Communications in Computational Chemistry
Commun. Comput. Chem., Vol. 7, 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