications in Computational Chemistry
Commun. Comput. Chem., Vol. 1, No. 1, 50–60
doi: 10.4208/cicc.2025.61.01
Regular Article

AI4NLO: An Integrated Data Platform for Machine Learning-Driven Exploration of Inorganic Nonlinear Optical Materials

Zhaoxi Yu[1], Shubo Zhang [2], Ding Peng [1], Zhan-Yun Zhang [1], Yue Chen [2,*] and Lin Shen [1,2,*]

[1] *Key Laboratory of Theoretical and Computational Photochemistry of Ministry of Education, College of Chemistry, Beijing Normal University, Beijing 100875, P. R. China;*

[2] *Yantai-Jingshi Institute of Material Genome Engineering, Yantai 265505, Shandong, P. R. China.*

[*] Corresponding authors: chenyue@xhtechgroup.com, lshen@bnu.edu.cn

Received on 02 March 2025; Accepted on 07 April 2025

Abstract: Nonlinear optical (NLO) materials, with their unique wavelength conversion capabilities, play a crucial role in a wide range of scientific and industrial applications. Despite significant progress, the development of novel NLO materials, particularly those in the deep ultraviolet and mid-infrared regions, remains a challenge. Recent advancements in machine learning (ML) technologies have injected new momentum into materials science research. In this work, we present an integrated data platform incorporating advanced ML techniques, designed to drive the discovery and exploration of inorganic NLO materials. The platform currently includes about 1000 entries with their structures and key properties. Users can apply built-in ML models developed in our group for immediate predictions of NLO properties or train their own models based on specific research needs. Additionally, the platform provides access to the results of deep generative models, allowing users to retrieve newly generated virtual crystal structures, thus expanding the chemical space for NLO materials exploration. This platform not only provides reliable data support for researchers but also holds the potential to accelerate the discovery of novel NLO materials.

Key words: nonlinear optical crystal, database, second harmonic generation, coefficient, birefringence, machine learning, generative artificial intelligence.

1. Introduction

With their unique capabilities of wavelength conversion, nonlinear optical (NLO) materials play a crucial role in a wide range of modern scientific and industrial applications [1-5]. In the past decades, significant breakthroughs have been made in the study of inorganic NLO crystals. Prominent examples such as $KBe_2BO_3F_2$ (KBBF), $Ba_3P_3O_{10}X$ (X=Cl, Br), and $NaNH_4PO_3F{\cdot}H_2O$ for deep ultraviolet (DUV) region, β-BaB_2O_4 (β-BBO), LiB_3O_5 (LBO), and $CsPbCO_3F$ for ultraviolet region, KH_2PO_4 (KDP), $KTiOPO_4$ (KTP), and $LiNbO_3$ (LN) for visible to near-infrared region, and $AgGaQ_2$ (Q=S, Se), $ZnGeP_2$ (ZGP), and $A_2BiI_5O_{15}$ (A=K, Rb) for mid-infrared (MIR) region [6-19]. These materials have been synthesized, characterized, and reported, marking significant progress in NLO crystal research.

The performance of NLO materials is primarily determined by three key properties: bandgap (E_g), second harmonic generation (SHG) coefficient (d_{ij}) and birefringence (Δn). Among them, bandgap not only determines the absorption cut edge of material, which directly impacts its efficiency in light conversion, but also is positively correlated with the laser damage thresholde [20]. The SHG coefficient of a NLO crystal is directly related to its SHG conversion efficiency, with larger SHG coefficient enabling high conversion efficiency. In principle, all noncentrosymmetric (NCS) materials with finite electronic bandgaps can exhibit SHG effects. Birefringence is a critical property to attain effective phase-matching (PM) in NLO crystals, which is essential for generating coherent light through SHG. In noncubic materials, PM can be achieved through appropriate birefringence at a given wavelength. In practice, an applicable NLO crystal is expected to possess a large E_g, a large d_{ij}, and a moderate birefringence. Specifically, for applications in DUV region, E_g of a NLO crystal is supposed to exceed 6.2 eV to achieve ultraviolet absorption below 200 nm, d_{ij} should be at least greater than 1 times KDP (d_{36} = 0.39 pm/V), and Δn is ideally in the range of 0.07-0.10 [21]. A good MIR NLO crystal requires an E_g greater than 3.0 eV (ideally beyond 3.5 eV), a d_{ij} at least 10 times KDP (ideally over 20 times), and a Δn in the range of 0.04-0.10 [22]. Considering the above fundamental requirements, along with experimental limitations such as challenges in crystal synthesizability, growth properties, and toxicity of certain elements, the availability of suitable NLO materials particularly in DUV and MIR regions remains limited. Therefore, the exploration of novel NLO materials with high performance is still one of the most challenging and promising frontiers in materials science.

As the demand for high-performance NLO materials grows, researchers are increasingly turning to data-driven approaches to accelerate the discovery of novel materials with optimal properties. With the advancement of data science and high-performance computing, researchers have successively developed a series of open general materials databases, such as Automatic FLOW (AFLOW) [23], Materials Project (MP) [24], and Open Quantum Materials Database (OQMD) [25]. These databases contain vast number of material entries, spanning a wide range of chemical systems and material types, with fundamental material properties including electronic structure, thermodynamics, magnetism, and elasticity provided. The availability of such data has played a crucial role in supporting and inspiring the design and discovery of novel materials. Focusing on the domain of NLO materials, Zhang and co-workers [26,27] established a screening pipeline based on first-principles high-throughput calculations and then conducted theoretical research on a large number of crystalline compounds mainly composed of borates and germanates. They subsequently released an open NLO materials database, which provides users access to DFT-calculated properties including E_g, d_{ij}, and Δn, thus supported the study of structure-property relationships in NLO materials. More recently, Yang, Pan, and co-workers [28] developed a prediction-driven database that includes thousands of NCS materials, along with theoretical values for their E_g and d_{ij}. This database not only encompasses NCS materials retrieved from existing general material databases but also includes numerous new thermodynamically stable and metastable structures obtained using evolutionary algorithms, thereby opening up opportunities for discovery of novel NLO materials with promising properties.

In recent years, the introduction of artificial intelligence (AI) technologies has provided researchers in the field of materials science with new perspectives and methodologies. By leveraging large data support and advanced algorithms, researchers can more efficiently predict material properties, identify novel materials and uncover complex relationships between structures and properties. Impressively, machine learning (ML) models trained on general datasets have made significant strides in predicting fundamental material properties [29-32]. For NLO materials, ML models has demonstrated reliable accuracy and efficiency in predicting key properties including E_g, d_{ij}, Δn, formation energy, and thermal conductivity [33-41]. At the same time, the application of generative AI in material design is leading a new paradigm. Deep generative models, such as crystal diffusion variational autoencoder (CDVAE) and MatterGen [42,43], enables researchers to probe uncharted chemical spaces by generating entirely new virtual crystal structures. These models work by learning patterns from existing material data and using obtained knowledge to create new materials with tailored properties and promising stability, which opens up new avenues for material discovery and design. Despite significant progress in reverse design of materials such as metal-organic frameworks, two-dimensional materials, superconductors, and perovskites [44-49], the application of deep generative models to NLO materials remains an underexplored frontier, offering new opportunities for research in this field.

Given the pressing need for more efficient discovery and design of NLO materials, coupled with the rapid development of AI technologies, there is an increasing demand for an integrated data platform of NLO materials that leverages ML-driven approaches. In this work, integrating data management solutions with advanced ML technologies, we develop the AI4NLO, an inorganic NLO materials genome data platform (www.bnucrystal.cn) which aims at facilitating the ML-driven exploration of novel inorganic NLO materials. The database currently contains about 1000 entries with

plans for continuous updates. The majority of these entries have been synthesized, characterized and reported in the literature. The platform includes data on key NLO-related properties, i.e., E_g, d_{ij}, and Δn for each entry, with detailed annotations on computational or experimental methods used to obtain these values. As an integrated platform, we have deployed one-click ML solutions for the rapid and accurate prediction of d_{ij}, and Δn. Users can either apply the built-in models developed in our group for immediate predictions or train their own ML models online based on specific research needs. Furthermore, the platform provides an interface to access results from deep generative models, enabling users to retrieve newly generated virtual crystal structures. This platform not only provides reliable data support for researchers in the field of NLO materials but also fosters a streamlined approach to material discovery, thereby contributing to the advancement of this rapidly evolving field.

2. Method

2.1. Data sources

We conducted a systematic literature survey to collect as much data as possible on the chemical compositions, properties and crystal structures of inorganic NLO materials. The key terms used for literature retrieval included nonlinear optical, second harmonic generation, birefringence, ultraviolet, infrared, noncentrosymmetry, and so on. The search was significantly expanded through cross-referencing within the literature. Additionally, relevant monographs on NLO materials also served as an important data source [27,50]. For each material entry, chemical composition was recorded including both the chemical formula and cation and anion group information. The three key NLO properties documented are E_g, d_{ij} and Δn. The source and type of each property are also indicated to differentiate between experimental measurements and calculations at different levels. Crystal structure data includes space group, lattice constants, and atomic positions. These data were sourced not only from supplementary information provided in the articles but also from well-established general material databases such as Inorganic Crystal Structure Database (ICSD) [51], Cambridge Crystallographic Data Centre (CCDC), MP [24,52], and SNU MATerial data center (SNUMAT) [53]. The structural data extracted from public databases were cross-verified with the original literature reports to ensure accuracy and reliability. All data entries are clearly referenced with their respective sources.

Given the continuous progress in the field of inorganic NLO materials, our database will be periodically updated to include new research findings and experimental results. Additionally, the database encourages collaborative contributions from users, who are granted certain rights to upload and edit entries, fostering an interactive and dynamic data-sharing environment.

2.2. Machine learning for predicting NLO properties

The ML functionalities for predicting NLO properties of d_{ij} and Δn based on the multilevel descriptors [35] is a core feature of the data platform. These descriptors consist of three parts, where the first level captures the fundamental properties of the constituent elements of the crystals, such as atomic mass, van der Waals radius, and Pauling electronegativity. These atomic properties are gathered from the PubChem database [54]. According to anionic group theory, the macroscopic NLO properties of a crystal are strongly influenced by the microscopic geometric arrangement of its anionic groups [55]. Inspired by this, focusing on the electronic structure properties of functional groups, the second level of descriptors was constructed to simulate this effect. Acid radicals (ARs) and metallic oxides (MOs) are extracted from the composition of crystals in the dataset. After structural optimization and calculations of polarizability and energy, properties including charge, multiplicity, HOMO-LUMO gap, dipole moment, and polarizability are collected for each group. For each crystal, based on chemical element and functional group composition, statistics such as the maximum, minimum, average, and summarization of these properties are calculated to form the first- and second-level features, respectively. The third level involves a few global crystallographic features including space group number, lattice parameters, and Wyckoff positions. Given that some crystal structure data may be unavailable or inaccurate, and previous work [35]has demonstrated that the crystallographic features at the third level do not significantly enhance the model performances, only the first two levels of descriptors are adopted in this work for crystal-structure-free representations of crystal entries. Detailed definitions of all features at the first and second level in multilevel descriptors are provided in the Tables S1 and S2, respectively.

For a specified property-criterion pair, a random forest (RF) binary classification model [56] is employed to identify and label positive samples that exceed the given criterion for the property. For example, when using a well-constructed RF model of d_{ij}-3.90 pm/V for prediction, crystal samples with large SHG coefficients greater than 3.90 pm/V are classified as positive (i.e., SHG-active for applications in the MIR region), otherwise they are classified as negative. The dataset for model training is composed of features generated from crystal entries from the database, which are then randomly split into a training set (90%) and a test set (10%). To enhance the efficiency and generalizability of the model, feature

selection is performed based on feature discreteness and correlation within the training set, reducing the original 81 features to 15. Additionally, the synthetic minority oversampling technique (SMOTE) [57] is applied when the majority class exceeds 60% in the training set to alleviate the potential impact of class imbalance on the prediction performance at certain classification criteria. Two hyperparameters of RF model, the number of trees and the minimum number of samples required for a split are optimized using 5-fold cross-validation grid search. The F_β score, as a combination of precision and recall, is used as the main evaluation metric during optimization, where β is set to 2.0 in order to retain as many true positive samples as possible during the preliminary screening. A detailed description of feature selection, model training, and evaluation is provided in the Supporting Information.

The platform offers two types of ML models: (1) built-in models, which have been trained on all the collected data and ready for use, allowing users to make immediate predictions, and (2) custom models, where users can customize datasets and property-criterion pairs to train their own models. For built-in models, different criteria are applied to assess birefringence and SHG activity, that is, Δn = 0.02, 0.04, or 0.08, and d_{ij} = 0.39, 1.00, or 3.90 pm/V. In total, six models have been trained and preloaded onto the platform, with their parameters optimized and fixed during development. For custom models feature selection and hyperparameters are optimized during each training process. The classification results for crystal activity enable these models to serve as preliminary filters to identify promising candidates prior to first-principles calculations and time-consuming experiments in NLO materials discovery.

2.3. Deep generative models

As an advanced deep generative model specifically designed for periodic material structures, CDVAE is capable of generating novel structures with promising thermodynamic stability and unique material properties by learning large datasets of existing materials [42]. Through its integration of variational autoencoders and diffusion models, CDVAE facilitates the generation of diverse crystal structures by sampling from latent spaces, and refining them through optimization processes. An overview of CDVAE model is provided in the Supporting Information.

A series of well-constructed CDVAE models have been deployed on the data platform, enabling users to acquire generated material structures without the need for additional computational resources. These models have been trained on general datasets of inorganic materials collected from the MP database [24]. The training sets were constrained and designed based on elemental composition, thermodynamic stability, and the maximum number of atomic sites per structure to ensure that these models learn structural patterns from stable materials, thereby generating novel materials with promising properties and representativeness.

3. Results and discussion

3.1. Data format and permissions

As of the submission, a total of 937 NLO crystal entries have been included in this database. For each recorded entry, the database contains detailed information on its chemical composition, properties, and data sources. Table 1 provides a brief description of the recorded content and specific examples.

Each entry in the database is associated with two IDs, the internal Database ID and the External ID, which are crucial for uniquely identifying and cross-referencing it with different data sources. The Database ID serves as the unique index for each entry within the database. Public and private entries are distinguished by their Database IDs, with the former labeled NLOP- (where P stands for public) and the latter labeled NLOS- (where S stands for secret). Public entries are assigned a unique Database ID, while private entries have distinct IDs under the respective user accounts. The database encourages collaborative contributions, allowing users to upload new crystal data. When uploading, users can choose whether the data should be made public or kept private. Public entries undergo a verification process by administrators with advanced permissions, ensuring the authenticity, uniqueness, and validity of the data before being made accessible to all users for viewing and downloading. On the other hand, private entries require minimal information and no manual verification and are only visible and editable by the uploader. This system ensures the traceability of each entry, as well as the independence and confidentiality of the data uploaded by different users. The External ID refers to the identifier assigned to each entry in widely recognized external databases including ICSD, CCDC, and MP. Each External ID is prefixed with the corresponding database abbreviation indicating its source. Each structure is linked to a single External ID that serves as a reference for checking and validating the data. The database also records detailed chemical composition and structural information for each entry, along with the corresponding crystallographic information files (CIF) available for download. Isolated AR and MO species extracted from the chemical formula are additionally labeled to support ML functionalities.

Table 1. Descriptions and examples of the database entry attributes.

Attribute	Description	Example
Database ID	Unique identifier for the entry in this database with the prefix NLOP- for public entries and NLOS- for private entries.	NLOP-256
External ID	Identifier of the entry in external databases (currently supports ICSD, CCDC and MP).	ICSD-238029
Formula	Chemical formula of the crystal entry.	$RbPbCO_3F$
Space Group Number	Space group number corresponding to the crystal structure of the entry.	187
Metallic Oxides	Metallic oxide species (charges omitted) extracted from the chemical formula, separated by semicolons.	Rb_2O; PbO
Acid Radical Ions	Acid radical ion species (charges omitted) extracted from the chemical formula, separated by semicolons.	CO_3; F
CIF	Availability of the CIF of the entry in the database (True or False).	True
E_g (eV)	Band gaps of the entry and the corresponding acquisition methods.	EXP: 4.1 GGA/PBEsol: 3.18 GGA/PBE: 3.343
Δn	Birefringence of the entry and the corresponding acquisition methods and wavelengths.	GGA/PBE: 0.165 GGA/PBE@1064nm: 0.171 GGA/PBE@600nm: 0.186
d_{ij} (pm/V)	Second-order nonlinear coefficients of the entry and the corresponding acquisition methods and wavelengths.	GGA/PBE: 5.48
References	Relevant references of the entry, such as DOI numbers, books or virtual crystal sources.	https://dx.doi.org/10.1021/ic500778n; https://dx.doi.org/10.1039/c6dt04196e
Submission Time	The time when the entry was reviewed and included in the database.	2025-01-06

For the three NLO-related properties, i.e., E_g, d_{ij}, and Δn recorded in the database, the entries present not only the collected data but also the corresponding source and acquisition method and wavelength (for d_{ij} and Δn). Experimental measurements and computational results from different methods are labeled accordingly, such as EXP, PBE, and HSE06, providing comprehensive sources of information and showcasing the potential of multi-fidelity ML applications. For consistency, d_{ij} reported as multiples of standard substances in some references have been converted to pm/V units, where representative standards are KDP (0.39 pm/V) and AGS (13.4 pm/V) [58]. Moreover, only the maximum value of d_{ij} in the tensor is presented in reported calculations for simplicity.

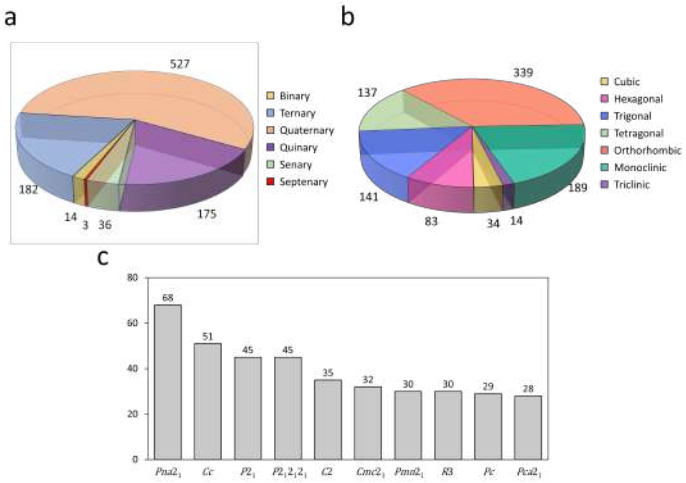

Figure 1. Data statistics from the database with pie chart (a) showing the distribution of compound types, pie chart (b) showing the distribution of crystal structures across the crystal systems and histogram (c) illustrating the distribution of the top-10 most prevalent space groups.

3.2. Data statistics

A comprehensive statistical summary of the chemical compositions and structures of the entries currently included in the database is provided in Figure 1. As shown in Figure 1a, the compounds in the database contain between two and seven elements. Among them, quaternary compounds are the most abundant, followed by ternary and quinary compounds, while compounds with very few or many elements are relatively scarce. Regarding crystal structures, as illustrated in Figure 1b, the entries span all seven crystal systems, with the orthorhombic system being the most represented, followed by the monoclinic system. In contrast, cubic and triclinic structures are less common. Specifically, the bar chart in Figure 1c lists the 10 most frequently occurring space groups and their respective counts. Among all crystal structures, the orthorhombic space group $Pna2_1$ (No.33) appears most frequently, followed by the monoclinic space group *Cc* (No.9).

Further analysis of the elemental composition of the crystals is presented in the form of element distribution heatmaps in Figures 2a and 2b. To distinguish between different categories, compounds containing oxygen and those without oxygen are analyzed separately. Of the 937 entries in the database, 680 are oxygenated compounds, representing a significant majority. As shown in Figure 2a, among all oxygenated compounds, 45.1% are borates, making them the most prevalent type. They are followed by phosphates and germanates, which account for 17.9% and 17.4%, respectively. Although the database includes compounds containing nearly all main group elements, carbonates, nitrates, silicates, and sulfates are relatively underrepresented. Meanwhile, the distribution of elements in oxygenated compounds reveals a balanced representation of alkali and alkaline earth metals. This reflects a common strategy in the design and synthesis of NLO materials: substituting metal cations in crystals with elements from the same group to tune properties such as bandgap. On the other hand, as shown in Figure 2b, the element distribution in non-oxygenated compounds exhibits similar patterns, with sulfides, selenides, and halides well-represented among these entries.

The scatter plots in Figures 2c-2e illustrate the distribution of three key properties for entries in the database. Specifically, d_{ij} and Δn prioritize experimental values measured at 1064 nm, selecting the maximum value if multiple measurements under the same conditions exist. On the other hand, E_g is derived from calculations based on the GGA method. The plots reveal that the entries span a broad range of values for these properties. Notably, oxygenated and non-oxygenated compounds, distinguishing by yellow and blue dots, exhibit distinctly different clustering patterns. Oxygenated compounds in the database generally exhibit smaller d_{ij} and larger E_g, forming a distribution trend that is entirely opposite to that of non-oxygenated compounds. In contrast, the distribution of Δn does not display a clear correlation with any specific element.

The above analysis results demonstrate the richness and diversity of the included entries, while also highlighting some extent the concentration and bias in the current NLO materials field. Therefore, this data platform not only provides comprehensive data of structures and properties for NLO materials but also offers researchers an insightful overview to guide further exploration in the field.

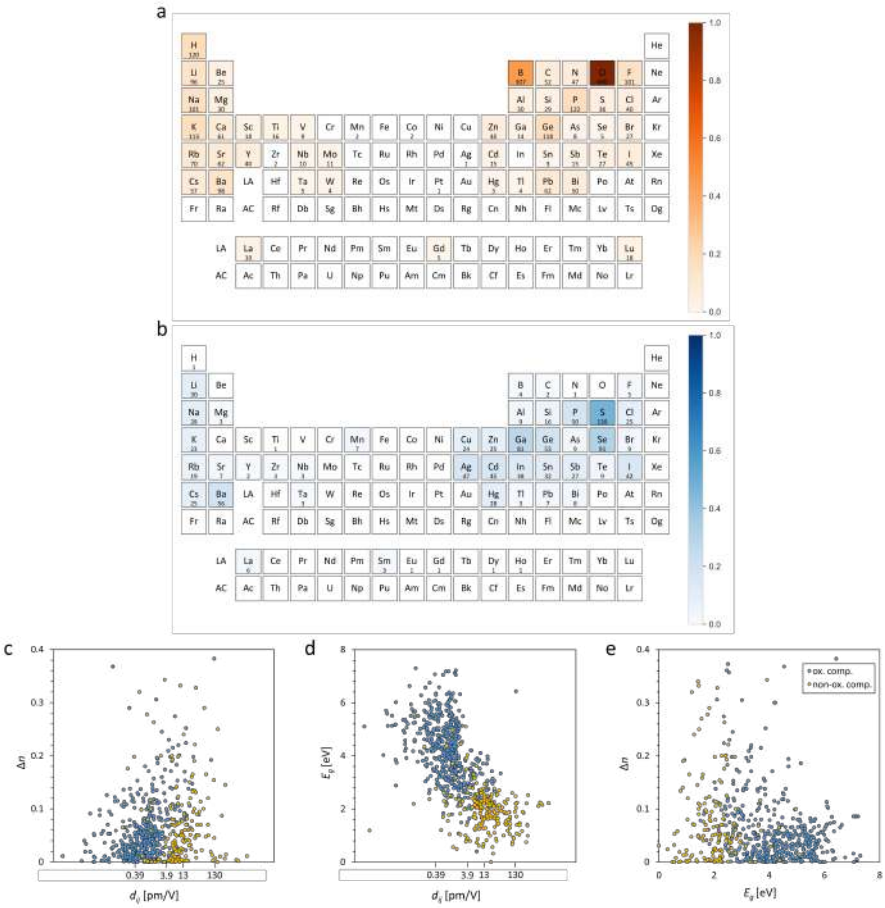

Figure 2. Elemental composition heatmaps of oxygenated (a) and non-oxygenated (b) compounds and distributions of NLO properties (c)-(e) of crystals in the database. Color intensity in (a)(b) represents the frequency of occurrence of each element, while elements that are not included in any entries are shown in white and unlabeled. Colors in (c)-(e) represent the data of oxygenated (in yellow) and non-oxygenated (in blue) compounds.

Table 2. Performances of built-in models for d_{ij} and Δn prediction.

Property	Criterion	Training set				Test set			
		Accuracy	Precision	Recall	F_2 score	Accuracy	Precision	Recall	F_2 score
d_{ij} [pm/V]	0.39	0.881	0.969	0.787	0.818	0.816	0.962	0.781	0.812
	1.00	0.855	0.904	0.794	0.814	0.793	0.854	0.788	0.801
	3.90	0.944	0.941	0.946	0.945	0.931	0.857	0.968	0.943
Δn	0.02	0.880	0.854	0.917	0.904	0.809	0.800	0.930	0.901
	0.04	0.850	0.812	0.947	0.916	0.809	0.765	0.839	0.823
	0.08	0.893	0.888	0.898	0.896	0.765	0.500	0.875	0.761

3.3. ML-driven exploration of NLO materials

There are two ways users can access and leverage ML functionalities available on the platform. The first is that users can directly apply the built-in models, which have been trained on all collected entries in the database during platform development, to obtain predictions of d_{ij} and Δn of crystals. The performances of these models are summarized in Table 2. The influence of random

splits on the datasets was analyzed by training five different models for each criterion, which was confirmed to be small (listed in Tables S3 and S4). The optimized hyperparameters and selected features of these models are listed in Tables S5 and S6. This method enables rapid screening of promising NLO materials without any additional model training or data processing.

Alternatively, users can choose to train and deploy custom ML models online with more flexibility. During this procedure, users only need to specify dataset range and classification criteria for specific research scenarios, and feature generation, hyperparameter optimization, and model training will be automatically performed on the platform. If the number of valid entries for model training is fewer than 150, the platform will prompt users to supplement the training samples to avoid potential issues caused by insufficient data. Once trained, these custom models are stored independently under individual accounts and can be applied to make predictions for entries of interest subsequently, providing a more comprehensive NLO property assessment. These solutions are designed to be user-friendly, ensuring that even those without an extensive ML background can easily navigate the platform and make full use of its capabilities.

For training these models, multilevel descriptors are utilized as the representations of the crystal entries. For entries with properly labeled ARs and MOs, the platform can immediately generate multilevel descriptors online. However, due to the optimization and convergence issues associated with certain isolated groups, entries involving any AR or MO outside the supported ranges (listed in Tables S7 and S8) will not generate valid features. The multilevel descriptor datasets for selected entries can be exported directly for use in external ML models, feature analysis or other data-driven research applications, or users can take advantage of the built-in ML tools for further investigation.

3.4. Preliminary results of deep generative models

The platform also provides an interface dedicated to showcasing virtual crystal structures generated by deep generative models. Considering that running generative models often requires substantial computational time, the generated structures available on the platform are not produced in real-time on the website server. Instead, these structures are pre-generated in a development environment and uploaded to the platform for users to access. Currently, up to 2000 newly generated valid structures can be retrieved daily. A note on the evaluation of structures generated by CDVAE models is provided in the Supporting Information.

Figure 3 illustrates several example crystal structures generated by these models, encompassing a variety of inorganic compounds including carbonates, borates, phosphates and compounds with mixed anions. The employed CDVAE model tends to generate structures of high complexity and low symmetry that are often beyond the original distribution of existing materials, making it a powerful tool for exploring unknown structural landscapes [42,47]. On the platform, users can search for these virtual crystal structures based on chemical composition and leverage built-in ML tools for rapid property prediction. Combined with first-principles calculations, promising candidates for novel NLO materials can be efficiently identified. These results can serve as starting points for further studies, including the exploration of additional material properties and potential applications.

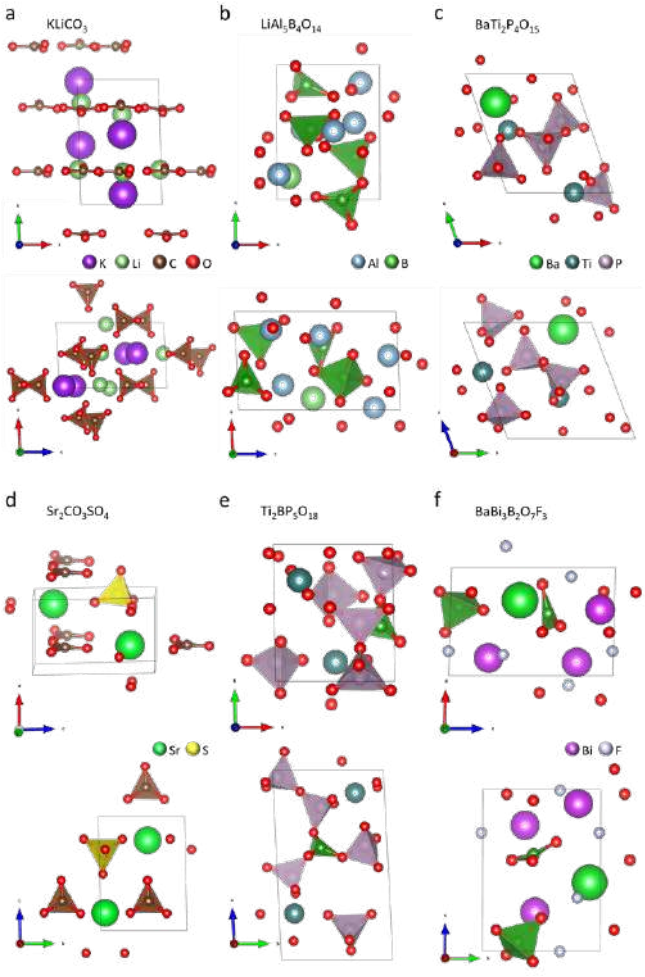

Figure 3. Example crystal structures generated by deep generative models.

Notably, the platform also allows for customization of generative models according to user preferences. For researchers focused on NLO materials containing certain specific elements,

we can curate a specialized dataset consisting primarily of compounds containing the desired elements. New generative models can then be trained on this dataset to generate a vast array of virtual inorganic compounds containing the target elements, contributing to the improvement of user-friendliness and the discovery of novel materials with targeted compositions. The current results of deep generative models are preliminary, and the development of models specifically designed for NLO materials is ongoing, with further improvements targeted at enhancing their reliability and applicability.

4. Conclusions

In this work, we have developed a ML-driven open data platform dedicated to exploring inorganic NLO materials. The database currently includes approximately 1000 entries of inorganic NLO crystals reported in the literature, along with their structural data and NLO properties. Notably, it is the first online platform to integrate ML functionalities for NLO materials discovery. On this platform, users can access publicly available data and are encouraged to contribute to the database through public or private data submissions. There are two types of ML models integrated into the platform: built-in models, trained on the entire dataset, and custom models, allowing users to define their dataset and classification criteria. By leveraging these models, users can achieve rapid and accurate predictions of NLO properties for selected entries, enabling high-throughput screening of potential candidates. Additionally, the platform provides access to results of deep generative models, offering a wealth of virtually generated inorganic crystal structures, which hold great potential for further exploration, significantly accelerating the discovery of novel NLO materials. These functionalities can be accessed and executed seamlessly on the website server of the platform, offering one-click solutions for all users, including those without a strong ML background, making it a powerful tool for accelerating discovery and exploration of inorganic NLO materials.

Supporting Information

The following supporting information can be downloaded at: https://global-sci.com/storage/self-storage/cicc-2025-61-1-r1-si.pdf

Definitions of multilevel descriptors, details of feature selection, model training and evaluation for the RF classification model, overview of CDVAE model, supplementary results of built-in ML models, supported AR and MO species, and evaluation of CDVAE-generated structures.

Acknowledgements

We are grateful for the financial support from the Natural Science Foundation of China (22193041). We also thank Zhihao Gu, Xianfeng Li, Siping Lin, Hui Rong, Pengxiang Sui, and Jianming Xie for their valuable contributions to the development of this data platform.

References

[1] C. Wu., G. Yang., M.G. Humphrey., C. Zhang. Recent advances in ultraviolet and deep-ultraviolet second-order nonlinear optical crystals. *Coord. Chem. Rev.*, **375** (2018), 459–488.

[2] B. Zhang., Z. Chen. Recent advances of inorganic phosphates with UV/DUV cutoff edge and large second harmonic response. *Chin. J. Struct. Chem.*, **42** (2023), 100033.

[3] Q. Zhang., R. An., X. Long., Z. Yang., S. Pan., Y. Yang. Exploiting deep-ultraviolet nonlinear optical material $Rb_2ScB_3O_6F_2$ originated from congruously oriented $[B_3O_6]$ groups. *Angew. Chem., Int. Ed.*, **64** (2025), e202415066.

[4] C. Huang., M. Mutailipu., F. Zhang., K.J. Griffith., C. Hu., Z. Yang., J.M. Griffin., K.R. Poeppelmeier., S. Pan. Expanding the chemistry of borates with functional $[BO_2]^-$ anions. *Nat. Commun.*, **12** (2021), 2597.

[5] D. F. Eaton. Nonlinear optical materials. *Science*, **253** (1991), 281–287.

[6] C. Chen., Y. Wang., Y. Xia., B. Wu., D. Tang., K. Wu., Z. Wenrong., L. Yu., L. Mei. New development of nonlinear optical crystals for the ultraviolet region with molecular engineering approach. *J. Appl. Phys.*, **77** (1995), 2268–2272.

[7] P. Yu., L.-M. Wu., L.-J. Zhou., L. Chen. Deep-ultraviolet nonlinear optical crystals: $Ba_3P_3O_{10}X$ (X = Cl, Br). *J. Am. Chem. Soc.*, **136** (2014), 480–487.

[8] L. Xiong., J. Chen., J. Lu., C.-Y. Pan., L.-M. Wu. Monofluorophosphates: a new source of deep-ultraviolet nonlinear optical materials. *Chem. Mater.*, **30** (2018), 7823–7830.

[9] J. Lu., J.-N. Yue., L. Xiong., W.-K. Zhang., L. Chen., L.-M. Wu. Uniform alignment of non-π-conjugated species

enhances deep ultraviolet optical nonlinearity. *J. Am. Chem. Soc.*, **141** (2019), 8093–8097.

[10] C.T. Chen., B.C. Wu., A.D. Jiang., G.M. You. A new-type ultraviolet SHG crystal-β-BaB_2O_4. *Sci. Sin., Ser. B*, **28** (1985), 235–243.

[11] C. Chen., Y. Wu., A. Jiang., B. Wu., G. You., R. Li., S. Lin. New nonlinear-optical crystal: LiB_3O_5. *J. Opt. Soc. Am. B*, **6** (1989), 616–621.

[12] G. Zou., L. Huang., N. Ye., C. Lin., W. Cheng., H. Huang. $CsPbCO_3F$: a strong second-harmonic generation material derived from enhancement via p–π interaction. *J. Am. Chem. Soc.*, **135** (2013), 18560–18566.

[13] D. Eimerl. Electro-optic, linear, and nonlinear optical properties of KDP and its isomorphs. *Ferroelectrics*, **72** (1987), 95–139.

[14] T.A. Driscoll., H.J. Hoffman., R.E. Stone., P.E. Perkins. Efficient second-harmonic generation in KTP crystals. *J. Opt. Soc. Am. B*, **3** (1986), 683–686.

[15] G.D. Boyd., R.C. Miller., K. Nassau., W.L. Bond., A. Savage. $LiNbO_3$: an efficient phase matchable nonlinear optical material. *Appl. Phys. Lett.*, **5** (1964), 234–236.

[16] D.S. Chemla., P.J. Kupecek., D.S. Robertson., R.C. Smith. Silver thiogallate, a new material with potential for infrared devices. *Opt. Commun.*, **3** (1971), 29–31.

[17] G. Boyd., H. Kasper., J. McFee., F. Storz. Linear and nonlinear optical properties of some ternary selenides. *IEEE J. Quantum Electron.*, **8** (1972), 900–908.

[18] G.D. Boyd., E. Buehler., F.G. Storz. Linear and nonlinear optical properties of $ZnGeP_2$ and CdSe. *Appl. Phys. Lett.*, **18** (1971), 301–304.

[19] Y. Huang., X. Meng., P. Gong., L. Yang., Z. Lin., X. Chen., J. Qin. $A_2BiI_5O_{15}$ (A = K^+ or Rb^+): two new promising nonlinear optical materials containing $[I_3O_9]^{3-}$ bridging anionic groups. *J. Mater. Chem. C*, **2** (2014), 4057–4062.

[20] D. Mei., W. Cao., N. Wang., X. Jiang., J. Zhao., W. Wang., J. Dang., S. Zhang., Y. Wu., P. Rao., Z. Lin. Breaking through the "3.0 eV wall" of energy band gap in mid-infrared nonlinear optical rare earth chalcogenides by charge-transfer engineering. *Mater. Horiz.*, **8** (2021), 2330–2334.

[21] L. Kang., M. Zhou., J. Yao., Z. Lin., Y. Wu., C. Chen. Metal thiophosphates with good mid-infrared nonlinear optical performances: a first-principles prediction and analysis. *J. Am. Chem. Soc.*, **137** (2015), 13049–13059.

[22] T.T. Tran., H. Yu., J.M. Rondinelli., K.R. Poeppelmeier., P.S. Halasyamani. Deep ultraviolet nonlinear optical materials. *Chem. Mater.*, **28** (2016), 5238–5258.

[23] S. Curtarolo., W. Setyawan., G.L.W. Hart., M. Jahnatek., R.V. Chepulskii., R.H. Taylor., S. Wang., J. Xue., K. Yang., O. Levy., M.J. Mehl., H.T. Stokes., D.O. Demchenko., D. Morgan. AFLOW: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.*, **58** (2012), 218–226.

[24] A. Jain., S.P. Ong., G. Hautier., W. Chen., W.D. Richards., S. Dacek., S. Cholia., D. Gunter., D. Skinner., G. Ceder., K.A. Persson. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.*, **1** (2013), 011002.

[25] S. Kirklin., J.E. Saal., B. Meredig., A. Thompson., J.W. Doak., M. Aykol., S. Rühl., C. Wolverton. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.*, **1** (2015), 1–15.

[26] B. Zhang., X. Zhang., J. Yu., Y. Wang., K. Wu., M.-H. Lee. First-principles high-throughput screening pipeline for nonlinear optical materials: application to borates. *Chem. Mater.*, **32** (2020), 6772–6779.

[27] J. Yu., B. Zhang., X. Zhang., Y. Wang., K. Wu., M.-H. Lee. Finding optimal mid-infrared nonlinear optical materials in germanates by first-principles high-throughput screening and experimental verification. *ACS Appl. Mater. Interfaces*, **12** (2020), 45023–45035.

[28] C. Xie., E. Tikhonov., D. Chu., M. Wu., I. Kruglov., S. Pan., Z. Yang. A prediction-driven database to enable rapid discovery of nonlinear optical materials. *Sci. China Mater.*, **66** (2023), 4473–4479.

[29] T. Xie., J.C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, **120** (2018), 145301.

[30] K. Choudhary., B. DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.*, **7** (2021), 1–8.

[31] Y. Lin., K. Yan., Y. Luo., Y. Liu., X. Qian., S. Ji. Efficient approximations of complete interatomic potentials for crystal property prediction. (2023), arXiv:2306.10045.

[32] À. Solé., A. Mosella-Montoro., J. Cardona., S. Gómez-Coca., D. Aravena., E. Ruiz., J. Ruiz-Hidalgo. A cartesian encoding graph neural network for crystal structures property prediction: application to thermal ellipsoid estimation. *Digital Discovery*, (2025), Advance Article.

[33] Z. Yu., P. Xue., B.-B. Xie., L. Shen., W.-H. Fang. Multi-fidelity machine learning for predicting bandgaps of

nonlinear optical crystals. *Phys. Chem. Chem. Phys.*, **26** (2024), 16378–16387.

[34] R. Wang., F. Liang., Z. Lin. Data-driven prediction of diamond-like infrared nonlinear optical crystals with targeting performances. *Sci. Rep.*, **10** (2020), 3486.

[35] Z.-Y. Zhang., X. Liu., L. Shen., L. Chen., W.-H. Fang. Machine learning with multilevel descriptors for screening of inorganic nonlinear optical crystals. *J. Phys. Chem. C*, **125** (2021), 25175–25188.

[36] J. Xiao., L. Yang., S. Wang., Z. He. Accurate prediction of second harmonic generation coefficients using graph neural networks. *Comput. Mater. Sci.*, **244** (2024), 113203.

[37] M. Wu., E. Tikhonov., A. Tudi., I. Kruglov., X. Hou., C. Xie., S. Pan., Z. Yang. Target-driven design of deep-UV nonlinear optical materials via interpretable machine learning. *Adv. Mater.*, **35** (2023), 2300848.

[38] Z. Fan., Z. Sun., A. Wang., Y. Yin., H. Li., G. Jin., C. Xin. Machine learning regression model for predicting the formation energy of nonlinear optical crystals. *Adv. Theory Simul.*, **6** (2023), 2200883.

[39] Z. Fan., S. Lian., G. Jin., C. Xin., Y. Li., B. Yuan. Predictive nonlinear optical crystal formation energy regression model based on convolutional neural networks. *CrystEngComm*, **26** (2024), 2652–2661.

[40] Q. Wu., L. Kang., Z. Lin. A machine learning study on high thermal conductivity assisted to discover chalcogenides with balanced infrared nonlinear optical performance. *Adv. Mater.*, **36** (2024), 2309675.

[41] Q. Liu., R. An., C. Li., D. Chu., W. Zhao., S. Pan., Z. Yang. Accelerating discovery of infrared nonlinear optical materials with high lattice thermal conductivity: combining machine learning and first-principles calculations. *Adv. Opt. Mater.*, (2025), 2403292.

[42] T. Xie., X. Fu., O.-E. Ganea., R. Barzilay., T. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. (2022), arXiv:2110.06197.

[43] C. Zeni., R. Pinsler., D. Zügner., A. Fowler., M. Horton., X. Fu., Z. Wang., A. Shysheya., J. Crabbé., S. Ueda., R. Sordillo., L. Sun., J. Smith., B. Nguyen., H. Schulz., S. Lewis., C.-W. Huang., Z. Lu., Y. Zhou., H. Yang., H. Hao., J. Li., C. Yang., W. Li., R. Tomioka., T. Xie. A generative model for inorganic materials design. *Nature*, (2025), 1–3.

[44] Z. Yao., B. Sánchez-Lengeling., N.S. Bobbitt., B.J. Bucior., S.G.H. Kumar., S.P. Collins., T. Burns., T.K. Woo., O.K. Farha., R.Q. Snurr., A. Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.*, **3** (2021), 76–86.

[45] J. Park., Y. Lee., J. Kim. Multi-modal conditional diffusion model using signed distance functions for metal-organic frameworks generation. *Nat. Commun.*, **16** (2025), 34.

[46] Y. Song., E.M.D. Siriwardane., Y. Zhao., J. Hu. Computational discovery of new 2D materials using deep learning generative models. *ACS Appl. Mater. Interfaces*, **13** (2021), 53303–53313.

[47] P. Lyngby., K.S. Thygesen. Data-driven discovery of 2D materials by deep generative models. *npj Comput. Mater.*, **8** (2022), 1–8.

[48] D. Wines., T. Xie., K. Choudhary. Inverse design of next-generation superconductors using data-driven deep generative models. *J. Phys. Chem. Lett.*, **14** (2023), 6630–6638.

[49] E.T. Chenebuah., M. Nganbe., A.B. Tchagang. A deep generative modeling architecture for designing lattice-constrained perovskite materials. *npj Comput. Mater.*, **10** (2024), 1–21.

[50] D. N. Nikogosyan. Nonlinear Optical Crystals: a Complete Survey. Springer-Verlag: New York, 2005.

[51] FIZ Karlsruhe Inorganic Crystal Structure Database. https://icsd.products.fiz-karlsruhe.de/.

[52] Cambridge Crystallographic Data Centre. https://www.ccdc.cam.ac.uk/.

[53] S. Kim., M. Lee., C. Hong., Y. Yoon., H. An., D. Lee., W. Jeong., D. Yoo., Y. Kang., Y. Youn., S. Han. A band-gap database for semiconducting inorganic materials calculated with hybrid functional. *Sci. Data*, **7** (2020), 387.

[54] S. Kim., J. Chen., T. Cheng., A. Gindulyte., J. He., S. He., Q. Li., B.A. Shoemaker., P.A. Thiessen., B. Yu., L. Zaslavsky., J. Zhang., E.E. Bolton. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49** (2021), D1388–D1395.

[55] C. Chen., Y. Wu., R. Li. The anionic group theory of the non-linear optical effect and its applications in the development of new high-quality NLO crystals in the borate series. *Int. Rev. Phys. Chem.*, **8** (1989), 65–91.

[56] L. Breiman. Random forests. *Mach. Learn.*, **45** (2001), 5–32.

[57] N.V. Chawla., K.W. Bowyer., L.O. Hall., W.P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16** (2002), 321–357.

[58] A. Abudurusuli., J. Li., S. Pan. A review on the recently developed promising infrared nonlinear optical materials. *Dalton Trans.*, **50** (2021), 3155–316