npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Machine learning for deep elastic strain engineering of semiconductor electronic band structure and effective mass

Evgenii Tsymbalov[1,4], Zhe Shi[2,4], Ming Dao[2,3], Subra Suresh[3]✉, Ju Li[2]✉ and Alexander Shapeev[1]✉

The controlled introduction of elastic strains is an appealing strategy for modulating the physical properties of semiconductor materials. With the recent discovery of large elastic deformation in nanoscale specimens as diverse as silicon and diamond, employing this strategy to improve device performance necessitates first-principles computations of the fundamental electronic band structure and target figures-of-merit, through the design of an optimal straining pathway. Such simulations, however, call for approaches that combine deep learning algorithms and physics of deformation with band structure calculations to custom-design electronic and optical properties. Motivated by this challenge, we present here details of a machine learning framework involving convolutional neural networks to represent the topology and curvature of band structures in **k**-space. These calculations enable us to identify ways in which the physical properties can be altered through "deep" elastic strain engineering up to a large fraction of the ideal strain. Algorithms capable of active learning and informed by the underlying physics were presented here for predicting the bandgap and the band structure. By training a surrogate model with ab initio computational data, our method can identify the most efficient strain energy pathway to realize physical property changes. The power of this method is further demonstrated with results from the prediction of strain states that influence the effective electron mass. We illustrate the applications of the method with specific results for diamonds, although the general deep learning technique presented here is potentially useful for optimizing the physical properties of a wide variety of semiconductor materials.

npj Computational Materials (2021)7:76 ; https://doi.org/10.1038/s41524-021-00538-0

INTRODUCTION

Elastic strain engineering (ESE) has emerged as a promising tool to enhance the performance of functional materials, whereby characteristics of semiconductor materials, such as carrier mobility, can be modulated solely through the introduction of strain[1]. With Moore's law approaching its widely anticipated limit and with an ever-accelerating search for improved device performance, tuning physical properties through controlled mechanical strains could offer a powerful pathway to advance the performance of semiconductors. Here, the magnitudes of elastic strains considered are significantly greater (deeper) than prior approaches adopted over the past several decades by the semiconductor industry, involving strained silicon with strain levels on the order of 1%[2–4].

To achieve "deep ESE" with significant performance enhancement in devices and to realize the optimal figure-of-merit (FoM), the required elastic strain state would have to significantly exceed what strained silicon technology has thus far achieved through epitaxy. Indeed, recent experiments[5–7] in free-standing geometry have revealed that several materials, at nanoscale dimensions typically used in semiconductor devices, are capable of withstanding large elastic strains at room temperature without inelastic shear relaxation, phase transformation, or fracture. For example, it has been demonstrated that even in the hardest material found in nature, diamond, the local tensile elastic strain can reach up to 10% in appropriately grown and oriented single-crystal nanoneedles[5,6] and microfabricated nanobridges[7]. In nanowires of silicon, a reversible tensile strain of 15% has been realized in uniaxial tension experiments[8]. These findings of ultra-large elastic deformation of semiconductor materials bookended by the ultra-wide bandgap (5.6 eV) diamond and the more manufacturing-friendly and ubiquitous silicon (with a bandgap of 1.1 eV) have opened up potential opportunities to design their performance characteristics for applications such as power electronics, nanophotonics, and quantum information processing.

The complexity of modulating the fundamental physical properties of materials, such as the electronic band structure and bandgap through ESE, calls for identifying preferred and actionable strain states within the general six-dimensional (6D) strain space, represented by the elastic strain tensor $\boldsymbol{\varepsilon} \equiv (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{33}, \varepsilon_{23}, \varepsilon_{13}, \varepsilon_{12})$. In order to achieve this through rigorous physics-informed computational predictions, first-principles calculations based on density functional theory (DFT) are necessary to screen for relevant properties and characteristics to realize targeted FoM. Appropriate mechanical boundary conditions would then have to be designed to impose optimal strain states at the targeted spatial regions at the device level. To picture the complexity of this task, consider for example a crystal's electronic band structure $E_n(\mathbf{k}; \boldsymbol{\varepsilon})$, which is a function of the 3-dimensional wave vector **k** and the six-dimensional homogeneous strain tensor $\boldsymbol{\varepsilon}$. Representing these band structures with a tabulation approach with 9 dependent variables would obviously require billions of first-principles calculations. Plain DFT calculations can also introduce systematic errors in the estimation of bandgap, and these errors have to be overcome with the extra computational cost incurred by employing more expensive calculation techniques such as many-body perturbation theory (so-called GW corrections[9]). Additional costs with representing and storing of ESE effects would arise when the elastic strain gradient $\nabla\boldsymbol{\varepsilon}$ is incorporated as a dependent variable (since $\nabla\boldsymbol{\varepsilon}$ can be on the order of 10^7/m in nanoscale devices),

[1]Skolkovo Institute of Science and Technology, Moscow, Russia. [2]Department of Materials Science and Engineering and Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Nanyang Technological University, Singapore, Republic of Singapore. [4]These authors contributed equally: Evgenii Tsymbalov, Zhe Shi. ✉email: SSuresh@ntu.edu.sg; liju@mit.edu; A.Shapeev@skoltech.ru

which could influence properties such as nonlinear optical response induced by the flexoelectric effect[10,11].

The modulation of bandgap and carrier mobility (which characterizes electrical conductivity based on the speed with which an electron or a hole moves through a semiconductor under the influence of an electric field) has long been examined using linear response perturbation theory[12]. While this approach is sufficiently powerful to guide the engineering of strained silicon under small strains, it loses validity for large deformation cases involving "deep" or nonlinear elastic strains so as to significantly enhance their electronic or optoelectronic performance characteristics[13]. To address the challenge of calculating, analyzing, archiving, and visualizing material characteristics from high-dimensional data, such as that related to $E_n(\mathbf{k}; \boldsymbol{\varepsilon})$, recent work has adopted a data fusion and transfer learning technique to integrate multiple data sources[13]. Specifically, we developed a machine learning (ML) method[13] that combined a dataset extracted from DFT calculations invoking GW correction with another dataset prepared with Perdew-Burke-Ernzerhof[14] (PBE) exchange functional correlations. This method was adopted to demonstrate how large elastic strains can be used to make a semiconductor such as silicon and an electrical insulator such as diamond become metal-like conductors with zero bandgap[15].

While our earlier deep ESE calculations are adequate for rapid data collection in a highly specialized model[13,15], they do not offer sufficient flexibility and accuracy for optimizing a broader consideration of physical characteristics such as the effective mass of electrons and holes, which is a second-derivative of $E_n(\mathbf{k}; \boldsymbol{\varepsilon})$ with respect to $\mathbf{k}$ and a strong sensitivity to noise. Therefore, it is appropriate at this stage of development of ML to incorporate a priori physics-informed neural network (NN) architectures into the calculations in such a way that various performance characteristics and FoM estimates could be much better optimized through a judicious combination of DFT and deep learning. Further details of motivation for this development are articulated in the succeeding sections. These recent advances enable multi-property optimization and Pareto-front type tradeoff analysis.

To accomplish these goals, we present here a physics-informed, convolutional neural network (CNN) technique that is more versatile, accurate, and efficient in its capability to facilitate autonomous deep learning of the electronic band structure of crystalline solids than the neural network architecture hitherto employed to address this class of problems. We propose more advanced algorithms and data representation schemes to provide markedly improved ML outcomes. The techniques described here enable detailed analysis of band structures in the general six-dimensional strain space to optimize select FoM of interest for specific performance targets. Moreover, our method achieves sufficient accuracy not only for the deep analysis of the bandgap and of the shape of the band structure, but also for capturing the curvature of the band and the effective mass.

RESULTS

Band structure and physics-informed ML

Our method seeks to develop accurate predictions of the band structure, by recourse to which many FoMs of technological interest could be directly estimated. Inspired by the wide adoption of deep learning in the field of computer vision[16], we draw an analogy between the color spectrum in a digital image and the band structure, regardless of whether it applies to electronic, phononic, or photonic band structure.

Using this analogy, we envision energy dispersions as stacked 3D "images", with the reciprocal coordinates $\mathbf{k} \equiv (k_1, k_2, k_3)$ representing the "voxels" (i.e., 3D "pixels" of a digital image) and with E_n denoting the spectrum and intensity of colors (similar to the RGB or grayscale of an image) at each voxel for a particular 3D image, where n is the particular band among a total of N bands. Energy bands are piecewise-smooth functions in the reciprocal space, and the information within the energy dispersion of a specific band includes *intraband* correlations with respect to $\mathbf{k}$. An illustration of this pictorial view of the band structure can be found in Fig. 1a. Note that previous ML schemes based on simple feed-forward NN treated an energy band as a flattened array of independent values[13]—thereby neglecting to account for intra-band correlation.

In prior work[13], different bands were analyzed separately by NN (Fig. 1a, b). Although this approach was sufficient to predict energy eigenvalues for a specific band or bandgap variations arising from strain, it could not capture interband physics accurately for the entire band structure because of limited data. The energy bands analyzed in the present method, however, are not "independent" of one another, as shown in Fig. 1, and they collectively describe the physical characteristics of the crystal. For example, consider a single electron in a periodic potential resulting from the interaction of the electron with the ions and other electrons. Solving the Schrödinger equation provides the solution for a series of Bloch waves, each of which has a predicted dispersive form. Through the first-principles method, all the quantized energy levels are determined. Specifically, the nth band is not calculated in isolation but is determined from the collective influence of its neighboring bands, including the adjacent $(n-1)$ th and $(n+1)$th bands, as well as other non-adjacent bands. In other words, information from *interband* correlation influencing the nth band is included in the band structure of the crystal.

To reveal the internal structure of the band data in our model, we incorporate CNN into our ML scheme. CNN is known for its capability to extract hierarchical patterns in digital images and to assemble complex patterns by integrating information from smaller datasets[17]. Utilizing the digital image analogy for the band structure, CNN is thus expected to serve as a useful tool for extracting useful patterns, or intraband/interband correlations.

Model description

The general setup of the proposed model is illustrated in Fig. 2a. It consists of a fully connected part followed by a CNN part. At the outset, the strain tensor $\boldsymbol{\varepsilon}$ is taken as the input and transformed into a feature vector through a series of fully connected layers, as depicted in Fig. 2a. This feature vector has a length of Nm^3, where m^3 is the number of $\mathbf{k}$-points sampled in the Brillouin zone, and N is the number of bands we want to represent. Depending on the $\mathbf{k}$-mesh density, the feature vector can be adopted as a rich representation of the intraband information for a band structure. Currently, this part has four hidden layers with a structure of $(6 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 512)_n$, where $512 = m^3$, for $n = 1, 2, \ldots, N$ separately, totaling ~1.1 million parameters. N is most often taken to be 4 in this work, sufficient for describing near-CBM/near-VBM properties of diamond for a particular strain state. Here, the band energy dispersion for the top valence band (VB, $n = n_{VB}$), the lowest conduction band (CB, $n = n_{CB}$), and their adjacent two bands ($n = n_{VB} - 1$ and $n = n_{CB} + 1$) could all be represented via 4 vectors each of which has a length of m^3. Stacking them together, we build an $m \times m \times m \times 4$ tensor representation of the band structure for any individual strain data, as illustrated in Fig. 2b. This process is similar to the decoding part of an autoencoder[18] whereby a representation as close as possible to the band structure is generated. The resulting tensor is then fed into the next block of convolution.

The convolutional part consists of several blocks that update this tensor representation until the final output is determined. Note that the output tensor retains the same dimension of the band structure, i.e., $m \times m \times m \times N$. This extraction process proceeds through many layers to deliver a band structure tensor

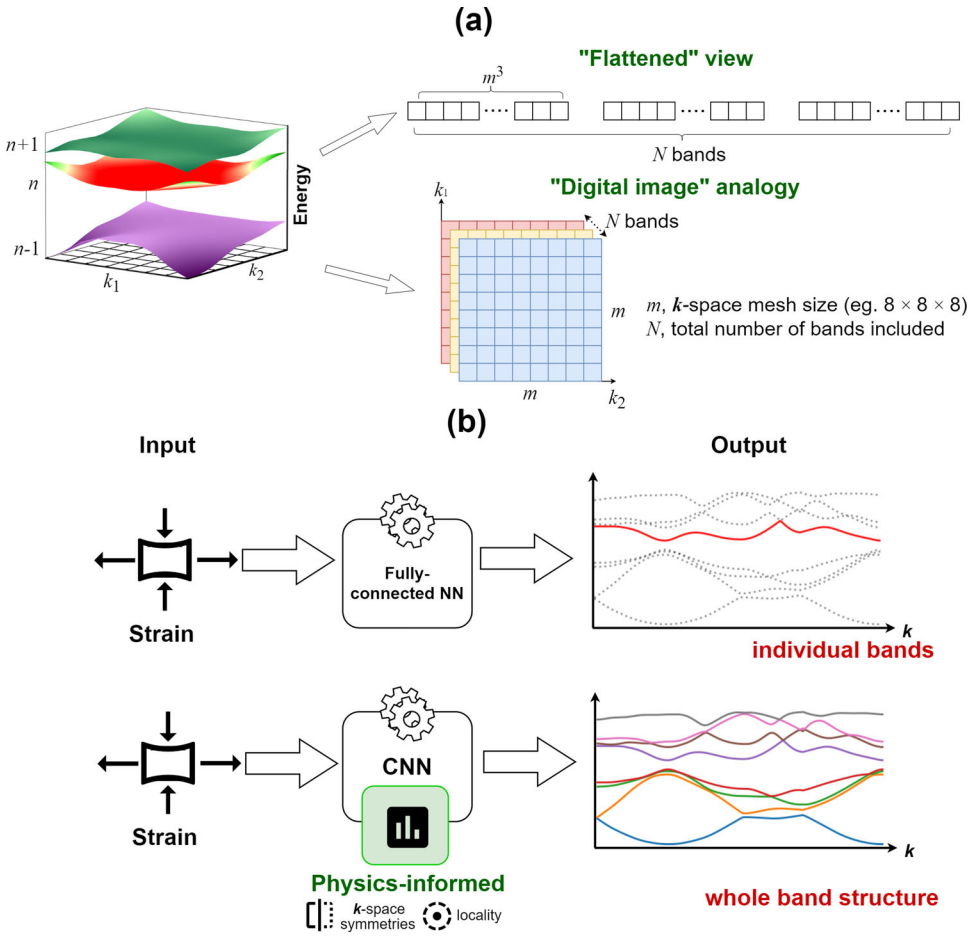

Fig. 1 Two different views of the band structure. a Different representation of a band structure. In the "flattened" view, a band structure is represented as N stacked flattened arrays (vectors) and processed like independent values. Each array is m^3 in length. In the "digital image" analogy, the band structure is envisioned as N different 3D images stacked together, each of which has a "voxel" dimension of $m \times m \times m$. The eigenvalues on an energy band can then be thought of as the "color-scale" of the voxels. **b** Comparison of the two different approaches to ML. We predict the eigenvalues for each energy band separately by utilizing the "flattened" band structure representation to obtain the entire band structure.

with features that capture deep intra-band and inter-band information. This output comprising the complete ML inference represents the band structure obtained by DFT calculations (Fig. 2a–b). In each convolutional block in the CNN part, the convolution is a two-step sequence. In the first step, a $3 \times 3 \times 3 \times 1$ kernel accounting for the intraband correlation (with periodical boundary conditions and symmetry) is used. In the second step, a $1 \times 1 \times 1 \times 3$ kernel accounting for the interband correlation is adopted. The convolution blocks can be stacked up at one's discretion. The model yielding the lowest error in our study has three CNN blocks, totaling ~276,000 parameters. One can also use a one-step convolution ($3 \times 3 \times 3 \times 3$) kernel instead of the aforementioned two-step convolution ($3 \times 3 \times 3 \times 1$)→convolution ($1 \times 1 \times 1 \times 3$) kernel, with more weights per block but better accuracy. Also, since $8 \times 8 \times 8$ is still a relatively coarse **k**-mesh when performing $\min_{\mathbf{k}}$, $\max_{\mathbf{k}}$ or **k**-derivative operations, we use polynomial interpolation on top of the floating-point $8 \times 8 \times 8$ representation, before carrying out such operations.

The power of this approach lies in the architecture of the proposed CNN model, which is tailored to the known physical structure and exploratory data analysis results (Fig. 2b and Supplementary Figs 1–2) in order to simplify training and to speed up inference. In particular, it takes advantage of:

i. *The time-reversal symmetry*, i.e., $E_n(-\mathbf{k}) = E_n(\mathbf{k})$ which holds for the diamond crystal. Corresponding tensor representation preserves this property.
ii. *The correlation between the same* **k***-point of different bands* (interband correlation). An interband convolution between the bands is applied at each **k**-point so that bands influence one another.
iii. *The correlation between the energy eigenvalues associated with adjacent* **k***-points of the same band* (intraband correlation), which ascertains that the band energy is a piecewise-smooth function of the **k**-space coordinates. The intraband convolutions are carried out over several cycles so that the underlying physics of how energy eigenvalues from adjacent **k**-points affecting one another are learned accurately.
iv. *Band structure calculations benefit from the periodic nature and symmetry of a crystal lattice*. The band structure plot resulting from restricting **k** to the first Brillouin zone, also known as the reduced zone scheme, is typically used. This reciprocal lattice periodicity is represented in our model using a special technique for the periodic boundary condition that follows the reduced zone scheme.

Model training

The training of our model is achieved in three parts: preliminary training, data fusion, and active learning. In the first part, preliminary training was performed on the large dataset (~35,000 strain samples) of the computationally cheap DFT-PBE

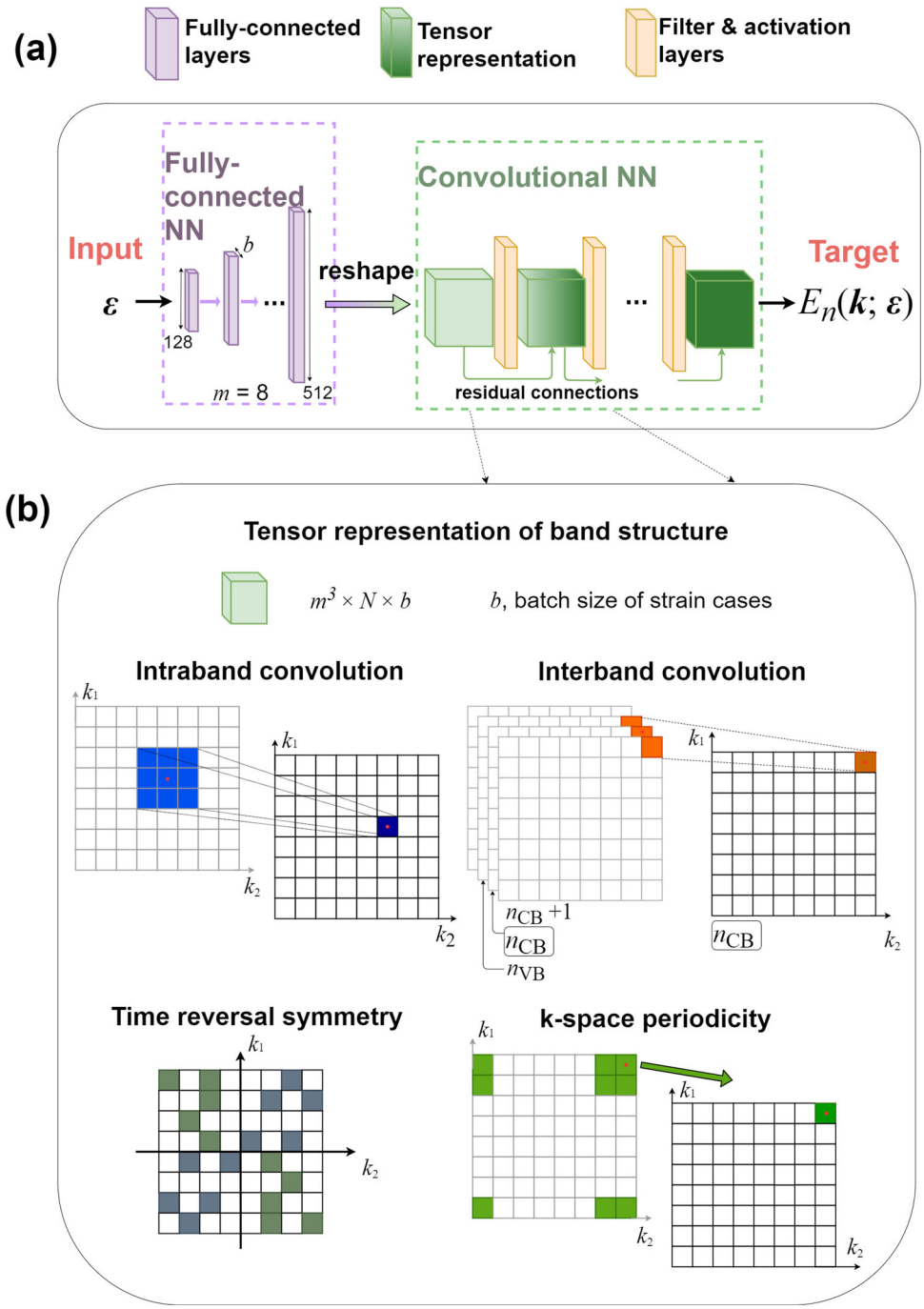

Fig. 2 ML model description. a CNN architecture for band structure prediction. The strain components are passed through fully connected layers, with the last layer reshaped into a rank-5 tensor. After a few convolutional layers with residual connections that improve convergence, the network produces the band structure as the output, which is fitted against the targeted DFT-computed band structure. A mesh comprising $8 \times 8 \times 8$ **k**-points are used. **b** Tensor representation and physical insights incorporated into the CNN model: time-reversal symmetry, **k**-space periodicity, and inter-band and intra-band convolution.

calculations. After a prescribed level of accuracy (less than 0.5% relative error) was achieved, in the second part, we performed training on a much smaller set (~6000 strain-samples) of the accurate many-body GW calculation, starting from the NN parameters learned in the previous stage. This approach is known as knowledge transfer, as some of the knowledge gathered by NN from the low-fidelity PBE data is exploited to ease the training on the relatively more costly and reliable GW data. See Fig. 3a for a schematic of this process and the "Methods" section for computational details.

Active learning

Another integral part of our training is active learning, which entails a class of ML algorithms for the automatic assembly of the training set. Here the goal is to reduce the uncertainty compared to that generated in a random sampling of strains. It is often convenient to begin with subsets of the data that offer uncertain levels of reliability and accuracy. Various uncertainty estimates have been proposed[19]. The particular choice of an uncertainty quantification procedure greatly influences performance in the active learning part.

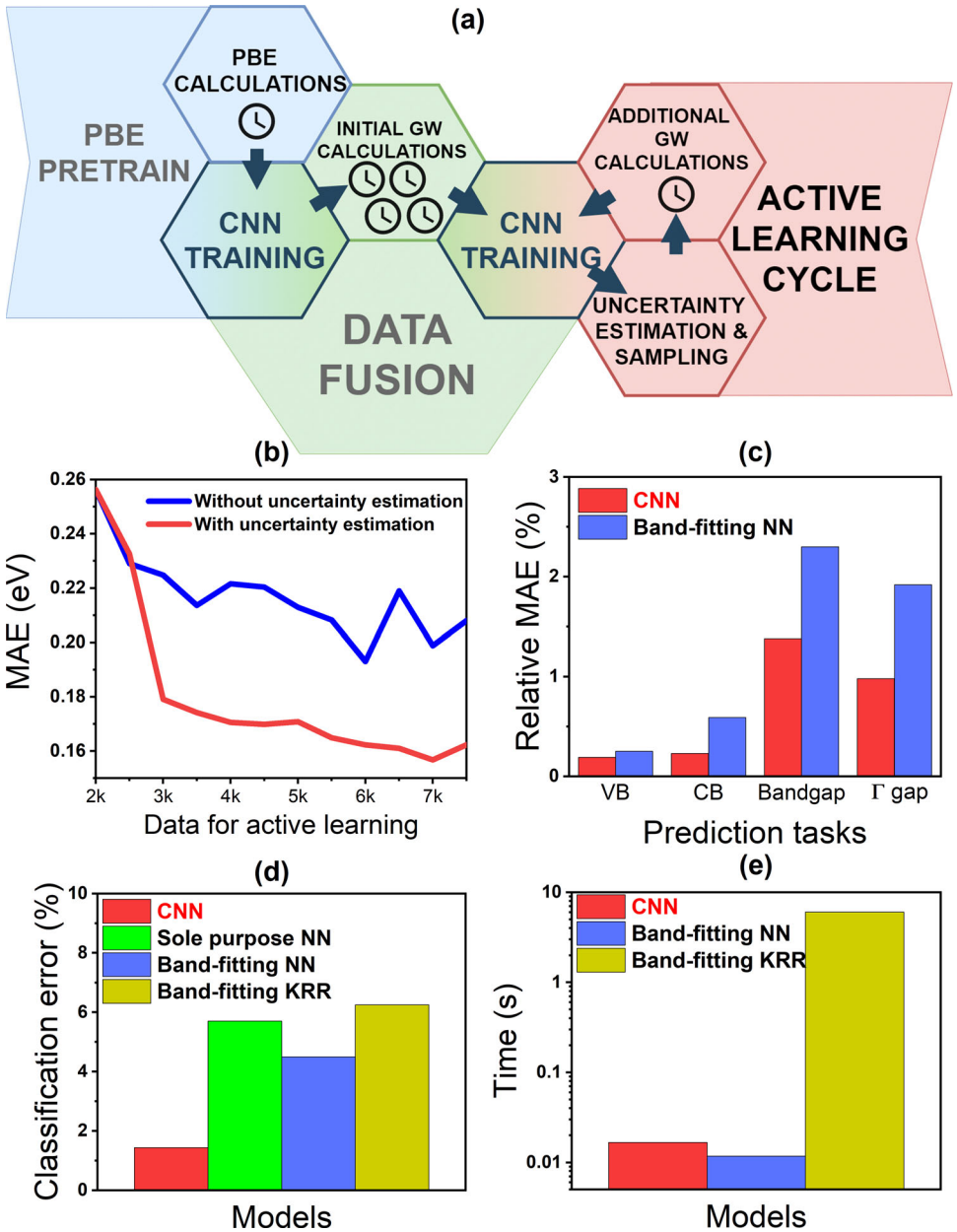

Fig. 3 ML accuracy and comparison of the different ML models. a The entire ML scheme involves pre-train, data fusion, and active learning. The solid arrows show the workflow, and clock symbols indicate the relative time required for ab initio calculations. **b** Steady improvement of model performance in terms of mean absolute error (MAE) during active learning with and without uncertainty estimation on PBE data. **c** Physics-informed CNN holds significant advantages over band-fitting NN while being able to accomplish prediction tasks, which the feed-forward NN and KRR do not offer. "Γ gap" is the difference between the conduction band (CB) and valence band (VB) at Γ and it usually does not coincide with E_g. **d** Accuracy of CNN and other models for CBM position classification task. **e** Inference time comparison. The CNN is much faster than its closest accuracy competitor band-fitting KRR model, providing a reasonable balance between time and accuracy capabilities.

There are three main routes to uncertainty estimation in NN: ensembling[20], variational inference, and dropout-based inference. Straightforward ensembling requires a few separate models to be trained, but it imposes additional computational costs to both training and inference procedures. On the other hand, variational inference requires the usage of Bayesian neural networks (which have probability distributions instead of real-valued weights), and they also lead to costly training and inference steps. Dropout can be seen as an intermediate solution: it can be applied in a simple way to the existing NNs with fully connected layers and also has a theoretical justification in the Bayesian framework[21].

Here, we use the dropout uncertainty estimation enhanced with the Gaussian processes for stability[22] to sample the most "uncertain" strain cases for further improvement of the model. Specifically, after the first round of the training on the GW data, we performed a calculation over a large set of random strains in 6D and chose a small amount of ~200 strain cases with the largest expected error as evaluated by this intermediate model (uncertainty measurement). These strain cases were added to the training set for the next round of training, as illustrated in Fig. 3a. Our study indicates that 5–10 cycles of the above active learning enable the trained CNN to reach the same level of accuracy with two to three times fewer data, thus considerably reducing the total amount of ab initio calculations without compromising the robustness of our ML model, see Fig. 3b. More details are provided in Supplementary Note 1 and Supplementary Fig. 3.

Model accuracy and performance

The ML framework outlined in Fig. 3a achieves high accuracy in a variety of tasks compared to existing ML methods. The CNN model outperforms previous simple feed-forward NN architecture, as well as an ensemble of kernel ridge regression (KRR) based models for band structure prediction, achieving a relative error no greater than 0.5%, as shown in Fig. 3c. The predictions of properties related to the band structure, such as the bandgap E_g (defined as the energy difference between the conduction band minimum (CBM) and valence band maximum (VBM) values), were treated in previous study[13] as an isolated ML regression problem with a direct fit to the scalar E_g and the estimation of CBM and VBM as two separate tasks with many repetitive ML runs. The present CNN model does not have this constraint. It is capable of simultaneously predicting intra-band and inter-band property/values, including E_g, CBM and VBM, and interband electron excitation and photon emission energy at every **k**-point (any vertical transition between any two bands), with a level of accuracy on par with or better than other models (Fig. 3c, Supplementary Fig. 4 and Supplementary Tables 1-2).

The current ML framework also achieves high reliability in locating the band edge **k**-points. Here, the present machinery surpasses all the other models by a significant margin, as shown in Fig. 3d for the specific case of finding the CBM position for diamond. Locating CBM is a demanding classification problem due to a large number of classes: there are seven possibilities for diamond CBM location under 6D elastic strain. Predicting the entire band becomes inevitable for wide-bandgap materials such as diamond to achieve a high classification accuracy. Thus, the present CNN model captures the subtle difference between two CB **k**-points.

The proposed framework is also shown to be sufficiently fast in terms of inference time to perform swift exploration and optimization in the 6D space of admissible strains. Though architecturally much more complex, the present model outcompetes KRR-based models by more than two orders of magnitude in computational speed, as shown in Fig. 3e. The CNN model has a time complexity comparable to simple NNs. In the next section, we discuss examples of ESE of a diamond crystal.

DISCUSSION

We now consider the optimization of band structure shape and curvature and effective electron mass of diamond at certain **k**-points. For this purpose, we explore the entire 6D strain space to identify energy-efficient pathways to metalize diamond by turning it into an electrical conductor with a 0-eV bandgap while preserving phonon stability. These results extend our deep learning analytical capabilities beyond those used previously to identify the conditions for the metallization of diamond using ESE[13,15].

Here we consider bandgap, arguably the most important band structure feature, as an example of the material property set as a target for deep ESE. The first objective is to identify the bandgap limits that can be reached by strained diamond within the phonon-stable region for ESE. We find the bandgap of the diamond can be increased to realize better performance in power electronics and optical applications. It can also be transformed to resemble the properties of any small-bandgap semiconductors and to exhibit a complete semiconductor-to-metal transition to become a metal-like electrical conductor at different strain states. The next objective is to determine the transitions between direct and indirect bandgap. Our study shows that the Γ point or the center of the Brillouin zone is associated with a direct bandgap, and it is achieved only when the proper shear strain components are imposed. Among all possible strains in the 6D strain space, the present model has identified a number of strain states that result in a direct bandgap in the diamond. These are illustrated in Fig. 4a. The present calculations explore the entire 6D strain state to identify optimal pathways for deep ESE within the full spectrum of theoretical possibilities. The power of ESE is demonstrated not only in tuning the bandgap value but also in facilitating the indirect-to-direct bandgap transition that benefits photon emission and absorption.

In ESE, there would be many possible choices of $\boldsymbol{\varepsilon}$ to reach a certain value of direct or indirect bandgap. Applications of these strain states, $\boldsymbol{\varepsilon}$'s, require different amounts of strain energy. The elastic strain energy density is defined as $h(\boldsymbol{\varepsilon}) \equiv \frac{E(\boldsymbol{\varepsilon})-E^0}{V^0}$, where V^0 is the undeformed supercell volume, E^0 and $E(\boldsymbol{\varepsilon})$ are the total energy of the undeformed and deformed supercell, respectively. The resultant distribution of available bandgap values E_g plotted against h represents the "density of states of bandgap"[13] as shown in Fig. 4a. There exist many strain states with an elastic strain energy density between 35 to 95 meV/Å^3 that can reach a direct 3 eV bandgap in the diamond. These strain states lie in the region bounded by the red dashed line. If one aims for the most energy-efficient strain case to achieve the goal, one should choose the left-most strains at a certain bandgap level. An upper-bound and lower-bound function can also be defined to describe the limits of reachable bandgap in strained diamond, as indicated by the black dotted lines in Fig. 4a. A complete ranking of the common crystal directions with respect to their role in reducing the bandgap can be found in Supplementary Note 2 and Supplementary Fig. 5. Similarly, an increase in the bandgap can be explored by following the upper-bound function (upper black dotted line in Fig. 4a). This line represents pure triaxial compression, i.e., $\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33} < 0, \varepsilon_{23} = \varepsilon_{13} = \varepsilon_{12} = 0$.

Strain cases resulting in the same value of bandgap form an isosurface[12] in the 6D space. For visualization purposes, we show only a 3D subspace by fixing three of the six strain components. Figure 4b–d illustrates the situation where only compressive and tensile normal strains are present ($\varepsilon_{23} = \varepsilon_{13} = \varepsilon_{12} = 0$). Key features of this bandgap isosurface in 3D include surfaces that are piecewise smooth ("carapaces"), ridgelines where two carapaces meet, and corners where three ridgelines meet. The multifaceted nature of the bandgap isosurface is attributed to the switch of the CBM **k**-space position. As a consequence of strain tensor and crystal symmetries, this isosurface has the following features:

- Three carapaces (the hard upper shell exoskeletons of turtles, tortoises, and crustaceans) labeled in red as Δ_1, Δ_2, and Δ_3 correspond to strain cases with the same value of indirect bandgap but different CBM positions: (0, 0.375, 0.375), (0.375, 0, 0.375), and (0.375, 0.375, 0), respectively.
- Three ridgelines labeled in green as r_1, r_2, and r_3 correspond to strain cases with relations $\varepsilon_{22} = \varepsilon_{33}$, $\varepsilon_{33} = \varepsilon_{11}$, and $\varepsilon_{11} = \varepsilon_{22}$, respectively.
- The corner μ labeled in purple is the intersection of r_1, r_2, and r_3 and is the most "tensile" hydrostatic strain point on the bandgap isosurface, i.e., $\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33}$.

The bandgap isosurface of strain cases where only shear strain components are present ($\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33} = 0$) is plotted in Fig. 4e. Besides three *different* indirect bandgap CBM positions at $X_1 : (0, 0.5, 0.5)$, $X_2 : (0.5, 0, 0.5)$, and $X_3 : (0.5, 0.5, 0)$, three-shear-strains can also give rise to direct bandgap in diamond where CBM is at the Γ point. The change from the carapace labeled X_1 to that labeled Γ thus indicates an indirect-to-direct bandgap transition in diamond (yellow arrow in Fig. 4e). The corresponding band structures for the indirect and direct bandgap are shown in Fig. 4f and g, respectively.

The effective mass of an electron is a key parameter that influences carrier mobility and electrical conductivity in semiconductor materials. If we denote the conduction band energy dispersion as $E_{n_{CB}}(\mathbf{k})$, then the corresponding effective mass tensor

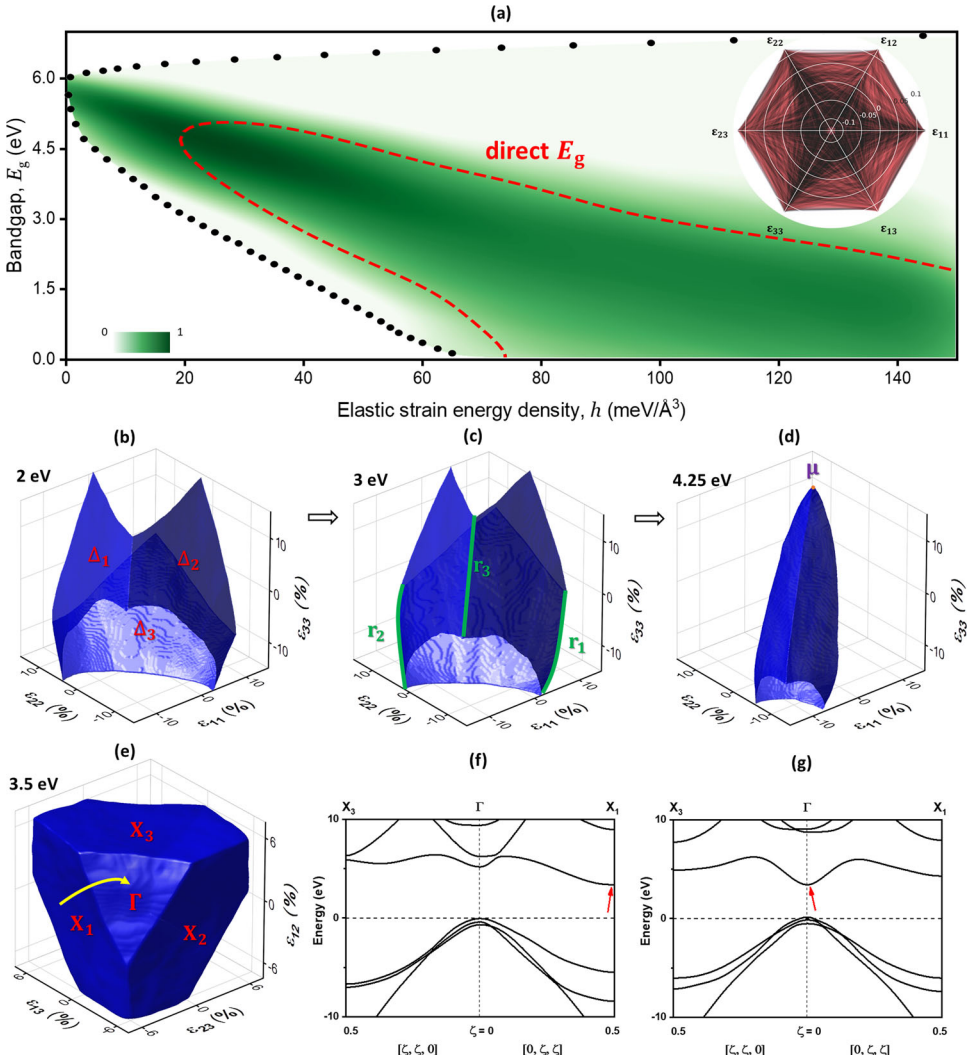

Fig. 4 Density of states of bandgap and bandgap isosurfaces. a Bandgap values achievable through ESE for various values of elastic strain energy density h within the strain space. The green shading of the region reflects the distribution of the available bandgap. The boundary of the strain region where a direct bandgap could occur is indicated by the red dashed line. The lowest h to achieve direct bandgap in diamond is about 20 meV/Å^3. Inset is the visualization of the direct bandgap strain cases in 6D. Every strain state is represented here as a hexagon with vertices on the $\varepsilon_{11}, \varepsilon_{13}, \varepsilon_{33}, \varepsilon_{23}, \varepsilon_{22}, \varepsilon_{12}$ axes. Black webs correspond to random 6D strains; brown webs correspond to the direct bandgap strains generated by our ML model. The most energy-efficient pathway to decrease the bandgap (i.e., the lower-bound function) and the upper bound of the attainable bandgap is denoted by the black dotted lines. **b**–**d** Bandgap isosurfaces in the $\varepsilon_{11}\varepsilon_{22}\varepsilon_{33}$ (normal only) strain space at 2 eV, 3 eV, and 4.25 eV levels, respectively. The carapaces (Δ_1, Δ_2, and Δ_3), ridgelines (r_1, r_2, and r_3), and corner (μ) are indicated in red, green, and purple letters, respectively. **e** Bandgap isosurface in the $\varepsilon_{23}\varepsilon_{13}\varepsilon_{12}$ (shear only) strain space at 3.5 eV. The yellow arrow indicates a change of carapaces on this isosurface pertaining to indirect-to-direct bandgap transition in the diamond. The corresponding change from the indirect bandgap structure to the direct bandgap structure of CBM **k**-space coordinates from X_1 (0, 0.5, 0.5) to Γ (0, 0, 0) is shown in band structure plots (**f**) and (**g**), respectively. Red arrows in both plots indicate the CBM.

can be defined in terms of the Hessian matrix $\mathbf{H}(E_{n_{CB}}(\mathbf{k}))$ consisting of second partial derivatives with respect to **k**. Based on the values drawn from our ML model, the electron effective mass $\boldsymbol{m}^*$ tensor for an undeformed diamond at CBM is extracted by fitting the band structure:

$$\boldsymbol{m}^* = \begin{bmatrix} 1.55m_e & 0 & 0 \\ 0 & 0.31m_e & 0 \\ 0 & 0 & 0.31m_e \end{bmatrix}, \quad (1)$$

where m_e is the free-electron mass. Given that $\boldsymbol{m}^*$ is a second-order derivative, it reveals not only the shape of an energy band but also its curvature, thereby providing more detailed information on band dispersion. The anisotropy at CBM is characterized by a longitudinal mass ($m_L^* = 1.55m_e$) along with the corresponding equivalent (100) reciprocal space direction and two transverse masses ($m_T^* = 0.31m_e$) in the plane perpendicular to the longitudinal direction. Our results for m_L^* and m_T^* are close to both the GW and experimental values (see Table 1), offering more evidence for the reliability of our electronic band structure representation. A plot that demonstrates the agreement between our model and GW calculations for effective mass components can be found in Supplementary Note 3 and Supplementary Fig. 6.

We also studied the 6D strain space to obtain the conduction-related properties and the elastic strain energy density as functions of ε. Here, we adopted our ML model to acquire the relation between the "conductivity effective mass" for the conduction electron $m^*_{cond}(\boldsymbol{\varepsilon})$ and $h(\boldsymbol{\varepsilon})$, as shown in Fig. 5a. The values of scalar m^*_{cond} are obtained by averaging individual longitudinal and transverse effective masses, as in ref. [23]. The blue shading in Fig. 5a reveals the distribution of the available m^*_{cond},

Table 1. Longitudinal and transverse electron effective masses at CBM in undeformed diamond (in units of free-electron stationary mass m_e).

	CNN (this work)	NN	GW_0 (this work)	LMTO	G_0W_0	Experiment
m_L^*	1.55	1.63	1.44	1.5	1.1	1.4
m_T^*	0.31	0.31	0.31	0.34	0.22	0.36
m_L^*/m_T^*	5.0	5.16	4.61	4.41	5.0	3.89

The results obtained through our CNN model are compared with experiments[33], our previous feed-forward NN model[13], and explicit calculations using existing methods including GW_0, linear muffin-tin-orbital (LMTO) model[34], and G_0W_0[35].

with darker shading implying more strains are able to reach a specific value of m^*_{cond} at a given h.

The cumulative "density of states" of conductivity effective mass can be defined as

$$c(m^*_{\mathrm{cond}}\prime; h') \equiv \int_{h(\boldsymbol{\varepsilon})<h'} d^6\boldsymbol{\varepsilon}\,\delta(m^*_{\mathrm{cond}}\prime - m^*_{\mathrm{cond}}(\boldsymbol{\varepsilon})) = \int d^6\boldsymbol{\varepsilon}\,\delta(m^*_{\mathrm{cond}}\prime - m^*_{\mathrm{cond}}(\boldsymbol{\varepsilon}))\Theta(h' - h(\boldsymbol{\varepsilon})), \tag{2}$$

where $\delta(\cdot)$ and $\Theta(\cdot)$ are the Dirac delta and unit step functions, respectively, $d^6\boldsymbol{\varepsilon} \equiv d\varepsilon_{11}d\varepsilon_{22}d\varepsilon_{33}d\varepsilon_{23}d\varepsilon_{13}d\varepsilon_{12}$ in the 6D strain space. The density of states of conductivity effective mass (g) at $h\prime$ can then be defined by the derivative of $c(m^*_{\mathrm{cond}}\prime;\ h\prime)$ with respect to $h\prime$:

$$g(m^*_{\mathrm{cond}}\prime; h') \equiv \frac{\partial c(m^*_{\mathrm{cond}}\prime; h')}{\partial h'} = \int d^6\boldsymbol{\varepsilon}\,\delta(m^*_{\mathrm{cond}}\prime - m^*_{\mathrm{cond}}(\boldsymbol{\varepsilon}))\delta(h' - h(\boldsymbol{\varepsilon})), \tag{3}$$

The meaning of g is explained by considering in the $(h - \frac{dh}{2}, h + \frac{dh}{2})$ interval all possible elastically strained states and the resultant distribution of m^*_{cond} arising from these states. Other plots of the density of states of individual effective mass tensor components are also available (see Supplementary Fig. 7). Moreover, the developed framework enables high-quality predictions of the $\boldsymbol{m}^*$ tensor components (as well as their averages) for every **k**-point at various deformation cases.

Direct bandgap together with a small effective mass within a semiconductor material is a preferable combination in the design of radiation detectors and photovoltaic cells that enables the combination of high conductivity and light yield. Moreover, lower elastic strain energy density means less effort for reaching the same property design in ESE. The three objectives, however, generally cannot be minimized simultaneously, to give the hands-down best solution; instead, there exists a set of Pareto-efficient solutions, which do not allow for any member of a triplet (E_g, m^*_{cond}, and h) to improve (i.e., decrease) without negatively affecting the other two members. The 3D Pareto front of minimized E_g, m^*_{cond}, and h, shown in Fig. 5b, indicates a compromise in simultaneously having a small bandgap and conductivity effective mass, where h could increase to more than 120 meV/Å^3. It is not possible to achieve, for example, a near-zero bandgap and $m^*_{\mathrm{cond}}<0.25m_e$ without paying a considerable penalty in h by straining diamond, as indicated by the "infeasible region" in Fig. 5b. Also, it is likely to find higher h values that correspond to the same combination of (E_g, m^*_{cond}). In Fig. 5b, such elastic strain energy density values are associated with strain cases in the "feasible region". In addition, Fig. 5c could serve as a blueprint to access all possible (E_g, m^*_{cond}) combinations achieved by straining diamond in order to find the smallest elastic strain energy density (h_{min}) for each combination. Note that it includes more (E_g, m^*_{cond}) combinations and is not a projection of Fig. 5b onto the $E_g - m^*_{\mathrm{cond}}$ plane.

In summary, by recognizing that the band dispersion is structured and highly correlated in continuous **k**, ε,and discrete n, the method presented in this work provides better approximation and less uncertainty in the estimation of key figures of interest in scientific and technological applications of semiconductors. This task is made possible through the implementation of physics-informed neural network architecture, synergistic PBE + GW data sampling, and active learning. Specifically, the CNN-based network structure we developed can handle the tasks of the fast query of properties of any electronic materials, including bandgap, band edges, and the energy difference between electron athermal (phonon-free) band transition, at accuracy on par with or better than their purpose-specific counterparts. Direct utilization of this fitting scheme on diamond reveals the strain levels where indirect-to-direct bandgap transition and insulator-to-metal transition take place.

To accomplish the task of band structure prediction, our network offers the capabilities of learning the complex intra-band and inter-band correlation in a self-directed manner while taking into account important physical characteristics, such as crystal periodicity and time-reversal symmetry. For example, the application of our method on computing the extremely sensitive energy dispersion-related properties such as the effective mass tensor demonstrates that the method can capture the second-order details of band structure within a level of precision comparable to that of the underlying calculation method. Multi-objective Pareto optimizations are also carried out aided by this model. The general ML framework we propose here thus effectively alleviates the heavy dependence upon DFT calculation, which takes up about 99% of the model construction time in an otherwise typical first-principles materials design project without ML. At the same time, it provides an avenue for deploying physics-informed deep learning. Finally, active learning technique coupled with data fusion provides smart and autonomous searching of the vast region of the 6D strain space for optimizing FoM.

METHODS

The models used in this work are described in detail in the "Results" section. Additional content regarding first-principles calculation settings and dataset construction is included here.

First-principles calculation

We used the projector augmented wave method (PAW)[24] in our DFT simulations implemented in the Vienna Ab initio Simulation Package (VASP)[25], with the exchange-correlation functional of PBE[14]. In all calculations, the electronic wavefunctions were expanded in a plane wave basis set with an energy cutoff of 600 eV. An 8 × 8 × 8 Monkhorst-Pack[26] **k**-mesh was used to conduct the Brillouin zone integration. The maximum residual force allowed for atoms after structural relaxation is 0.0005 eV Å^{-1}. Computations that invoke GW corrections were conducted on top of the above PBE settings. We chose to sample the strain cases in a range of $\{-0.15 \le \varepsilon_{ii} \le 0.15,\ -0.1 \le \varepsilon_{ij} \le 0.1, (i,j = 1,\ 2,\ 3)\}$ that yields stable structures, ie. without imaginary phonon frequencies. To identify the phonon stability boundaries, we performed phonon calculations for densely sampled strain cases in the 6D space. These calculations were primarily carried out using the VASP-Phonopy package[27]. 2 × 2 × 2 supercells of 16 carbon atoms were created, and phonon calculations were conducted with a 3 × 3 × 3 **q**-mesh. We also took full advantage of the known symmetries to further reduce the computations needed when collecting the strain data.

Dataset construction

In the data generation step of database construction, we took the Latin-Hypercube-sampled[28] strain points and adopted the above parameters in our ab initio calculations to acquire the bandgap, band structures, and related properties for every deformed structure. To validate our calculation, we compared it with accessible values obtained in experiments. Specifically, the undeformed diamond properties are widely available, and we validated our many-body G_0W_0 calculation settings by matching our result at zero-strain with the experimental lattice constant, elastic properties, dielectric constant, and most importantly, bandgap and band structure of diamond. Since we have adopted phonon calculations to

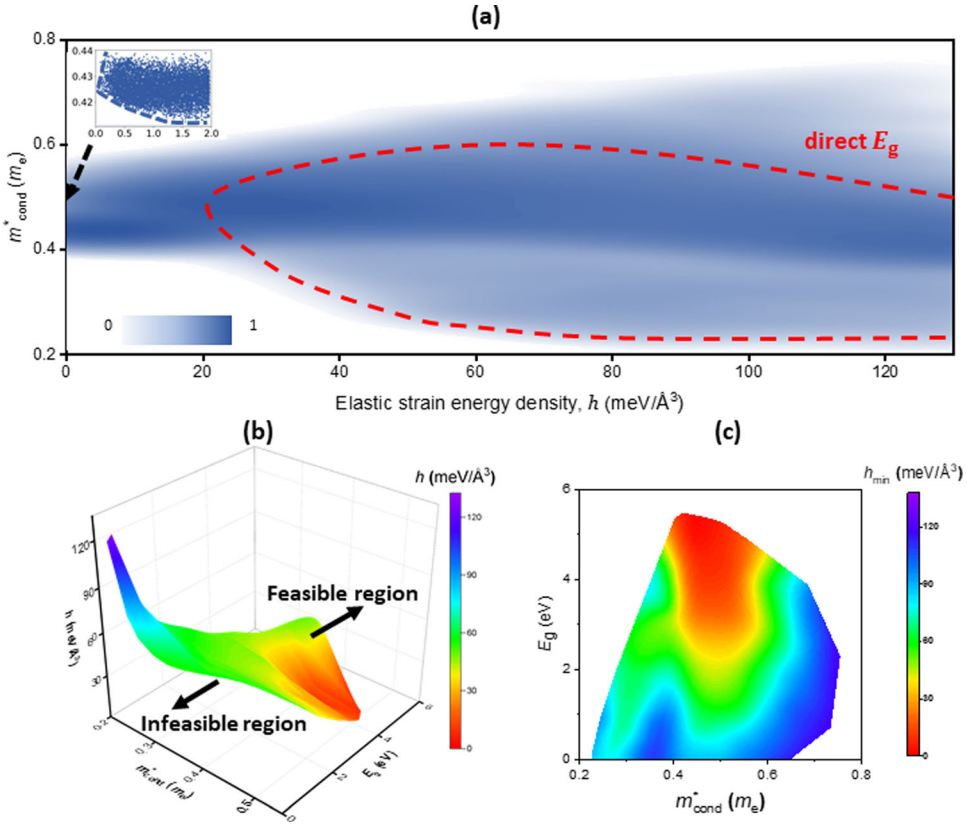

Fig. 5 ML-based exploration of electron effective mass tensor. a The density of states of conductivity effective mass. A darker shading implies more strains can reach a specific value of m^*_{cond} at a given h. The red dashed line indicates the region of the possible direct bandgap configurations. Inset is the zoomed-in plot near $h = 0$ of the m^*_{cond} distribution. **b** Pareto front for minimizing m^*_{cond}, bandgap, and h. The color contours denoted different h values. The (E_g, m^*_{cond}, h) triplets within the Pareto front are feasible, meaning that there exist strain cases that can realize the three properties. **c** Color contours of the smallest elastic strain energy density (h_{min}) required for achieving any combinations of bandgap and m^*_{cond}. **c** is not a 2D projection of **b** in which only optimized (minimized) E_g and m^*_{cond} exist.

eliminate the cases where phase transitions (such as graphitization[15,29]) could happen and focused our search on the elastic regime, the diamond structures which we conducted high-throughput computations are all of the sp^3 hybridization types. Therefore, unlike the Materials Project database construction[30] where separate DFT settings and experimental references had to be employed for different classes/phases of materials across a much larger chemical space, it would be enough for us to use the undeformed diamond as the reference to benchmark the calculations. In addition, for strain cases of greater interest (such as the near metallization and direct-bandgap strain cases), we went beyond the single-shot G_0W_0 method and used partially self-consistent GW_0 calculation settings (allowing Green's function iterations to acquire more accurate bandgap) that is known to obtain results better than calculations with hybrid-functional DFT[31] and comparable with experimental measurement for many semiconductor materials[32].

DATA AVAILABILITY

The authors declare that all data supporting the findings of this study are available within the paper and its Supplementary Information file. Further information is available upon reasonable request.

Received: 12 December 2020; Accepted: 8 April 2021;
Published online: 28 May 2021

REFERENCES

1. Li, J., Shan, Z. & Ma, E. Elastic strain engineering for unprecedented materials properties. *MRS Bull.* **39**, 108–114 (2014).
2. Jain, S. C., Maes, H. E. & Van Overstraeten, R. Semiconductor strained layers. *Curr. Opin. Solid State Mater. Sci.* **2**, 722–727 (1997).
3. Bedell, S. W., Khakifirooz, A. & Sadana, D. K. Strain scaling for CMOS. *MRS Bull.* **39**, 131–137 (2014).
4. Sun, Y., Thompson, S. E. & Nishida, T. Physics of strain effects in semiconductors and metal-oxide-semiconductor field-effect transistors. *J. Appl. Phys.* **101**, 104503 (2007).
5. Banerjee, A. et al. Ultralarge elastic deformation of nanoscale diamond. *Science* **360**, 300–302 (2018).
6. Nie, A. et al. Approaching diamond's theoretical elasticity and strength limits. *Nat. Commun.* **10**, 1–7 (2019).
7. Dang, C. et al. Achieving large uniform tensile elasticity in microfabricated diamond. *Science* **371**, 76–78 (2021).
8. Zhang, H. et al. Approaching the ideal elastic strain limit in silicon nanowires. *Sci. Adv.* **2**, e1501382 (2016).
9. Aryasetiawan, F. & Gunnarsson, O. The GW method. *Rep. Prog. Phys.* **61**, 237–312 (1998).
10. Lu, H. et al. Mechanical writing of ferroelectric polarization. *Science* **336**, 59–61 (2012).
11. Wang, Z. et al. Non-linear behavior of flexoelectricity. *Appl. Phys. Lett.* **115**, 252905 (2019).
12. Bardeen, J. & Shockley, W. Deformation potentials and mobilities in non-polar crystals. *Phys. Rev.* **80**, 72–80 (1950).
13. Shi, Z. et al. Deep elastic strain engineering of bandgap through machine learning. *PNAS* **116**, 4117–4122 (2019).
14. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
15. Shi, Z. et al. Metallization of diamond. *PNAS* **117**, 24634–24639 (2020).
16. Ciregan, D., Meier, U. & Schmidhuber, J. Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 3642–3649 (IEEE Computer Society, Eric Mortensen, 2012).

17. Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A. & Tsunoda, T. DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture. *Sci. Rep.* **9**, 11399 (2019).
18. Kingma, D. P. & Welling, M. *An Introduction to Variational Autoencoders*. (Now Publishers, 2019).
19. Shapeev, A. et al. Active learning and uncertainty estimation. Machine Learning Meets Quantum Physics, 309-329. (Springer International Publishing, 2020).
20. Cohn, D. A., Ghahramani, Z. & Jordan, M. I. Active learning with statistical models. *J. Artif. Intell. Res.* **4**, 129–145 (1996).
21. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. Vol. 48, 1050–1059 (eds Balcan, M. F. & Weinberger, K. Q.) (JMLR.org, 2016).
22. Tsymbalov, E., Makarychev, S., Shapeev, A. & Panov, M. Deeper Connections between Neural Networks and Gaussian Processes Speed-up Active Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. 3599–3605. http://ijcai.orgijcai.org (Sarit Kraus, 2019).
23. Zeghbroeck, B. J. V. *Principles of Semiconductor Devices*. (Bart Van Zeghbroeck, 2011).
24. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
25. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
26. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
27. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
28. McKay, M. D., Beckman, R. J. & Conover, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239–245 (1979).
29. Regan, B. et al. Plastic deformation of single-crystal diamond nanopillars. *Adv. Mater.* **32**, 1906458 (2020).
30. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
31. Kaczkowski, J. Electronic structure of some Wurtzite semiconductors: hybrid functionals vs. ab initio many body calculations. *Acta Phys. Pol. A* **121**, 1142–1144 (2012).
32. Hybertsen, M. S. & Louie, S. G. Electron correlation in semiconductors and insulators: band gaps and quasiparticle energies. *Phys. Rev. B* **34**, 5390–5413 (1986).
33. Nava, F., Canali, C., Jacoboni, C., Reggiani, L. & Kozlov, S. F. Electron effective masses and lattice scattering in natural diamond. *Solid State Commun.* **33**, 475–477 (1980).
34. Willatzen, M., Cardona, M. & Christensen, N. E. Linear muffin-tin-orbital and k·p calculations of effective masses and band structure of semiconducting diamond. *Phys. Rev. B* **50**, 18054–18059 (1994).
35. Löfås, H., Grigoriev, A., Isberg, J. & Ahuja, R. Effective masses and electronic structure of diamond including electron correlation effects in first principles calculations using the GW-approximation. *AIP Adv.* **1**, 032139 (2011).

ACKNOWLEDGEMENTS

The computations involved in this work were conducted on the computer cluster at Skolkovo Institute of Science and Technology (Skoltech) CEST Multiscale Molecular Modelling group and Massachusetts Institute of Technology (MIT) Nuclear Science Engineering department. E.T., Z.S., A.S., and J.L. acknowledge support by the Skoltech-MIT Next Generation Program 2016-7/NGP. E.T. and A.S. acknowledge support by the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy Office of Science by Los Alamos National Laboratory (Contract 89233218CNA000001) and Sandia National Laboratories (Contract DE-NA-0003525). M.D. acknowledges support from MIT J-Clinic for Machine Learning and Health. S.S. acknowledges support from Nanyang Technological University through the Distinguished University Professorship.

AUTHOR CONTRIBUTIONS

Z.S. and E.T. contributed equally to this work. E.T. and Z.S. performed the research, including first-principles calculations, machine learning model development, and data analysis. E.T., Z.S., M.D., S.S., J.L., and A.S. participated in writing and revising the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00538-0.

Correspondence and requests for materials should be addressed to S.S., J.L. or A.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.