Zhang *et al. J Mater Inf* 2024;4:13
DOI: 10.20517/jmi.2024.20

**Journal of
Materials Informatics**

Research Article

Open Access

Data-driven exploration and first-principles analysis of perovskite material

Lei Zhang[*], **Jiacheng Zhou, Xuexiao Chen**

Department of Materials Physics, School of Chemistry and Materials Science, Nanjing University of Information Science & Technology, Nanjing 210044, Jiangsu, China.

[*]**Correspondence to:** Prof. Lei Zhang, Department of Materials Physics, School of Chemistry and Materials Science, Nanjing University of Information Science & Technology, 219 Ning Liu Road, Nanjing 210044, Jiangsu, China. E-mail: 002699@nuist.edu.cn

How to cite this article: Zhang L, Zhou J, Chen X. Data-driven exploration and first-principles analysis of perovskite material. *J Mater Inf* 2024;4:13. https://dx.doi.org/10.20517/jmi.2024.20

Received: 29 Jun 2024 **First Decision:** 13 Aug 2024 **Revised:** 7 Sep 2024 **Accepted:** 18 Sep 2024 **Published:** 26 Sep 2024

Academic Editor: Ming Hu **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

In this study, we employ data-driven and first-principles methods (machine learning, density-functional theory and language model) to comprehensively explore crystal structures, electronic properties and applications of an emerging perovskite material, gadolinium scandate ($GdScO_3$), which is an intriguing material that demonstrates potentials in electronics and optics. Using advanced machine learning algorithms based on genetic programming, we have discovered new crystal structures of $GdScO_3$ that have not been previously reported, which are further examined via density functional theory (DFT) calculations and language models to provide detailed insights into their electronic and optical properties and potential applications. Our findings reveal novel new stable phases of $GdScO_3$ and highlight the intricate influence of structural variations on its electronic band structures and light absorption properties. A subsequent domain-specific language model analysis indicates its possibility for photovoltaics pending further efforts to engineer defects revealed in the first-principles calculations. The integration of machine learning with first-principles calculations demonstrates a feasible approach for accelerating the exploration and analysis of materials. This work enriches the understanding of $GdScO_3$ and establishes a robust framework for exploration and ontological analysis of new functional materials combining diverse data-driven techniques (e.g., language model and genetic programming) and first-principles methods.

Keywords: Materials informatics, language model, crystal structure prediction, first-principles, genetic algorithm

INTRODUCTION

In recent years, the integration of artificial intelligence (AI) and data-driven approaches has significantly advanced materials science, aligning with the goals of the Materials Genome Initiative (MGI)[1-5]. By utilizing large datasets that encompass material structures, properties, synthesis methods, and performance metrics, machine learning can predict material properties and, inversely, propose novel compositions with desired functionalities. Moreover, utilizing natural language processing (NLP) machine learning algorithms such as Word2Vec and large language models such as ChatGPT and MatBERT[6,7], researchers have been able to effectively extract latent scientific information from vast amounts of literature and propose novel material compositions with tailored properties[3,8,9].

Crystal structure prediction plays a critical role in physics and materials science by enabling the discovery of new structures of novel materials[10-15]. Software tools such as ab initio random structure searching (AIRSS)[16,17] and crystal structure analysis by particle swarm optimization (CALYPSO)[18,19] are exemplar approaches used for crystal structure prediction. AIRSS employs a random search algorithm within the framework of first-principles calculations to explore the structural landscape and identify the most stable configurations of materials. CALYPSO utilizes particle swarm optimization to efficiently search for global energy minima in complex potential energy landscapes, thus enabling the prediction of new and metastable structures. The crystal structure prediction process is often time-consuming and computationally expensive. However, the use of genetic algorithms has significantly accelerated this process, providing an efficient alternative for discovering new materials. The machine learning and graph theory-assisted universal structure searcher (MAGUS) software[20] utilizes advanced computational algorithms such as genetic algorithms and machine learning to predict the stable crystal structures of materials, and the integrated approach holds promise for accelerating the discovery of functional materials and optimizing their performance across various applications in electronics, energy storage, and beyond.

Gadolinium scandate ($GdScO_3$), a rare earth scandate perovskite material, is of interest for applications in electronics[21-25]. However, the availability of its crystal structures is rather limited, impeding further understanding on its structure-property relationships. In electronics, the high dielectric constant and low loss characteristics of $GdScO_3$ have led to its utilization in capacitors and dielectric memory devices. Additionally, its ion transport properties and chemical stability have positioned it as a frontrunner in solid oxide fuel cells and solid-state electrolytes for energy-related applications. Beyond its electronic properties, typical $GdScO_3$ shows promise in optical waveguides, where its transparency and structural properties make it suitable for guiding light. Through density functional theory (DFT) calculations, researchers have unraveled the complex interplay between the atomic arrangement of oxide perovskites and their functional characteristics. For example, DFT simulations have elucidated the intricate oxygen vacancy formation energies, migration pathways, and their impact on their ionic conductivity, essential for applications in solid oxide fuel cells and solid-state electrolytes[26-29]. Moreover, DFT has been instrumental in predicting the thermodynamic stability and phase transitions of $GdScO_3$ under various temperature and pressure conditions, guiding experimental synthesis efforts. Additionally, DFT simulations have shed light on the dopant incorporation mechanisms and their effects on the electronic band structure, aiding in tailoring its electronic properties for semiconductor applications. Nevertheless, there exists a gap in the exploration of the crystal structure of $GdScO_3$, with a scarcity of data in materials databases. The lack of study on exploring alternative crystal structures of $GdScO_3$ limits the further explorations of the perovskite material, and there is a pressing need for research dedicated to the generation of crystal structures for $GdScO_3$.

In this study, we employ genetic algorithms to predict novel crystal structures of $GdScO_3$ and analyze its structural and electronic properties; three new crystal structures of $GdScO_3$ not reported elsewhere are discovered. The approach leverages genetic algorithms coupled with DFT to perform crystal structure generation [Figure 1]. The genetically predicted new structures of $GdScO_3$ are further examined by post-hoc DFT calculations for detailed characterization of structural, electronic and optical properties, including band structures, partial density of states (PDOS) and ultraviolet-visible (UV-vis) absorption spectra. The strain effects on the electronic band structure are examined based on an exemplary new crystal structure. Additionally, large language model analysis, combined with dimensionality reduction and clustering techniques, is utilized to further explore the $GdScO_3$ material from a literature data-driven perspective, offering an alternative viewpoint based on insights from the model. By combining machine learning methods with first-principles calculations, we uncover new crystal structures of $GdScO_3$ with distinctive electronic structures that can be finely tuned by strain, and ontologically analyze the material comprehensively using a language model starting from 1.18 million scientific articles.

METHODS

Machine learning and genetic algorithm (crystal structure generation)

The MAGUS software is employed to generate new crystal structures of $GdScO_3$ using a combination of machine learning and density functional theories by leveraging evolutionary strategies. The genetic algorithm [Supplementary Materials] to obtain new crystal structures of $GdScO_3$ is initiated with (1) a set of preliminary random crystal structures and (2) different Vienna ab-initio simulation package (VASP)[30] input parameters (INCAR) to fully account for the impacts of software geometrical optimization details and intrinsic crystal structures on the resulting crystal structures. The starting population undergoes iterative cycles of selection, with operations including crossover and mutation to evolve new generations of crystal structures. At each iteration, the crystal structures are evaluated based on their energy and stability using first-principles calculation, with the lower-energy configurations being preferentially selected for subsequent generations. The following hyperparameters are used: initSize = 20; popSize = 20; numGen: 10; saveGood = 3, min_n_atoms = 5; spacegroup = 2-230; d_ratio = 0.6; volume_ratio = 3. The preliminary DFT calculation involves Perdew-Burke-Ernzerhof (PBE) functional, 380 eV cutoff energy and convergence criteria (forces) of 0.01 eV/Å for atoms and 0.001 eV/Å for electrons. Four distinct k-spacing values expressed in different VASP input files are considered during the genetic process: KSPACING = 1.256000, KSPACING = 0.942, KSPACING = 0.618, and KSPACING = 0.314. These different input files act as the starting population to facilitate the genetic process. Full structural relaxation including unit cell parameters is performed, allowing the atomic positions and lattice parameters to adjust until the forces on the atoms are minimized and the total energy converges to stable values. Higher quality DFT calculations are executed based on these initial optimized structures. Through this combined approach of machine learning-driven genetic algorithms and DFT calculations, a diverse set of $GdScO_3$ structures are generated and their electronic structures are further analyzed in detail.

New structure screening

The obtained structures based on the previous step are ranked according to their relative energy and direct visual inspection in the structural integrity. The top ten candidates with decent stabilities are compared with entries stored in publicly available crystallographic databases, including Materials Project, Jarvis, Open Quantum Materials Database (OQMD) and Atomly[31-33]. This process eliminates repetitive and existing $GdScO_3$ crystal structures, resulting in three new crystal structures **1**-**3** [Table 1 and Figure 2]. The data associated with this article is provided and is publicly available. The optimized atomic structures are provided at https://github.com/Zhang-NJ-Lab/GdScO3_CSP/tree/main/Crystal_Structures/OptimizedStructures. The initial input files for these four k-spacing values are available on the GitHub repository at: https://github.com/Zhang-NJ-Lab/GdScO3_CSP/tree/main/inputFold. The MAGUS input files (scripts) are provided at https://github.com/Zhang-NJ-Lab/GdScO3_CSP/blob/main/input.yaml (MAGUS yaml scripts) and https://github.com/Zhang-NJ-Lab/GdScO3_CSP/tree/main/inputFold (VASP

Table 1. New crystal structures of $GdScO_3$ predicted by machine learning

	Lattice	Space group	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
1	Monoclinic	*Cm*	6.372	3.485	8.819	90	138	90
2	Hexagonal	*P63/mmc*	3.612	3.612	12.396	90	90	120
3	Orthorhombic	*Pnma*	5.776	7.977	5.534	90	90	90

$GdScO_3$: Gadolinium scandate.

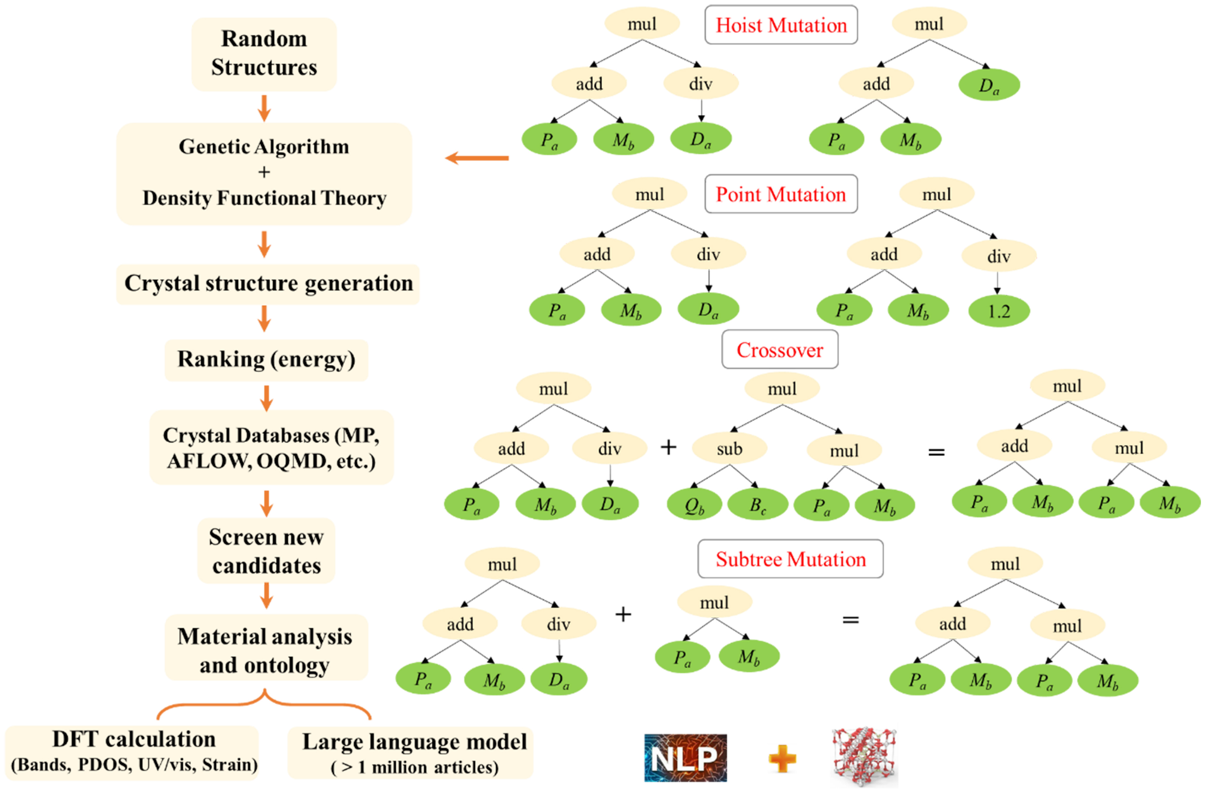
Random Structures
Genetic Algorithm + Density Functional Theory
Crystal structure generation
Ranking (energy)
Crystal Databases (MP, AFLOW, OQMD, etc.)
Screen new candidates
Material analysis and ontology
DFT calculation (Bands, PDOS, UV/vis, Strain)
Large language model (> 1 million articles)
Hoist Mutation
Point Mutation
Crossover
Subtree Mutation
mul
add
div
sub
P_a
M_b
D_a
Q_b
B_c
1.2
+
=
NLP

Figure 1. General workflow of the integrated machine learning (genetic algorithm and language model) and first-principles process in this study to explore crystal structures and ontologies of $GdScO_3$. $GdScO_3$: Gadolinium scandate.

input files). The code and the NLP model are provided on the GitHub website: https://github.com/Zhang-NJ-Lab/NJmatNLP/blob/main/NLP.py and https://figshare.com/articles/software/NJmatML/24607893?file=45863628.

Post-hoc first-principles electronic structure calculation

DFT calculations are performed using the CASTEP[34,35] software package. Structural optimizations and property calculations are performed based on the three new crystal structures identified in the previous genetic+DFT step. The convergence criteria in the post-hoc DFT initial geometry optimization step are set to 1.0×10^{-5} eV for the energy, 0.01 eV/Å for the force, 0.02 GPa for the stress and 5.0×10^{-4} Å for the displacement, with PBE functional, 530 cutoff energy, ultrasoft pseudopotential and a 5 × 5 × 2 k-point set. Subsequently, a 9 × 9 × 4 k-point set is adopted to obtain the electronic properties such as its band structures, density of states (DOS) and UV-vis absorption spectra. The optimized crystal structure and the Magus parameter files are provided on the GitHub website (publicly available): https://github.com/Zhang-NJ-Lab/GdScO3_CSP. For example, the three crystal structures (after geometrical optimization) are

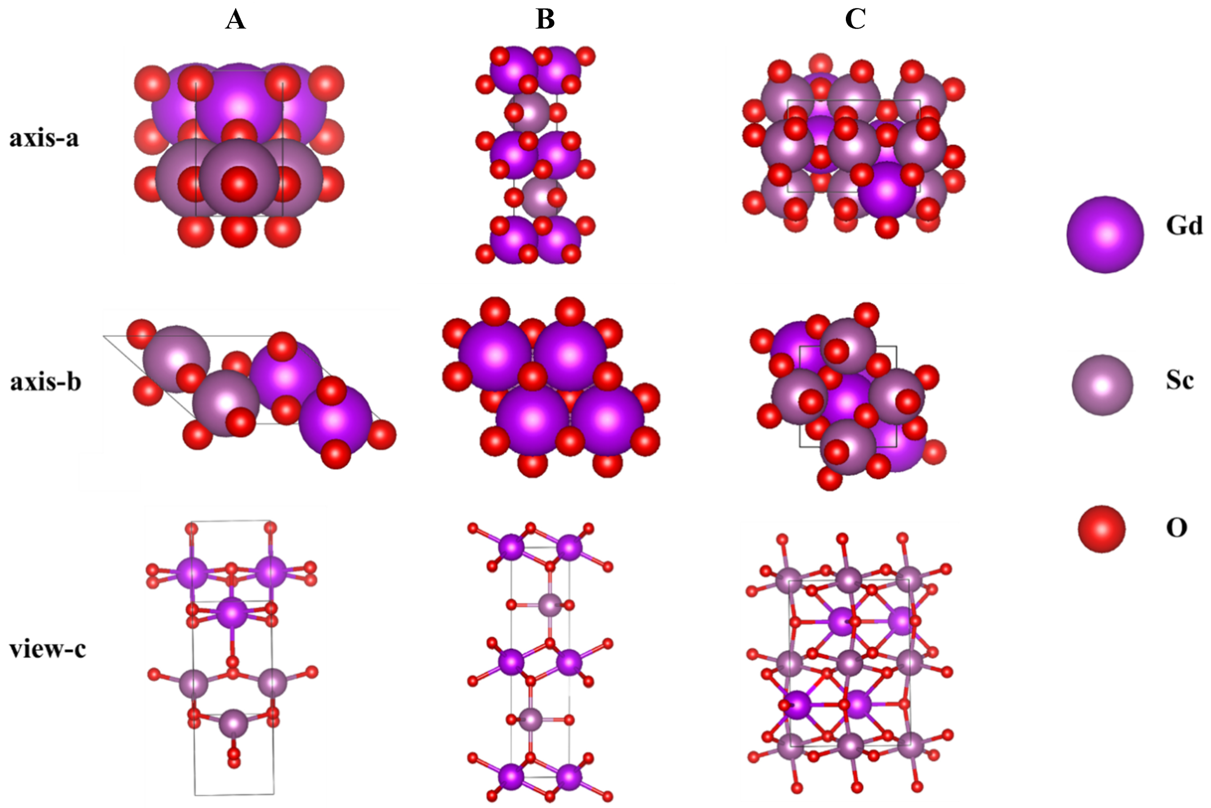

Figure 2. Crystal structures of three new predicted $GdScO_3$ with different viewpoints. (A) The monoclinic structure **1**; (B) The hexagonal structure **2**; (C) The orthorhombic structure **3**. $GdScO_3$: Gadolinium scandate.

provided at https://github.com/Zhang-NJ-Lab/GdScO3_CSP/tree/main/Crystal_Structures/OptimizedStructures.

Literature language model

The NLP workflow to construct the language model involves several essential steps: preparing the literature database, preprocessing, training the literature model, and performing dimensional reduction and clustering for visualization purposes. The construction of the language model analyzes the ontology of $GdScO_3$ from an alternative machine learning point-of-view based on the abstracts of 1.18 million scientific articles published in Springer Nature from 1960 to 2024 [Supplementary Materials], which corresponds to a concise summary of scientific information. This makes them a versatile data source despite some information loss compared to full texts. The searching criteria are associated with subjects in the domain of materials science, physics and chemistry. Additional preprocessing tasks include sentence splitting, tokenization, custom dictionary creation, spell checking, part-of-speech tagging, lemmatization, stemming, and tokenization. These steps are executed using the Natural Language Toolkit (NLTK) and verified with ChemDataExtractor for appropriate named entity recognition. The literature model is trained using Word2Vec with the following parameters: vector_size set to 100, window size of 10, sg set to 1, sample set to 1×10^{-3}, and trained over five epochs. Additionally, t-distributed stochastic neighbor embedding (t-SNE) is employed to reduce the dimensionality of the word vectors, with hyperparameters set to n_components = 2 and random_state = 42. The cosine similarity is calculated to demonstrate the evaluation of the potential applications of the perovskite material as follows:

$$cosine\ similarity = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

Here, A_i and B_i represent the word vectors of the material formula and the target vocabulary in the language model. Cosine similarity ranges from -1 to 1, where a cosine similarity of 1 indicates that the angle between the two vectors is zero (more likelihood that the material can be used for certain applications), a cosine similarity of 0 indicates that the angle is 90°, and a cosine similarity of -1 indicates complete dissimilarity with an angle of 180°.

RESULTS AND DISCUSSION

Crystal structure prediction

Three new crystal structures of $GdScO_3$ (**1**-**3**) are predicted by the machine learning crystal structure prediction method assisted via genetic algorithm: **1** has a monoclinic lattice and *Cm* space group (a = 6.372 Å, b = 3.485 Å, c = 8.819 Å, α = 90°, β = 138°, γ = 90°); **2** has a hexagonal lattice and $P6_3/mmc$ space group (a = 3.612 Å, b = 3.612 Å, c = 12.396 Å, α = 90°, β = 90°, γ = 120°); **3** has an orthorhombic crystal lattice and *Pnma* space group (a = 5.776 Å, b = 7.977 Å, c = 5.534 Å, α = 90°, β = 90°, γ = 90°). The atoms in the three crystal structures exhibit tilted octahedrons [Figure 2] that correspond to Jahn-Teller distortions with longer equatorial bonds and shorter vertical bonds. In **1**, the Sc–O bond lengths range from 2.04 to 2.09 Å, while the Gd–O bond lengths vary from 2.27 to 2.41 Å. In **2,** the Sc–O bond lengths span from 2.07 to 2.09 Å, and the Gd–O bond length is 2.33 Å. In the structure, the planar Gd–O layers are interconnected by ScO_6 octahedrons, forming a quasi-layered architecture that may benefit the solar-rechargeable battery application[36,37]. In addition, **3** exhibits Sc–O bond lengths from 2.10 to 2.13 Å, while the Gd–O bond lengths range from 2.28 to 2.67 Å. To sum up, the crystal structures of **1**-**3** differ from those reported in available crystal databases and their electronic and optical properties will be further evaluated.

Electronic structures

The band structures of the three new crystal structures of $GdScO_3$ predicted via genetic algorithms are examined to understand their electronic properties [Figure 3 and Supplementary Figure 1]. The band structures reveal a semimetallic pattern with a minority of band lines crossing the Fermi level near the Γ point before bending back. However, the semimetal feature is not significant and they can be considered as shallow defects that minimally influence the electronic excitation between valence band and conduction band in many cases[38,39]. Several semimetal materials have been demonstrated to deliver bulk photovoltaic effects and generate shift and injection photocurrents allowed by noncentrosymmetry[40]. Specifically, the monoclinic phase **1** characterized by the *Cm* space group exhibits an electronic band structure with a forbidden gap of 0.48 eV between 0.259 and 0.735 eV; this refers to a single band line ranging from -0.136 to 0.259 eV passing through the Fermi energy, and a nearby band line ranging from 0.735 to 0.974 eV is available. As a result, the band structure shows an energy gap between the neighboring bands near the Fermi level and the defect is considered to be shallow. The semimetallic behavior of the monoclinic phase suggests potential applications in spintronics and thermoelectric devices, where high electrical conductivity and thermoelectric efficiency are desirable. The hexagonal phase **2** of $GdScO_3$, characterized by the $P6_3/mmc$ space group, exhibits a similar band structure feature showcasing the semimetallic behavior. For **2**, multiple band lines pass through the Fermi energy level, such as one ranging from -0.064 to 0.086 eV and a second band line ranging from -0.058 to 0.122 eV. Meanwhile, there is no distinctive energy gap near the Fermi energy level, and the metallic behavior of **2** is suggested to be stronger than **1**. In addition, the orthorhombic phase **3** of $GdScO_3$ characterized by the *Pnma* space group exhibits a band structure indicating slight semimetallic behavior; e.g., a forbidden gap of 0.851 eV is available for **3** between 0.056 and 0.907 eV and

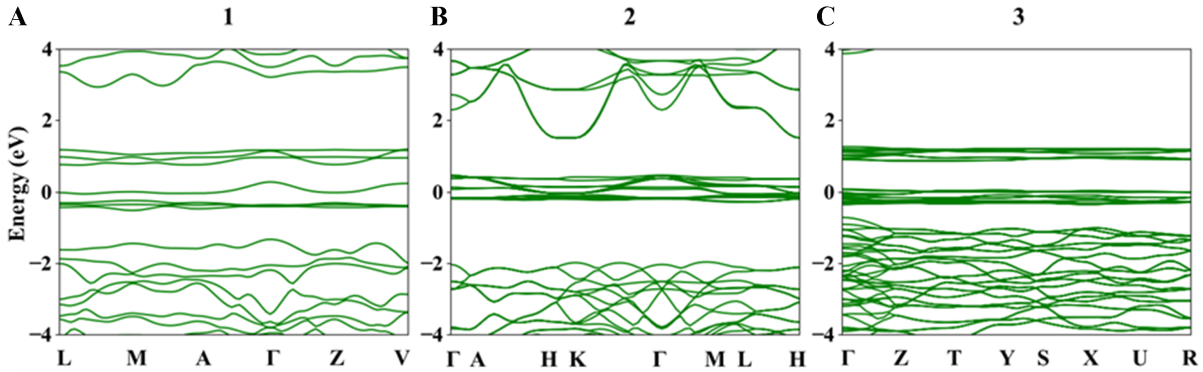

Figure 3. Band structures of new $GdScO_3$ crystal structures, including (A) the monoclinic structure **1**; (B) the hexagonal structure **2**; and (C) the orthorhombic structure **3**. The Fermi level corresponding to the highest occupied state is set at 0 eV. $GdScO_3$: Gadolinium scandate.

the empty states at 0.056 eV is considered to be shallow defects[41-43]. The semimetal feature observed in these structures can be considered as shallow defects. This is based on the fact that these defects, akin to those observed in perovskite materials, introduce energy levels close to the band edges. As a result, they have a negligible effect on electronic excitation processes, which is crucial for maintaining efficient charge transport in applications such as photovoltaics. The shallow nature of these defects ensures that they do not act as major recombination centers, aligning with the established behavior of similar defects in perovskite materials. To sum up, the genetic algorithm identifies three distinct crystal structures of $GdScO_3$ with minor semimetallic features corresponding to shallow defects toward optoelectronic applications.

The PDOS spectra of **1**-**3** [Figure 4] further illustrate the details of the electronic properties of the genetic-predicted new crystal structures of $GdScO_3$. An energy gap is available for **1**-**3** near 0.5 eV. The semimetallic feature is more prominent for **2** because of the additional empty states near 0.4 eV. In contrast, the valence band and conduction band of **1** and **3** are more separated, suggesting a reduced possibility of charge recombination after light excitation. For **1** and **3**, the Gd elements mainly contribute to the energy states near 0 eV while the Sr-d orbitals mainly contribute to the conduction band for **2**. In addition, the oxygen-p orbitals are universally present in the valence bands of **1**-**3** and contribute strongly to the bands from -4 to -2 eV. It is expected that charge transfers such as Gd → Gd, Gd → Sr, O → Gd and O → Sc may occur upon light excitation. In the X-ray photoelectron spectroscopy (XPS) and DFT analysis in the literature[44], the O 2p states are observed in the region from -3 to -5.5 eV, which is consistent with the present research. The Gd 4f states appear in a more localized region, and this spatial and energetic proximity suggests interactions between the O 2p and Gd 4f orbitals, indicating a potential charge transfer from oxygen to gadolinium. Additionally, the presence of Gd 5p states and Sc 3p states provides further evidence of charge distribution involving Gd and Sc atoms. The consistency between these XPS features and DFT calculations supports the proposed charge transfer pathways, particularly O → Gd and O → Sc. As a result, the optoelectronic performance of these new crystals of $GdScO_3$ is considered to be decent, and further structural engineering is required to further optimize the optoelectronic properties of these crystals.

The simulated UV-vis absorption spectra of **1**-**3** [Figure 5] demonstrate distinctive absorption bands in the UV-vis region. The absorption intensity is higher for **3** between 450 and 700 nm in the visible region (reaching 30,000 at the peak wavelength of 525 nm). However, **1** displays a strong absorption band near 800 nm, which is distinctive among the three. The absorption intensity of **2** in the visible region is comparatively inferior; however, **2** exhibits a large UV absorption intensity at 300 nm in the UV region. To

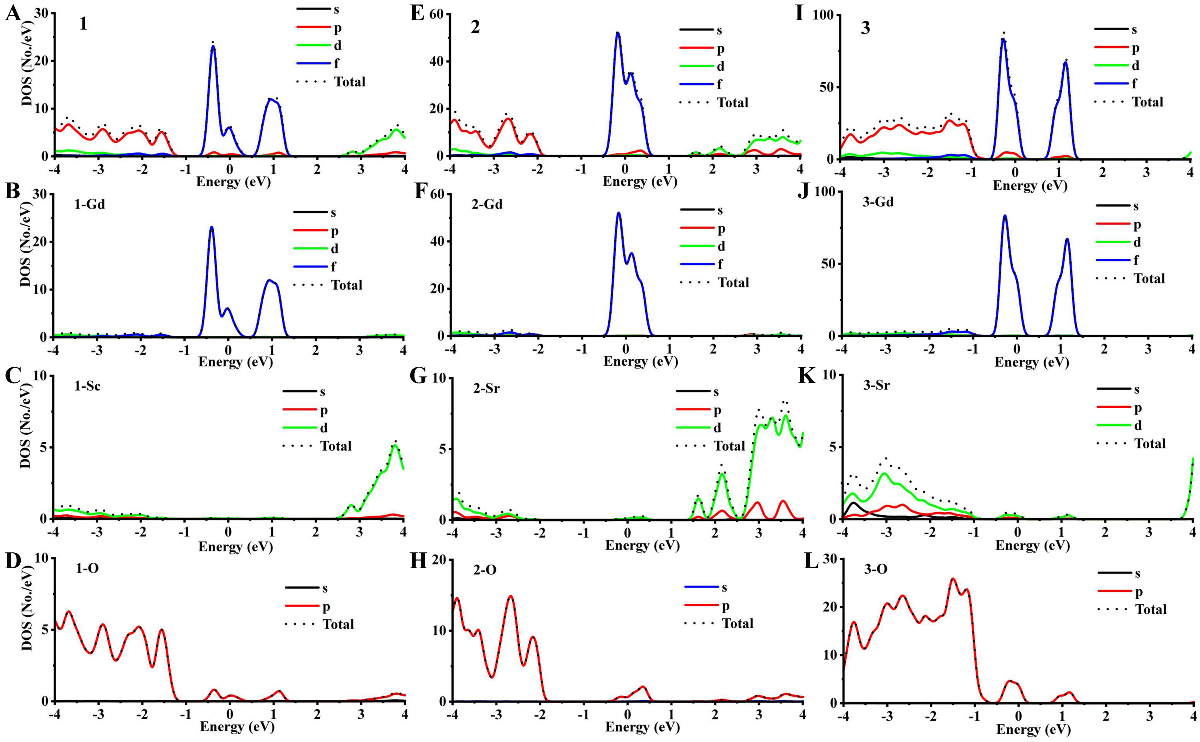

Figure 4. PDOS spectra of the $GdScO_3$. (A-D) Total DOS (total) and projected DOS (s, p and d) of **1**; (E-H) Total DOS and projected DOS of **2**; (I-L) Total DOS and projected DOS of **3**. PDOS: Partial density of states; $GdScO_3$: gadolinium scandate; DOS: density of states.

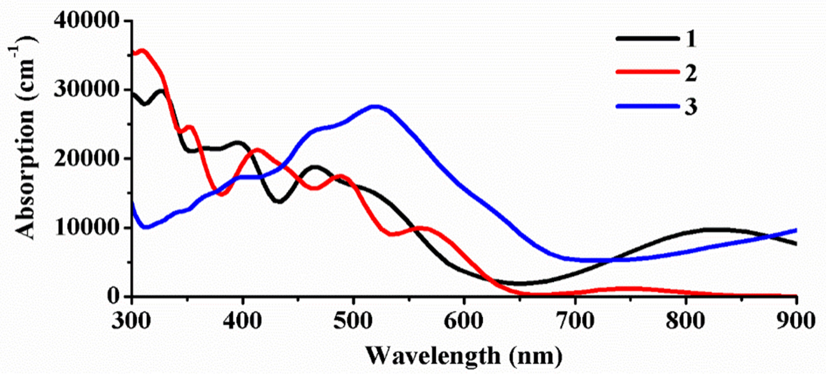

Figure 5. Simulated UV-vis absorption spectra of $GdScO_3$ crystal structures. UV-vis: Ultraviolet-visible; $GdScO_3$: gadolinium scandate.

conclude, the new $GdScO_3$ perovskite structures demonstrate decent absorption in the UV-vis region toward optoelectronic applications.

Strain effects

The strain effects on the band structure of material 1 are evaluated by applying strains ranging from -10% to 10%, in increments of 2% [Figure 6]. This range includes both compressive (negative) and tensile (positive) strains, allowing for a comprehensive analysis of how different strain levels influence the material's electronic properties. The band structures exhibit intricate changes as the strain is varied from compression (-10%) to tension (+10%). The symmetry of the band structures changes with strain; under compressive strain, the band lines show higher symmetry compared to those under tensile strain. In addition, the

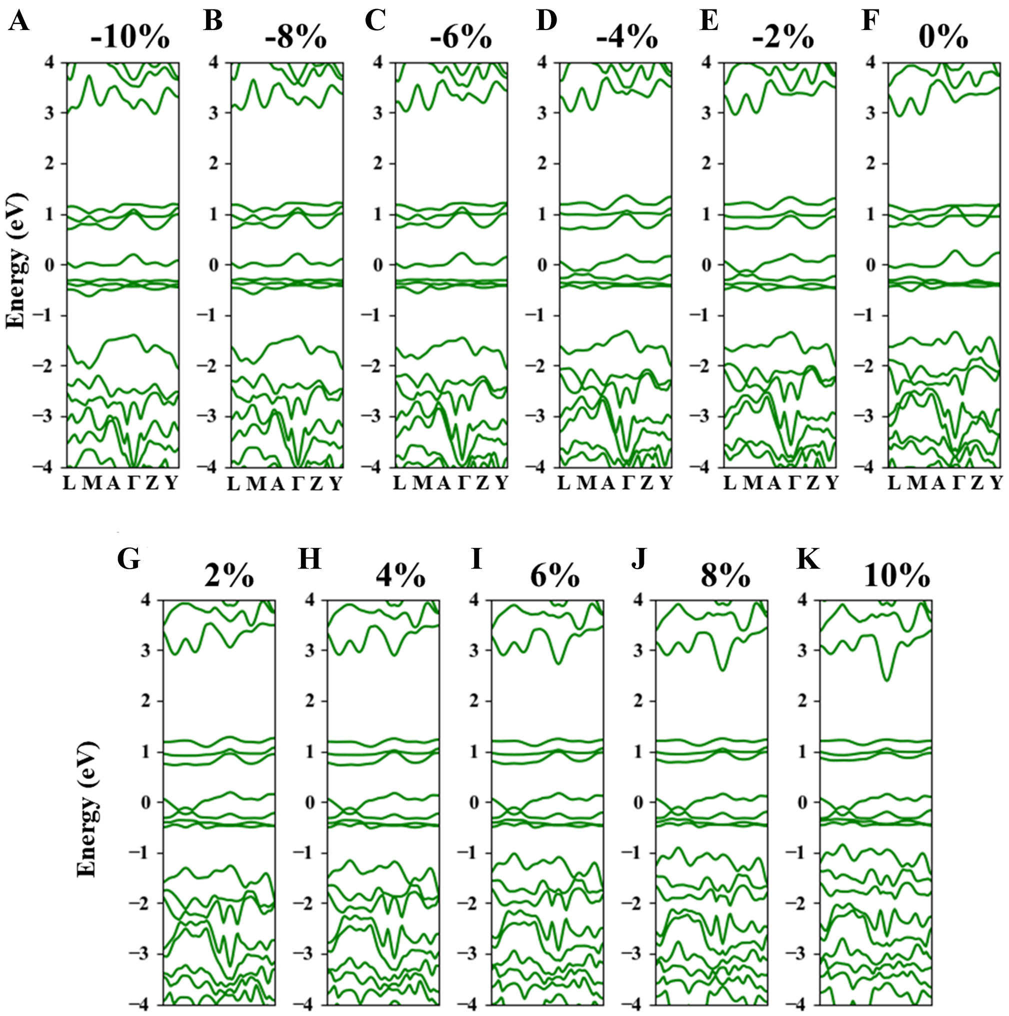

Figure 6. Band Structure of the monoclinic structure **1** with strain level from -10% (compressive strain) to 10% (tensile strain): (A) -10%, (B) -8%, (C) -6%, (D) -4%, (E) -2%, (F) 0%, (G) 2%, (H) 4%, (I) 6%, (J) 8%, (K) 10%.

bandwidth changes with the strain values, with certain bands widening or narrowing significantly as the strain varies, highlighting the intricate sensitivity of the material's electronic band structure to mechanical strain. (1) Applying compressive strain to $GdScO_3$ causes variation in the band lines that cross the Fermi level, resulting in a less pronounced crossing at the *Γ* and Y points in the Brillouin zone and potentially a reduction in the level of defects near the Fermi level toward optoelectronic application. In addition, this influences the effective mass of charge carriers as well as charge carrier mobility and electrical conductivity. The negative strain causes smaller energy separation between the neighboring empty band near 1 eV and the valence bands, leading to less pronounced semimetallic behavior under compressive strain; (2) Conversely, under tensile strain, the band line that crosses the Fermi level shifts upwards near the special points including M and A. Nevertheless, the level of defects near the highest occupied states is not prominent and the separation between the conduction band near 1 eV and the valence band near 0 eV is distinctive; this is associated with a shallow defect and desirable band gap formation in topological materials. Apart from that, the application of tensile and compressive strain often witnesses a reduction in the amount of empty states above the highest occupied states, signifying a possible minimization of undesirable defects near the valance band toward optoelectronic applications. These observations are crucial for photovoltaic applications where the material may be subjected to varying mechanical stresses, such as in flexible electronics or strain-engineered semiconductors.

The simulated UV-vis absorption of **1** further displays the impacts of mechanical strain on the optoelectronic properties of $GdScO_3$ [Figure 7]. Three distinctive regions in the UV-vis spectra are clearly influenced by the strain: a band in the UV region centered at 350 nm; a second band in the visible region centered at 590 nm; a third band in the near-infrared region centered at 850 nm. For the first band in the UV region, shifting from the negative strain to the tensile strain (from -10% to 10%) results in a monotonic absorption intensity reduction (bleaching effects). Nevertheless, both bands peaked at 590 and 850 nm demonstrate an initial reduction in the absorption intensity but a subsequent upward shift in the intensity when the tensile strain is more prominent. The transition occurs at strain = -4% for the band peaked at 590 nm and strain = -2% for the band peaked at 850 nm. As a result, both tensile and compressive strains may lead to enhanced absorption intensity in the visible and near-infrared region, while the absorption intensity is higher for the compressive strain and lower for the tensile strain in the UV region. To sum up, the light absorption properties of the predicted $GdScO_3$ can be strongly influenced by external mechanical stimuli.

Language model analysis of $GdScO_3$

The ontology of $GdScO_3$ material is comparatively insufficiently investigated and a holistic investigation is carried out in this study using language model employing 1.18 million scientific articles. The textual words are vectorized to evaluate the latent relationships between materials and their potential applications [Figure 8]. The NLP workflow for constructing the language model involves several key steps, including preparing the literature database, preprocessing, training the model using Word2Vec with a skip-gram approach, and performing dimensionality reduction for visualization [Supplementary Materials]. The literature database, built using the SpringerNature API, spans physics and chemistry publications from 1960 to 2020, covering abstracts from 1.18 million articles. In the skip-gram model, a shallow neural network is employed in an unsupervised manner to convert each word into a 200-dimensional vector. This model optimizes the probability $P(w_c|w_t)$ of a context word w_c given a target word w_t, by minimizing the loss function J, where w_t is the center word and w_{t-1}, w_{t-2}, w_{t+1}, w_{t+2} are the surrounding context words. To visualize the high-dimensional word vectors, t-SNE is used, which reduces the dimensionality by preserving the local structure of data, effectively mapping similar high-dimensional points to nearby points in lower dimensions. This method minimizes the divergence between probability distributions of points in the high-dimensional space and their corresponding low-dimensional representations. Finally, cosine similarity is calculated to evaluate the potential applications of materials, and they are ranked for photovoltaic applications. Known solar cell materials such as $FAPbI_3$, CIGS, $MAPbI_3$, GaAs, CdTe, and InP are identified. $GdScO_3$, with comparable cosine similarity, suggests potential as a promising photovoltaic material, as corroborated by the clustering effects observed after t-SNE dimensional reduction. Currently, the applications of $GdScO_3$ are mainly limited to high-k dielectric material in electronic devices, thin film capacitors and as a substrate for the epitaxial growth of materials in advanced electronic devices. However, the language model suggests that the $GdScO_3$ material is also recommended for various alternative applications related to semiconductors, photoconductivity, composites, photoemission, interfaces, devices and photovoltaics. For example, the language model recommends $GdScO_3$ for photovoltaic application; after t-SNE, $GdScO_3$ clusters with CIGS, CIGSe, GaAs and $MAPbI_3$ that are typical solar cell materials (dark region) [Figure 9 and Supplementary Figure 2]. In addition, the cosine similarity between $GdScO_3$ and photovoltaic is high, and the similarity value is comparable to that of typical solar cell materials such as CdTe and InP. This suggests the possibility of this perovskite oxide material for optoelectronic application. This, to some extent, contradicts the previous DFT calculations regarding the defects and semimetallic features of $GdScO_3$. However, the following two points are suggested to justify the suitability of $GdScO_3$ for

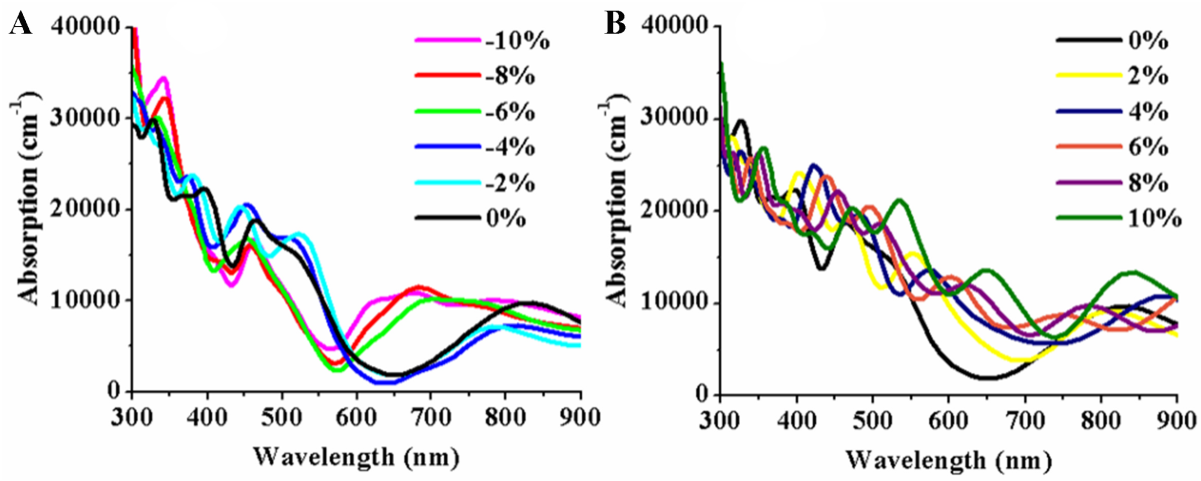

Figure 7. Strain effects on the simulated UV-vis absorption spectra of the monoclinic structure **1**. (A) Negative strain effects; (B) Positive strain effects. UV-vis: Ultraviolet-visible.

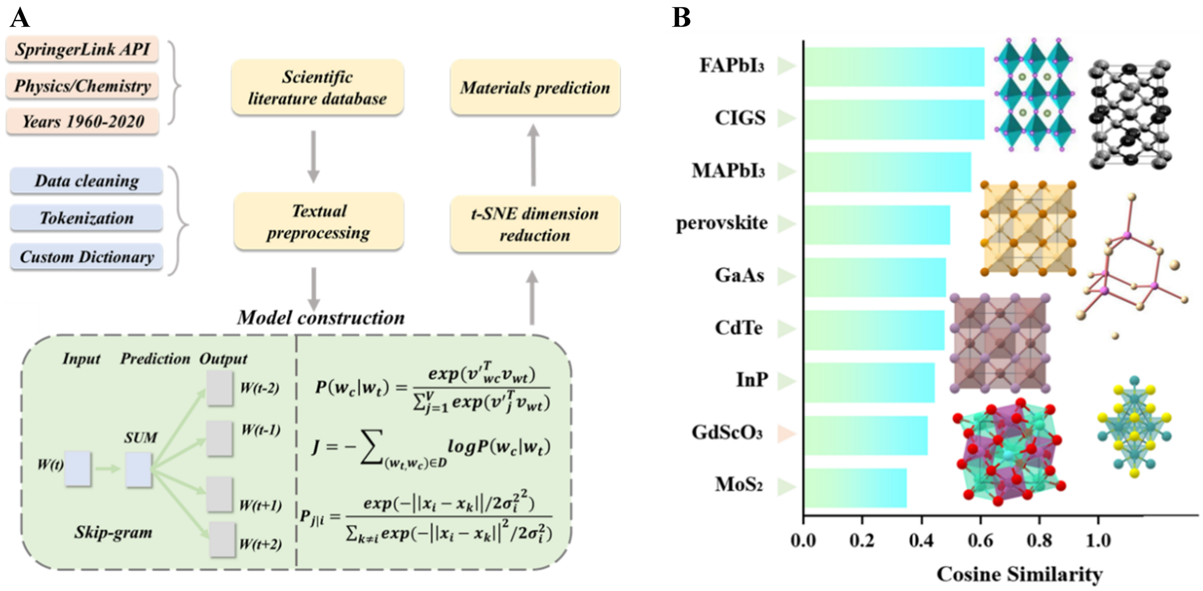

Figure 8. Language model analysis for $GdScO_3$. (A) Workflow of the language model construction and materials analysis, starting from scientific literature database and followed by preprocessing, model construction, dimension reduction and materials prediction. Here, the skip-gram method to predict textual words; (B) Ranking of exemplar materials identified by the language model toward photovoltaic application based on cosine similarity. $GdScO_3$: Gadolinium scandate.

photovoltaic applications. (1) The semimetallic features of **1-3** are not completely detrimental for photovoltaics evidenced by the apparent forbidden gap next to the band lines near the Fermi level, and the defects are considered to be shallow toward the light excitation; (2) Photovoltaic effects have been observed in several semimetal materials, including the colossal mid-infrared bulk photovoltaic effect in a Type-I Weyl semimetal, and optically induced thermal gradients via the Seebeck effect are suggested to benefit the current production provided with careful balance of the optical, electronic and thermal material properties[40,45,46]. In addition, for perovskite-based materials (where the target word is perovskite and the cosine similarity is calculated between the material formula entity and the application entity), the $GdScO_3$ material clusters with $MAPbI_3$, $FAPbI_3$ and $MAPbBr_3$ that are typical hybrid perovskite materials toward

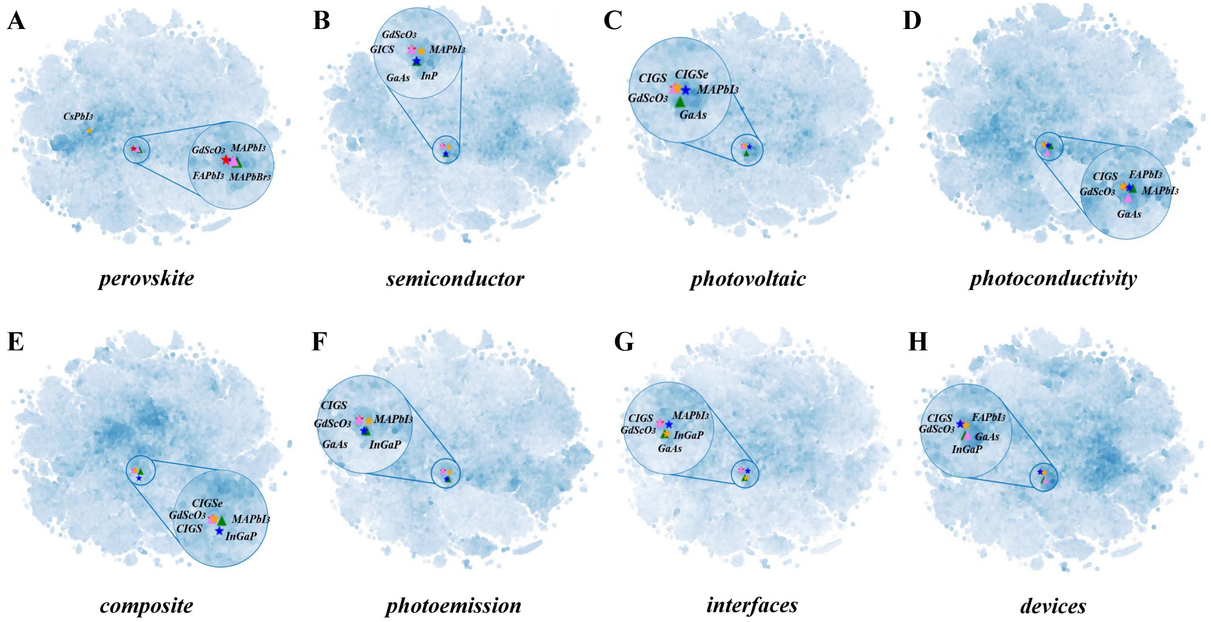

Figure 9. Visualization of language model for analyzing $GdScO_3$ toward various potential applications, which are based on t-SNE dimension reduction from a language model using 1.18 million scientific articles. The following targets are evaluated: (A) perovskite, (B) semiconductor, (C) photovoltaic, (D) photoconductivity, (E) composite, (F) photoemission, (G) interfaces and (H) devices. $GdScO_3$: Gadolinium scandate; t-SNE: t-distributed stochastic neighbor embedding.

photovoltaic applications. Meanwhile, the spatial distribution of $GdScO_3$ is far away from the comparatively inferior halide perovskite $CsPbI_3$ material toward the photovoltaic application, manifesting the potential of $GdScO_3$ toward this particular application. Moreover, the $GdScO_3$ material clusters well with CIGS, GaAs and InP for semiconductor application. Similarly, higher cosine similarity is observed between $GdScO_3$ and various applications such as photoconductivity, composite, photoemission, interfaces and devices. To sum up, the language model obtained via 1.18 million scientific articles suggests that $GdScO_3$ is a possible material for photovoltaic applications.

CONCLUSION

A hybrid approach combining genetic algorithms and DFT is employed to explore novel crystal structures of $GdScO_3$ and elucidate the electronic properties based on monoclinic (*Cm*), hexagonal (*P63/mmc*) and orthorhombic (*Pnma*) crystal systems. The post-hoc DFT calculations provide detailed insights into the crystal structures and electronic and optical properties of these new phases, suggesting the possibility for optoelectronic applications despite the existence of shallow defects. The electronic properties can be fine-tuned by mechanical strain. Additionally, leveraging language model analysis with dimensionality reduction and clustering techniques enables a data-driven ontological exploration of $GdScO_3$, and further suggests the possibility of $GdScO_3$ toward photovoltaic applications. This integrated approach enhances the understanding of $GdScO_3$ and establishes a robust framework for exploring novel functional materials combining DFT calculations and machine learning methods.

DECLARATIONS

Authors' contributions

Conceptualization, resources, supervision, writing, funding acquisition: Zhang L

Software and methodology: Zhou J
Visualization: Chen X

Data availability statement

The data that supports the findings of this study are available within the article.

Financial support and sponsorship

This work was supported by the Jiangsu Provincial Science and Technology Project Basic Research Program (Natural Science Foundation of Jiangsu Province) (No. BK20211283).

Conflict of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Scheffler M, Aeschlimann M, Albrecht M, et al. FAIR data enabling new horizons for materials research. *Nature* 2022;604:635-42. DOI PubMed
2. Gupta V, Choudhary K, Mao Y, et al. MPpredictor: an artificial intelligence-driven web tool for composition-based material property prediction. *J Chem Inf Model* 2023;63:1865-71. DOI PubMed PMC
3. Zhang L, He M, Huang E, et al. Overcoming language barrier for scientific studies via unsupervised literature learning: case study on solar cell materials prediction. *Solar RRL* 2024;8:2301079. DOI
4. Shao S, Yan L, Zhang L, et al. Data-driven exploration of terbium-doped tungsten oxide for ultra-precise detection of 3H-2B: implications for gas sensor applications. *Chem Eng J* 2024;487:149680. DOI
5. Wang S, Huang Y, Hu W, Zhang L. Data-driven optimization and machine learning analysis of compatible molecules for halide perovskite material. *npj Comput Mater* 2024;10:1297. DOI
6. Zhang J, Zhang L, Sun Y, Li W, Quhe R. Named entity recognition in the perovskite field based on convolutional neural networks and MatBERT. *Comput Mater Sci* 2024;240:113014. DOI
7. Trewartha A, Walker N, Huo H, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* 2022;3:100488. DOI PubMed PMC
8. Luu RK, Buehler MJ. BioinspiredLLM: conversational large language model for the mechanics of biological and bio-inspired materials. *Adv Sci* 2024;11:e2306724. DOI PubMed PMC
9. Gupta T, Zaki M, Krishnan NMA, Mausam. MatSciBERT: a materials domain language model for text mining and information extraction. *npj Comput Mater* 2022;8:784. DOI
10. Li Y, Feng X, Liu H, et al. Route to high-energy density polymeric nitrogen *t*-N via He-N compounds. *Nat Commun* 2018;9:722. DOI PubMed PMC
11. Tong Q, Gao P, Liu H, et al. Combining machine learning potential and structure prediction for accelerated materials design and discovery. *J Phys Chem Lett* 2020;11:8710-20. DOI PubMed
12. Marchenko EI, Fateev SA, Petrov AA, et al. Database of two-dimensional hybrid perovskite materials: open-access collection of crystal structures, band gaps, and atomic partial charges predicted by machine learning. *Chem Mater* 2020;32:7383-8. DOI
13. Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput Mater* 2020;6:138. DOI
14. Omee SS, Wei L, Hu M, Hu J. Crystal structure prediction using neural network potential and age-fitness Pareto genetic algorithm. *J Mater Inf* 2024;4:1. DOI
15. Fu N, Hu J, Feng Y, Morrison G, Loye HZ, Hu J. Composition based oxidation state prediction of materials using deep learning language models. *Adv Sci* 2023;10:e2301011. DOI PubMed PMC
16. Pickard CJ, Needs RJ. Ab initio random structure searching. *J Phys Condens Matter* 2011;23:053201. DOI PubMed

17. Pickard CJ, Needs RJ. High-pressure phases of silane. *Phys Rev Lett* 2006;97:045504. DOI PubMed
18. Wang Y, Lv J, Zhu L, Ma Y. CALYPSO: a method for crystal structure prediction. *Comput Phys Commun* 2012;183:2063-70. DOI
19. Wang Y, Lv J, Zhu L, et al. Materials discovery via CALYPSO methodology. *J Phys Condens Matter* 2015;27:203203. DOI PubMed
20. Wang J, Gao H, Han Y, et al. MAGUS: machine learning and graph theory assisted universal structure searcher. *Natl Sci Rev* 2023;10:nwad128. DOI PubMed PMC
21. Song Q, Zhang N, Liu J, et al. Efficient continuous wave and broad tunable lasers with the Tm:GdScO$_3$ crystal. *Opt Lett* 2023;48:640-3. DOI PubMed
22. Cai E, Du L, Zhao J, et al. Sub-100 fs pulses lasers from Yb:GdScO$_3$ crystal based on semiconductor saturable absorber mirrors. *Infrared Phys Technol* 2024;139:105309. DOI
23. Eremeev K, Loiko P, Zhao C, et al. Growth, spectroscopy and laser operation of Tm,Ho:GdScO$_3$ perovskite crystal. *Opt Express* 2024;32:13527-42. DOI PubMed
24. Dong J, Li J, Wang Q, et al. Crystal growth and spectroscopic analysis of Ho,Eu:GdScO$_3$ crystal for 3 μm mid-infrared emission. *J Lumin* 2023;254:119515. DOI
25. Guo R, Wang F, Wang S, et al. Exploration of the crystal growth and crystal-field effect of Yb^{3+} in orthorhombic GdScO$_3$ and LaLuO$_3$ crystals. *Cryst Growth Des* 2023;23:3761-8. DOI
26. Stegmaier S, Voss J, Reuter K, Luntz AC. Li$^+$ defects in a solid-state Li ion battery: theoretical insights with a Li$_3$OCl electrolyte. *Chem Mater* 2017;29:4330-40. DOI
27. Moradabadi A, Kaghazchi P. Effect of strain on polaron hopping and electronic conductivity in bulk LiCoO$_2$. *Phys Rev Appl* 2017;7:064008. DOI
28. He Y, Galli G. Perovskites for solar thermoelectric applications: a first principle study of CH$_3$NH$_3$AI$_3$ (A = Pb and Sn). *Chem Mater* 2014;26:5394-400. DOI
29. Talapatra A, Uberuaga BP, Stanek CR, Pilania G. A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides. *Chem Mater* 2021;33:845-58. DOI
30. Kresse G, Hafner J. Ab initio molecular dynamics for liquid metals. *Phys Rev B Condens Matter* 1993;47:558-61. DOI PubMed
31. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* 2013;65:1501-9. DOI
32. Liu M, Meng S. Atomly.net materials database and its application in inorganic chemistry. *Sci Sin Chim* 2023;53:19-25. DOI
33. Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater* 2013;1:011002. DOI
34. Clark SJ, Segall MD, Pickard CJ, et al. First principles methods using CASTEP. *Z Krist Cryst Mater* 2005;220:567-70. DOI
35. Segall MD, Lindan PJD, Probert MJ, et al. First-principles simulation: ideas, illustrations and the CASTEP code. *J Phys Condens Matter* 2002;14:2717. DOI
36. Kumar A, Thakur P, Sharma R, Puthirath AB, Ajayan PM, Narayanan TN. Photo rechargeable Li-ion batteries using nanorod heterostructure electrodes. *Small* 2021;17:e2105029. DOI PubMed
37. Boruah BD, Wen B, De Volder M. Light rechargeable lithium-ion batteries using V$_2$O$_5$ cathodes. *Nano Lett* 2021;21:3527-32. DOI PubMed PMC
38. Dey K, Roose B, Stranks SD. Optoelectronic properties of low-bandgap halide perovskites for solar cell applications. *Adv Mater* 2021;33:e2102300. DOI PubMed
39. Zhang Y, Zhang J, Gao W, et al. Near-edge band structures and band gaps of Cu-based semiconductors predicted by the modified Becke-Johnson potential plus an on-site Coulomb U. *J Chem Phys* 2013;139:184706. DOI PubMed
40. Ahn J, Guo GY, Nagaosa N. Low-frequency divergence and quantum geometry of the bulk photovoltaic effect in topological semimetals. *Phys Rev X* 2020;10:041041. DOI
41. Xu L, Li Y, Shi J, et al. Suppressing shallow defect of printable mesoscopic perovskite solar cells with a N719@TiO$_2$ inorganic–organic core–shell structured additive. *Solar RRL* 2020;4:2000042. DOI
42. Zhang L, Li S, Hu W. First-principles investigation on adsorption of anchors on two-dimensional halide perovskite material. *Appl Surf Sci* 2022;604:154527. DOI
43. Cohen B, Alafi R, Beinglass J, et al. In-gap states and carrier recombination in quasi-2D perovskite films. *Solar RRL* 2023;7:2300813. DOI
44. Raekers M, Kuepper K, Bartkowski S, et al. Electronic and magnetic structure of *R*ScO$_3$ (R = Sm,Gd,Dy) from x-ray spectroscopies and first-principles calculations. *Phys Rev B* 2009;79:125114. DOI
45. Bao X, Ou Q, Xu Z, Zhang Y, Bao Q, Zhang H. Band structure engineering in 2D materials for optoelectronic applications. *Adv Mater Technol* 2018;3:1800072. DOI
46. Osterhoudt GB, Diebel LK, Gray MJ, et al. Colossal mid-infrared bulk photovoltaic effect in a type-I Weyl semimetal. *Nat Mater* 2019;18:471-5. DOI PubMed