Applied Physics Letters ARTICLE pubs.aip.org/aip/apl

Inverse design of high-performance piezoelectric semiconductors via advanced crystal representation and large language models

Cite as: Appl. Phys. Lett. **126**, 111901 (2025); doi: 10.1063/5.0247738
Submitted: 8 November 2024 · Accepted: 4 March 2025 ·
Published Online: 17 March 2025

Chen Zhang,[1] Siyuan Lv,[1] Haojie Gong,[1] Qianxi Cheng,[1] Junwei Guo,[1] Zheng Duanmu,[1,a)] and Hang Xiao[2,a)]

AFFILIATIONS

[1]School of Instrument Science and Opto-Electronics Engineering, Beijing Information Science and Technology University, Beijing, China
[2]School of Interdisciplinary Studies, Lingnan University, Tuen Mun, Hong Kong SAR, China

[a)]Authors to whom correspondence should be addressed: ariesdmz@bistu.edu.cn and hangxiao@ln.edu.hk

ABSTRACT

The inverse design of solid-state materials with targeted properties represents a significant challenge in materials science, particularly for piezoelectric semiconductors where both structural symmetry and electronic properties must be carefully controlled. Here, we employ the simplified line-input crystal-encoding system representation combined with the MatterGPT framework for discovering potential piezoelectric semiconductors. By training on a curated dataset of 1556 piezoelectric materials from the Materials Project database, our model learns to generate crystal structures with targeted piezoelectric properties through an autoregressive sampling process. Starting from approximately 5000 generated structures, we implemented a comprehensive screening workflow incorporating structural validity, thermodynamic stability, and property verification. This approach identified several promising candidates from 4100 reconstructed structures, each representing compounds unrecorded in existing databases. Among these, the most notable material demonstrated a piezoelectric stress coefficient of 25.9 C/m^2 in the e[1,6] direction. Additionally, these materials demonstrate suitable bandgaps ranging from 1.63 to 3.61 eV, suggesting potential applications in high-sensitivity sensors and high-temperature electronics. Our work demonstrates the effectiveness of combining crystal structure language encoding with generative models for accelerating the discovery of functional materials with targeted properties.

The piezoelectric effect, a reversible physical process where materials generate electrical charge under applied stress and deform in response to electric fields, plays a crucial role in numerous technological applications. This phenomenon occurs in materials lacking inversion symmetry, enabling polarization under external forces.[1] Piezoelectric semiconductor materials, which combine charge transport capabilities with electromechanical coupling, are particularly attractive for applications ranging from pressure sensors and actuators to nanogenerators and other energy harvesting devices.[2] The coupling of their mechanical, electrical, and optical properties makes them promising candidates for next-generation sensing technologies and smart materials.[3] The development of high-performance piezoelectric semiconductors could enable transformative advances in these applications. However, discovering such materials with optimal properties remains a significant scientific and technological challenge.[4]

Traditional approaches to materials discovery rely heavily on empirical knowledge and iterative optimization through high-throughput screening. These forward design methods start from candidate structures and progressively refine them to achieve desired properties.[5] While this approach has led to successes, it faces several fundamental limitations. The process is both time-consuming and computationally intensive,[6,7] often encounters discrepancies between simulations and experiments,[8] and struggles to simultaneously optimize multiple competing properties.[9,10] Forward design methods are inherently constrained by human intuition and known structural motifs, limiting exploration of the vast chemical space needed to discover high-performing piezoelectric materials, where even subtle structural features can significantly impact performance.[11]

Inverse design, which starts from desired properties and works backward to determine the corresponding structure and composition,

offers a promising alternative approach.[12,13] This paradigm aims to directly generate material structures that satisfy target property constraints, potentially enabling more efficient exploration of the materials space.[14] However, inverse design of crystalline materials presents unique challenges due to the complex interplay between atomic composition, crystal structure, and material properties.[15] The challenge is particularly acute for functional materials like piezoelectric semiconductors, where both structural symmetry and electronic properties must be carefully controlled.

Recent advances in artificial intelligence, particularly generative models, have opened possibilities for inverse materials design.[16] Various architectures, including variational autoencoders (VAEs),[17] generative adversarial networks (GANs),[18] and diffusion models,[19] have been explored in this context. However, existing approaches face significant limitations in handling crystalline materials. For instance, crystal representations using 3D images or voxels, as in the image-based materials generator network, show some success in reversible crystal generation but suffer from high computational costs, low symmetry invariance, and limited applicability and reliability.[20–22] While crystal graph convolutional neural networks can effectively encode structural information and predict properties, they still lack the critical ability to generate potential structures.

The simplified line-input crystal-encoding system (SLICES),[23] inspired by the SMILES representation widely used in molecular design,[24] offers a promising solution to these challenges. By efficiently encoding crystal structures into string representations, SLICES enables both reversible structure generation and maintains important invariances.[23] Recent work has shown that when combined with large language models, the SLICES crystal structure encoding approach enables efficient inverse design of solid-state materials. MatterGPT can generate stable and potential materials with targeted properties across the periodic table.[25] This combination of robust crystal representation and advanced large language models enables comprehensive learning of the structure–property relationships between encoded crystal structures and crystal properties, providing an ideal framework for tackling challenging inverse design problems such as piezoelectric material discovery.

In this work, we leverage the SLICES-based MatterGPT architecture for inverse design of potential piezoelectric semiconductor materials. By training on a curated database of piezoelectric materials, the model learns both the SLICES representation of crystal structures and the correlation between structures and piezoelectric properties in a one-shot manner. During generation, we condition the model on target piezoelectric modulus values to produce candidate structures through an autoregressive sampling process, where crystal structures are generated token by token through the transformer decoder. These generated SLICES strings are then converted back to crystal structures for evaluation. Through multi-step screening, including structural validity check, stability assessment via DFT calculations, and piezoelectric property verification, several promising piezoelectric semiconductor materials were discovered. Through this systematic process, we designed several promising high-performance piezoelectric semiconductors, with performance metrics exceeding those of many existing materials. These results not only demonstrate the effectiveness of our approach but also highlight its potential for discovering functional materials for applications in medical devices, electronic systems, and environmental monitoring.

The piezoelectric effect is mathematically described by a third-order tensor relating strain (or stress) to the electric field. This tensor describes the response of any piezoelectric bulk material to an applied electric field or mechanical load. In the Material Project, the piezoelectric stress coefficient e^{T}_{ijk} is calculated using the DFPT[26] and is given in units of C/m^2, which can be defined using the thermodynamic derivative.[27] Written in Voigt notation, it is given by

$$e^{T}_{ijk} = \left(\frac{\partial D_i}{\partial \varepsilon_{jk}}\right)_{E,T} = -\left(\frac{\partial \sigma_{jk}}{\partial E_i}\right)_{\varepsilon,T}, \tag{1}$$

where D, E, ε, σ, and T represent the electric displacement field, the electric field, the strain tensor, the stress tensor, and the temperature, respectively.

It should be noted that the piezoelectric strain constant, usually denoted by d^{T}_{ijk}, is the more common piezoelectric constant found in the (experimental) literature. If the elastic compliance s^{ET}_{lmjk} of the material (at a constant electric field and temperature) is known, it can be related to the piezoelectric stress constant e_{ilk} using the following formula:

$$d^{T}_{ijk} = e_{ilm} s^{ET}_{lmjk}. \tag{2}$$

A high piezoelectric stress constant along certain crystallographic directions is a key indicator for potential applications in high-performance electronic devices, as it determines the material's ability to efficiently convert between mechanical and electrical signals. The piezoelectric modulus of the material designed in this paper refers to the maximum value in the piezoelectric stress coefficient tensor e^{T}_{ijk}.

Our training dataset was derived from the Materials Project (MP) database,[28] which provides piezoelectric tensor components calculated using density functional perturbation theory (DFPT).[26] These piezoelectric constants arise from the combined electronic and ionic contributions to the material's response to mechanical deformation.

To construct a high-quality training dataset, we implemented several filtering criteria:

1. We excluded low-dimensional structures (0D, 1D, and 2D) as the quotient graph method employed in our reconstruction scheme is only applicable to three-dimensional crystals.[29]
2. We limited our dataset to structures containing elements with atomic numbers below 86 due to constraints in the GFN-FF used in the reconstruction algorithm of SLICES.
3. We restricted our analysis to semiconductors and insulators by removing metallic structures, which were identified based on their electronic band structures.[30] This step ensures the dataset focuses on materials most relevant for piezoelectric applications.

Starting from an initial set of 2804 materials with known piezoelectric constants in the MP database, we extracted their crystallographic information files (CIF) using pymatgen.[31] After applying the above filtering criteria, 1788 structures remained. The filtered dataset includes 98 elements, with 63 distinct symmetry space groups, predominantly P1, Cmc2_1, F-43m, and R3m.The distribution of piezoelectric modulus is shown in Fig. 1, with a mean value of 1.31 C/m^2, a median of 0.54 C/m^2, and a standard deviation of 4.29 C/m^2. Notably, the number of samples with a piezoelectric modulus greater than 5 C/m^2 is relatively small, highlighting the significance of exploring materials with high piezoelectric modulus. These structures were then

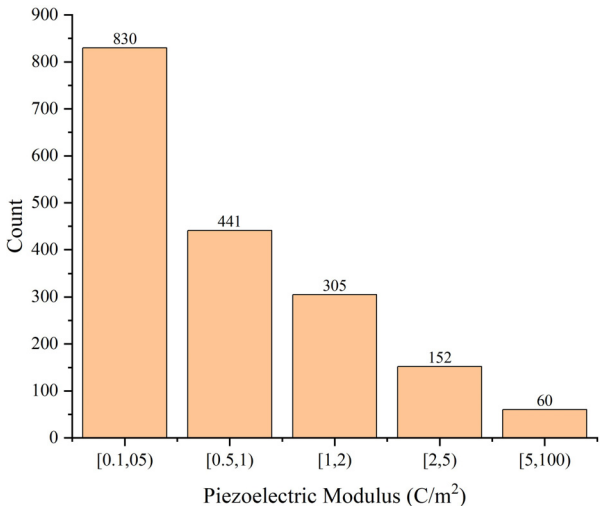

FIG. 1. Distribution of piezoelectric constants for the filtered dataset.

encoded into SLICES representations, resulting in 1556 valid strings that formed our final training dataset.

SLICES is an invertible, invariant crystal representation that encodes three key aspects of crystal information: atomic symbols, node indices, and edge labels. The string representation begins with atomic symbols that encode the chemical composition of the corresponding crystal structure. The edges are explicitly represented in the form "u v a b c," where "u v" are node indices and "a b c" denotes the location of the unit cell to connect to. In Fig. 1, the label of edge e_4 is "001," indicating that node A_0 connects to the copy of A_1 shifted one unit along the z axis. These edge labels, which specify the translational periodicity of edges, enable the construction of suitable initial guess structures derived from graph theory.

The encoding of crystal structures to SLICES strings involves: (1) parsing the structure file with Pymatgen[31] into a structure object, (2) constructing a structure graph using the crystal near-neighbor (crystalnn) algorithm,[32] and (3) extracting the chemical composition, bonding connectivity, and translation vectors to generate the corresponding SLICES string.

Decoding SLICES strings to crystal structures uses the SLI2Cry algorithm,[23] which comprises three steps:

(1) Initial structure generation using Eon's topology-based method.
(2) Optimization based on geometry predicted by a modified polarizable force field.
(3) Structural refinement using the universal deep learning force field.

This reconstruction scheme reconstructs 94.95% of original crystal structures, demonstrating SLICES' high invertibility while maintaining invariance to translation, rotation, and permutation. It provides a solid foundation for studying the crystal structures and properties of piezoelectric semiconductors.

We employed the SLICES-based MatterGPT framework for the inverse design of piezoelectric semiconductor materials. Our workflow consists of four main stages, as illustrated in Fig. 2:

(1) Training: The MatterGPT model was trained on a curated dataset from the Materials Project database to learn both the SLICES representation and the relationship between crystal structures and piezoelectric properties. The property information was incorporated through an embedding layer concatenated with the SLICES embedding.
(2) Generation: The trained model generates candidate SLICES strings conditioned on target property values using classifier-free guidance. The sampling temperature and nucleus sampling parameters were tuned to balance generation diversity and validity.
(3) Structure Reconstruction: The generated SLICES strings are converted back to crystal structures using the SLI2Cry algorithm.
(4) Screening: The candidate structures undergo multi-step screening to identify promising piezoelectric semiconductors.

For property-guided generation, we conditioned on both formation energy and piezoelectric modulus. The generated structures were evaluated against the following criteria:

(1) Chemical novelty: The composition must not exist in the Materials Project database.
(2) Structural uniqueness: The crystal structure should exhibit low similarity to known materials, as determined by structural matching algorithms.
(3) The energy above hull should be less than 0.5 eV/atom to ensure synthetic accessibility and reasonable stability.[33]
(4) Electronic properties: The material should possess an appropriate bandgap characteristic of semiconductors.

Our model implementation follows the MatterGPT architecture illustrated in Fig. 2(b), which adopts a ChatGPT-like autoregressive

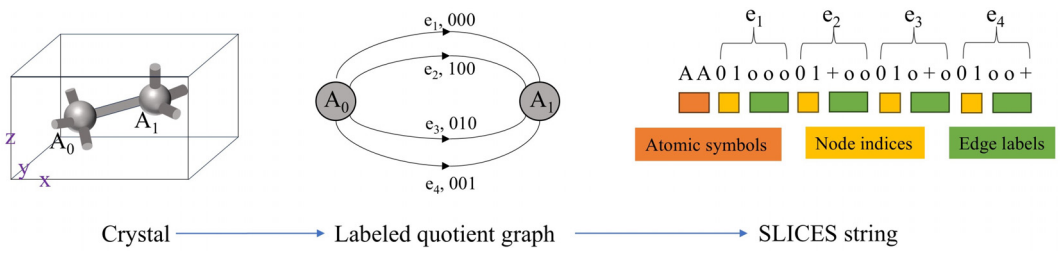

FIG. 2. Principle of representing a crystal code as a string.

TABLE I. Generative performance for targeted properties.

Formation energy (eV/atom)	Piezoelectric modulus (C/m^2)	Validity (%)	Uniqueness (%)	Novelty (%)	MAPE (%)
−2	7.5	82.9	99.9	64.7	42.4

decoder-only transformer architecture.[25] The model consists of n blocks of transformer decoders, where each block contains a masked multi-head self-attention module followed by a feed-forward network. The self-attention mechanism employs the "scaled dot product attention,"

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{3}$$

where Q, K, and V denote the query, key, and value matrices, respectively. For self-attention, these matrices are derived from the same input sequence, while in cross-attention, Q comes from the input sequence while K and V are obtained from different sources. The masked attention mechanism ensures autoregressive generation by restricting each token to attend only to previous tokens in the sequence. The model employs multi-head attention, where multiple attention heads operate in parallel to capture diverse features and dependencies in the sequence, with their outputs concatenated and linearly transformed.

To enhance generation diversity during sampling, we employ Gumbel-Softmax sampling with a temperature parameter $T = 1.2$, which introduces controlled noise into the predicted logit distributions. We also utilize nucleus sampling (top-p sampling) with $p = 0.9$, ensuring that only tokens from the top 90% of the cumulative probability mass are considered during generation, thereby maintaining both coherence and diversity. The model was trained using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 60. The architecture features an embedding dimension of 768, 12 transformer decoder layers, and 12 attention heads, totaling approximately 80×10^6 trainable parameters. All hyperparameters were optimized through grid search.

Thus, based on the SLICES crystal language, the MatterGPT framework utilizes the advanced transformer architecture, leveraging large language models to comprehensively learn the structure-property relationships between the SLICES-encoded crystal structures and their properties.

The performance of our crystal generative model is evaluated using four key metrics:

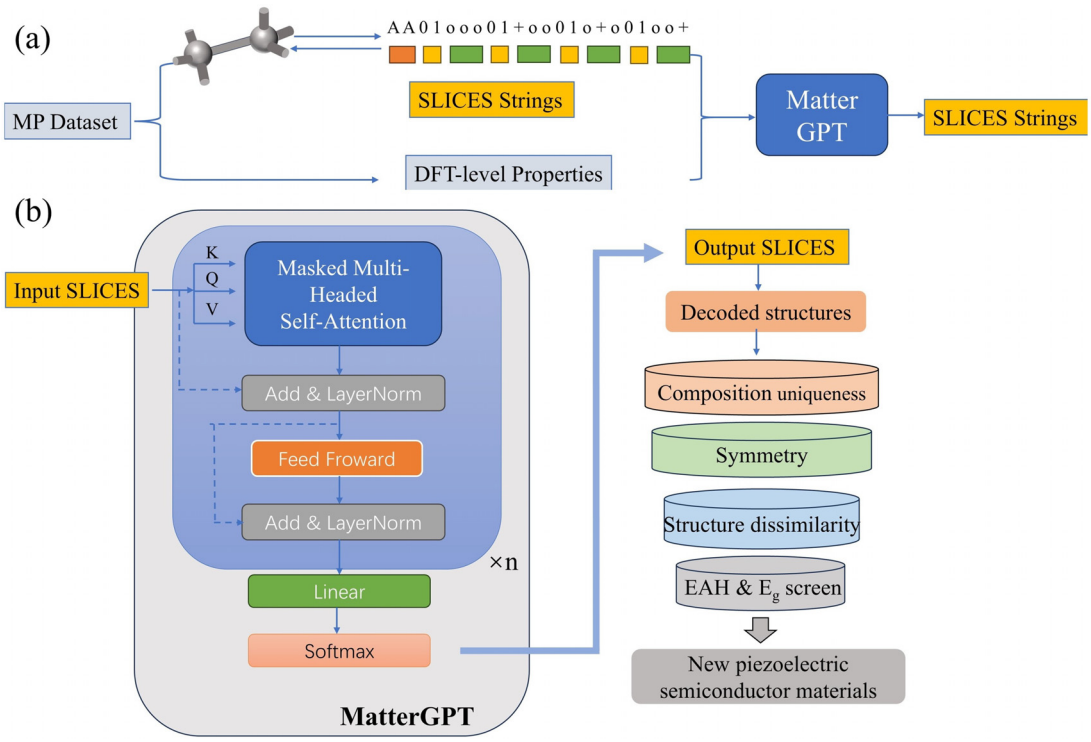

FIG. 3. The architecture of MatterGPT. (a) Pipeline for training and sampling using the conditional MatterGPT model. (b) The preprocessing of SLICES strings and model architecture utilizing an autoregressive decoder-only transformer with masked multi-head self-attention.

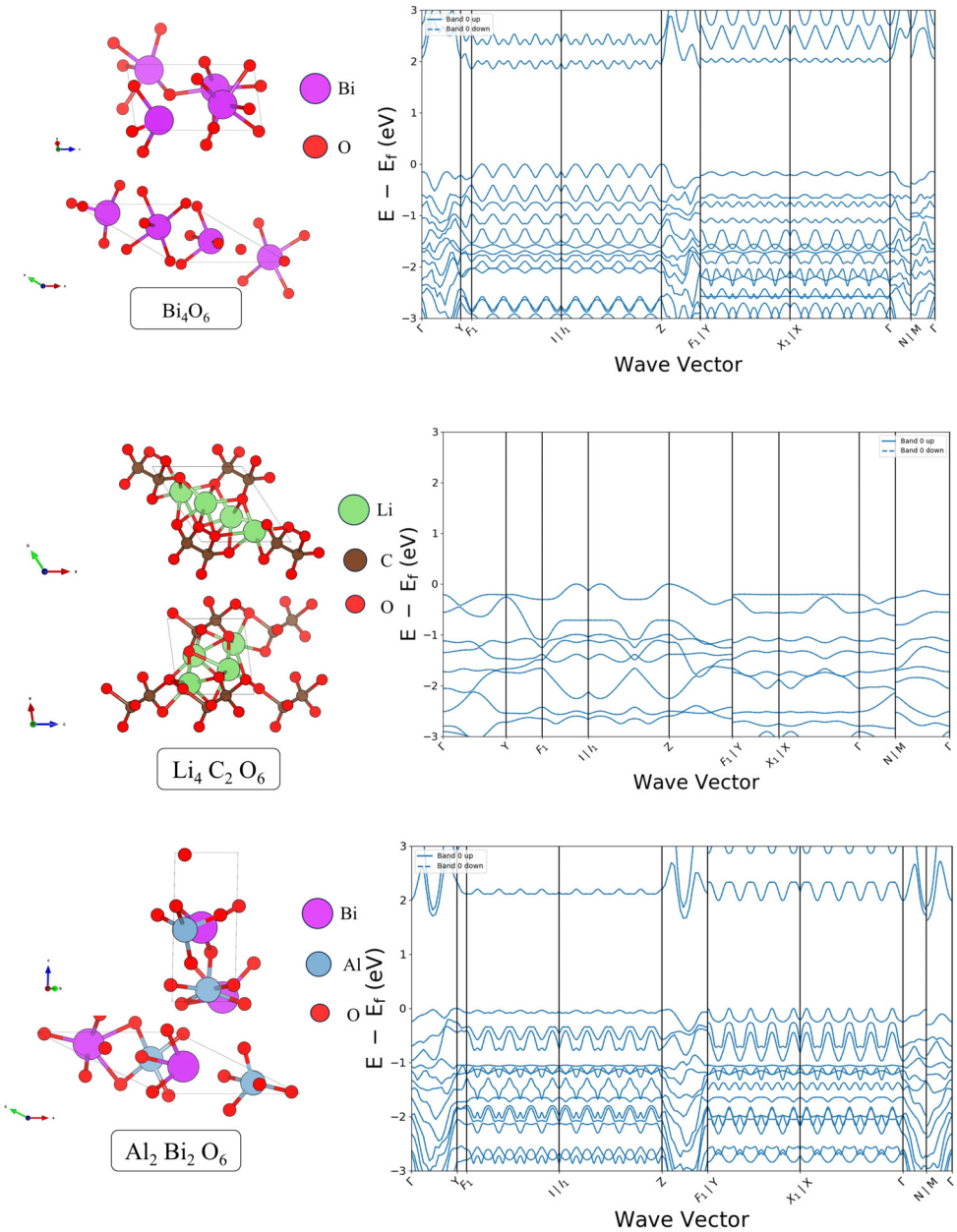

FIG. 4. Crystal structures and electronic band structures of representative generated materials.

(1) Validity: The percentage of generated SLICES strings that satisfy all grammatical rules of the representation.
(2) Uniqueness: The percentage of non-duplicate structures among valid samples, determined by converting samples into canonical SLICES strings and removing duplicates. This metric reflects the model's capability to generate diverse structures.
(3) Novelty: The percentage of unique generated structures that do not exist in the training dataset. We assess novelty by comparing the atomic arrangements between structures with identical compositions using the StructureMatcher utility in pymatgen,[31] with default parameters (ltol = 0.2, stol = 0.3, angle_tol = 5).
(4) Mean absolute percentage error (MAPE): A measure of how closely the generated structures match the target property values, calculated as

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{P_i - T}{T}\right| \times 100\%, \tag{4}$$

where n is the number of samples, P_i represents the DFT-calculated property value of the ith generated structure, and T is the target property value.

Using our trained MatterGPT model, we generated approximately 5000 SLICES strings. Of these, 4100 were reconstructed into crystal structures, while the remaining failed due to edge redundancy. The candidate structures underwent a multi-step screening process. First, we eliminated structures with compositions already present in the Materials Project database, reducing the pool to 1538 candidates. We then applied additional filtering criteria: structural dissimilarity threshold of 0.75 relative to known structures, formation energy lower than 0 eV/atom, energy above hull lower than 0.5 eV/atom, and semiconductor-like bandgaps predicted by the ALIGNN model. This screening process yielded 28 promising candidates.

The inverse design calculations were performed on a workstation equipped with an Intel I9-9960 processor (16 × 2 cores, 3.1 GHz) and an NVIDIA RTX2080Ti GPU. The model's performance metrics are summarized in Table I.

Following DFT structural relaxation and band structure calculations at the PBE level, we identified several potential piezoelectric semiconductor materials that satisfied our design criteria (Fig. 3). Most notably, Bi_4O_6 exhibits a significant piezoelectric response with a stress coefficient of 25.9 C/m^2 in the e[1, 6] component, corresponding to shear deformation in the yz plane coupled to polarization along the x_1 direction. This material also possesses a bandgap of 1.85 eV, well within the semiconductor range. The calculated band structure, referenced to the Fermi level, shows a clear bandgap with no electronic states, confirming its semiconducting nature (Fig. 4).

We also discovered two other promising candidates. $Li_4C_2O_6$ shows a wide bandgap of 3.61 eV and a substantial piezoelectric response of −5.05 C/m^2 in the e[3,6] direction, making it particularly interesting for high-temperature electronic applications. Additionally, $Al_2Bi_2O_6$ demonstrates potential with a bandgap of 1.63 eV and a piezoelectric modulus of −3.07 C/m^2 in the e[2] direction.

Structural similarity analysis was performed using the StructureMatcher algorithm implemented in pymatgen,[31] comparing these structures with existing Bi_2O_3, Li_2CO_3, and $AlBiO_3$ compounds documented in the Materials Project database. The default parameters used in the StructureMatcher utility are: lattice tolerance (ltol) = 0.2, site tolerance (stol) = 0.3, and angle tolerance (angle_tol) = 5. These default settings are sufficient for structural similarity comparison.

With tolerance parameters, no structural matches were identified between our compounds and the previously reported materials of identical compositions. These findings indicate that these generated materials represent crystal structures not previously documented in the Materials Project database.

The discovered piezoelectric semiconductor materials exhibit exceptional performance, with piezoelectric stress coefficients up to 25.9 C/m^2. To explore additional properties, we analyzed the Born effective charges (BEC) that describe the piezoelectric performance of the crystal structure using density functional perturbation theory (DFPT), as in Fig. 5. For the Born effective charge (BEC) tensor of different ions, the BEC value of the Bi ion (ion 1) is relatively high (up to 5.79 e), indicating a large displacement under an external electric field. Additionally, the charge distribution is asymmetric, suggesting the lack of inversion symmetry, which is a necessary condition for piezoelectricity. The differences in BEC values between different ions (such as the charge variations of ions 2 and 3 in different directions) facilitate the formation of larger dipoles, further enhancing the piezoelectric response of the material. Moreover, the non-zero BEC values for all ions indicate that the piezoelectricity of this material is a bulk property, rather than a localized effect, supporting its potential as an efficient piezoelectric material.

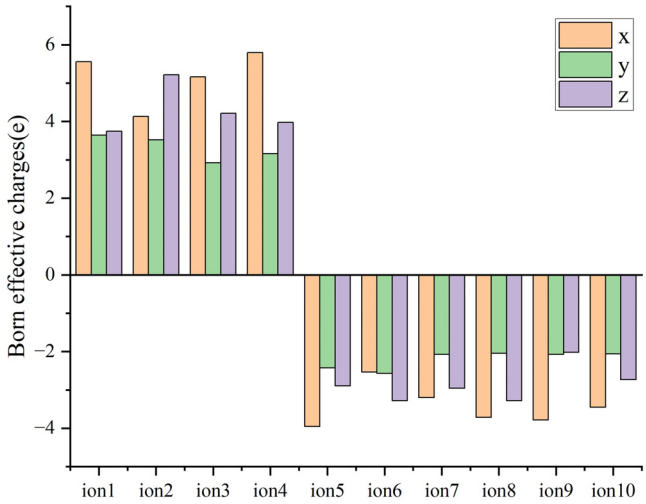

FIG. 5. Born effective charges (e) for Bi_4O_6 Ions in different directions.

Meanwhile, in order to further validate the stability of the structure, we employed the finite displacement method based on first-principles calculations and used the atomate2 forcefields framework to compute the phonon spectrum of Bi_4O_6, with phonopy utilized to derive the phonon frequencies.[34] The calculation results show that no imaginary frequencies are present throughout the Brillouin zone, indicating that this structure is dynamically stable. This result further confirms the stability of Bi_4O_6 and provides in-depth physical insight into its vibrational properties. The corresponding results are shown in Fig. 6.

These materials are particularly promising for applications in high-sensitivity sensors, where their ability to efficiently convert mechanical stress into electrical signals enables precise detection of

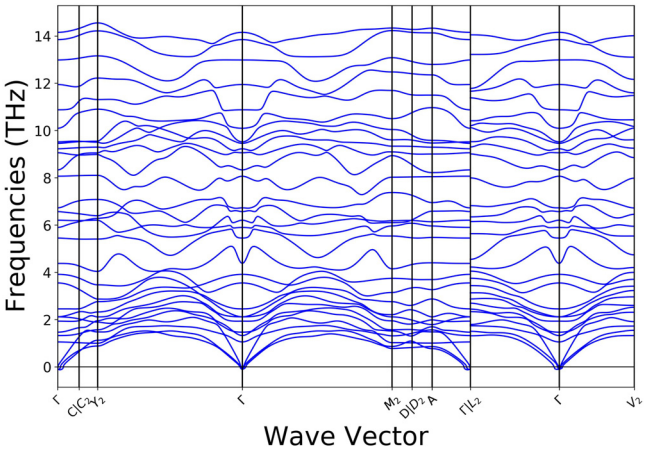

FIG. 6. Phonon band structure of Bi_4O_6.

both static pressure changes and dynamic acoustic waves. Their combined piezoelectric and semiconducting properties make them ideal candidates for next-generation devices in precision medicine, industrial monitoring, and acoustic sensing applications that demand high sensitivity and signal fidelity.

The integration of SLICES representation with transformer-based models represents a significant advance in computational materials discovery. By leveraging the power of large language models through the MatterGPT architecture, we have demonstrated an efficient approach to the inverse design of functional materials. The model's ability to learn complex structure–property relationships from data enables targeted generation of materials with desired properties, offering possibilities for materials discovery and optimization.

While our computational approach shows promise, important challenges remain in experimental realization of the predicted materials. These include controlling synthesis conditions, ensuring phase purity, and optimizing processing parameters. Furthermore, device integration of these potential materials requires addressing additional challenges in process compatibility, performance optimization, and scalable manufacturing. Bridging the gap between computational prediction and experimental implementation remains a critical direction for future research.

The inverse design of crystalline materials with targeted properties remains a fundamental challenge in materials science. Although our SLICES-based approach demonstrates significant progress, its potential is currently limited by the scarcity of training data. Unlike molecular design, where SMILES-based models can leverage millions of structures, crystalline materials databases such as the Materials Project, JARVIS, and Alexandria contain only tens of thousands of entries, with even fewer samples having specific property measurements. This data scarcity highlights the urgent need for expanded materials databases to advance inverse design capabilities.

Using the SLICES-based MatterGPT framework, we report an effective computational strategy for discovering piezoelectric semiconductor materials. Through this property-guided generation approach, we identified previously unrecorded materials with piezoelectric stress coefficients up to 25.9 C/m^2 in the e[1, 6] direction, showing promising potential for high-sensitivity sensing applications. In particular, we discovered three promising candidates: Bi_4O_6 with excellent piezoelectric response, $Li_4C_2O_6$ with a wide bandgap suitable for high-temperature applications, and $Al_2Bi_2O_6$ showing balanced electronic and piezoelectric properties. The development in identifying these materials with targeted properties demonstrates the capability of advanced crystal representation and generative models in accelerating functional materials discovery.

This research was supported by the National Key R&D Program of China (Nos. 2021YFB3201700 and 2021YFB3201705).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Chen Zhang: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal). **Siyuan Lv:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Writing – original draft (equal). **Haojie Gong:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Software (equal). **Qianxi Cheng:** Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Software (equal). **Junwei Guo:** Conceptualization (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal). **Zheng Duanmu:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Project administration (equal); Resources (equal); Supervision (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Hang Xiao:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (lead); Project administration (equal); Resources (equal); Software (lead); Supervision (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

REFERENCES

[1]Y. Chen, J. Liu, J. Yu, Y. Guo, and Q. Sun, "Symmetry-breaking induced large piezoelectricity in Janus tellurene materials," Phys. Chem. Chem. Phys. **21**, 1207–1216 (2019).
[2]R. Shao, R. Ma, X. An, C. Wang, and S. Sun, "Challenges and emerging opportunities in transistor-based ultrathin electronics: Design and fabrication for healthcare applications," J. Mater. Chem. C **10**, 2450–2474 (2022).
[3]Z. Xiao, W. Liu, S. Xu, J. Zhou, Z. Ren, and C. Lee, "Recent progress in silicon-based photonic integrated circuits and emerging applications," Adv. Opt. Mater. **11**, 2301028 (2023).
[4]K. Jeronimo, V. Koutsos, R. Cheung, and E. Mastropaolo, "PDMS-ZnO piezoelectric nanocomposites for pressure sensors," Sensors **21**, 5873 (2021).
[5]Z. Jihong, Z. H. W. Chuang, Z. L. Y. Shangqin, and W. Zhang, "A review of topology optimization for additive manufacturing: Status and challenges," Chin. J. Aeronaut. **34**, 91–110 (2021).
[6]A. S. Fuhr and B. G. Sumpter, "Deep generative models for materials discovery and machine learning-accelerated innovation," Front. Mater. **9**, 865270 (2022).
[7]S. Liu and C. Yang, "Machine learning design for high-entropy alloys: Models and algorithms," Metals **14**, 235 (2024).
[8]X. Jinghua, F. Xueqing, C. Jun, and Z. Shuyou, "Precision forward design for 3D printing using kinematic sensitivity via Jacobian matrix considering uncertainty," Int. J. Adv. Manuf. Technol. **110**, 3257–3271 (2020).
[9]M. Ashby, "Multi-objective optimization in material design and selection," Acta Mater. **48**, 359–369 (2000).
[10]K. Alberi, M. B. Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M. L. Green *et al.*, "The 2019 materials by design roadmap," J. Phys. D **52**, 013001 (2019).
[11]Y. Meng, G. Chen, and M. Huang, "Piezoelectric materials: Properties, advancements, and design strategies for high-temperature applications," Nanomaterials **12**, 1171 (2022).
[12]M. Dijkstra and E. Luijten, "From predictive modelling to machine learning and reverse engineering of colloidal self-assembly," Nat. Mater. **20**, 762–773 (2021).
[13]J. Noh, G. H. Gu, S. Kim, and Y. Jung, "Machine-enabled inverse design of inorganic solid materials: Promises and challenges," Chem. Sci. **11**, 4871–4881 (2020).
[14]V. Fung, J. Zhang, G. Hu, P. Ganesh, and B. G. Sumpter, "Inverse design of two-dimensional materials with invertible neural networks," npj Comput. Mater. **7**, 200 (2021).

[15]T. Long, Y. Zhang, N. M. Fortunato, C. Shen, M. Dai, and H. Zhang, "Inverse design of crystal structures for multicomponent systems," Acta Mater. **231**, 117898 (2022).
[16]L. Chen, W. Zhang, Z. Nie, S. Li, and F. Pan, "Generative models for inverse design of inorganic solid materials," J. Mater. Inf. **1**, 4 (2021).
[17]D. P. Kingma, "Auto-encoding variational Bayes," arXiv:1312.6114 (2013).
[18]I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems 27 (2014).
[19]J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in neural information processing systems 33* (Neural Information Processing Systems (NeurIPS), 2020), pp. 6840–6851, see https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.
[20]T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**, 145301 (2018).
[21]J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, and Y. Jung, "Inverse design of solid-state materials via a continuous representation," Matter **1**, 1370–1384 (2019).
[22]C. Qin, J. Liu, S. Ma, J. Du, G. Jiang, and L. Zhao, "Inverse design of semiconductor materials with deep generative models," J. Mater. Chem. A **12**, 22689–22702 (2024).
[23]H. Xiao, R. Li, X. Shi, Y. Chen, L. Zhu, L. Wang, and X. Chen, "An invertible, invariant crystallographic representation for inverse design of solid-state materials using generative deep learning," ChemRxiv (2023).
[24]D. Weininger, "SMILES, A chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).
[25]Y. Chen, X. Wang, X. Deng, Y. Liu, X. Chen, Y. Zhang, L. Wang, and H. Xiao, "MatterGPT: A generative transformer for multi-property inverse design of solid-state materials," arXiv:2408.07608 (2024).
[26]S. Baroni, P. Giannozzi, and A. Testa, "Green's-function approach to linear response in solids," Phys. Rev. Lett. **58**, 1861 (1987).
[27]J. F. Nye, *Physical Properties of Crystals: Their Representation by Tensors and Matrices* (Oxford University Press, 1985).
[28]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).
[29]J.-G. Eon, "Euclidian embeddings of periodic nets: Definition of a topologically induced complete set of geometric descriptors for crystal structures," Acta Crystallogr. A **67**, 68–86 (2011).
[30]S. G. Jung, G. Jung, and J. M. Cole, "Automatic prediction of band gaps of inorganic materials using a gradient boosted and statistical feature selection workflow," J. Chem. Inf. Model. **64**, 1187–1200 (2024).
[31]S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," Comput. Mater. Sci. **68**, 314–319 (2013).
[32]N. E. Zimmermann and A. Jain, "Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity," RSC Adv. **10**, 6063–6081 (2020).
[33]G. Hautier, A. Miglio, D. Waroquiers, G.-M. Rignanese, and X. Gonze, "How does chemistry influence electron effective mass in oxides? A high-throughput computational analysis," Chem. Mater. **26**, 5447–5458 (2014).
[34]A. M. Ganose, H. Sahasrabuddhe, M. Asta, K. Beck, T. Biswas, A. Bonkowski, J. Bustamante, X. Chen, Y. Chiang, D. Chrzan, J. Clary, O. Cohen, C. Ertural, M. Gallant, J. George, S. Gerits, R. Goodall, R. Guha, G. Hautier, M. Horton, A. Kaplan, R. Kingsbury, M. Kuner, B. Li, X. Linn, M. McDermott, R. S. Mohanakrishnan, A. Naik, J. Neaton, K. Persson, G. Petretto, T. Purcell, F. Ricci, B. Rich, J. Riebesell, G.-M. Rignanese, A. Rosen, M. Scheffler, J. Schmidt, J.-X. Shen, A. Sobolev, R. Sundararaman, C. Tezak, V. Trinquet, J. Varley, D. Vigil-Fowler, D. Wang, D. Waroquiers, M. Wen, H. Yang, H. Zheng, J. Zheng, Z. Zhu, and A. Jain, "Atomate2: Modular workflows for materials science," ChemRxiv (2025).