Deep elastic strain engineering of materials electronic properties by machine learning

By

Zhe Shi

B.A.Sc., University of Toronto, 2016

Submitted to the Department of Materials Science and Engineering in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

June 2021

Signature of Author ....
Zhe Shi
Department of Materials Science and Engineering

Certified by ....
Ju Li
Battelle Energy Alliance Professor of Nuclear Science and Engineering and Professor of Materials Science and Engineering

Accepted by ....
Frances M. Ross
Ellen Swallow Richards Professor of Materials Science and Engineering
Chair, Departmental Committee on Graduate Studies

Abstract

The introduction of elastic strains has become an appealing strategy for providing unique and exciting electronic properties in nanostructured materials. Recent successes in diamond and silicon deformation experiments further extended the applicable strain levels in these materials, harbingering a new stage of elastic strain engineering of semiconductor fundamental electronic properties and device performance. However, it is not generally easy to know from experiments how much an electronic property change is for materials undergoing bending or even uniaxial tension, let alone designing the optimal combination of functional properties for the material in a vast and more complex six-dimensional (6D) strain space. The complexity of controllably engineering materials properties in such a 6D strain space necessitates high-fidelity high-efficiency computer screening for a desirable figure-of-merit and then designing a proper straining pathway to guide future experiments.

To address this challenge, we developed in this thesis a general framework that combines machine learning and a limited amount of *ab initio* calculations to guide strain engineering whereby basic electronic properties are designed. Our method invokes deep neural networks, convolutional neural networks, data fusion, and active learning algorithms, allowing for accurate and efficient prediction of strain-dependent fundamental electronic properties such as band structure, bandgap, band extrema location, and effective mass, as well as other properties with minor modifications. It is also used for discovering indirect-to-direct bandgap transition that would benefit photon emission and absorption in a semiconductor such as silicon by scanning the entire strain space.

Integrating this method with finite-element simulations, we predicted energy-efficient strain pathways that would reversibly transform an ultrawide-bandgap material such as diamond to a metalized state in an experimentally feasible geometry. The fast and reliable inference of the proposed framework opens a path beyond analyzing and scrutinizing electronic band structures. In particular, an application of this framework in the studies of phonon band structure and phonon stability of diamond yielded a visualization and theoretical understanding of the deep elastic strain engineering boundary in the vast 6D strain space. We also applied the machine learning models to investigate the strain-induced variations of defect ionization energy and predicted deep-to-shallow defect level transition in diamond, offering a theoretical possibility to make strain-controlled switchable devices with doped diamond.

We illustrate the applications of the method with results for silicon and diamond, although the general technique presented here is potentially useful for optimizing figures-of-merit for a variety of semiconductors, providing guidance for experimentally tailoring materials properties via deep elastic strain engineering for electronic, photonic, and energy applications.

Acknowledgment

PhD research is intrinsically challenging but de facto interesting for me, due to the great help and support from many people whom I would like to sincerely thank.

First and foremost, I want to express my greatest thanks to my PhD advisor Professor Ju Li, who is a great scientist, a fantastic mentor and a man of wisdom. Without guidance from a giant like him, it is hard to believe I can make any achievement in this and many other academic works.

I am also graciously indebted to my committee members, Professor Krystyn Van Vliet and Professor C. Cem Tasan. They offered great insights and encouraged me to look at my research from various perspectives. Their comments and suggestions during my PhD study proved very useful, especially when I tried to convey my research at conference talks. Professor Tasan's course on mechanical properties of materials is the best I have taken at MIT which apparently provided me with the knowledge I required to conduct this thesis work. I feel deeply honored to have them on my thesis committee.

I would like to express my deep gratitude to Professor Subra Suresh and Professor Ming Dao. Busy though he is, Professor Suresh spent a lot of effort providing guidance and offering help to my work, including even helping the detailed narratives of my writing. Similarly, Professor Dao not only provided knowledge and viewpoint but also generously offered a huge amount of his time in my work. I enjoyed the many chats I had with Professor Dao.

Special thanks go to my intimate collaborator Dr. Evgenii Tsymbalov from Skoltech, who is an outstanding data scientist. Together we published two important papers during our PhD studies. Also, I would like to thank Evgenii's advisor Professor Alexander Shapeev, a bright and prolific mathematician, for his kind help.

I would like to thank all the members of the Li group. I will always miss the luxury having hot pot with Ziqiang and Wenbin, the optimistic (sometimes pessimistic) discussion with Weijiang and Zhi on work and life, the interesting conversations with Yucheng about philosophy, the fun I had in speculating stocks with Yunwei, the hospitality of Tianquan, Fei, Xiaohui, Jian and Liumin, and the chats with many others I had the pleasure knowing (many of them are professors now), including: Yang Yang, Cong Su, Yu-Chieh Lo, Yuming Chen, Zhichu Ren, Rui Gao, Chen Wang, Hua Wang, Yanhao Dong, So Yeon Kim, Yao Li, Haowei Xu, David Yu, Shitong Wang, Wei Fan, Yimeng Huang, David Bloore, Baoming Wang, Guiyin Xu and so on. This list will not be complete without mentioning our staff member Rachel Bastista, who deserves a big thank-you from me for her kindness and helpfulness.

Haozhe Wang and Keran Rong are peers I often resort to for help beyond academic life. They, along with many others, are the buddies whom I am lucky to meet at MIT. During my PhD research, I have the fortune collaborating with or learning from researchers in many institutions, including Jiaping Wang of Tsinghua University, Fang Kong and Tingting Yin of Nanyang

Technological University, Qiubo Zhang of UC Berkeley, and Professor Chandra Veer Singh of the University of Toronto. Professor Singh was my advisor in my undergrad years, and I am forever grateful to him for his mentorship.

Last but not the least, to my parents who are always there to be supportive. They are the invisible “co-first authors” of everything I have achieved.

List of Figures

List of Tables

Table of Contents

Chapter 1. Introduction

1.1. Motivation for deep elastic strain engineering

Over the past two decades, experiments on nanostructured materials have repeatedly demonstrated the "smaller is stronger" phenomenon, the concept of which was first introduced for micro-structured material systems dating back to the 1950s [1–4]. As materials in the nanoscale are mechanically much stronger than their microscale or bulk counterparts, one can exert even greater tensile, compressive or shear strains to alter their physical-chemical properties for a sufficient amount of time without inelastic relaxation brought up by fracture or plasticity. This is a significant improvement, because unlike high-pressure physics where exotic properties are realized only through hydrostatic (non-deviatoric) strains (e.g., de-metallization/insularization of solid sodium under a high-pressure environment in a diamond anvil cell), the shear and tensile strain states involved here typically exceed 1% sample-wide, giving rise to a class of ultra-strength materials. These materials, thanks to their nanostructures, successfully delay the onset of what otherwise could be ~0.2% deviatoric strain limit in conventional materials.

The electronic, optical, thermal, and chemical properties of crystals are functions of the 6D strain tensor $\boldsymbol{\varepsilon}$ (a 3 × 3 symmetric tensor with six independent strain components), which provides a continuously tunable set of variables analogous to the chemical composition of a septenary alloy, given that the behavior of a deformable body and composition field are both governed by the generalized mass and momentum conservation laws. The total deformation of a material point is described by the summation of elastic strain ($\boldsymbol{\varepsilon}_{\mathrm{e}}$) and inelastic strain ($\boldsymbol{\varepsilon}_{\mathrm{ie}}$). The field of $\boldsymbol{\varepsilon}_{\mathrm{e}}$ is a local description for Bravais lattice vectors distortion of otherwise unperturbed crystals and can be experimentally measured by crystallography performed inside a transmission electron microscope (TEM), such as selected area electron diffraction. $\boldsymbol{\varepsilon}_{\mathrm{ie}}$, on the other hand, is accompanied by bond switching (bond breaking and reformation with no net reduction of total bond count), bond loss (a net reduction of atom coordination number as in brittle solids), or phase transformation changes. Suppressing the incipient plasticity including defect nucleation and subsequent evolution, for instance, enables the material to experience only elastic deformation, making possible the rational tailoring of materials properties – a term also called "elastic strain engineering (ESE)".

1.1.1. Conventional and deep elastic strain engineering

There are ideal and realistic limits to ESE. The former one is the ideal strain, $\boldsymbol{\varepsilon}_{\mathrm{ideal}}$, which is the maximum strain achievable in a perfect crystal at zero Kelvin without giving rise to lattice dynamical instability. Several criteria have been proposed for predicting lattice instability, including the Born criterion [5], the Li-Van Vliet's Λ criterion [6,7], the Miller-Rodney criterion [8], and the soft phonon criterion [9]. Among these criteria, soft phonon offers both necessary and sufficient conditions when an ideal lattice with Born-Von Karman periodical boundary conditions undergoes a uniform displacement/strain-controlled loading mode, and the instability mode of the lattice can be characterized by the instable wave vector together with the eigenvectors of dynamical matrix. The realistic ESE limit, $\boldsymbol{\varepsilon}_{\mathrm{realistic}}$, is much more conservative, as it takes into account temperature, time and pre-existing defects. $\boldsymbol{\varepsilon}_{\mathrm{realistic}}$ corresponds to a threshold stress (more commonly known as the material strength) beyond which the plasticity or phase transformation takes place. It is very clear that achieving such theoretically ideal strain is virtually impossible in any real-world experimental settings, since the requirement of being defect-free (not even material surface is allowed to present) is not attainable. However, $\boldsymbol{\varepsilon}_{\mathrm{ideal}}$ still gives us a good indication of the upper bound for designing strain magnitude when practicing ESE to a certain material. Also, the ratio between $\boldsymbol{\varepsilon}_{\mathrm{realistic}}$ and $\boldsymbol{\varepsilon}_{\mathrm{ideal}}$ tells us whether we are in the realm of common ESE or "deep" ESE.

As a convenient rule of thumb, if $\boldsymbol{\varepsilon}_{\mathrm{realistic}}$ is on the order of 1/10 of the $\boldsymbol{\varepsilon}_{\mathrm{ideal}}$ of a material, one is practicing conventional ESE. The past two decades have witnessed many works belonging to this kind, including but not limited to using strains to promote conductivity in $SrTiO_3$-based systems [10–13], to enable chemical reactions on metals and oxides [14], to improve the performance of known ferroic oxides ($BiFeO_3$, $EuTiO_3$, etc.) [15], to tailor exciton dynamics in ZnO nanowires [16], to facilitate Mott transition in VO_2 [17], to enhance light emission in germanium for better laser designs [18,19] and so on. Besides research interests, there also exists one huge commercial success in the field of conventional ESE: the strain silicon technology where a biaxial or uniaxial elastic strain of the order of 1% applied to a thin complementary metal-oxide-semiconductor (CMOS) channel of silicon enhances the mobility of charge carriers by more than 50% and increases central processing unit (CPU) clock speed correspondingly. Such elastic strain is also implemented in the more recent device architecture such as the fin field-effect transistors (FinFETs) to continue the strain scaling for CMOS (See Table 1.1 for examples of conventional ESE of semiconductors). From a business perspective, the strained silicon technology, though belongs to conventional ESE (within ~1/10 of $\boldsymbol{\varepsilon}_{\mathrm{ideal}}$), is arguably the most successful technology in the entire field of nanotechnology.

Table 1.1. Strained silicon technologies in the realm of conventional ESE.

Materials	Geometry	Reachable elastic strain	Loading mode	FET device made
Si [20]	Si on SiGe	0.55%-1%	Biaxial tension	55 nm n-type FET
Si [21]	Si sandwiched between two SiGe contacts	-1%	Uniaxial compression	45 nm p-type FET
Si [21]	Si stretched on two ends by Si_3N_4	1%	Uniaxial tension	45 nm n-type FET
Si [22]	Si sandwiched between two SiGe contacts	-1.26%	Uniaxial compression	22 nm p-type FET
Si [22]	Si stretched by two Si:C sides	1.25%	Uniaxial tension	22 nm n-type FET
Si [23]	Stressed by substrate and drain/source	1.08% and 0.57%	Mixed uni- and biaxial tension	n-type FET
Si [24,25]	Fin-shaped, stressed by drain/source	-1.6% to -0.4%	Vertical compression, longitudinal and transverse tension [25]	n-type FinFET

If one is able to let $\boldsymbol{\varepsilon}_{\text{realistic}}$ reach a considerable fraction of or even approach $\boldsymbol{\varepsilon}_{\text{ideal}}$ and sustain such deformation for sufficient time, then one is entering into the so-called deep ESE regime ("deeply" into the elastic regime). It is necessary to emphasize these ultra-high strain levels should be retained sample-wide or at least a significant portion of the functional body instead of just in a local region, since it is very common to have only a tiny volume of material, such as near crack tip, reached abnormally high strain with other regions barely deformed, which is generally not useful for practical ESE. Also, the requirement for having "sufficient time" is relative and depends on specific materials application scenarios. For instance, one may want to hinder vacancy diffusion and dislocation nucleation so that a 10%+ elastic strain in a FinFET chip made of silicon can be sustained at operating temperature for the service life of 3-5 years of a smartphone, which is quite different from the relevant timescale for evaluating plastic flows in a suspension bridge steel wire rope cable or glaciers.

1.1.2. What has been done for deep elastic strain engineering experiments

With the proliferation of ultra-strength nanostructured materials that can sustain a wide range of non-hydrostatic and potentially dynamically varying stresses, and various miniaturization-enabled means of applying $\boldsymbol{\varepsilon}$ [26], a historical window of opportunity has now opened up to scan a vast unexplored deep strain space for the development of materials and devices with desirable properties [27,28]. For example, while it is well known that unstrained Si has an electronic bandgap (E_g, the energy gap between conduction and valence energy bands) of 1.1 eV, we know that, when subjected to a strain of say 9%, it would have a different bandgap. Furthermore, a 9% tensile strain (significantly greater/deeper than prior approaches adapted by conventional ESE) on Si would produce a different bandgap from a 9% shear strain. At large strains, all these differently strained pure Si crystals would not behave as the unstrained "typical silicon." An added benefit is that with deep ESE, it is in principle possible to dynamically change the mechanical actuation, and switch between these differently strained materials.

Indeed, recent experiments [29–32] in free-standing geometries 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 (Figure 1.1a) [29] and nanowires (Figure 1.1b) [30], and the sample-wide elastic strain can achieve up to 9%+ in diamond microbridge arrays (Figure 1.1c) [31]. In nanowires of silicon, as shown in Figure 1.1d, a reversible tensile strain of 15% has been realized in uniaxial tension experiments [32]. These findings of deep elastic deformation of semiconductor materials bookended by the ultra-wide bandgap (5.6 eV) diamond and the more manufacturing-friendly and ubiquitous silicon have opened up potential opportunities to design their performance characteristics with ESE for applications such as power electronics, nanophotonics and quantum information processing.

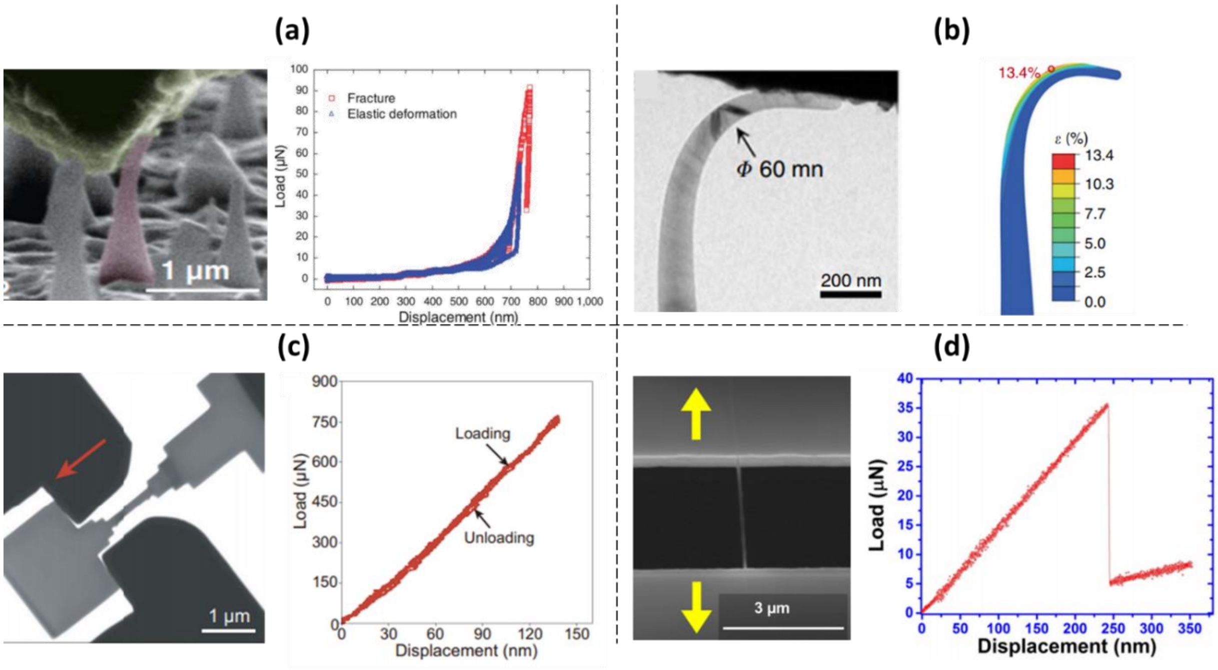


Figure 1.1 *In-situ* near-$\boldsymbol{\varepsilon}_{\text{ideal}}$ deformation experiments of diamond and silicon. (a) Ultralarge diamond nanoneedle bending experiment. The load-displacement curves were measured by pushing the nanoindenter tip onto the nanoneedle for the fully reversible elastic deformation and the final fracture run. Taken from Ref. [29]. (b) Bending of a diamond nanoneedle by pushing toward a rigid wall. Finite-element simulation reproducing the needle geometry is also shown. Taken from Ref. [30]. (c) Uniaxial tension of microfabricated diamond bridge samples. The load-displacement curve is also shown. Taken from Ref. [31]. (d) Uniaxial tension of single crystalline silicon nanowire clamped at two sides. The load-displacement curve for a complete run is present. Taken from Ref. [32].

1.1.3. Four mainstays of deep elastic strain engineering

With the recent success in exploiting the ability of Si and diamond to deform up to near $\boldsymbol{\varepsilon}_{\text{ideal}}$ magnitudes under certain conditions, it is natural to study the physical property or figure-of-merit (FoM) change in the deformed materials which is obviously the next step to show the impact of this strategy on designing semiconductors with improved functional properties through deep ESE. In fact, there exist much greater possibilities than what is realized in engineering Si or diamond for a wide variety of electronic, optoelectronic, and photonic materials employed in communication, information, and energy applications that impact every aspect of modern life [33]. However, in order to have the next decade witnessed explosive growth in the application of deep ESE, collective advances on four major scientific fronts must be made.

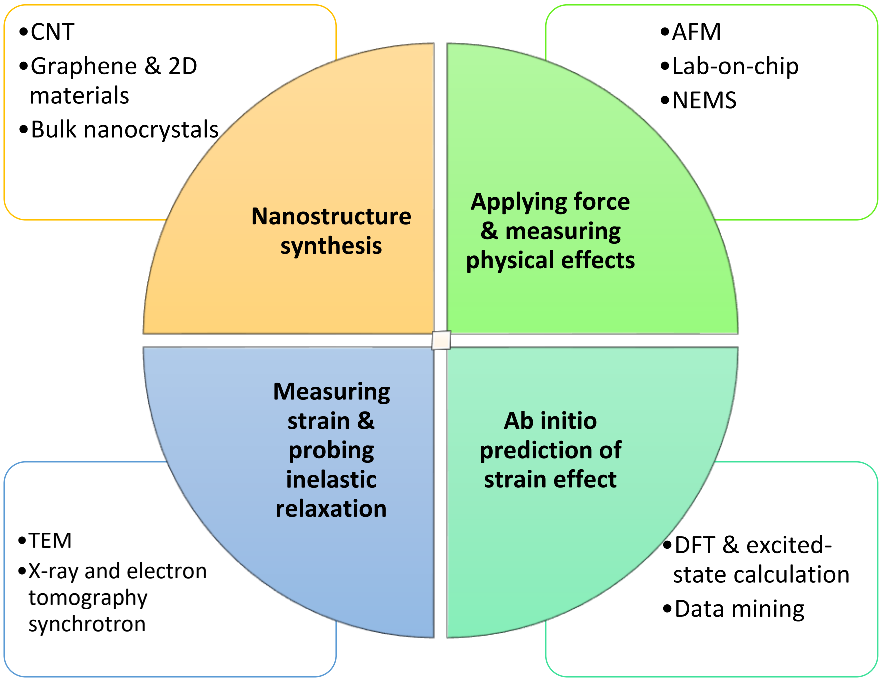

•CNT
•Graphene & 2D materials
•Bulk nanocrystals
•AFM
•Lab-on-chip
•NEMS
Nanostructure synthesis
Applying force & measuring physical effects
Measuring strain & probing inelastic relaxation
Ab initio prediction of strain effect
•TEM
•X-ray and electron tomography synchrotron
•DFT & excited-state calculation
•Data mining

Figure 1.2 Four mainstays of deep ESE.

As shown in Figure 1.2, the first mainstay is about synthesizing or fabricating deep elastic strain-bearing nanostructured materials, including but not limited to carbon nanotubes (CNTs), graphene, MoS_2 and other 2D materials, and nanocrystals (Si/diamond nanowires/pillars/needles/bridges). The second mainstay is about applying the deep elastic strain and measuring physical and chemical effects at the nanoscale, the means of which involves nanoindentation/bending by atomic force microscope (AFM), nano/microelectromechanical systems (NEMS/MEMS) loading, etc. These are the “tiny hands” conveying the strain as envisioned by Richard Feynman in his talk in 1959 [26]. The third mainstay is about measuring strain and understanding the different mechanisms for the occurrence of inelastic relaxations including plasticity, fracture or phase transition in the material. Note that the reason we study inelastic deformation is not to introduce it, which may cause retardation of functional properties of our loading-bearing device, but to avoid it for an extended time period. Tools capable of doing so are TEM [34], X-ray and electron tomography synchrotron, etc. Sometimes molecular dynamics (MD) simulations can also be conducted to support experimental findings. The last mainstay, which is also the main focus of this thesis, is about using *ab initio* means to reliably predict how materials physicochemical properties or FoMs may be altered by pure ESE assuming microstructural evolution and any other defect generation have been got rid of. First-principles density functional theory (DFT) and excited-state calculations are frequently used. Data science techniques may be utilized. Density functional perturbation theory (DFPT) calculations are also conducted in this stage to determine the $\boldsymbol{\varepsilon}_{\mathrm{ideal}}$.

1.1.4. Why focusing on engineering electronic properties

Deep ESE is both a new and broad field. There are a plethora kinds of properties that deserved to be studied. In this thesis, the author mainly focuses on understanding how deep ESE alters the *fundamental* electronic properties of perfect semiconductors. It is the electronic band structure of a material that dictates almost all physical and chemical properties [35]. In particular, the electronic bandgap (a scalar value retrievable from the band structure) is related to “chemical hardness” [35,36] and forms a basis for various properties. Since the electronic structure tends to be changed dramatically (such as the Herzfeld-Mott metal/insulator transition [37]) near the onset of spontaneous relaxation ($\boldsymbol{\varepsilon}_{\mathrm{ideal}}$), a material stressed near the ideal strain surface $\boldsymbol{\varepsilon}$ (a 5D surface $f(\boldsymbol{\varepsilon}) = 0$) in 6D results in drastic alteration of the electrical, thermal, optical, and magnetic characteristics [28].

Therefore, not only does ESE help to tailor the value of a figure-of-merit (FoM) as well as its character (e.g., direct or indirect bandgap), it can also push chemical [38] and physical behavior [39,40] toward extremes. However, it is not until a time when great advances are made in the fourth mainstay (i.e. *ab initio* prediction of strain effects) can we see great deep ESE applications in reality for many electronic materials.

There are various groups of semiconductors whose electronic properties and FoMs deserve the deep ESE treatment, including but not limited to elemental semiconductors (Si, diamond, Ge), oxides, compound semiconductors (GaN, SiC), and 2D materials. The following is a non-exhaustive list:

Silicon

Being the most versatile semiconductor, silicon has been used in virtually all commercial chips that have affected many aspects of modern life. The strained silicon technology is the "poster child" for ESE. However, this is only a "tip-of-the-iceberg" in terms of how much Si and diamond can deform. As shown in Section 1.1.1, recent experiments by our collaborators [29–32] harbinger a new age of deep ESE of the band structure and device performance of electronic materials.

If one can apply into a real device a significant fraction of the amount of what our collaborators have achieved in Si and diamond single crystalline nanoneedles and nanowires, we can expect to realize order-of-magnitude enhancement in transport properties. If given industrial limitations, it is difficult to achieve considerable strain magnitude along one particular loading direction, one can also resort to combining tension/compression/shear modes (with affordably small magnitude for each of the 6D strain components) to achieve the equivalent or even better improvement in key FoMs. Either way, a fundamentally much greater improvement in devices can be expected if we are able to realize controlled deep ESE in Si or diamond.

Diamond

Though silicon will continue to dominate semiconductor manufacturing for years and probably decades to come, more and more chip designers and engineers are paying attention to alternative semiconductors, known as the 3rd-gen semiconductors. Among these materials, diamond, with its exceptionally high hardness and stiffness, has extremely advantageous electronic properties that enable ultrawide-bandgap applications [33,41]. According to experimental measurements [42], diamond has an outstanding charge carrier mobility of ~3000 cm^2 V^{-1} s^{-1}, charge carrier saturation velocity >0.8 × 10^7 cm s^{-1} (high-frequency application), dielectric breakdown field in excess of 10 MV cm^{-1} (high-voltage applications). Diamond's Johnson's FoM is about 8200, ensuring a high power-frequency product for device applications. The Baliga's FoM of diamond is much higher than that of silicon. It also has a second-to-none thermal conductivity (>2000 W m^{-1} K^{-1}) among all known materials with a high Keyes' FoM, making it good at heat removel in power electronics. Some defect centers found in diamond, such as the NV-center, also have a great potential being used as a platform of qubit for quantum information applications.

Despite hard to achieve near-$\boldsymbol{\varepsilon}_{\text{ideal}}$ strains in epitaxy architecture of diamond, recent studies have confirmed they are achievable in free-standing structures. As introduced in Section 1.1.1, the authors' collaborators have shown that nanoscale needles and wires of diamond can be bent to a local maximum tensile elastic strain of more than 9% [29,30]. More recently, the author's group has helped in realizing micrometer-length ultralarge elastic deformation of diamond ribbons of 8%+ and sustaining it for a sufficiently long time [31], indicating a vital early step in potentially achieving deep ESE by reversible loading for diamond systems.

Therefore, as noted in Ref. [28] many possibilities remain to be investigated as to what pure silicon can do as the most versatile electronic material and what an ultrawide bandgap material such as diamond, with many appealing functional FoMs, can offer after overcoming its present commercial immaturity.

Perovskites

ESE for ferroics was triggered by the illustration that a non-ferroelectric oxide can be made ferroelectric at room temperature purely through mechanical strains. For example, 1% biaxial tensile strain can elevate the ferroelectric transition temperature of $SrTiO_3$ to near room temperature [43]. $EuTiO_3$ could be transmuted from a normal dielectric into the strongest ferroelectric ferromagnet by epitaxy strains on the order of 1% [44]. In addition, experiments and *ab initio* calculations have also shown that biaxial strain exerted upon $SrTiO_3$ can tailor transport properties [12] and facilitate accessible switching of electronic defect type [10].

Going beyond misfit strain from epitaxy, Chi et al have recently designed a three-pointing bending apparatus with *in-situ* impedance measurement capabilities [45]. Combining external (strain) fields with the existing substrate effect would further tune the strain level in material to achieve greater performance enhancement and more evident physical phenomena exploration. With an ever-increasing $\boldsymbol{\varepsilon}_{\text{realistic}}$ imparted, the prospect of deep ESE for thin film materials is very good.

2D materials

2D semiconductors/semimetals such as transition metal dichalcogenides (TMDs) and graphene are another material class suitable for deep ESE. AFM experiments have been conducted to measure their intrinsic strength and elastic properties (Figure 1.3a and b) [46,47]. Although 2D materials only have three in-plane strain components, there are also three degrees of freedom for elastic bending. Per the Cauchy-Born rule, the positions of individual atoms in a crystal follow the crystal lattice when strained. This approximation holds in general for Bravais lattice. However, the 2D hexagonal lattice considered here is a composite of two mutually shifted sublattices and the Cauchy-Born rule is not obeyed. Therefore, the internal degrees of freedom between the sublattices are also considered in our calculations. In Figure 1.3c, more than 40% or even orders of magnitude improvement can be already achieved in a variety of 2D materials by strains only in the conventional ESE level. It is expected deep ESE would trigger even more fascinating results on the electronic properties of these materials.

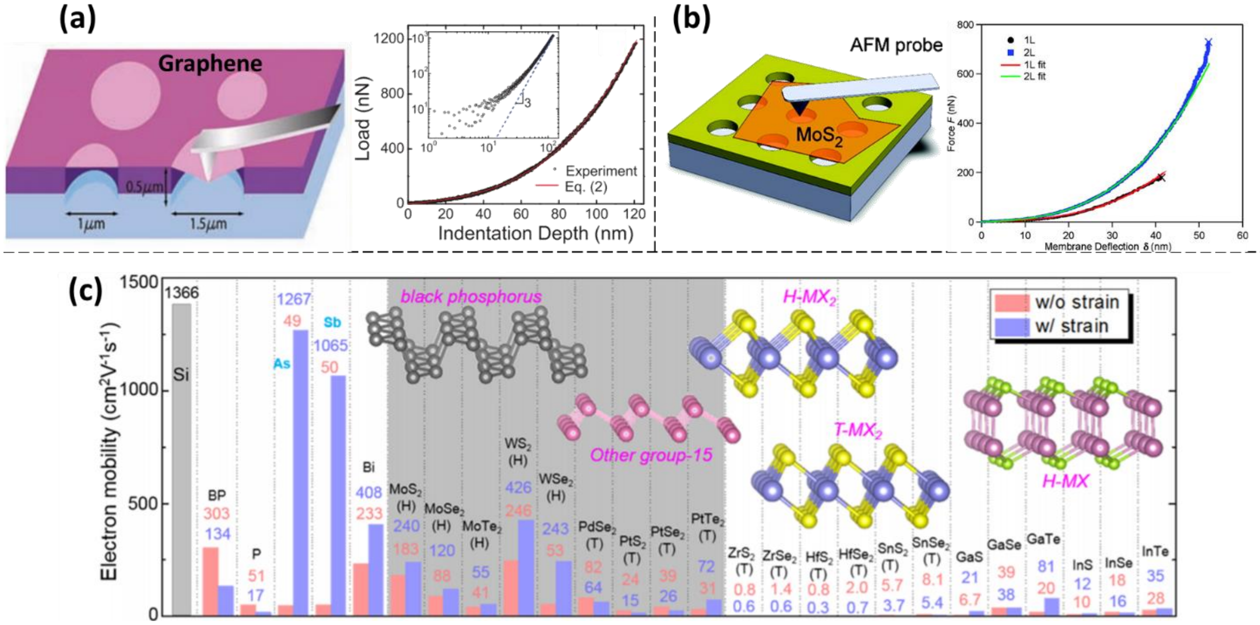


Figure 1.3 Ultralarge deformation of 2D materials and strain effect in functional 2D materials. (a-b) The AFM indentation experimental setups for graphene and MoS_2 as well as the associated load-indentation depth curve are shown. Exactly how much elastic strain the free-standing 2D material withstood before fracture was not measured in each experiment, but one can expect the $\boldsymbol{\varepsilon}_{\text{realistic}}$ in these 2D materials is quite high. Taken from Refs. [46,47]. (c) Strain effect on room temperature intrinsic electron mobility of common 2D semiconductors. Results of materials with a 3% tensile strain are in blue. The image is taken from Ref. [48].

1.2. Motivation for *ab initio* calculation and machine learning

Deep ESE deals with material states that are far from equilibrium for optimizing functional properties and performance. A strained material is in a state of higher energy than when it is in a stress-free state, characterized by the strain energy density (h, in units of meV/Å^3) [28]. Therefore, addressing the following question is at the heart of deep ESE: What is the energy cost (h) to achieve the desired property change? Consider the challenges of increasing the electron mobility of Si by 1000 cm^2 V^{-1} s^{-1}, converting germanium from an indirect to direct bandgap semiconductor, or transmuting diamond from an ultrawide-bandgap material into a medium- or even small-bandgap material so that its potentially appealing characteristics for microelectronics and optoelectronics could be realized. To achieve the above transitions in the most efficient manner, it is important to design $\boldsymbol{\varepsilon}$ through the most optimal combination of its normal and shear components [28].

To address the foregoing question, we resort to deep ESE which exploits first-principles modeling and the latest advances in artificial intelligence [28]. To set the scene, consider a situation where it is desirable to examine all possible combinations of the components of $\boldsymbol{\varepsilon}$, over a range of potential interest, say between −10% and +10% in each of the six independent strain components. Here, say that the objective is to determine the least energetically expensive route to decrease the bandgap of channel Si material by 1 eV to realize enhanced performance.

We consider a strategy of conducting a limited number of experiments that provide theoretical simulations with valuable parameters to benchmark, followed by batch runs of simulations that guide further deep ESE. Later the loading case which yields the optimal carrier mobility as predicted by simulations should be confirmed in experimental measurement. This way, we can do high-throughput computations rather than high-throughput experiments to save time and cost. Indeed, the idea of discovering and deploying advanced materials twice as fast and at a fraction of the cost is the cornerstone of the ambitious Materials Genome Initiative. This example for exploring the materials space offers a close integration between computation and experiment and significantly reduces the risks involved in an Edisonian way to innovation characterized by prolonged hunt-and-try cycles.

In our particular work, when practicing deep ESE by simulations we are required to describe and predict the properties of solids such as Si and diamond based on our understandings of the fundamentals of nature down to the atomic scale. When scaling microscopic theories up to explain actual materials, the difficulty emerges from the sheer number of particles (on the order of $O(10^{23})$ and the complexity of interactions among them. Luckily, advances in algorithms/theories and abundance in computational resources have enabled the prediction of a wide range of properties in recent decades without turning to parameter fitting or empirical modeling.

However, although the *ab initio* calculations such as those involving many-body corrections can provide an accurate evaluation of physical properties [49], the scope of such calculations is somewhat limited to about 10,000 strain points because of high computational cost [28]. In the abovementioned case study, by discretizing $\boldsymbol{\varepsilon}$ with a regular grid comprising 20 nodes separated at each 1% strain interval (a very coarse grid in the practice of deep ESE) over the strain range of −10% to +10%, the computational model would entail about 10^7 bandgaps. Even if we can use preexisting knowledge about crystal/strain symmetries to cross out some redundant cases, the amount of strain cases to be evaluated is still around millions, up to two orders of magnitude higher computational requirement than what can be reasonably achieved presently [28].

To overcome these difficulties, proper exploitation of the produced information is of paramount importance, a task that can be accomplished by machine learning (ML). The ML machinery has made significant inroads into many fields of materials science as a powerful tool for accurately and acceleratedly mapping out-of-equilibrium phases of matters [50], solving quantum many-body problems [51], decoding crystal structures [52,53], fitting interatomic potentials [54–56], mechanical properties extraction [57,58], and so forth. It helps to harvest a basic understanding of the physical factors underlying materials structures and properties and may give rise to the revelation of Matthiessen-like rules.

ML works that focus on the electronic properties of *undeformed* materials include bandgap and band structure fitting [59,60] for different material families (ABX_3 perovskites [61,62], elpasolite compounds [63], general inorganic solids [64–68], etc), Brillouin zone exploration [69], and bandgap prediction in 2D hybridized structures [70]. To the best of the author's knowledge, no ML efforts have been made so far in the field of deep ESE for semiconductor electronic properties in the 6D strain hyperspace.

In this thesis, we would present methods that combine ML and *ab initio* calculations to investigate how elemental semiconductors such as silicon and diamond alter their bandgap and band structure under general 6D deep elastic strains and identify energy-efficient pathways to engineer targeted FoMs. These methods invoke deep neural networks (NNs) and convolutional neural networks (CNNs) to assess, to a reasonable degree of accuracy, material properties as functions of strain based on a limited amount of *ab initio* data.

1.3. Thesis structure overview

Generally speaking, the ML model and workflow we aim to develop should work for any semiconductor. For the purpose of model development, we need to choose 1-2 representative material systems in which ESE is practiced. In the scope of this thesis, we primarily focus on two semiconductors, namely silicon and diamond, not only because they are commercialized or important, but also due to the near-$\boldsymbol{\varepsilon}_{\mathrm{ideal}}$ deformations already realized in the nanostructures of these semiconductors in real-world experiments (Figure 1.1), making them the most likely ones whose electronic properties can be engineered by deep elastic strains in the near future.

The rest of the thesis is as follows: we begin by briefly review in Chapter 2 the fundamental *ab initio* theories and ML methods we relied our study upon. Then, we introduce our deep NN model and CNN model in Chapter 3 and Chapter 4, respectively. In these two chapters, the new behavior and physics we discovered in silicon and diamond by using these models are discussed in detail. Next, we present in Chapter 5 joint machine learning-finite element simulation (ML-FEM) studies of electronic property engineering in experimentally feasible loading scenarios, followed by Chapter 6 where applications of the models in studying phonon and defect related properties are introduced. We conclude the thesis in Chapter 7 by identifying several the limitation of ML models and calling for close collaboration with experiments while practicing deep ESE.

Chapter 2. Fundamental Theory and Methodology

2.1. Deformation and strain measures

The semiconductors we studied are three-dimensional (3D) deformable bodies. We can denote the coordinate of a point in such an undeformed and homogeneously deformed body as $\mathbf{X}$ and $\mathbf{Y}$, respectively. A corresponding displacement vector $\mathbf{u}$ referenced to $\mathbf{X}$ can then be defined as $\mathbf{u} = \mathbf{Y} - \mathbf{X}$. Then the transformation of the material point from the undeformed to deformed state can be injectively (uniquely) described by a second-order deformation gradient tensor $\mathbf{F}$ as:

$$\mathbf{F} = \frac{\partial(\mathrm{X} + \mathbf{u})}{\partial \mathrm{X}} = \mathbf{I} + \frac{\partial \mathbf{u}}{\partial \mathrm{X}}. \tag{1}$$

This F eliminates the rigid-body translation upon which the material properties do not change. Through polar decomposition, $\mathbf{F}$ can be factored as:

$$\mathbf{F} = \mathbf{RU} = \mathbf{VR} \tag{2}$$

where $\mathbf{R} \in SO_3$ describes rigid body rotation, and $\mathbf{U}$ (or $\mathbf{V} = \mathbf{RUR^T}$) is the right (left) Cauchy-Green tensor.

It is noted that there are multiple strain ($\boldsymbol{\varepsilon}$) measures to evaluate a given deformation. There is a family of generalized strain measures, namely the Seth-Hill family of strain measures, that can be associated with the same $\mathbf{F}$:

$$\boldsymbol{\varepsilon}^{(m)} = \begin{cases} \frac{1}{m}(\mathbf{U}^m - \mathbf{I}), \text{if } m \neq 0; \\ \ln\mathbf{U}, \text{if } m = 0. \end{cases} \tag{3}$$

Some strain measures of special interest included in this family are:

i) $m = 0, \boldsymbol{\varepsilon}^{(\mathbf{0})} = \ln\mathbf{U}$ (the logarithmic strain, or more commonly known as the true strain)

ii) $m = 1, \boldsymbol{\varepsilon}^{(\mathbf{1})} = \mathbf{U} - \mathbf{I}$ (the Biot strain, or more commonly known as the nominal strain)

iii) $m = 2, \boldsymbol{\varepsilon}^{(\mathbf{2})} = \frac{1}{2}(\mathbf{U^2} - \mathbf{I}) = \frac{1}{2}(\mathbf{F^TF} - \mathbf{I})$ (the Green-Lagrange strain)

iv) $m = -2, \boldsymbol{\varepsilon}^{(-\mathbf{2})} = \frac{1}{2}(\mathbf{I} - \mathbf{U^{-2}}) = \frac{1}{2}(\mathbf{I} - \mathbf{F^{-T}F^{-1}})$ (the Euler-Almansi strain).

At small deformations, all the above strain measures are close to each other. At each configuration along a path of deformation, any infinitesimal increment of work (or power) per unit volume, dw (or $d\dot{w}$) by the traction on this volume can be expressed in terms of differential strain:

$$dw \;=\; \boldsymbol{\tau} : d\boldsymbol{\varepsilon} \text{ or } d\dot{w} = \boldsymbol{\tau} : d\dot{\boldsymbol{\varepsilon}} \tag{4}$$

where different choices of strain ($\boldsymbol{\varepsilon}$) measures in the above formulation give different work/power-conjugate stress ($\boldsymbol{\tau}$) measures. For example, the Green-Lagrange strain has the second Piola-Kirchhoff stress as its conjugate pair.

Furthermore, if there exist certain constraints that allow one to uniquely obtain **F** from $\boldsymbol{\varepsilon}$, then the deformation space becomes 6D [71]. For a deformed lattice, if the loading direction is along x_1 for uniaxial tension, or along x_1 on a plane orthogonal to x_3 for pure shear, we can further simplify the **F** tensor as an upper triangular matrix:

$$\mathbf{F} = \begin{bmatrix} \mathrm{F}_{11} & \mathrm{F}_{12} & \mathrm{F}_{13} \\ 0 & \mathrm{F}_{22} & \mathrm{F}_{23} \\ 0 & 0 & \mathrm{F}_{33} \end{bmatrix} \tag{5}$$

Then the Green-Lagrange strain $\boldsymbol{\varepsilon}^{(\mathbf{2})}$ can be written as:

$$\boldsymbol{\varepsilon}^{(\mathbf{2})} = \frac{1}{2}\begin{bmatrix} \mathrm{F}_{11}^2 - 1 & \mathrm{F}_{11}\mathrm{F}_{12} & \mathrm{F}_{11}\mathrm{F}_{13} \\ \mathrm{F}_{11}\mathrm{F}_{12} & \mathrm{F}_{12}^2 + \mathrm{F}_{22}^2 - 1 & \mathrm{F}_{12}\mathrm{F}_{13} + \mathrm{F}_{22}\mathrm{F}_{23} \\ \mathrm{F}_{11}\mathrm{F}_{13} & \mathrm{F}_{12}\mathrm{F}_{13} + \mathrm{F}_{22}\mathrm{F}_{23} & \mathrm{F}_{12}^2 + \mathrm{F}_{22}^2 + \mathrm{F}_{33}^2 - 1 \end{bmatrix} \tag{6}$$

Note there can be other unique deformation spaces [71]. For instance, **F** is usually constrained to be symmetric and has 6 degrees of freedom in performing atomistic simulation at constant stress [71]. In this case, **F** can be uniquely determined through $\mathbf{F} = \sqrt{1 + 2\,\boldsymbol{\varepsilon}^{(\mathbf{2})}}$.

2.2. Band structure and bandgap

Solutions to the Schrödinger equation for an (ideal) crystal such as pure silicon or diamond, can be expressed as a periodic modulation of a plane wave, according to the Bloch theorem:

$$\psi_{n\mathbf{k}}(\mathbf{r}) \;= e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}), \tag{7}$$

where $u_{n\mathbf{k}}(\mathbf{r}+\mathbf{R}) = u_{n\mathbf{k}}(\mathbf{r})$ for all **R** of the Bravais lattice is the periodic function. The corresponding energies $E_n(\mathbf{k})$ for each band index *n* are continuous functions in wave vector **k** and constitute the energy bands. The range of energy in a solid where no electronic states can exist is called bandgap (E_g). The value of E_g varies a lot across solids, resulting in different electrical, thermal, optical and magnetic characteristics.

Upon straining, the original O_h crystal point group of a diamond cubic lattice no longer holds and $E_n(\mathbf{k})$ and E_g vary accordingly. For example, the point group turns into D_{2h} under a three-normal-strain. The Brillouin zone for deformed Si, in this case, is shown in Figure 2.1. It is no longer a regular truncated octahedron with equilateral hexagonal and square faces. The reciprocal space lattice vectors are adjusted by the inverse transpose of the deformation gradient tensor in real space, i.e., $\mathbf{F}^{-\mathrm{T}}$, as a result of the deformation. We keep the Γ label for the center of a Brillouin zone. In undeformed Si or diamond, the centers of the square and regular hexagonal surfaces on the Brillouin zone boundary are degenerate and labeled as X and L, respectively. For comparison simplicity, we follow a similar notation scheme and still denote the 'X'-type points as the centers of the tetragon surfaces and 'L'-type points as the centers of the regular/non-regular hexagonal surfaces. The lines that connect the Γ point to the 'X'-type points are labeled as 'Δ'-type. This way, the six 'X'- and 'L'-type points, though non-degenerate, would keep the correct fractional coordinates of the ⟨0.5,0,0.5⟩- and ⟨0.5,0,0⟩-type, and the **k**-points along the Γ-'X' line would all have the ⟨ζ, 0, ζ⟩-type coordinates. These notations will be used repeatedly in the rest of this thesis.

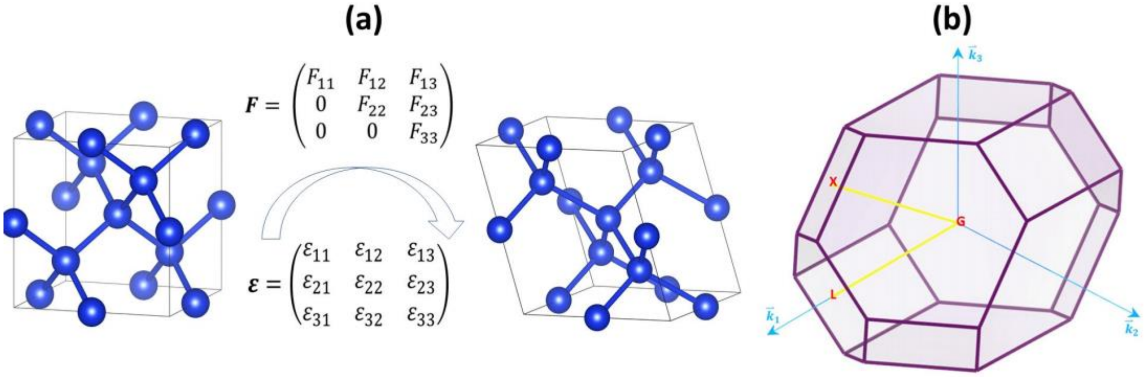

(a)
(b)
$F = \begin{pmatrix} F_{11} & F_{12} & F_{13} \\ 0 & F_{22} & F_{23} \\ 0 & 0 & F_{33} \end{pmatrix}$
$\varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{pmatrix}$
$\vec{k}_3$
X
G
L
$\vec{k}_1$
$\vec{k}_2$

Figure 2.1 Primitive cell under elastic strain. (a) Deep ESE is achieved by applying a reduced deformation gradient tensor to the undeformed diamond cubic lattice of Si or C in the real space. (b) Brillouin zone of diamond cubic crystal under a general strain. It is a tetradecahedron with 8 hexagonal and 6 quadrilateral faces. The discussions in the subsequent sections of this thesis will incorporate the same labels and **k** coordinates as shown here. The so-called ⟨ζ, 0, ζ⟩-type coordinates include ⟨ζ, 0, ζ⟩, ⟨ζ, ζ, 0⟩, and ⟨0, ζ, ζ⟩.

2.3. First-principles calculations

The evaluation of the properties of semiconductors in our study greatly relies on *ab initio* calculations. This section briefly reviews the DFT and many-body GW approximation method which we used for high-throughput high-fidelity electronic structure calculations.

2.3.1. Density functional theory

In order to understand many of the microscopic properties of a system, one can jot down the total Hamiltonian of all the nuclei and electrons in the Hartree atomic unit system as:

$$H = \left(\frac{1}{2} \sum_{j \neq j'} \frac{Z_j Z_{j'}}{|\mathbf{R}_j - \mathbf{R}_{j'}|} + \frac{1}{2} \sum_{i \neq i'} \frac{1}{|\mathbf{r}_i - \mathbf{r}_{i'}|} + \frac{1}{2} \sum_{i,j} \frac{-Z_j}{|\mathbf{r}_i - \mathbf{R}_j|} \right) + \left(\sum_j \frac{\mathbf{P}_j^2}{2M_j} + \sum_i \frac{\mathbf{p}_i^2}{2} \right) \tag{8}$$

in which Z and M denote nucleus charge and mass, $\mathbf{R}$ and $\mathbf{P}$ denote nucleus position and momentum, and $\mathbf{r}$ and $\mathbf{p}$ denote electron position and momentum. i and j enumerate all the electrons and nuclei, respectively. The Hamiltonian is expressed on the right-hand side as the summation of potential (first bracket) and kinetic energy (second bracket) of all particles. Specifically, the first three potential terms, from left to right, represent the totaling of nucleus-nucleus, electron-electron, and nucleus-electron pairwise electrostatic interactions, respectively. The last two terms describe the kinetic energy of all the nuclei and electrons, respectively. Applying the adiabatic or Born-Oppenheimer approximation, on the ground that the mass of a nucleus is much larger than the electron mass and the nuclei can be considered as immobile relative to the electrons, (8) can be simplified to:

$$H_{\text{elec}} = \frac{1}{2} \sum_{i \neq i'} \frac{1}{|\mathbf{r}_i - \mathbf{r}_{i'}|} + \sum_i \sum_j \frac{-Z_j}{|\mathbf{r}_i - \mathbf{R}_j|} + \sum_i \frac{\mathbf{p}_i^2}{2} \tag{9}$$

which satisfies the time-independent Schrodinger equation with N-electron wave function $\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$:

$$H_{\text{elec}}\Psi = E\Psi. \tag{10}$$

It is noted that the formulation above neglects relativistic effects and one can resort to the Dirac equation for the incorporation of spin orbit effects.

According to the Hohenberg-Kohn theorems, the external potential, and hence the total energy, for a system is a unique functional of the electron density $\rho(\mathbf{r})$. Specifically, the total energy E can be expressed as a functional of ρ as

$$E[\rho(\mathbf{r})] = E_{\text{T}}[\rho(\mathbf{r})] + E_{\text{H}}[\rho(\mathbf{r})] + \int d\mathbf{r} \rho(\mathbf{r}) V_{\text{ext}}(\mathbf{r}) + E_{\text{xc}}[\rho(\mathbf{r})], \tag{11}$$

where E_{T} is the kinetic energy, $E_{\text{H}} = \frac{1}{2} \int d\mathbf{r} \int d\mathbf{r}' \frac{\rho(\mathbf{r})\rho'(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}$ is the Coulomb (or Hartree) energy, V_{ext} is the external potential (if the Hamiltonian is separated into ion core and valence electrons, V_{ext} represents the external potential due to the ion), and E_{xc} is the exchange-correlation energy.

Compared to the wave function $\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$, $\rho(\mathbf{r})$ is only a single-variable function that is much more trackable. Also, H_{elec} and ground state energy can be determined through minimization of $E[\rho(\mathbf{r})]$. In other words, ground state energy can be acquired variationally: the density that minimizes the total energy is the ground state density.

To evaluate $E[\rho]$, Kohn and Sham developed a practical approach [72]. In their formulation, the ground state charge density of a fictitious system with non-interacting particles moving in an “effective” potential ($V_{\mathrm{eff}}(\mathbf{r})$) is constructed as

$$\rho(\boldsymbol{r}) = \sum_{i} |\varphi_i(\mathbf{r})|^2, \tag{12}$$

where $\varphi_i(\mathbf{r})$ are a set of Kohn-Sham orbitals. Varying $V_{\mathrm{eff}}(\mathbf{r})$ so that $\rho(\mathbf{r})$ minimizes the total ground state energy of the interacting system subject to the constraints on these $\varphi_i(\mathbf{r})$ leads to a set of one-electron Schrodinger equations, known as the Kohn-Sham equations, that can be solved for $\varphi_i(\mathbf{r})$:

$$\left(-\frac{\nabla^2}{2} + V_{\mathrm{eff}}(\mathbf{r})\right)\varphi_i(\mathbf{r}) = E_i\varphi_i(\mathbf{r}). \tag{13}$$

It is an eigenvalue equation with E_i being the eigenvalues for the corresponding $\varphi_i(\boldsymbol{r})$ and

$$V_{\mathrm{eff}}(\mathbf{r}) = \int d\boldsymbol{r}' \frac{\rho'(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} + V_{\mathrm{ext}}(\mathbf{r}) + \frac{\delta E_{\mathrm{xc}}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})}. \tag{14}$$

The functional derivative of E_{xc} is the exchange-correlation potential.

Many options of density functional approximations have been proposed for E_{xc} (Jacob’s ladder). The very first approach is the local density approximation (LDA) where $E_{\mathrm{xc}}[\rho]$ is expressed in terms of the exchange-correlation energy per particle of a uniform electron gas ($\epsilon_{\mathrm{XC}}^{\mathrm{gas}}$) with electron density $\rho(\mathbf{r})$:

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = \int d\mathbf{r}\rho(\mathbf{r})\epsilon_{\mathrm{XC}}^{\mathrm{gas}}(\rho(\mathbf{r})). \tag{15}$$

Building upon the LDA approximation to incorporate the local density gradient in expressing E_{xc} is the generalized gradient approximation (GGA):

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[\rho] = \int d\mathbf{r}\rho(\mathbf{r})\epsilon_{\mathrm{XC}}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})). \tag{16}$$

This thesis adopted the GGA method developed by Perdew, Burke, and Ernzerhof (PBE) for most of the ground state DFT calculations and as the starting point for excited-states calculations.

We can see from (14) that the Hartree term and the exchange-correlation term rely on $\rho(\mathbf{r})$ and hence on the Kohn-Sham orbitals by (12). But these $\varphi_i(\mathbf{r})$ are what we are looking for when trying to solve (13). Therefore, an iterative process, known as the self-consistent field (SCF) procedure, is adopted as follows (Figure 2.2):

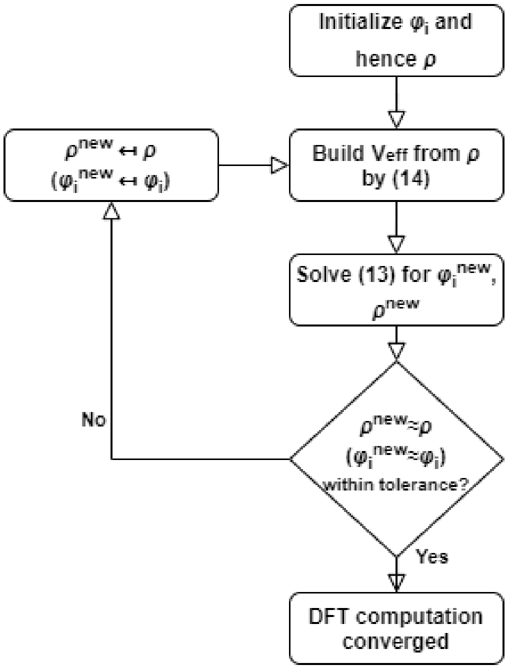

Initialize φ_i and hence ρ
$\rho^{new} \leftarrow \rho$ $(\varphi_i^{new} \leftarrow \varphi_i)$
Build V_{eff} from ρ by (14)
Solve (13) for φ_i^{new}, ρ^{new}
No
$\rho^{new} \approx \rho$ $(\varphi_i^{new} \approx \varphi_i)$ within tolerance?
Yes
DFT computation converged

Figure 2.2 The iterative process for SCF calculation

2.3.2. The GW approximation

It is important to recognize that the *N* single-electron states φ_i correspond to made-up particles having the same charge density as the interacting system and should not be assigned to a particular physical meaning. When an electron is moved to an unoccupied state with higher energy, a quasiparticle is formed. The misinterpretation of these E_i as the quasiparticle energies is the cause of the electronic bandgap underestimation problem [73]. This underestimation in DFT calculation within GGA can be as much as 0.4-0.5 eV for silicon. Considering the intrinsic bandgap of silicon is measured to be 1.1 eV in experiments, this error brought by ground-state DFT is too large for our ESE studies, and we adopt a theory based on Green's function to describe electron excitations in solids. This theory can better account for electron excitation than plain DFT since its formulation involves the excited state energies of the interacting electron systems with *N*-1 and *N*+1 electrons.

We can introduce a self-energy ($\sum$) to determine such a Green's function. With the system described as a polarizable medium, GW approximation [49] (G for Green's function and W for screened Coulomb interaction) can be used to express $\sum$. In this thesis, we primarily use the single-shot G_0W_0 approach to evaluate $\sum$:

$$\sum \approx i\mathrm{GW}. \tag{17}$$

This is the lowest order term in an expansion of $\sum$ in W. The GW approximation is very popular for the calculation of band structures in solids [74], which is the main focus of our research. We did not in our work adopt the fully self-consistent solution not only because it is computationally demanding, but it also sometimes does not yield the closest quasiparticle energy compared with experimental results.

2.4. Machine learning

2.4.1. Basic theory and concept

A basic pursuit of science is the formulation of theories from existing knowledge to make predictions that are empirically verified to be accurate. In most cases, the predictive theory can be expressed by a mathematical mapping G which takes in a group of input x to fit some output t. Here, G can be fairly simple, as in using the Coulomb potential for charged defects assigned with Kröger-Vink formal charges, or it may be slightly delicate with more information incorporated, as in the same model but with an account for the Debye screening, or it may be demandingly complex, as in models that describe a nuclear fusion. There are no restrictions on the dimensions of x and t. The collection of the known outputs and inputs is the *training set*. The goal of a researcher is to propose a function relationship that can determine the output value for another set of x and t not used during the training. If the values of t are from a discrete set (failure modes, phonon stability, material phases, etc.), the process of proposing a G is called *classification*. If the value of t is a continuous variable (bandgap, effective mass, strain energy density, etc.), the process is then known as *regression*. Supervised machine learning is nothing different from the conventional scientific query except that the predictive theory (or the constitutive equation) is learned and determined by a computer rather than a human, which of course, saves the latter much time.

In most, if not all, situations, a learner may exhaust the hypothesis space and still cannot come up with a function or a model that describes all the current data 100% accurately and meanwhile predicts any other newly incoming data 100% accurately. In these cases, the predictions are expressed as $t = c + G(x)$ where G is the function guessed (or, in a more scientific term, summarized) by the learner - be it a machine or a human researcher - to describe the materials science phenomena at hand and c is a random error or loss. Therefore, the goal of any model development and certainly our development of a strain model boils down to minimize the loss. That is to say, we would propose a model whenever be used to predict a material-related value under a particular deformation, say 10% uniaxial compression, a reasonably small deviation from the ground true value computed by the first-principles method can be achieved.

There are certain ML outcomes that we surely want to avoid in numerical research. They are *underfitting* and *overfitting*, both of which mean the model cannot generalize well. Underfitting (high bias) is the case in which c is still very large, indicating the model has yet learned much from the training dataset. Overfitting (high variance) is the case where c is extremely small, but the F is unreliable. The aim is to find the trend rather than fit a line to cut through all the data points. This is due to the model learning “too much” from the training dataset. One can think of a normal computer program as utterly overfitted, since it only provides answers robotically to the inputs they are trained (programmed) for but not others. Many numerical *regularization* techniques have been developed by the ML community to suppress these issues and have the right bias-variance trade-off, among which we used dropout and weight regularizations in our study.

Active learning and uncertainty estimation

Active learning entails a class of ML algorithms for the automatic assembly of the training set. Despite stunning accuracies in training, deep learning algorithms may not be good at quantifying the uncertainty of model predictions. Such failure may cause problems especially in risk-sensitive areas [75]. In this thesis, 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 [76]. The particular choice of an uncertainty quantification procedure greatly influences performance in the active learning part. There are three main ways to perform an uncertainty estimation for the NN: ensembling [77], variational inference [78], and dropout-based inference [79].

Ensembling is a mere stacking of a few similar models, which are trained starting from different parameters initialization or on the different subsets of data; various modifications exist [80,81]. This method is not model-specific and yet achieves near state-of-the-art performance in some applications [82]. The main drawback is that one needs to train a number of models and the memory consumption, training and inference time are scaling proportionally: an ensemble of 10 models will take 10 times more computational resources, which are used only to produce an uncertainty estimate.

Variational inference is a standard Bayesian technique, which relies on the stochasticity incorporated within the model in a form of the model's parameters being the random variables [78]. This method produces robust and theoretically bounded uncertainty estimates; however, the model needs to be constructed and trained in a special way, which increases the training time. Moreover, the variational inference procedure is infeasible for the case of a large number of parameters and large datasets without special assumptions and approximations [83].

The utilization of dropout [84,85] as an “engineering” way of the model regularization led to better results in the number of ML areas. One of the ways to interpret the dropout is Bayesian [79]. Dropout-based inference may be seen as a Bayesian approximation of the variance of the ML output: in our study, to get the uncertainty estimate for a given strain one needs to enable the dropout during the inference time and then calculate the variance of the few stochastic passes of the same sample through the model.

Data fusion

As defined in the monograph “Data Fusion Lexicon” by Franklin E. White [86], data fusion is “a process dealing with the combination of data and information from multiple sources to achieve refined position and identity estimates.” In our study, through comparing the DFT-PBE calculation results with experimental measurement, we realized the need for additional data obtained from a different and more accurate level of theory (the GW calculation) to achieve an improved dataset. The joint force of both datasets allows us to achieve improved information with less error. The superiority brought by fusion is obvious, as will be shown in detail in Chapter 3 and Chapter 4.

2.4.2. A general formulation for elastic strain engineering by supervised learning

In formulating the problem from a statistical learning point of view, we can define an input data – strain tensor $\boldsymbol{\varepsilon} \in D \subset R^6$, as well as output data – electronic band structure represented as energy eigenvalues in a predefined **k**-mesh $\epsilon \in R^d$. One part of the problem is to predict the mapping

$$f: D \rightarrow R^d, \tag{18}$$

using some training data $S_{\text{train}} = \{(\boldsymbol{\varepsilon}^{(1)}, \epsilon^{(1)}), (\boldsymbol{\varepsilon}^{(2)}, \epsilon^{(2)}), \dots, (\boldsymbol{\varepsilon}^{(n)}, \epsilon^{(n)})\}$ and approximation model $\hat{f}: D \rightarrow R^d$. We can access the quality of the approximation via the loss function

$$L(\hat{f}|S) = \frac{1}{|S|} \sum_{\boldsymbol{\varepsilon}^{(j)} \in S} ||\hat{f}(\boldsymbol{\varepsilon}^{(j)}) - f(\boldsymbol{\varepsilon}^{(j)})||_2 , \tag{19}$$

and thus formulate the fitting problem as an optimization problem

$$L(\hat{f}|S_{\text{train}}) \rightarrow \text{min.} \tag{20}$$

Theoretically, the mapping (1) is piecewise smooth and is known to contain some symmetries. However, the practical evaluation of the function f available to us via various DFT approximations f_{PBE} and f_{GW} differs by accuracy and evaluation time. In general, GW calculations are more precise due to the advanced level of many-body perturbation theory involved and we thus refer to them as the “ground truth”. Meanwhile, PBE approximation turns out to be useful in the early training stages, since we can obtain large amounts of reasonable data to pre-train our model on. The output dimensionality d depends on the accuracy one possesses. We take advantage of the $8 \times 8 \times 8$ Monkhorst-Pack **k**-point mesh and only 4 energy bands: valence band (VB), conduction band (CB), and their nearest-neighbor bands, respectively, if we were to describe the energy dispersions near the Fermi level of a semiconductor.

Band structure calculations take time. Since we can access the answer $f_{\text{GW}}(\boldsymbol{\varepsilon})$ for an arbitrary strain $\boldsymbol{\varepsilon}$, we can take advantage of it and perform a smart design. We refer to the initial dataset we obtained as $S_{\text{init}} = \{(\boldsymbol{\varepsilon}^{(1)}, \epsilon^{(1)}), (\boldsymbol{\varepsilon}^{(2)}, \epsilon^{(2)}), \dots, (\boldsymbol{\varepsilon}^{(N_{\text{init}})}, \epsilon^{(N_{\text{init}})})\}$, and the unlabeled pool set as $S_{\text{pool}} = \{\boldsymbol{\varepsilon}^{(1)}, \boldsymbol{\varepsilon}^{(2)}, \dots, \boldsymbol{\varepsilon}^{(N_{\text{pool}})}\} \subset D$. We do not know the answers $\epsilon^{(1)}, \epsilon^{(2)}, \dots, \epsilon^{(N_{\text{pool}})}$ from the S_{pool} in advance but we can obtain them using the oracle model f_{GW} in exchange for computational resources spent. We will also need a separate test set $S_{\text{test}} = \{(\boldsymbol{\varepsilon}^{(1)}, \epsilon^{(1)}), (\boldsymbol{\varepsilon}^{(2)}, \epsilon^{(2)}), \dots, (\boldsymbol{\varepsilon}^{(N_{\text{test}})}, \epsilon^{(N_{\text{test}})})\} \subset D$ to access the quality of the approximation.

The idea of active learning is to select some samples $S_{\text{selected}} \subset S_{\text{pool}}$ using an uncertainty estimator

$$UE(\hat{f}, \cdot): D \to R_{+}, \tag{21}$$

which ranks input data from the S_{pool}. Those rankings are processed via an acquisition function

$$A(\hat{f}, \cdot, UE): S_{\text{pool}} \to R_{+}. \tag{22}$$

In most cases, we just select the samples with the largest uncertainty estimates: $A(\hat{f}, \cdot, UE): S_{\text{pool}} \to UE(S_{\text{pool}})$. However, to save time spent on the *ab initio* calculations, we would like to sample as few points as possible. If we define the desirable accuracy as δ, then we can formulate an active learning problem as an optimization one:

$$|S_{\text{selected}}| \to \min \quad \text{subject to } L(\hat{f} | S_{\text{test}}) < \delta. \tag{23}$$

While the process of obtaining the solution of the optimization problem (19) is straightforward for given data S_{train}, a rigorous solution to (22) is time-consuming since one needs to explore all the possible subsets of S_{train}. Here we will make use of a greedy approach, training the model on S_{train} and then taking a small subset from S_{pool} to calculate using an *ab initio* model. Uncertainty estimation procedure helps us to select the samples that are somewhat hard for the model; together with the calculated answers we add additional data to S_{train} and the process starts again, with an increased training set. This procedure aims at minimization of the selected dataset S_{selected} in a greedy way, checking an actual error on the test set S_{test} after each iteration.

2.4.3. Models included in this work

This section reviews the various ML models we adopted in our ESE work, including feed-forward NN, CNN, ensemble-tree methods, and k-nearest neighbor classification method.

Feed-forward NN & CNN

As arguably the most successful and powerful method of ML, artificial NNs have been permeating the field of materials science introducing a research paradigm where large data are exploited to meet the satisfaction of empirical modeling. As shown in Figure 2.3a, NN is based on a collection of linked entities called neurons. The connections between neurons can send “signals” from one to another, similar to biological synapses. The input signal to a specific neuron may come from multiple other neurons and is mathematically represented as a weighted sum $z = w^T x + b$, where b is called *bias*. After being rectified by an activation function (in Figure 2.3a, a step function), the signal is passed on to the next neuron. This is the simplest NN architecture known as the *perceptron*. In a typical NN training for predicting strain effect, there is a network of neurons like this forming intermediate hidden layers, as depicted in Figure 2.3b.

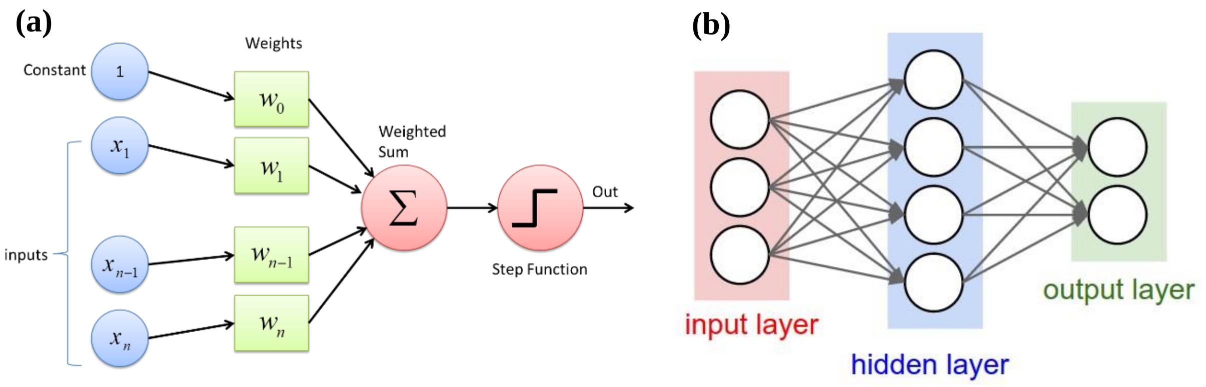


Figure 2.3 (a) Basic architecture of a perceptron. (b) Illustration of an NN with a hidden layer (multilayer perceptron).

In order to effectively minimize the loss of the NN model, i.e., to reduce the *c* in equation $t = c + G(x)$, the weights of the NN are sequentially adjusted by a backpropagation algorithm to make signal transfer along some neuron pathways stronger than the other. This learning process is bound to yield a well-optimized model capable of accounting for huge numbers of salient features in its hidden layers as introduced in other structure-property fitting works [56,87–89].

It is noticed that in a multilayer perceptron each neuron is fully connected to all neurons in the subsequent layer (Figure 2.3), making it prone to overfitting (Section 2.4.1). Instead of “barbering” the connectivity through dropout, CNNs adopt a different regularization approach by exploiting the hierarchical pattern in data. In feed-forward NN, the hidden layers are all the middle layers between the input and output layer, whereas in CNN hidden layers refer to the layers that conduct convolutions, as depicted in Figure 2.4. CNN is regarded by the ML community as a powerful tool for extracting features out of an image-type object. In our ESE of properties of materials, similar hierarchies exist in band structures, which can be treated as a stacked 3D image (details can be found in Chapter 4).

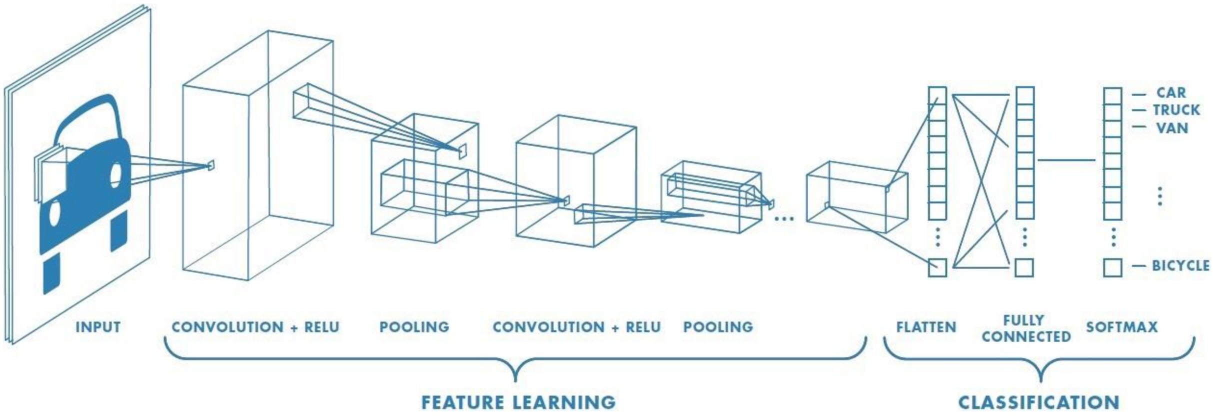


Figure 2.4 A general CNN architecture for classification. Adapted from Ref. [90].

Ensemble-based methods

Two popular ensemble methods, random forests regression (RFR) and gradient boosting regression (GBR), are also used in learning the bandgap variations under large elastic strains. They predict (regression or classification) by combining the outputs from individual trees, as shown in Figure 2.5. These two methods differ in the way the trees are built.

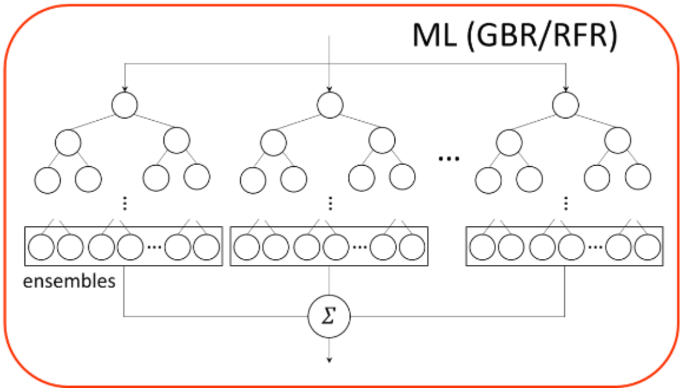


Figure 2.5 A schematic of the ensemble-tree architecture of the adopted GBR/RFR.

Chapter 3. Deep elastic strain engineering of semiconductor electronic properties by neural networks

3.1. Chapter introduction

In previous studies, bandgap engineering for semiconductors (mainly Si) used in functional devices was conducted largely by tuning only one or two strain components [91]. However the optimal FoM of a material may be located in a more general place in the 6D hyperspace, where a combination of normal and shear components exist. The author and his collaborators aim to explore this 6D strain space by analyzing highly nonlinear relations between electronic properties and the strain tensor.

In the subsequent sections, a deep NN model capable of learning the electronic band structure and bandgap of silicon from ML through a limited amount of calculations is introduced. The resultant accuracies of and techniques incorporated into the ML model are presented in Section 3.2. Different strains may result in the same bandgap, and in seeking a specific bandgap, or any other material FoM, one should choose the strain with a minimal effort required given the non-uniqueness of choice of a given target property or FoM. For this purpose, the density of states of bandgap envelope introduced in Section 3.3 of this chapter is important in understanding and utilizing deep ESE. In this work, the elastic strain energy density is used as a scalar metric or "norm" of the strain tensor for rationally choosing the ESE route that requires the least energy penalty and corresponds to the safest deformation manner in principle. For example, the model is demonstrated to locate the most energy-efficient pathway in the 6D strain hyperspace to transform silicon from a semiconductor to metal or specific deforming ways in a low-dimensional strain space to convert silicon's bandgap from indirect to direct (Section 3.4).

Note: some argumentation and figures/tables in this chapter are directly taken from the author's publication of Ref. [28]: Shi et al, *Deep Elastic Strain Engineering of Bandgap through Machine Learning*, Proc. Natl. Acad. Sci. **116**, 4117 (2019).

3.2. Machine learning models and outcomes

We aim to describe the electronic bandgap and band structure as functions of strain by training ML models on DFT data. This approach leads to reasonably accurate training with much fewer computed data than fine-grid *ab initio* calculations and a fast evaluation time. The DFT calculations were conducted in two settings: a large computationally inexpensive DFT-PBE dataset obtained for fitting and a small but accurate GW dataset for correction.

As depicted in Figure 3.1a, the strain tensor and/or the **k**-point coordinates are fed into different ML models as input to fit or make predictions about energy eigenvalues or bandgap. Figure 3.1b demonstrates the accuracy of these models on the PBE data, the best of which is attained by the NN. The data fusion technique [92,93] is adopted to further improve the learning outcome of bandgap, namely the most technically important property for electronic material. More specifically, given $E_{\mathrm{g}}^{\mathrm{PBE}}$ computed using an approximate baseline level of theory (PBE) at a particular query strain case, a related $E_{\mathrm{g}}^{\mathrm{GW}}$ value corresponding to a more accurate and more demanding target level of theory (GW) can be estimated as a function of both $E_{\mathrm{g}}^{\mathrm{PBE}}$ and $\boldsymbol{\varepsilon}$. Therefore, the $E_{\mathrm{g}}^{\mathrm{GW}}$ consistent with the query strain case is learned using exclusively $\boldsymbol{\varepsilon}$ and $E_{\mathrm{g}}^{\mathrm{PBE}}$ as input, as illustrated in Figure 3.1b and Table 3.1. The resulting data fusion model reduces the mean absolute error (MAE) in the prediction of bandgap by more than half for kernel-based ensemble methods and allows the bandgap predicted by NN be reach an extremely high accuracy of 8 meV, as shown in Figure 3.1b and Table 3.2. We also acknowledge that features such as bond length and bond angle can be used as the ML input. However, the change in lengths and angles depends on the orientation of such bonds relative to the strain directions, and the combination over all the orientations becomes a more complex descriptor (higher-dimensional) than the 6D strain. The current approach would work in general regardless of the number of atoms in the unit cell of any material system, as it depends on the number of bands and the **k**-mesh density only. For these reasons, we did not need to directly featurize bond information in the present work.

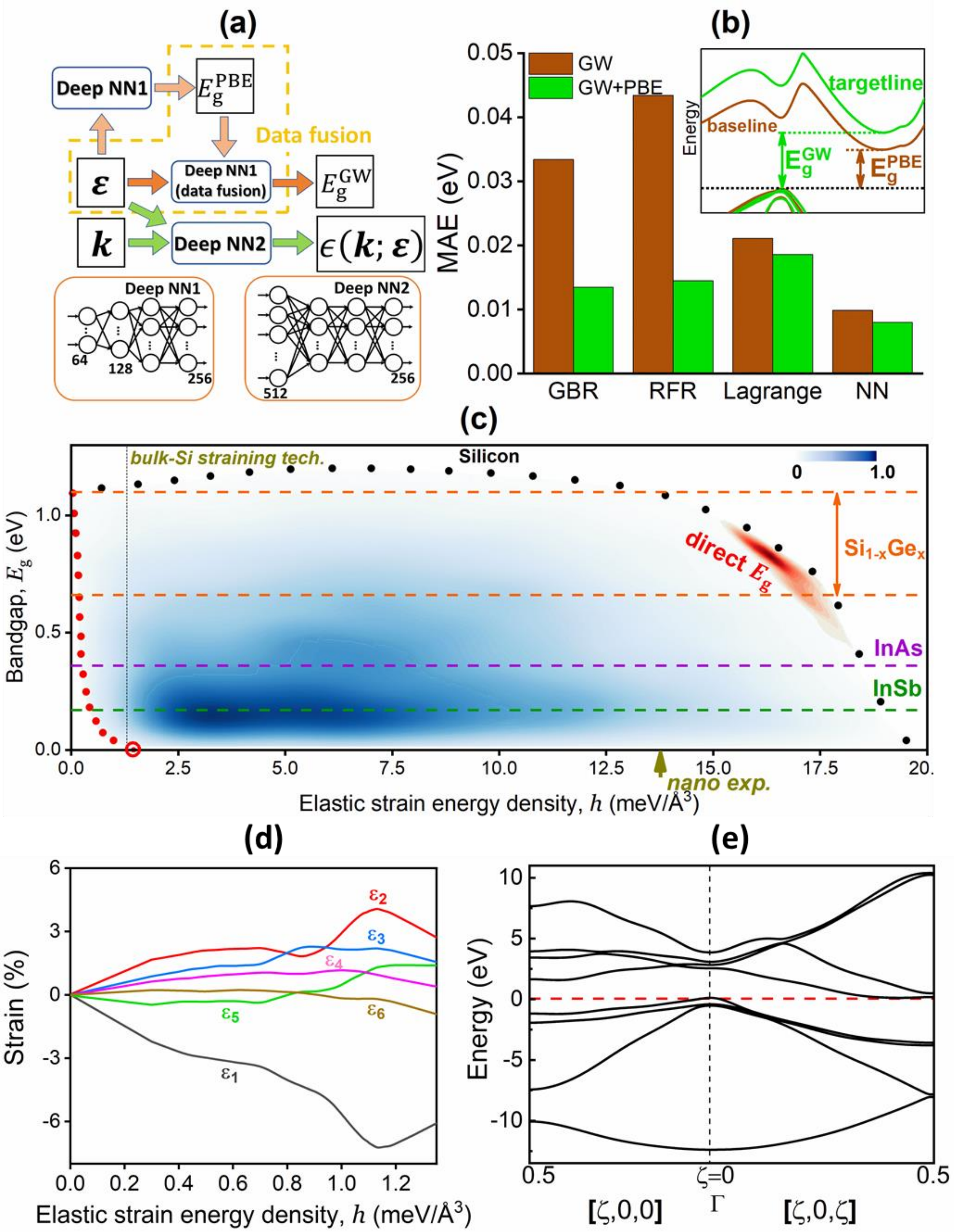


Figure 3.1 ML model and outcomes. (a) ML workflow for NN fitting. For a typical bandgap prediction task, the input contains the strain information and the target is either E_g^{PBE} or E_g^{GW}, depending on which dataset we use. In the data fusion process, the bandgap predicted from fitting the PBE dataset is also taken in as an input to fit the GW bandgap. For the whole band structure fitting task, the input contains both strain information and the $\mathbf{k}$-point coordinates and the target is the energy dispersion $\epsilon_n(\mathbf{k};\boldsymbol{\varepsilon})$, where n is the band index, $\mathbf{k}$ is the wavevector and $\boldsymbol{\varepsilon}$ is the crystal strain tensor. The hidden layer structures of the two associated deep NNs are also depicted. (b) Better bandgap fitting results measured by MAE are yielded by data fusion compared to the sole use of $\boldsymbol{\varepsilon}$ as input to fit

GW data. Inset: data fusion-based learning of the difference between $E_{\mathrm{g}}^{\mathrm{PBE}}$ and $E_{\mathrm{g}}^{\mathrm{GW}}$. Ensemble methods on decision tree classifiers including gradient boosting regression (GBR) and random forest regression (RFR), Lagrange interpolation and NN are adopted for ML fitting. (c) Reachable bandgap values for various h within the whole deformation space for silicon. The region where the strained silicon has a direct bandgap is colored in red. The red circle on the horizontal axis indicates the lowest energy penalty for the semiconductor-to-metal transition. The arrow on the horizontal axes in (c) indicates reachable h by *in-situ* experiment. (d) The most energy-efficient strain pathway ($\varepsilon_1 \equiv \varepsilon_{11}, \varepsilon_2 \equiv \varepsilon_{22}, \varepsilon_3 \equiv \varepsilon_{33}, \varepsilon_4 \equiv \varepsilon_{23}, \varepsilon_5 \equiv \varepsilon_{13}, \varepsilon_6 \equiv \varepsilon_{12}$) to reach the zero-bandgap state, i.e. the lower-envelope function $E_{\mathrm{g}}^{\mathrm{lower}}(h)$ in silicon corresponding to the red-dotted line in (c). (e) GW band structure associated with the 0-eV bandgap state. The fractional coordinates for the three high-symmetry points along the selected $\mathbf{k}$-path are (0.5, 0, 0), (0, 0, 0) and (0.5, 0, 0.5), respectively.

Table 3.1 MAE and RMSE (in units of eV) for ML algorithms for bandgap prediction with or without data fusion. Here, the Lagrange polynomial of degree 8 is used. Relative error: the norm of the difference between the true value and the prediction divided by the norm of the true value.

ML algorithms	GW		GW+PBE	
	MAE	RMSE	MAE	RMSE
Lagrange	0.0211	0.0274	0.0186	0.0241
GBR	0.0334	0.0521	0.0135	0.0209
RFR	0.0434	0.0596	0.0145	0.0215
NN	0.0099	0.0144	0.0080	0.0118
NN relative error	1.72%	2.78%	1.38%	2.05%

Table 3.2 Root mean squared error (RMSE) for various ML algorithms for the bandgap and band structure prediction tasks from PBE data for silicon (in units of eV). $\boldsymbol{\varepsilon}^{\mathrm{normal}}$ and $\boldsymbol{\varepsilon}^{\mathrm{6D}}$ denote three-normal-strains and general deformation cases, respectively. Lagrange polynomial of degree 9 is used. For all the details on ML and DFT settings, see Section 3.5 Technical details.

ML input		ML algorithms				ML target
		Lagrange	GBR	RFR	NN	
$\boldsymbol{\varepsilon}^{\mathrm{normal}}$		0.0150	0.0367	0.0247	0.0049	E_{g}
$\boldsymbol{\varepsilon}^{\mathrm{6D}}$		-	0.0743	0.0781	0.0264	
$\mathbf{k}$ and $\boldsymbol{\varepsilon}^{\mathrm{6D}}$	VB	-	0.1125	0.1078	0.0131	$\epsilon_n(\mathbf{k};\boldsymbol{\varepsilon})$
	CB	-	0.1593	0.1555	0.0184	

We next show that our NN-based surrogate models can successfully learn from several datasets and assimilate them. This capability is becoming increasingly important with the spread of materials property databases that collect data from different studies [94,95]. The incremental training of the NN starts from the same weights but is done on the extended dataset with the additional data included. We also increase the learning rate of the stochastic gradient descent algorithm and regularizers (dropout rate and weight regularization [96]) to circumvent limitations arising from the same local minima of the loss function established during the training on the initial dataset. This allows the model to not only handle additional training on the incoming data appended to a database but do it much faster than from scratch.

Table 3.3 Si bandgap prediction errors, RMSE and MAE (in units of eV), for the incremental fitting scenario on reduced datasets. The error in both metrics is reduced for both $\boldsymbol{\varepsilon}^{\mathrm{normal}}$ and $\boldsymbol{\varepsilon}^{\mathrm{6D}}$ datasets after the incremental fitting.

	$\boldsymbol{\varepsilon}^{\mathrm{normal}}$		$\boldsymbol{\varepsilon}^{\mathrm{6D}}$	
	before	after	before	after
RMSE	0.0403	0.0069	0.0264	0.0253
MAE	0.0167	0.0052	0.0179	0.0167

Numerical experiments conducted on the NN model demonstrate that incremental fitting of the models effectively reduces the error on a new dataset, see Table 3.3. Such incrementally fitted models are, thus, equally applicable to the bandgap approximation and various optimization tasks. Moreover, these models may be reused when shifting to other materials such as Ge or GaAs, since the implicit insights about symmetries, transitions and extreme cases are stored in the parameters of NN. Training the model for the other material starting from the weights for Si would significantly reduce the time and amount of data needed due to knowledge transfer, also referred to as transfer learning [97], leading to the rapid development of versatile surrogate models for deep ESE.

3.3. “Density of states” of bandgap

In ESE experiments, the objective is to identify the highest or lowest bandgap that can be achieved through the expenditure of a certain elastic strain energy density (h) defined as:

$$h(\boldsymbol{\varepsilon}) \equiv \frac{E(\boldsymbol{\varepsilon}) - E^0}{V^0}, \tag{24}$$

where $E(\boldsymbol{\varepsilon})$ is the total energy of the cell deformed by strain $\boldsymbol{\varepsilon}$, and E^0 and V^0 are the total energy and volume of the undeformed cell. Here, we data-mine the 6D deformation by machine learning the bandgap distribution and the elastic strain energy density against $\boldsymbol{\varepsilon}$. The many-to-many relation between $h(\boldsymbol{\varepsilon})$ and the bandgap $E_{\mathrm{g}}(\boldsymbol{\varepsilon})$ is shown in Figure 3.1c. In the stress-free equilibrium state, silicon has a bandgap of 1.1 eV; with an increase in strain energy density, a variety of possible bandgaps emerge. Even silicon with strain energy density as small as 0.5 meV/Å^3 can become quite a different material from the stress-free silicon. As h further increases, the largest allowable bandgap drops and an “envelope” forms, as evidenced by the change of maximal and minimal bandgap reachable under a fixed h. The shading of the envelope regions in Figure 3.1c reflects the distribution of the available bandgap. A darker shading qualitatively indicates that the number of possible strains to achieve a specific bandgap at a given h is higher. Outside the envelope the shading color is white, meaning that the corresponding bandgap is not attainable. Mathematically, we can define the cumulative “density of states” of bandgap as:

$$
\begin{aligned}
c(E_{\mathrm{g}}{}'; h') \ \equiv \ & \int_{h(\boldsymbol{\varepsilon})<h\prime} d^6\boldsymbol{\varepsilon}\delta(E_{\mathrm{g}}{}' - E_{\mathrm{g}}(\boldsymbol{\varepsilon})) \\
& = \int d^6\boldsymbol{\varepsilon}\delta(E_{\mathrm{g}}{}' - E_{\mathrm{g}}(\boldsymbol{\varepsilon}))\Theta(h' - h(\boldsymbol{\varepsilon}))
\end{aligned}
\tag{25}
$$

where $d^6\boldsymbol{\varepsilon} \equiv d\varepsilon_1 d\varepsilon_2 d\varepsilon_3 d\varepsilon_4 d\varepsilon_5 d\varepsilon_6$ in the 6D strain-space, $\delta(\cdot)$ is the Dirac delta function, and $\Theta(\cdot)$ is the Heaviside step function. We then define the "density of states" of bandgap at h' by taking the derivative of $c(E_{\mathrm{g}}{}'; h')$ with respect to h':

$$
\rho(E_{\mathrm{g}}{}'; h') \ \equiv \ \frac{\partial c(E_{\mathrm{g}}{}'; h')}{\partial h'} \ = \ \int d^6\boldsymbol{\varepsilon}\delta(E_{\mathrm{g}}{}' - E_{\mathrm{g}}(\boldsymbol{\varepsilon}))\delta(h' - h(\boldsymbol{\varepsilon})) \tag{26}
$$

The meaning of density of states of bandgap can be described by considering elastically strained states within the $(h - \frac{dh}{2}, h + \frac{dh}{2})$ energy interval, and the resultant distribution of bandgaps arising from these states. The density of states of bandgap function $\rho(E_{\mathrm{g}}; h)$ offers a blueprint for determining which bandgaps are accessible at what energy cost. One can use the definitions above not only for bandgap, but also generally for any scalar property that will provide an essential road map for deep ESE such as the thermoelectric FoM zT, Baliga's FoM [98], Curie temperature [27], etc. An upper-envelope function $E_{\mathrm{g}}^{\mathrm{upper}}(h)$ and lower-envelope function $E_{\mathrm{g}}^{\mathrm{lower}}(h)$ can also be defined based on $\rho(E_{\mathrm{g}}; h)$:

$$
E_{\mathrm{g}}^{\mathrm{upper}}(h) \ \equiv \ \max \operatorname{supp}_{E_{\mathrm{g}}}(\rho(E_{\mathrm{g}}; h)), \quad E_{\mathrm{g}}^{\mathrm{lower}}(h) \ \equiv \ \min \operatorname{supp}_{E_{\mathrm{g}}}(\rho(E_{\mathrm{g}}; h)) \tag{27}
$$

which are rendered as black and red dotted lines in Figure 3.1c, so the non-zero density of bandgaps falls within $(E_{\mathrm{g}}^{\mathrm{lower}}(h), E_{\mathrm{g}}^{\mathrm{upper}}(h))$. In deep ESE, $E_{\mathrm{g}}^{\mathrm{lower}}(h)$ also indicates the path to obtain the fastest change in E_{g}. For instance, if the goal is to reduce the bandgap of silicon from 1.1 eV as fast as possible, with the least cost of elastic energy, the red-dotted line in Figure 3.1c (which is further detailed in Figure 3.1d) offers the best design of the strain tensor $\boldsymbol{\varepsilon}$ to achieve this goal.

It is seen from Figure 3.1c that, with the application of a relatively small amount of mechanical energy, the overall distribution of Si bandgap shifts downward. This means that by modulating the strain (shear/tension/compression combinations) in multiple directions, strained silicon becomes capable of absorbing a different part of the electromagnetic spectrum than when it is in a stress-free state. It was also found that the bandgap of Si can vanish, corresponding to the semiconductor-to-metal transition in the 6D strain space (see Figure 3.1e for the band structure). Figure 3.1d illustrates that silicon's "fastest path to metallization" is actually a curved path in the strain space: the initial fastest-descent direction for E_{g} (at $h = 0$) is quite different from when E_{g} hits zero and linear perturbation theory such as the deformation potential theory [99] is not expected to work well in deep-strain space. In the case of diamond, deep ESE provides an opportunity to reduce its bandgap to a level comparable to that of InAs. Our results thus demonstrate that by straining diamond in the most optimal way, it can be transformed to mimic the properties of a lower bandgap semiconductor while almost preserving its uniqueness such as high strength and thermal conductivity, thereby paving the way for designing hitherto unexplored combinations of material characteristics.

Another important issue for optical applications pertains to whether the bandgap is direct or indirect. This direct bandgap envelope is a subset of the density of states of bandgap. We define the density of direct bandgaps in parallel to (25), (26) and (27), but with $E_{\mathrm{direct}\,E_{\mathrm{g}}}$ instead of E_{g}, to obtain density of states of direct bandgap $\rho_{\mathrm{d}}(E_{\mathrm{direct}\,E_{\mathrm{g}}}; h)$ and its bounds $E^{\mathrm{upper}}_{\mathrm{direct}\,E_{\mathrm{g}}}(h)$, $E^{\mathrm{lower}}_{\mathrm{direct}\,E_{\mathrm{g}}}(h)$. Obviously, if direct bandgaps exist at any strain, for that strain there will be

$$(E^{\mathrm{lower}}_{\mathrm{direct}\,E_{\mathrm{g}}}(h), E^{\mathrm{upper}}_{\mathrm{direct}\,E_{\mathrm{g}}}(h)) \;\subseteq\; (E^{\mathrm{lower}}_{\mathrm{g}}(h), E^{\mathrm{upper}}_{\mathrm{g}}(h)). \tag{28}$$

Our deep ESE model found within experimentally accessible strain range that the indirect-to-direct bandgap transition takes place in silicon in the high h region and a minimum strain energy density $h^{\mathrm{min}}_{\mathrm{d}}$ exists for the direct bandgap to appear (the red region in Figure 3.1c):

$$h^{\mathrm{min}}_{\mathrm{d}} \;=\; \min\ \mathrm{supp}_h(E^{\mathrm{upper}}_{\mathrm{direct}\,E_{\mathrm{g}}}(h) - E^{\mathrm{lower}}_{\mathrm{direct}\,E_{\mathrm{g}}}(h)). \tag{29}$$

The conventional way to modulate electronic properties in semiconductors is the so-called compositional grading technique. Through varying the stoichiometry of a semiconductor, as for example by molecular beam epitaxy, a graded bandgap can be produced [103]. This means of tweaking the material property is conceptually based on traditional chemical alloying, whereby the chemical composition is tuned in an alloy melt to produce desirable strength or ductility. Invoking this approach, conventional bandgap engineering resorted to chemical alloying such as $\mathrm{GaAl_{1-x}As_x}$ or $\mathrm{Ga_{1-x}In_x\,As}$. However, we have demonstrated here that the stress-free situation is usually not the optimal state for a FoM, and elastic strains allow the bandgap to exhibit many more possible values so that each pure material candidate should occupy a much larger hyperspace enabled through the achievable 6D strain space. The more general bandgap engineering approach should utilize gradients in both composition and strain to achieve the desired band alignment.

3.4. Exploring bandgap ridgelines in strain space

Here we choose the most widely-used semiconductor material, Si, as an example to demonstrate the generality and flexibility of our method. Since the full 6D strain space does not allow for easy visualization, we restrict ourselves to tensile and compressive normal strains only ($\varepsilon_4 =$

$\varepsilon_5 = \varepsilon_6 = 0$) for illustration purposes. Note that combinations of tensile and compressive strains can be used to generate shear strains in the material even though not all shear strains are considered. Figure 3.2 illustrates the isosurface for Si bandgap, i.e., the set of points in the strain space where the bandgap equals some given value, for different E_g levels obtained by our high-throughput NN model. The most striking visual feature of this E_g-isosurface in $\varepsilon_1\varepsilon_2\varepsilon_3$ space is its piecewise smoothness. There are cusp singularities of a different order: ridgelines where two smooth pieces of the E_g-isosurface meet, and corners where three ridgelines meet. These singularities are characterized by discontinuities in the slope (but not value) in the strain space due to band cross-over or even band topology change. Such cuspy features also exist in E_g-isosurface in the general-$\varepsilon_1\varepsilon_2\varepsilon_3\varepsilon_4\varepsilon_5\varepsilon_6$ space, although they are more difficult to visualize directly. One can mathematically define these non-smooth features on the 5D isosurface (embedded in 6D) as n^{th}-order ridges (E_g) if they are differentiable in 5-n directions, while sustaining a change in slope in the other n directions in the strain-space.

Since both the crystal structure and deformation tensor have symmetries, and the bandgap as a function of strain is invariant with respect to some of them, the "paleolith"-like E_g-isosurface (in analogy to the Tresca yield surface in the strength of materials) has the following symmetry structure:

(i) The points μ (the most "compressive" hydrostatic strain point on the E_g-isosurface) and χ (most "tensile" hydrostatic strain point on the E_g-isosurface) lie on the $\varepsilon_1 = \varepsilon_2 = \varepsilon_3$ line. We thus denote their strain space coordinates by (a, a, a) and (b, b, b), respectively. At small or moderate E_g, χ splits and gives rise to a topologically new triangular region $\chi_1\chi_2\chi_3$ as shown in Figure 3.2. It will later be shown these χ-type points form the direct bandgap region on the E_g-isosurface.

(ii) The points α_j $(j = 1,2,3)$ form a regular triangle which lies in a plane orthogonal to the $\varepsilon_1 = \varepsilon_2 = \varepsilon_3$ line. Their coordinates are denoted by $(c, d, d), (d, c, d)$ and (d, d, c), respectively.

(iii) The points β_j $(j = 1,2,3)$ also form a regular triangle which lies in a plane orthogonal to the $\varepsilon_1 = \varepsilon_2 = \varepsilon_3$ line. Their coordinates are denoted by $(e, e, f), (f, e, e)$ and (e, f, e), respectively.

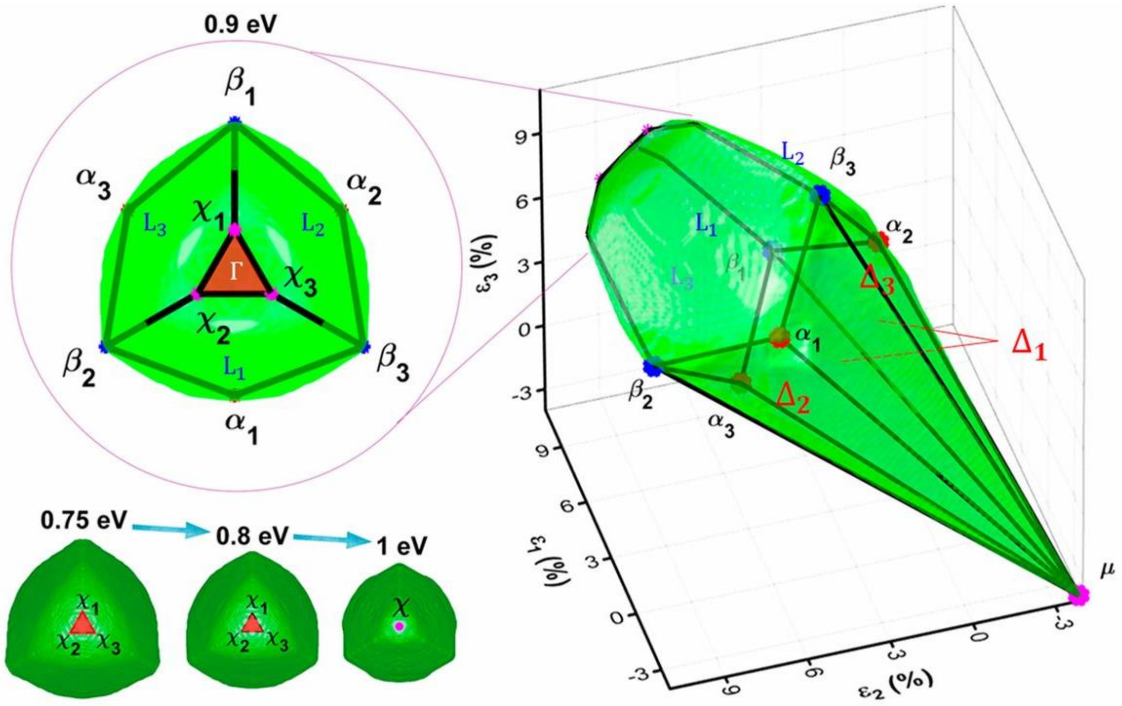


Figure 3.2 Bandgap isosurfaces for silicon in the $\varepsilon_1\varepsilon_2\varepsilon_3$ strain space appear to have the paleolith shape for every E_g level. The main corners ($\chi, \mu, \alpha_j, \beta_j$) of an isosurface at $E_g = 0.9$ eV are indicated by different colors. The red triangular faces indicate the direct bandgap region at different E_g levels. As bandgap increases, the area for the red triangle eventually shrinks to a single χ point. GW model used as a reference.

The shape of the isosurface is similar for both PBE and GW bandgaps, although the specific strain values may differ for the same PBE and GW bandgap levels. It was found that the easiest way (with the least $h(\boldsymbol{\varepsilon}^{\text{normal}})$) to obtain the 0-eV bandgap without any shear strain is to apply a normal strain of -3.86% and 4.36% along any two of the three $\langle 100 \rangle$ directions while leaving the third $\langle 100 \rangle$ direction undeformed. Therefore, there are six strain cases that are equivalent, as indicated by red dots in Figure 3.3b. The position of the vertices of the E_g-isosurface in the strain space is the function of selected bandgap value, and the detailed relationship between the bandgap and the strains is shown in Figure 3.3c. According to our PBE+GW model, the maximum bandgap reachable by strained silicon is 1.24 eV under a hydrostatic tensile strain of 6.5%. It should be noted that silicon strained to such an extent almost reaches the maximum theoretical efficiency, known as the Shockley-Queisser limit [104], of a single p-n junction solar cell. This demonstrates the theoretical feasibility of the application of deep ESE for performance improvement in solar energy conversion devices.

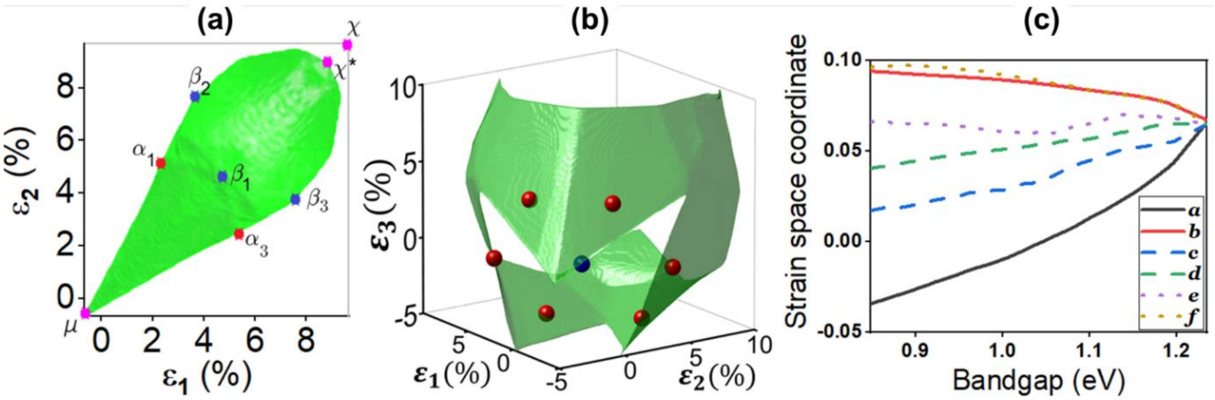


Figure 3.3 (a) Bandgap isosurface shown through the $\varepsilon_1 - \varepsilon_2$ projection of Si at 1 eV level with GW data. The χ point corresponds to the direct bandgap case and it splits into three at small E_g as shown in Figure 3.2. (b) 0-eV bandgap isosurface in the strain space based on GW data. The blue point corresponds to the strain-free state; red points are cases with the least h on this isosurface. (c) Strain space coordinates of the bandgap isosurface corners (defined as in Figure 3.2) as a function of the bandgap level. In the cases where three χ-type points exist, b equals the average coordinate of them.

The formation of the E_g-isosurfaces, such as the ones in Figure 3.2, is due to the relative position of valence band maxium (VBM) and conductiona band minimum (CBM). Despite different shape variations of the two energy bands, modulating elastic strain provides possibilities for the VBM and CBM to differ by the same amount. For undeformed silicon with a bandgap of 1.1 eV, the VBM is located at the Γ-point and the CBM lies on the straight line (the Δ-line in the **k**-space) and is positioned at about 85% of the way from the Brillouin zone center to the zone boundary [105]. Under the three-normal-strain, the cubic crystal symmetry of Si is lifted and we follow the **k-**point labeling scheme explained in Section 2.1 to describe band extrema positions. It is found that VBM remains at Γ irrespective of deformation whereas the position of CBM can be greatly affected by external strains. Using the geometry of the E_g-isosurface as a visualization tool, we identify four types of **k**-space transition in CBM that may happen across the ridgelines on the isosurface.

Starting with the strain points on the lower faces separated by $\mu - \alpha_j$ ridgelines of the E_g-isosurface in Figure 3.2, we found that the CBM remains the relative position along the 'Δ'-type line as in the undeformed case, and that crossing the ridgelines only switches CBM among $(k_1, k_1, 0)$, $(0, k_1, k_1)$, and $(k_1, 0, k_1)$, where $k_1 \approx 0.425$. We term this transition occurring in the small strain region as the *'Δ'-switching*. In this case, the linear deformation potential theory can be used to describe the strain effects on the band extremum [99]. However, investigation of the large deformation points on its upper faces in Figure 3.2 reveals that the CBM would not retain its location and major changes would happen.

Our ML model captures the occurrence of *'L-Δ' transition* across the $\beta_i - \alpha_j$ ridgelines where the CBM changes to 'L' points in **k**-space, see Figure 3.4a-b for an example. The large, non-perturbative deformation makes the conventional theory ineffective in predicting it. Moving further toward χ in the strain space, CBM would remain at 'L' and a cross-over of the $\chi_2 - \beta_j$ ridgelines is referred to as an L-*switching*. *Indirect-to-direct bandgap transition* occurs near the upper tip of the paleolith-like isosurface where CBM appears at Γ, as shown in Figure 3.4c. This can be explained by the competition between drops of different band edges. In general, as strain increases the band edge at both Γ and 'L' would decrease.

As a result of high strains, the energy decrease at Γ is faster and eventually the bandgap becomes direct, as shown in Figure 3.4d. When the strained Si turns into a direct bandgap semiconductor, it would exhibit a significant enhancement in its optical transitions around the fundamental adsorption edge compared to an undeformed Si, due to the elimination of phonon involvement to facilitate adsorption or emission. Furthermore, as absorbance increases exponentially with thickness in a material, a solar cell based on direct bandgap Si with a high adsorption coefficient would require much less thickness to absorb the same amount of light, paving the way for the design of light-weight high-efficiency solar cells. Table 3.4 summarizes all the details of the **k**-space transitions, thus resolving the conduction band properties exhaustively for a wide range of strains.

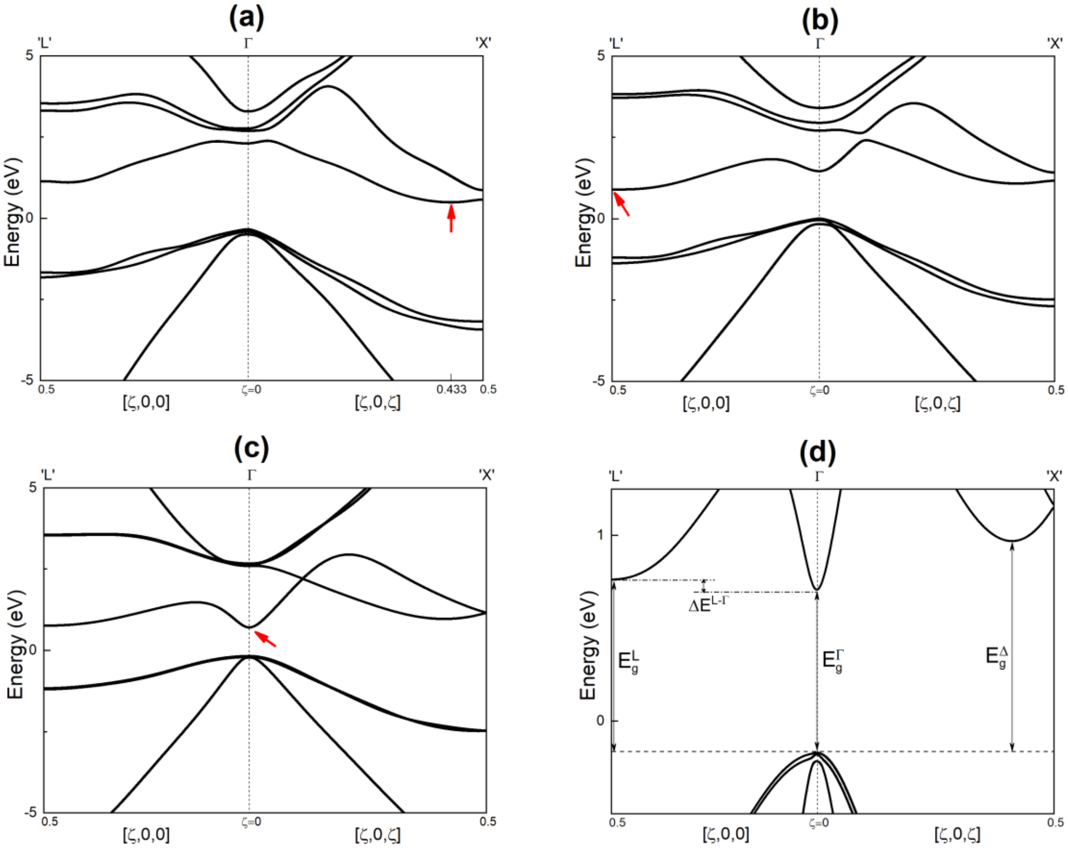
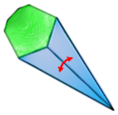
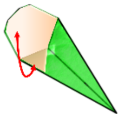
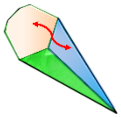
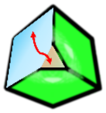


Figure 3.4 Illustration of **k**-space transition in Si predicted by deep ESE. All the transitions are verified by GW calculations. (a-b) represents the 'Δ-L' transition and (b-c) shows the indirect-to-direct transition. The CBM (red arrows) locates at **k**-point (0.433, 0.433, 0), (0.5, 0, 0), and (0, 0, 0) respectively. (d) The enlarged band structure around Fermi energy shows the competition of the three possible CBM positions. The three-normal-strain cases for (a-c) correspond to points on different faces of the bandgap isosurface in Figure 3.2.

Table 3.4 **k**-space CBM transitions. Each of 12 separating ridgelines of the iso-bandgap body tabulated. The constants k_1 and k_2 are approximately equal to 0.425 and 0.5, corresponding to points on Δ and L, respectively.

Type	Change of "carapace"		**k**-coordinate of CBM
'Δ'-switching		$\Delta_1 \leftrightarrow \Delta_2$	$(0, k_1, k_1) \leftrightarrow (k_1, 0, k_1)$
		$\Delta_2 \leftrightarrow \Delta_3$	$(k_1, 0, k_1) \leftrightarrow (k_1, k_1, 0)$
		$\Delta_3 \leftrightarrow \Delta_1$	$(k_1, k_1, 0) \leftrightarrow (0, k_1, k_1)$
'L'-switching		$L_1 \leftrightarrow L_2$	$(k_2, 0, 0) \leftrightarrow (0, k_2, 0)$
		$L_2 \leftrightarrow L_3$	$(0, k_2, 0) \leftrightarrow (0, 0, k_2)$
		$L_3 \leftrightarrow L_1$	$(0, 0, k_2) \leftrightarrow (k_2, 0, 0)$
'L-to-Δ' transition		$L_1 \leftrightarrow \Delta_2$	$(k_2, 0, 0) \leftrightarrow (k_1, 0, k_1)$
		$L_1 \leftrightarrow \Delta_3$	$(k_2, 0, 0) \leftrightarrow (k_1, k_1, 0)$
		$L_2 \leftrightarrow \Delta_1$	$(0, k_2, 0) \leftrightarrow (0, k_1, k_1)$
		$L_2 \leftrightarrow \Delta_3$	$(0, k_2, 0) \leftrightarrow (k_1, k_1, 0)$
		$L_3 \leftrightarrow \Delta_1$	$(0, 0, k_2) \leftrightarrow (0, k_1, k_1)$
		$L_3 \leftrightarrow \Delta_2$	$(0, 0, k_2) \leftrightarrow (k_1, 0, k_1)$
Indirect-to-direct bandgap transition		$L_1 \leftrightarrow \Gamma$	$(k_2, 0, 0) \leftrightarrow (0, 0, 0)$
		$L_2 \leftrightarrow \Gamma$	$(0, k_2, 0) \leftrightarrow (0, 0, 0)$
		$L_3 \leftrightarrow \Gamma$	$(0, 0, k_2) \leftrightarrow (0, 0, 0)$

3.5. Technical details

First-principles calculation details

We used the PBE [106] exchange-correlation functional and the projector augmented wave method (PAW) [107] in our DFT simulations implemented in the Vienna Ab initio Simulation Package (VASP) [108] with spin-orbit coupling incorporated. A plane wave basis set with an energy cutoff of 520 eV was adopted to expand the electronic wavefunctions. The Brillouin zone integration was conducted on a 13 × 13 × 13 Monkhorst-Pack **k**-mesh [109] (6 × 6 × 6 for GW calculations). Atomic coordinates in all the structures were relaxed until the maximum residual force was below 0.0005 eV Å^{-1}. We focused on the strain range of $\{-5\% \leq \varepsilon_j \leq 10\%,\ j = 1, \ldots, 6\}$ for silicon.

Since the material properties do not change upon rotations of the crystal, we eliminated the rotational degrees of freedom by adopting an upper triangular deformation gradient tensor to map out deformation cases, as outlined in Section 2.1. This treatment ensures a one-to-one correspondence between the applied strains and deformation cases, and we have exploited symmetry in this setting for general 3D and 6D cases and implemented it in our study to avoid repetitive calculations. For example, consider the following four arbitrary strain cases:

$$\varepsilon_{\mathrm{A}} = (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{33}, \varepsilon_{23}, \varepsilon_{13}, \varepsilon_{12})$$
$$\varepsilon_{\mathrm{B}} = (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{33}, \varepsilon_{23}, -\varepsilon_{13}, -\varepsilon_{12})$$
$$\varepsilon_{\mathrm{C}} = (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{33}, -\varepsilon_{23}, \varepsilon_{13}, -\varepsilon_{12})$$
$$\varepsilon_{\mathrm{D}} = (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{33}, -\varepsilon_{23}, -\varepsilon_{13}, \varepsilon_{12})$$

They ought to be equivalent and we can have three times more additional data. Such symmetry acts as a regularizer and helps to reduce overfitting if our goal is to train a bandgap model or finding the density-of-states of bandgap distribution against varied elastic strain energy densities.

Machine learning details

NN fitting is implemented within the Tensorflow [110] framework. To predict the bandgap we used deep NNs with four hidden layers with a (64→128→256→256) structure in the case of three-normal-strains ($\boldsymbol{\varepsilon}^{\mathrm{normal}}$) and a (512→256→256→256) structure for the general case with shear strains ($\boldsymbol{\varepsilon}^{\mathrm{6D}}$). For the more complicated task of band energy prediction at a single $\mathbf{k}$-point, the architecture of (512→256→256→256) was used. The leaky rectified linear unit was chosen as an activation function. We used the Adam stochastic optimization method [111], the orthogonal weight initialization [96] and the dropout technique to prevent overfitting. The tree-based ensemble algorithms were implemented in Scikit-learn [112]. For our regression task, we used two types of ensembling on decision trees: the random forest regression [113] and the gradient boosting regression [114]. The architecture is shown in Figure 3.1a. Hyper-parameters tuning was executed by using cross-validation on a training set to enhance the fitting process.

Chapter 4. Deep elastic strain engineering of semiconductor electronic properties by convolutional neural networks

4.1. Limitations of feed-forward neural networks

Chapter 3 demonstrates a framework that, in principle, can be used for tailoring properties such as energy gaps between different electronic bands of any other material by recourse to deep ESE and deep learning. But is it capable of tailoring any other electronic properties, such as those related to not only the energy band levels but also the energy band curvatures?

While the feed-forward NN models presented in Chapter 3 are adequate for rapid data collection in a highly specialized model [28,115], 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 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. These recent advances enable multi-property optimization and Pareto-front type tradeoff analysis.

To accomplish these goals, this chapter introduces a physics-informed 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 NN architecture hitherto employed to address this class of problems. More advanced algorithms and data representation schemes are involved to provide markedly improved ML outcomes. The techniques described here enable detailed analysis of band structures in the general 6D 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 bandgap and of the shape of band structure, but also for capturing the curvature of the band and the effective mass.

Some argumentation and figures/tables in this chapter are directly taken from the author's own publication of Ref. [116]: E. Tsymbalov*, Z. Shi*, M. Dao, S. Suresh, J. Li and A. Shapeev, *Machine learning for deep elastic strain engineering of semiconductor electronic band structure and effective mass*, npj Computational Materials, In press, (2021). *Equal contribution.

4.2. Digital-image view of band structures

Inspired by the wide adoption of deep learning in the field of computer vision [117], an analogy is drawn 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, energy dispersions is envisioned 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 Figure 4.1a. Note that previous ML schemes based on simple feed-forward NN treated an energy band as a flattened array of independent values [28] - thereby neglecting to account for intraband correlation.

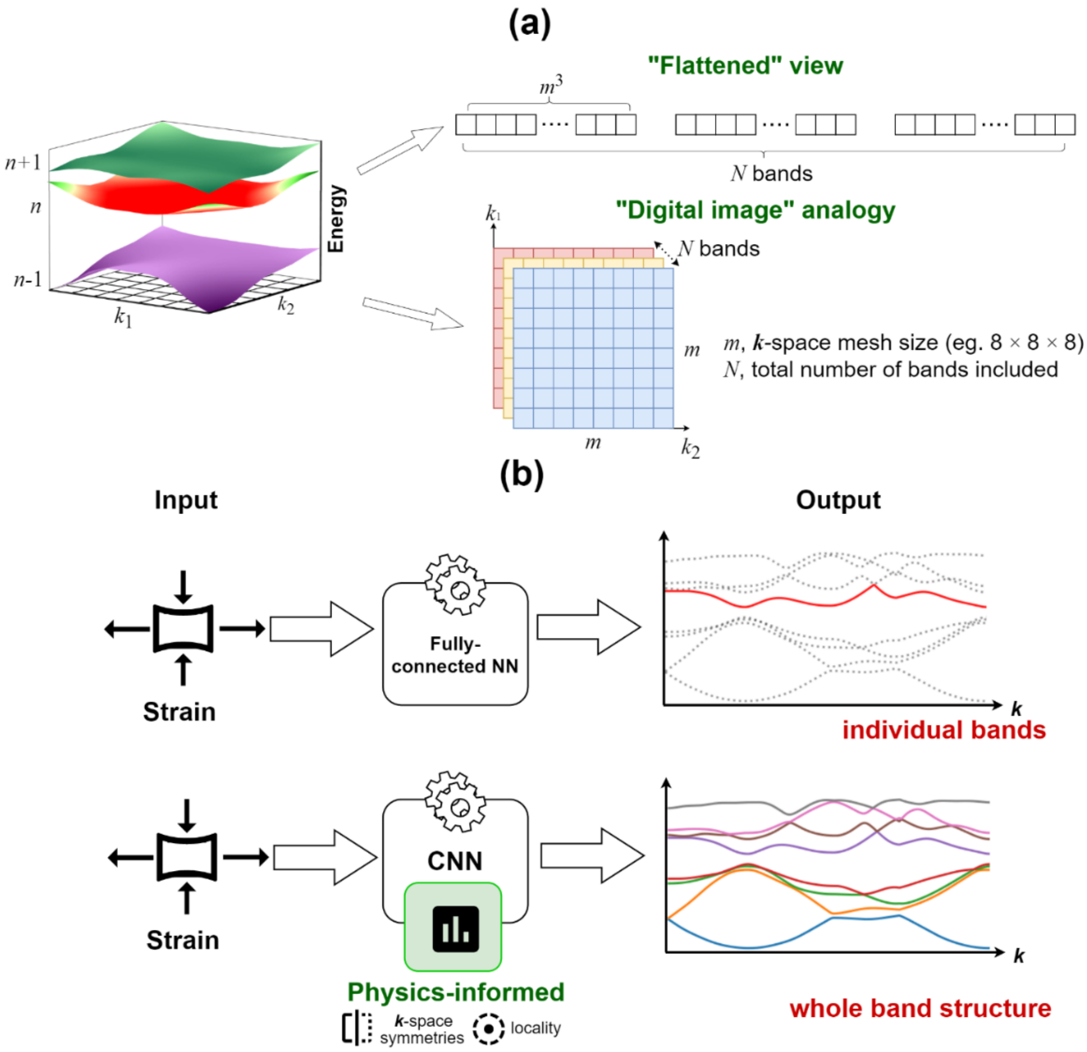


Figure 4.1 Two different views of band structure. (a) Different representations of a band structure. In the "flattened" view, a band structure is represented as N stacked flattened arrays (vectors) and processed like independent values. Each array is m^3 in length. In the "digital image" analogy, the 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 machine learning. We predict the eigenvalues for each energy band separately by utilizing the "flattened" band structure representation to obtain the entire band structure.

In prior work [28], different bands were analyzed separately by NN (Figure 4.1a, 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 Figure 4.1, 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 n^{th} band is not calculated in isolation, but is determined from the collective influence of its neighboring bands, including the adjacent $(n - 1)^{\text{th}}$ and $(n + 1)^{\text{th}}$ bands as well as other non-adjacent bands. In other words, information from *interband* correlation influencing the n^{th} 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 [118]. 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.

4.3. Model description, training, and active learning

The general setup of the proposed model is illustrated in Figure 4.2a. 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 Figure 4.2a. 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 $(6\rightarrow128\rightarrow256\rightarrow512\rightarrow512)_n$, where $512 = m^3$, for $n = 1, 2, ..., N$ 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 ($n = n_{\text{VB}}$), the lowest conduction band ($n = n_{\text{CB}}$), and their adjacent two bands ($n = n_{\text{VB}}$ - 1 and $n = n_{\text{CB}}$ + 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 Figure 4.2b. This process is similar to the decoding part of an autoencoder [119] 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 with features that capture deep intra- and inter-band information. This output comprising the complete ML inference represents the band structure obtained by DFT calculations (Figure 4.2a-b). In each convolutional block in the CNN part, the convolution is a two-step sequence. In the first step, a 3 × 3 × 3 × 1 kernel accounting for the intraband correlation (with periodical boundary conditions and symmetry) is used. In the second step, a 1 × 1 × 1 × 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 × 3 × 3 × 3) kernel instead of the aforementioned two-step convolution (3 × 3 × 3 × 1)→convolution (1 × 1 × 1 × 3) kernel, with more weights per block but better accuracy. Also, since 8 × 8 × 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 × 8 × 8 representation, before carrying outs such operations.

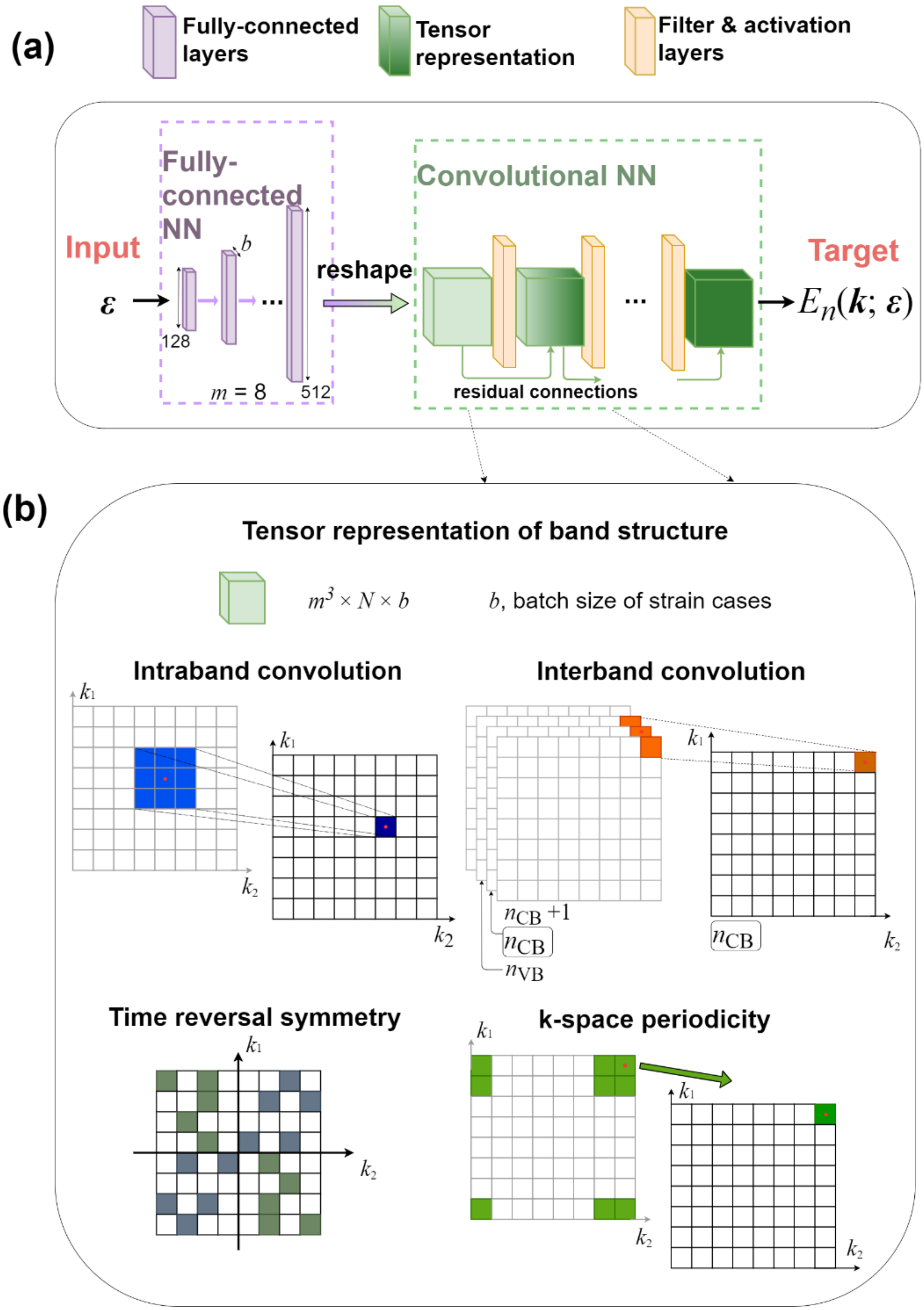

(a)
Fully-connected layers
Tensor representation
Filter & activation layers
Input
ε
Fully-connected NN
b
128
m = 8
512
reshape
Convolutional NN
residual connections
Target
$E_n(\boldsymbol{k};\, \boldsymbol{\varepsilon})$
(b)
Tensor representation of band structure
$m^3 \times N \times b$
b, batch size of strain cases
Intraband convolution
k_1
k_2
Interband convolution
$n_{\mathrm{CB}}+1$
n_{CB}
n_{VB}
Time reversal symmetry
k-space periodicity

Figure 4.2 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[28] 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 × 8 × 8 **k**-points is used. (b) Tensor representation and physical insights incorporated into the CNN model: time-reversal symmetry, **k**-space periodicity, and inter- and intra-band convolution.

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 (Figure 4.2b and Section 4.8 Figures A4.1 and A4.2) 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 that 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.

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 inexpensive DFT-PBE 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 (~6,000 strain-samples) of the accurate 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 Figure 4.3 for a schematic of this process and Section 4.8 for computational details.

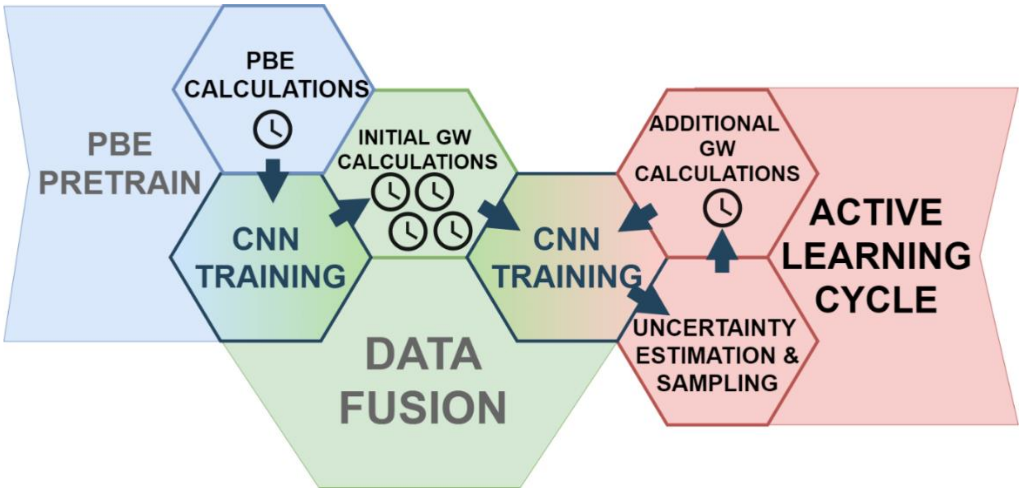


Figure 4.3 The entire ML scheme involves pre-train, data fusion, and active learning. The solid arrows show the workflow, and clock symbols indicate the relative time required for *ab initio* calculations.

Another integral part of our training is active learning (Figure 4.3), which entails a class of machine-learning 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 [76]. The particular choice of an uncertainty quantification procedure greatly influences performance in the active learning part. There are three main routes to uncertainty estimation in NN: ensembling [77], 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 NNs (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 [120].

Here, we use the dropout uncertainty estimation enhanced with the Gaussian processes for stability [121] 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 Figure 4.3. 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 Figure 4.4a-b.

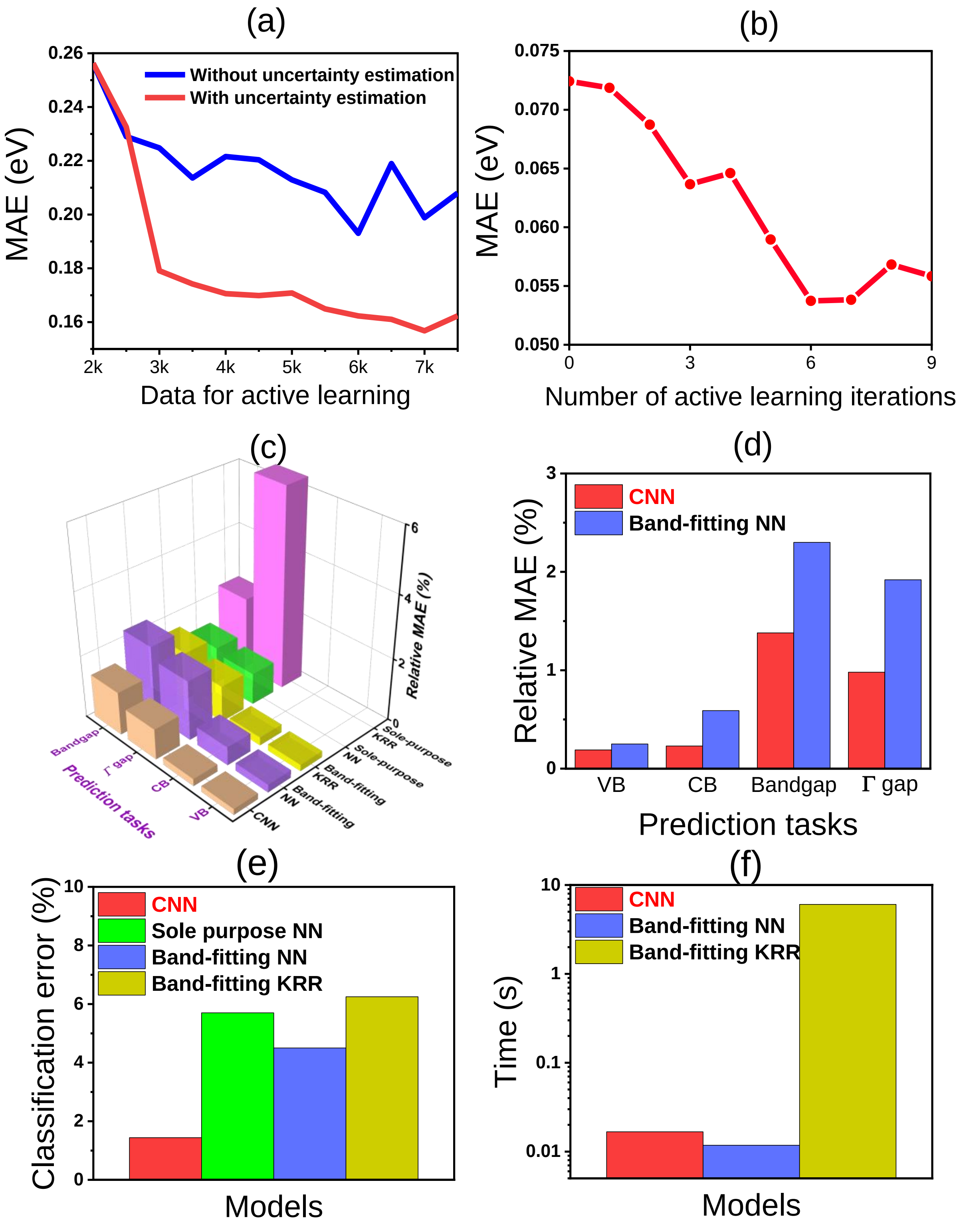


Figure 4.4 ML accuracy and comparison of the different ML models. (a) Steady improvement of model performance in terms of MAE during active learning with and without uncertainty estimation on PBE data. (b) MAE of the bandgap estimation reduces with active learning iterations. 200 strain values were sampled at each step. The last three iterations did not contribute to the error reduction. (c) Physics informed CNN holds advantages against band-fitting NN and band-fitting KRR in every front while being able to accomplish predictions tasks the

sole-purpose NN and KRR cannot do. The “Γ gap” is the difference between CB and VB at Γ and usually does not coincide with E_{g}. (d) 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}. (e) Accuracy of CNN and other models for CBM position classification task. (f) 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.

4.4. Model accuracy and performance

The ML framework outlined in Figure 4.3 achieves high accuracy in a variety of tasks compared to existing ML methods. The CNN model outperforms our 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 Figure 4.4d. The predictions of properties related to the band structure, such as the bandgap E_{g} (defined as the energy difference between the CBM and VBM values), were treated in a previous study in Chapter 3 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- 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 (Figure 4.4c-d and Section 4.8 Tables A4.1 and A4.2).

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 Figure 4.4e 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 Figure 4.4f. The CNN model has a time complexity comparable to simple NNs. In the next section, we discuss examples of elastic strain engineering of a diamond crystal.

4.5. Results and 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 metallize diamond by turning it into an electrical conductor with zero 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 (See Chapter 5 for more details).

4.5.1. Density of states of bandgap and bandgap isosurface

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 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 diamond. These are illustrated in Figure 4.5a. 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. We take the same elastic strain energy density as defined in Chapter 3: $h(\boldsymbol{\varepsilon}) \equiv \frac{E(\boldsymbol{\varepsilon})-E^0}{V^0}$, 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" as shown in Figure 4.5a. There exist many strain states with an elastic strain energy density that can reach a direct 3-eV bandgap in 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- 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 Figure 4.5a. Similarly, an increase in the bandgap can be explored by following the upper-bound function (upper black dotted line in Figure 4.5a). This line represents pure triaxial compression, i.e., $\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33} < 0, \varepsilon_{23} = \varepsilon_{13} = \varepsilon_{12} = 0$.

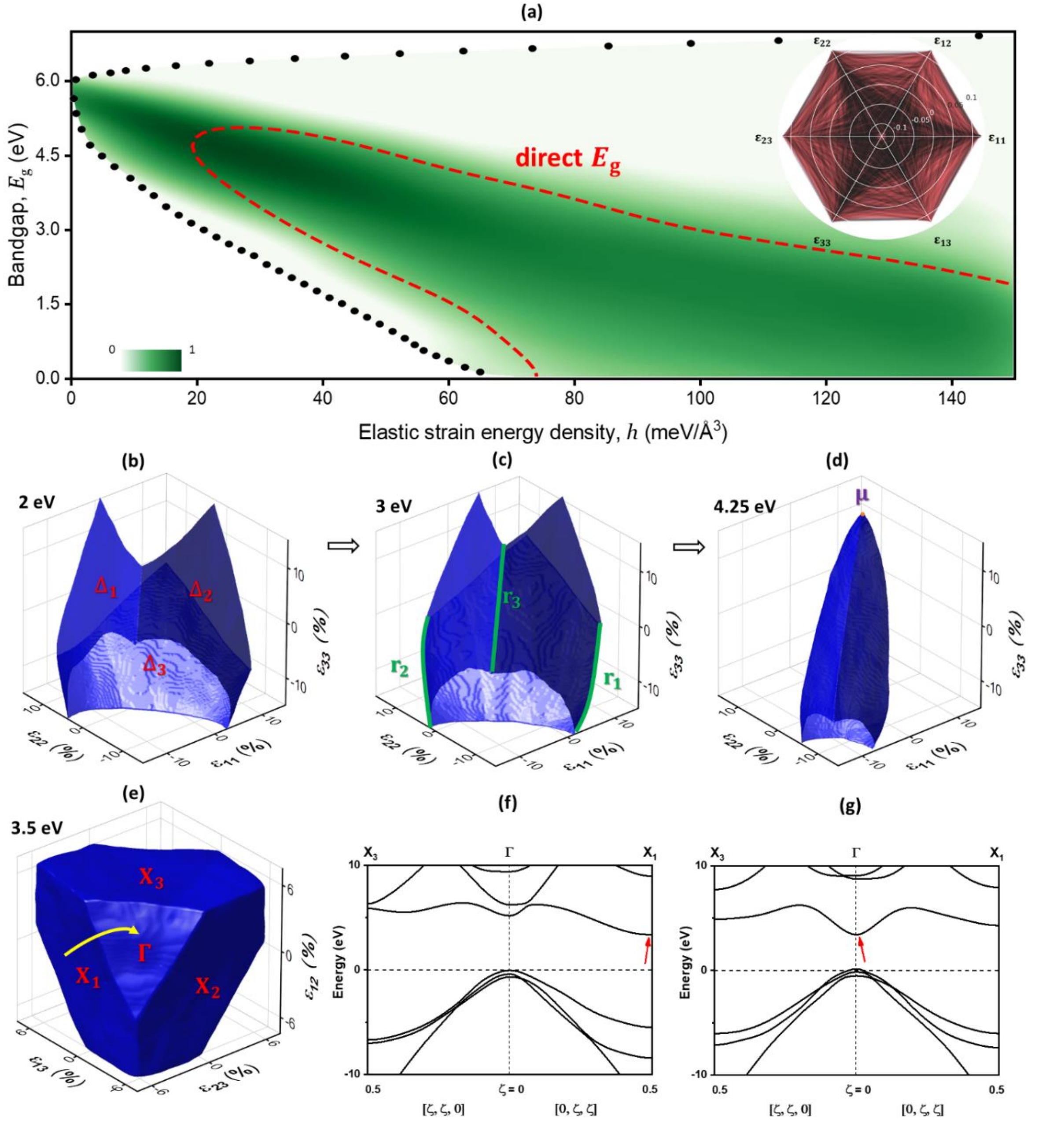


Figure 4.5 Density of states of bandgap and bandgap isosurfaces. (a) Bandgap values achievable through elastic strain engineering for various values of elastic strain energy density h within the strain space. The green shading of the region reflects the distribution of the available bandgap. The boundary of the strain region where a direct bandgap could occur is indicated by the red dashed line. 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 ($\mathrm{r}_1, \mathrm{r}_2$, 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 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.

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 4.5b-d illustrate 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 $\mathbf{k}$-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 $\mathrm{r}_1, \mathrm{r}_2$, and r_3 correspond to strain cases with relations $\varepsilon_{22} = \varepsilon_{33}$, $\varepsilon_{33} = \varepsilon_{22}$, and $\varepsilon_{11} = \varepsilon_{22}$, respectively.
- The corner μ labeled in purple is the intersection of $\mathrm{r}_1, \mathrm{r}_2$, 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 Figure 4.5e. Besides three *different* indirect bandgap CBM positions at $\mathrm{X}_1: (0, 0.5, 0.5)$, $\mathrm{X}_2: (0.5, 0, 0.5)$, and $\mathrm{X}_3: (0.5, 0.5, 0)$, 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 Figure 4.5e). The corresponding band structures for the indirect and direct bandgap are shown in Figure 4.5f and g, respectively.

4.5.2. Bandgap reduction capability ranking

We also acknowledge that it is not straightforward yet to achieve the identified optimal 6D complex strain states introduced in Figure 3.1d and e in a commercial device of today. Nevertheless, the development of new experimental methods is not beyond the realm of research and development possibilities in the future that are afforded by MEMS and NEMS and various instruments and tools available to design and fabricate complex geometries and shapes [122–125]. Also, precise strain control can be achieved through the fabrication of micro and nanodevices, and various quantitative force-displacement probes that can measure down to pico-Newton forces and nanometer displacements. To provide additional information that may be able to guide future experiments, we demonstrate in Figure 4.6 the ranking of common diamond or silicon crystal orientations to attain the same target bandgap through uniaxial tensile or compressive straining (i.e., constrained straining without allowing for the Poisson effect) can differ at different strain levels. For example, in order to achieve a 5-eV bandgap in diamond, uniaxial tensile straining along <100> direction requires a smaller strain magnitude than along <111> direction; whereas to achieve 4 eV bandgap in diamond, uniaxial tensile straining along <111> direction requires a smaller strain magnitude than along <100> direction, as depicted in Figure 4.6a. It is also found that allowing internal atomic relaxation during straining results in evident structural reconfiguration, especially in large deformation cases. Some of the diamond straining cases may even facilitate graphitization [115,126]. This section is a complement and correction to the study in the SI Appendix Note S3 and Figure S3 of Ref. [28], where only non-relaxed results were given for silicon.

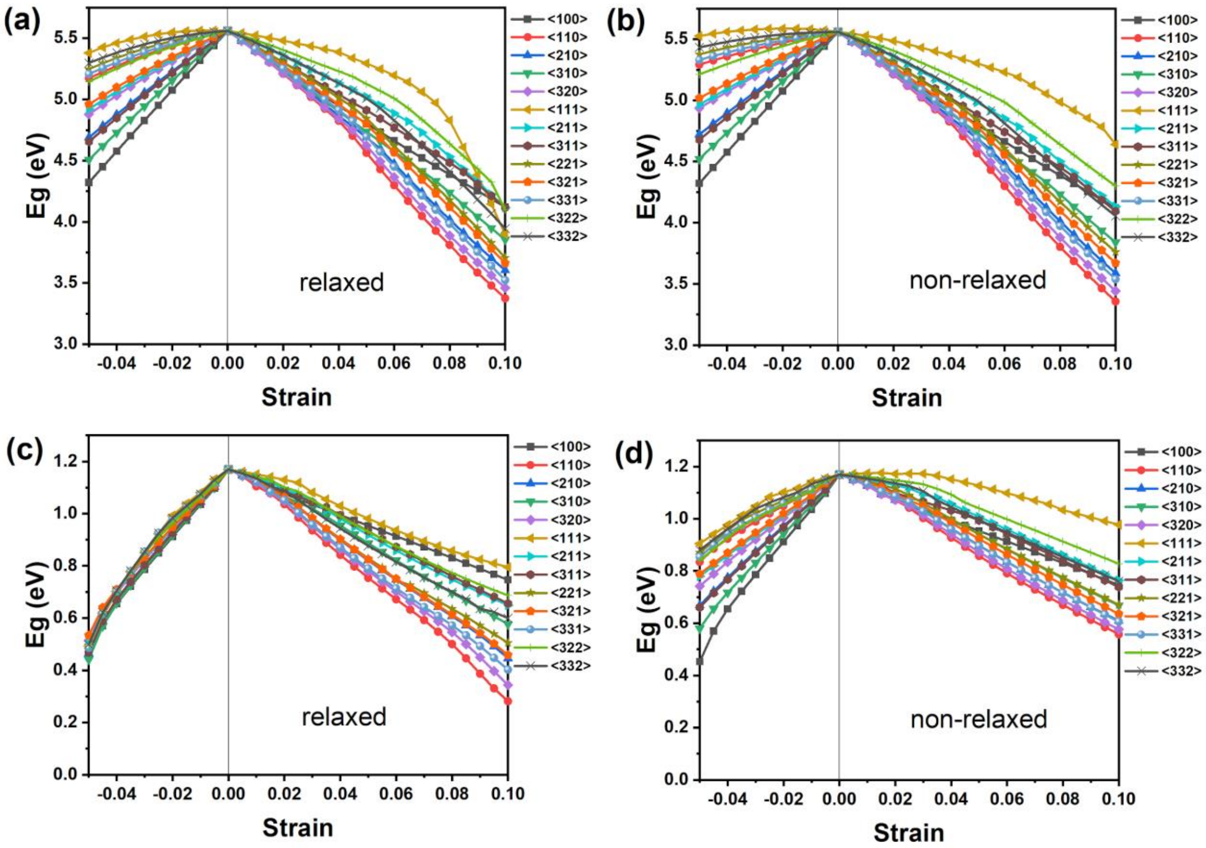


Figure 4.6 Bandgap change as a function of strain for uniaxial straining along different crystal orientations in (a-b) diamond and (c-d) silicon with relaxed and non-relaxed atomic structures, respectively.

4.5.3. Machine learning for effective mass

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 $E_{n_{\mathrm{CB}}}(\mathbf{k}) = E_{n_{\mathrm{CB}}}(k_1, k_2, k_3)$, then the corresponding electron effective mass tensor can be defined in terms of the Hessian matrix $\mathbf{H}(E_{n_{\mathrm{CB}}}(\mathbf{k}))$ consisting of second partial derivatives with respect to $\mathbf{k}$. These partial derivatives are approximated for $\boldsymbol{m}^*$ at a particular $\mathbf{k}$-point (such as CBM).

Based on the values drawn from our ML model, $\boldsymbol{m}^*$ for an undeformed diamond at CBM is extracted by fitting the band structure:

$$\boldsymbol{m}^* = \begin{bmatrix} 1.55m_{\mathrm{e}} & 0 & 0 \\ 0 & 0.31m_{\mathrm{e}} & 0 \\ 0 & 0 & 0.31m_{\mathrm{e}} \end{bmatrix},$$

where m_{e} is the free-electron mass. Given that $\boldsymbol{m}^*$ is a second 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 ($m_{\mathrm{L}}^* = 1.55m_{\mathrm{e}}$) along with the corresponding equivalent (100) reciprocal space direction and two transverse masses ($m_{\mathrm{T}}^* = 0.31m_{\mathrm{e}}$) 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 4.1), 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 is shown in Figure 4.7.

Table 4.1 Longitudinal and transverse electron effective masses at CBM in undeformed diamond (in units of free-electron stationary mass m_{e}). The results obtained through our CNN model are compared with experiments [127], our previous NN model [28], and explicit calculations using existing methods including GW_0, linear muffin-tin-orbital (LMTO) model [128] and G_0W_0 [129].

	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
$m_{\mathrm{L}}^* / m_{\mathrm{T}}^*$	5.0	5.16	4.61	4.41	5.0	3.89

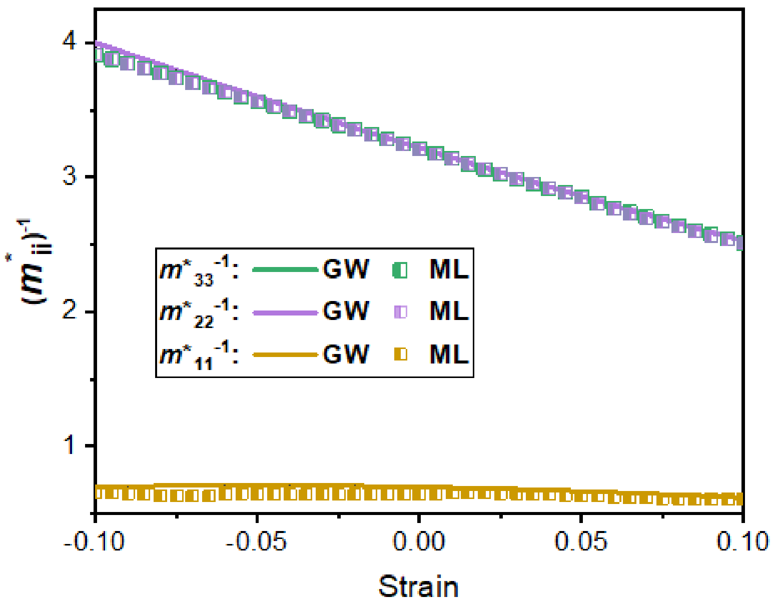


Figure 4.7 The reciprocal of the effective mass tensor components at the CBM as a function of hydrostatic strain. Predictions made by our ML model are shown in comparison to values obtained from GW calculations.

We also studied the 6D strain space to obtain the conduction-related properties and the elastic strain energy density as functions of $\boldsymbol{\varepsilon}$. Here, we adopted our ML model to acquire the relation between the "conductivity effective mass" for the conduction electron $m^*_{\text{cond}}(\boldsymbol{\varepsilon})$ and $h(\boldsymbol{\varepsilon})$, as shown in Figure 4.8a. The values of scalar m^*_{cond} are obtained by averaging individual longitudinal and transverse effective masses, as in Ref. [130]. The blue shading in Figure 4.8a reveals the distribution of the available m^*_{cond}, 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

$$\begin{aligned} c\left(m^*_{\text{cond}}{}'; h'\right) &\equiv \int\limits_{h(\boldsymbol{\varepsilon})<h'} d^6\boldsymbol{\varepsilon}\,\delta\left(m^*_{\text{cond}}{}' - m^*_{\text{cond}}(\boldsymbol{\varepsilon})\right) \\ &= \int d^6\boldsymbol{\varepsilon}\,\delta\left(m^*_{\text{cond}}{}' - m^*_{\text{cond}}(\boldsymbol{\varepsilon})\right)\Theta\left(h' - h(\boldsymbol{\varepsilon})\right), \end{aligned} \tag{30}$$

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' can then be defined by the derivative of $c(m^*_{\text{cond}}{}';\ h')$ with respect to h':

$$
\begin{aligned}
g\left(m_{\text{cond}}^{*}{}';h'\right) &\equiv \frac{\partial c\left(m_{\text{cond}}^{*}{}';h'\right)}{\partial h'} \\
&= \int d^{6}\boldsymbol{\varepsilon}\delta\left(m_{\text{cond}}^{*}{}' - m_{\text{cond}}^{*}(\boldsymbol{\varepsilon})\right)\delta\left(h' - h(\boldsymbol{\varepsilon})\right).
\end{aligned}
\tag{31}
$$

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 in

Figure 4.8. 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.

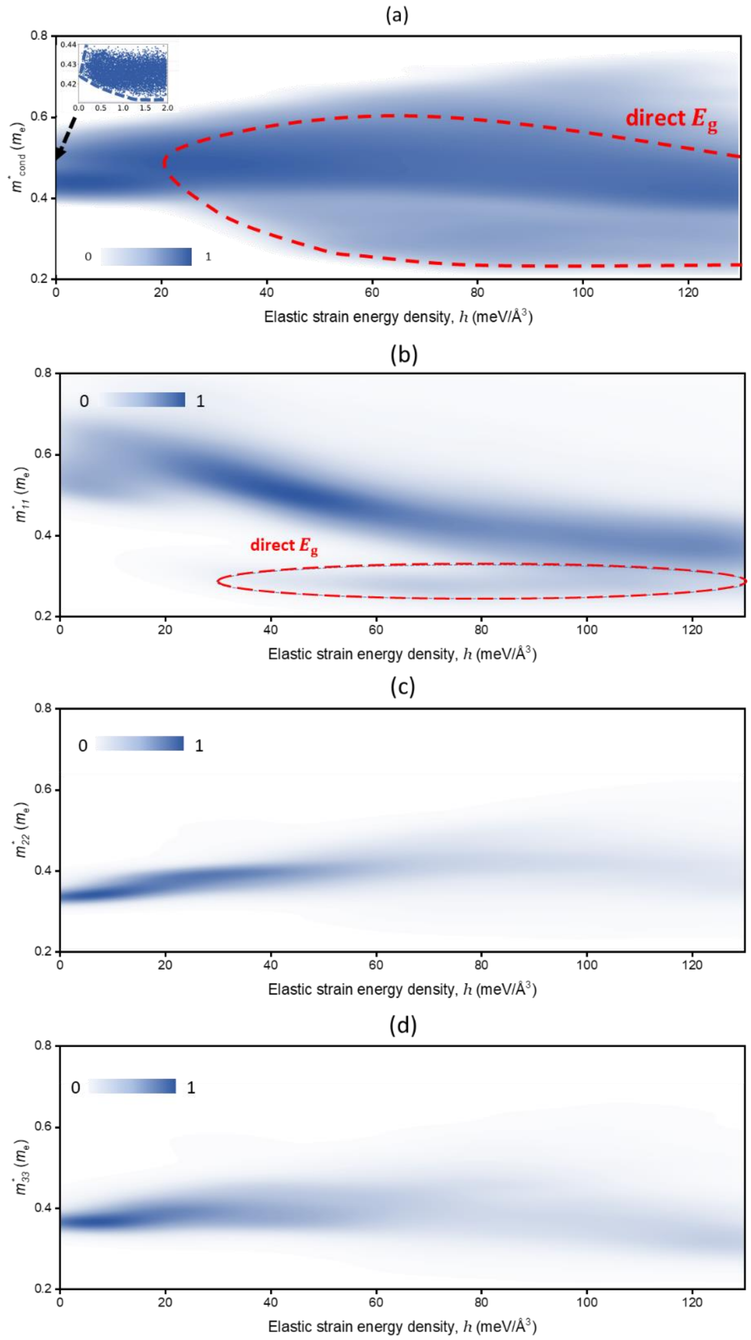


Figure 4.8 ML-based exploration of electron effective mass tensor. (a) Density of states of conductivity effective mass. A darker shading implies more strains can reach a specific value of m^*_{cond} at a given h. The red dashed line indicates the region of the possible direct bandgap configurations. Inset is the zoomed-in plot near h = 0 of the m^*_{cond} distribution. Distribution of effective mass tensor components (m^*_{11}, m^*_{22}, and m^*_{33}) for various h are shown in (b), (c), and (d), respectively.

4.5.4. Multi-objective optimization

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 Figure 4.9a, indicates a compromise in simultaneously having a small bandgap and conductivity effective mass. It is not possible to achieve, for example, a near-zero bandgap and $m_{\mathrm{cond}}^{*} < 0.25m_{\mathrm{e}}$ without paying a considerable penalty in h by straining diamond, as indicated by the "infeasible region" in Figure 4.9a. Also, it is likely to find higher h values that correspond to the same combination of (E_{g}, m_{cond}^{*}). In Figure 4.9a, such elastic strain energy density values are associated with strain cases in the "feasible region". Additionally, Figure 4.9b 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 Figure 4.9a onto the $E_{\mathrm{g}} - m_{\mathrm{cond}}^{*}$ plane.

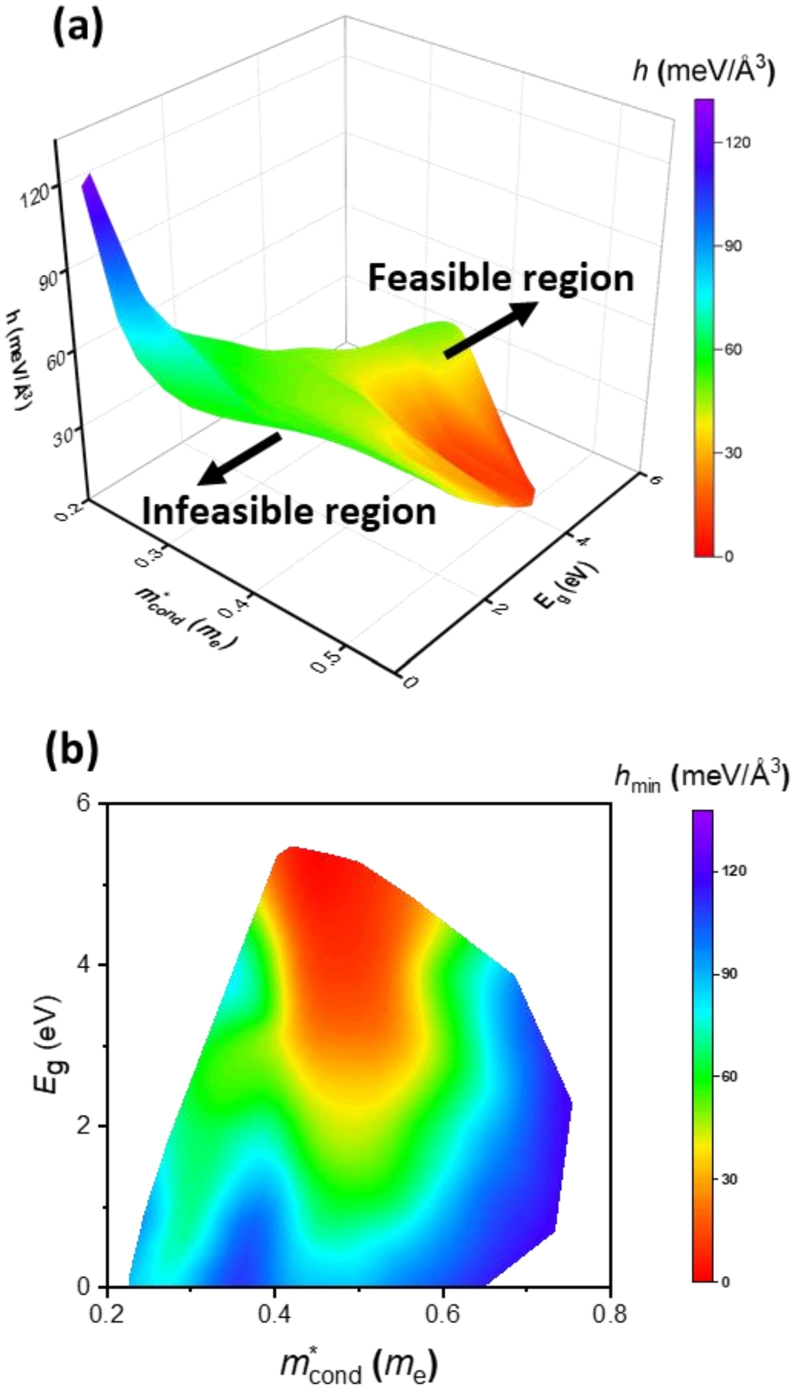


Figure 4.9 Multi-objective optimization. (a) Pareto front for minimizing m^*_{cond}, bandgap and h. The color contours denoted different h values. The (E_{g}, m^*_{cond}, h) triplets within the Pareto front are feasible, meaning that there exist strain cases that can realize the three properties. (b) Color contours of the smallest elastic strain energy density (h_{min}) required for achieving any combinations of bandgap and m^*_{cond}. (b) is not a 2D projection of (a) in which only optimized (minimized) E_{g} and m^*_{cond} exist.

4.6. Chapter conclusion

In summary, by recognizing that the band dispersion is structured and highly correlated in continuous $\mathbf{k}$, $\boldsymbol{\varepsilon}$, 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 NN 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 or insulator-to-metal transition takes place.

To accomplish the task of band structure prediction, our network offers the capabilities of learning the complex intra- and inter-band correlation in a self-directed manner while taking into account important physical characteristics, such as crystal periodicity and time-reversal symmetry. As an 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 is capable of capturing 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.

4.7. Technical details

First-principles calculation details

We used the PAW [107] in our DFT simulations implemented in the VASP [108], with the exchange-correlation functional of PBE [106]. In all calculations, the electronic wavefunctions were expanded in a plane wave basis set with an energy cutoff of 600 eV. An $8 \times 8 \times 8$ Monkhorst-Pack [109] $\mathbf{k}$-point 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-PAW 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 yield stable structures, ie. without imaginary phonon frequencies.

Database construction and validation

In the data generation step of database construction, we took the Latin-Hypercube-sampled [131] 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 data, we compared them 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 eliminate the cases where phase transitions (such as graphitization [115,132]) 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 type. Therefore, unlike the Materials Project database construction [95] 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 one 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 [133] and comparable with experimental measurement for many semiconductor materials [134].

Despite high reliability in the accuracy with our GW dataset, it is still far from large enough to train the NN in an adequate way without overfitting, due to the formidable computational expense of carrying out tens of thousands of GW calculations. To tackle this issue, we firstly adopted data fusion to take the quantitative advantage of the PBE dataset (~35000 strain points) and the qualitative advantage of the GW dataset (~6000 strain points). Also, we introduced active learning cycles to reduce errors compared to complete random sampling. After the first round of the training on the GW data, we performed an estimation over a large set of random strains in 6D and chose a small amount of ~100 strain cases with the largest expected error as evaluated by this temporary model. These strain cases were added to the training set for the next round of training. Our study indicates that 5-10 cycles of the above active learning may enable the trained CNN to reach the same level of accuracy with twice or three times less additional data, thus considerably reducing the total amount of *ab initio* calculations without compromising the robustness of our machine learning model.

4.8. Chapter appendix

Figure A4.1. Average Pearson correlation coefficient between the energies in two points separated by a given Manhattan distance in **k**-space fractional coordinate. Unsurprisingly, the correlation is strongest in the case of two adjacent **k**-points. This is exploited by the convolution layers in our network, which introduces the correlation of adjacent **k**-points (i.e., the intraband correlation).

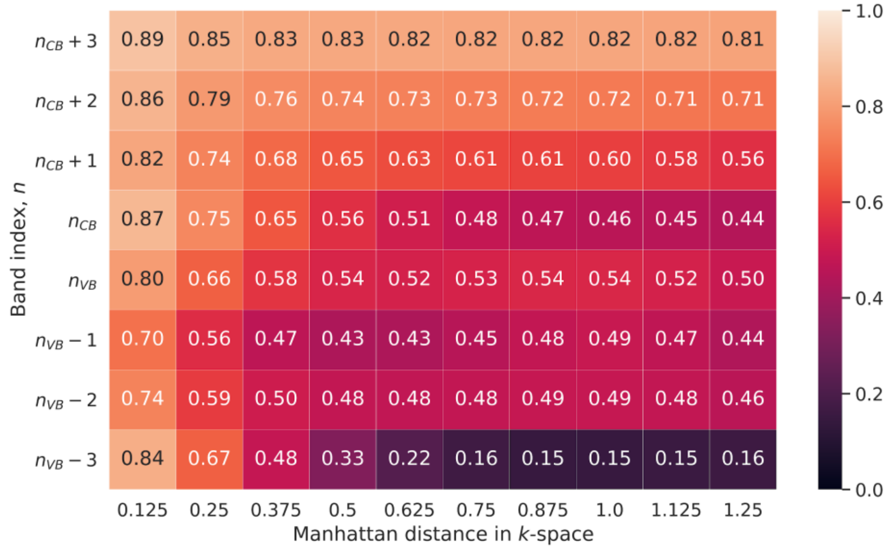


Figure A4.2. Average Pearson correlation coefficient between the energies in the same points in **k**-space but different bands. The correlation is strong in the case of adjacent bands: top VB ($n = n_{\mathrm{VB}}$), lowest CB ($n = n_{\mathrm{CB}}$) and its adjacent band ($n = n_{\mathrm{CB}} + 1$). This is exploited by the convolutions over the band dimension in the CNN layers, which introduces the interband correlation.

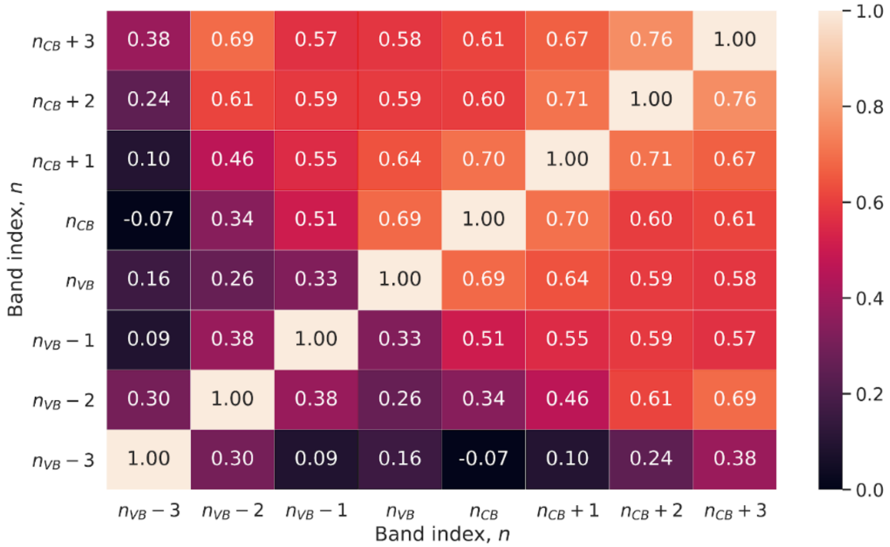


Table A4.1. Accuracy comparison among specialized models. This means that all the models (except for the CNN) were trained for the selected task only.

	CNN	NN	KRR
Bandgap prediction			
RMSE, eV (relative error, %)	0.108214 (2.22%)	0.096223 (1.83%)	0.168535 (3.21%)
MAE, eV (relative error, %)	0.072424 (1.38%)	0.062605 (1.19%)	0.122125 (2.33%)
Γ gap prediction			
RMSE, eV (relative error, %)	0.088265 (1.62%)	0.085658 (1.57%)	0.439139 (8.05%)
MAE, eV (relative error, %)	0.053497 (0.98%)	0.055520 (1.02%)	0.347457 (6.37%)
CBM prediction (classification problem)			
Accuracy	98.56%	94.30%	66.20%*
Inference time			
Time	14.4 ms ± 59.2 µs	1.29 ms ± 14.6 µs	48.8 ms ± 287 µs

* Linear model was used instead of radial basis function kernel

Table A4.2. Accuracy comparison among ensemble models. This means that all the models (except for CNN) were trained for the VB and/or CB prediction; other measures were inferenced as in the proposed model.

	CNN	NN	KRR
VB prediction			
Mean RMSE, eV (relative error, %)	0.038464 (0.23%)	0.052195 (0.31%)	0.043643 (0.26%)
Mean MAE, eV (relative error, %)	0.031379 (0.19%)	0.042052 (0.25%)	0.035710 (0.21%)
CB prediction			
Mean RMSE, eV (relative error, %)	0.045981 (0.30%)	0.111479 (0.72%)	0.059352 (0.38%)
Mean MAE, eV (relative error, %)	0.035453 (0.23%)	0.091714 (0.59%)	0.042620 (0.27%)
Inferenced bandgap prediction			
Mean RMSE, eV (relative error, %)	0.108214 (2.22%)	0.158525 (3.02%)	0.101998 (2.13%)

Mean MAE, eV (relative error, %)	0.072424 (1.38%)	0.120696 (2.30%)	0.082020 (1.56%)
Inferenced Γ gap prediction			
Mean RMSE, eV (relative error, %)	0.088265 (1.62%)	0.149067 (2.73%)	0.097539 (1.79%)
Mean MAE, eV (relative error, %)	0.053497 (0.98%)	0.105048 (1.92%)	0.063617 (1.17%)
CBM prediction (classification problem)			
Accuracy	98.56%	95.50%	93.75%
Inference time			
Time	14.4 ms ± 59.2 μs	2.48 ms ± 32.2 μs	25 s ± 105 ms

Chapter 5. Semiconductor electronic bandgap mapping in experimentally feasible loading geometry

5.1. Chapter introduction

Experimental realization of ultra-large elastic deformation in nanoscale diamond and machine learning of its band structures have created opportunities to address new scientific questions. Can diamond, with an ultrawide bandgap of 5.6 eV, be completely metallized, solely under mechanical strain without phonon instability, so that its bandgap fully vanishes?

The answer is yes, based on the results shown in Figure 4.5a. This is a quick answer given by first-principles modelers though. In previous chapters, models capable of predicting any band structure related properties are proposed. One can resort to the density of states of bandgap plot for not only knowing it is possible to metallize diamond, but also finding the most energy-efficient strain pathway to achieve metallization. However, is this strain pathway achievable in experimentally feasible loading geometries of diamond? After all, a sample in the real-world cannot be arbitrarily deformed by strain states such as that in Figure 3.1d.

The immediate intuition is to build up such a geometry and conduct simulations to get out the properties. However, the first-principles simulation capable of accurately evaluate the electronic bandgap cannot deal with such a large system with tons of atoms (the DFT calculation time scales by $O(N^3)$, where N is the number of atoms). Finite-element (FEM) simulations, on the other hand, can easily deal with geometry as such and evaluate the strain distributions inside the material, but it does not have any functionality to compute bandgap. How to integrate the power of both computational tools to map out the distribution of bandgap, or any other electronic FoM, in a deformed geometry?

In this chapter, we try to address this question. Through first-principles calculations, FEM validated by experiments, and deep learning, we show here that metallization/de-metallization, as well as indirect-to-direct bandgap transitions, can be achieved reversibly in diamond below threshold strain levels for phonon instability. We identify the pathway to metallization within 6D strain space for different sample geometries. We also explore conditions that promote phase transition to graphite. These findings offer opportunities for tailoring properties of diamond via strain engineering for electronic, photonic, and quantum applications.

In the subsequent sections, we focus on addressing the following scientific questions, respectively:

- Is it possible to metallize diamond at room temperature and pressure, from its natural unstrained state with an ultrawide electronic bandgap of 5.6 eV to full metallization with 0-eV bandgap, without phonon instability or structural transformation such as graphitization, solely through the imposition of strain? (Section 5.2)

- What are the conditions that trigger indirect-to-direct bandgap electronic transition, or a competing graphitization phase change, in diamond under straining? (Section 5.2)
- How much of such “safe” metallization can be realized within deformation conditions that have already been shown to be achievable experimentally? (Section 5.3)
- How do crystallographic and geometric variables influence the metallization of diamond? (Section 5.3)
- What are the strain states and a viable low strain energy density path to achieve such “safe” bandgap metallization in other loading geometries? (Section 5.4)

Some argumentation and figures/tables in this chapter are directly taken from the author’s own publication of Ref. [115]: Z. Shi*, M. Dao*, E. Tsymbalov, A. Shapeev, J. Li and S. Suresh, *Metallization of diamond*, Proc. Natl. Acad. Sci. **117**, 24634 (2020).

5.2. “Safe” metallization of diamond

In this section, we demonstrate that it is possible to achieve 0-eV electronic bandgap in diamond exclusively through the imposition of reversible elastic strains, without triggering phonon instability or phase change [9,135]. This discovery implies that reversible metallization/de-metallization is feasible through the design of mechanical loading conditions and geometry in nanoscale diamond. We further show that “safe” metallization can be achieved at elastic strain energy density values comparable to what has been demonstrated in experiments of reversible deformation of diamond nanopillars [9,135]. Our results also reveal that even simple bending of low-index <110> oriented monocrystalline diamond nanoneedles can effectively reduce the bandgap from 5.6 eV down to 0 eV without phonon instability, at about 10.8% local compressive elastic strain. Further bending the nanoneedle can however induce phonon instabilities [9] that lead to irreversible $sp^3 \rightarrow sp^2$ (diamond to graphite) phase transition or fracture. Indeed, plasticity induced by such $sp^3 \rightarrow sp^2$ phase-transition has recently been observed in the large bending of a single-crystalline diamond pillar [132], substantially agreeing with our calculations. Similar graphitization transition is also seen in nanoindentation experiments [126]. Navigating the treacherous elastic strain space above 80 meV/Å^3 or at > 9% local compressive or tensile principal elastic strain to induce complete metallization in diamond without encountering phonon instabilities is an important demonstration for power electronics, optoelectronics and quantum sensing systems, the pursuit of which is a primary objective of this investigation.

Whether mechanically strained or not, the absence of imaginary phonon frequency for the wavevector in the entire Brillouin zone is the hallmark of a locally stable crystal lattice [6,9,136]. If a strained perfect crystal lattice has a stable phonon band structure, then at T = 0 K and in the absence of defects such as free surfaces, interfaces and dislocations, this lattice is guaranteed to avoid spontaneous phase transition or defect nucleation. Consequently, phonon stability is the minimal requirement for lattice stability and loading reversibility [9]. If such a phonon-stable diamond can have zero electronic bandgap, $E_g = 0$ eV (reduced from E_g = 5.6 eV at zero strain), then this extreme electronic material [33] is expected to demonstrate unprecedented functional flexibility, from ultrawide bandgap semiconductor to the far-infrared and even metallic, in one material, without any change in chemical composition and possibly under dynamic loading. The electronic band structures of diamond under tensorial strain can be predicted with high accuracy based on *ab initio* DFT followed by GW calculations [49]. However, because GW calculations are computationally expensive, it is necessary to invoke a stress-strain constitutive law for modeling large elastic deformation of diamond in any arbitrary sample geometry, along with fast proxy models for the electronic and phonon band structures. In this work, we employ ML algorithms of band structures (as introduced in previous chapters of this thesis), so as to perform coupled *ab initio* and finite element calculations with constitutive laws based on NNs (see Methods for details).

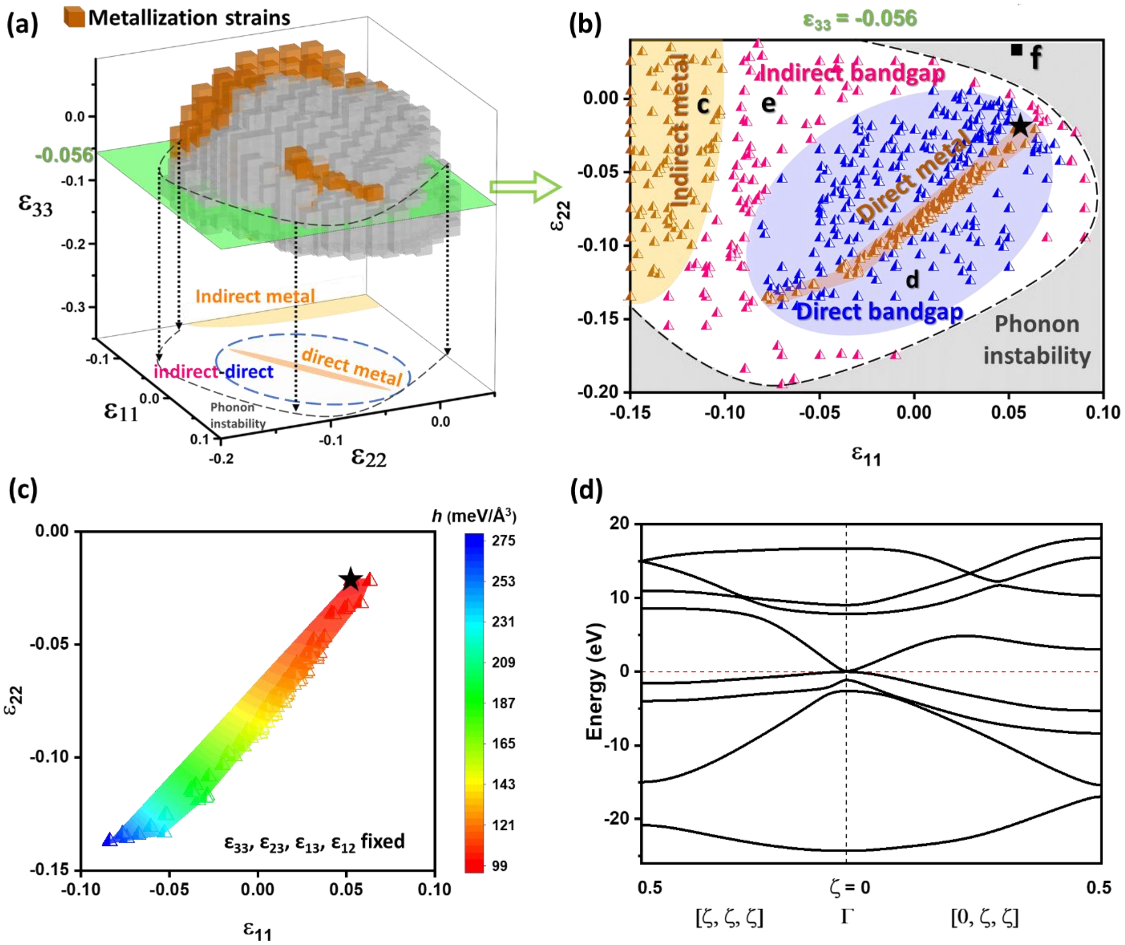


Figure 5.1 Metallization of diamond. (a) Stratification of the strain hyperspace into regions of metallization and bandgap transition in diamond. Metallization in elastically strained diamond for different values of normal strain components ε_{11}, ε_{22} and ε_{33}, with the other three strain components held fixed. The plane with fixed ε_{33} cuts the 3D volume and results in a projection onto the ε_{11}-ε_{22} 2D plane. (b) Detailed characterization of the ε_{11}-ε_{22} strain space includes a region of direct metal (brown) strains within the region of direct bandgap (blue) strains and a

region of indirect metal (brown) strains within the nonzero indirect bandgap strains (white zone with magenta symbols). Alternative visualization of the metallization strains in Figure 5.1a is presented in Section 5.6 Figure A5.2. (c) Elastic strain energy density (h) analysis of the direct metal region in (a). The axes in (c) are absolute strain component values of ε_{11} and ε_{22}, with the other four strain components fixed. Color contours indicate regions of constant *h* for different deformation states. (d) GW band structure showing complete closure of bandgap leading to metallization of diamond which is subjected to deformation at a 6D strain state in the [100][010][001] coordinate frame. This particular strain case corresponds to the black star symbol in (d).

We first present some 6D strain states in Figure 5.1 which make the bandgap of diamond vanish without phonon instability or graphitization. In the crystallographic [100][010][001] coordinate frame, our calculations revealed one such complete and "safe" metallization region as shown in Figure 5.1a. It illustrates a region of "safe" metallization of diamond without phonon instability and demonstrates reversible indirect-to-direct bandgap transitions under large elastic strains. Possible strain states in the 3D space of normal strains ε_{11}, ε_{22} and ε_{33} within which "safe" metallization is induced (highlighted in brown color) are shown. Regions of metallization are also plotted in Figure 5.1b in the 2D strain space of ε_{11} versus ε_{22}, with the other four strain components held fixed (i.e. formed as a result of 2D projection out of 3D strain region tessellated by cubes onto the plane $\varepsilon_{33} = -0.056$ in Figure 5.1a). The triangle data points of different colors in Figure 5.1b represent results of computational simulations of the effect of mechanical strain on bandgap and band structure. Two types of "safe" metallization, direct metal and indirect metal (where the bandgap that closes is indirect, i.e. from two different **k**-points), are identified.

The 2D region of direct metal, shaded in brown, is embedded within the strain space of direct bandgap (blue region, Figure 5.1b). The contours of strain energy density are plotted in 2D strain space in Figure 5.1c. Figure 5.1d is a plot of the GW band structure for diamond deformed to this particular strain state, resulting in a direct metal (see Section 5.6 Figure A5.1 for comparison of GW band structure with that for DFT). Note that the strains and strain energy density values in Figure 5.1 are comparable to the values achieved experimentally [29,30] in reversible ultra-large elastic bending of diamond nanoneedles or pillars.

In Figure 5.2a, the GW band structure is plotted to illustrate such indirect-metal state at point c (Figure 5.1b) inside this zone of "safe" metallization. Examples of nonzero direct and indirect bandgap cases indicated by the band structure plots are shown in Figure 5.2b-c, respectively. The area shaded in gray outside of the dashed lines is the region of large elastic strains and unstable metallization where phonon instability leading to defect nucleation and/or phase transition occurs [9]. Figure 5.2d reveals pronounced reduction in phonon frequency and the occurrence of soft mode associated with strain point f in Figure 5.1b where phonon instability and associated phase transition from diamond to graphite takes place. The location of the special strain region containing metallization is not unique in a general 6D strain hyperspace and such stratified regions may exist in a broad range of semiconductors. Our findings offer a systematic strategy in the search for strain-engineered semiconductor to metal transition, indirect-to-direct bandgap transition, as well as phase transition.

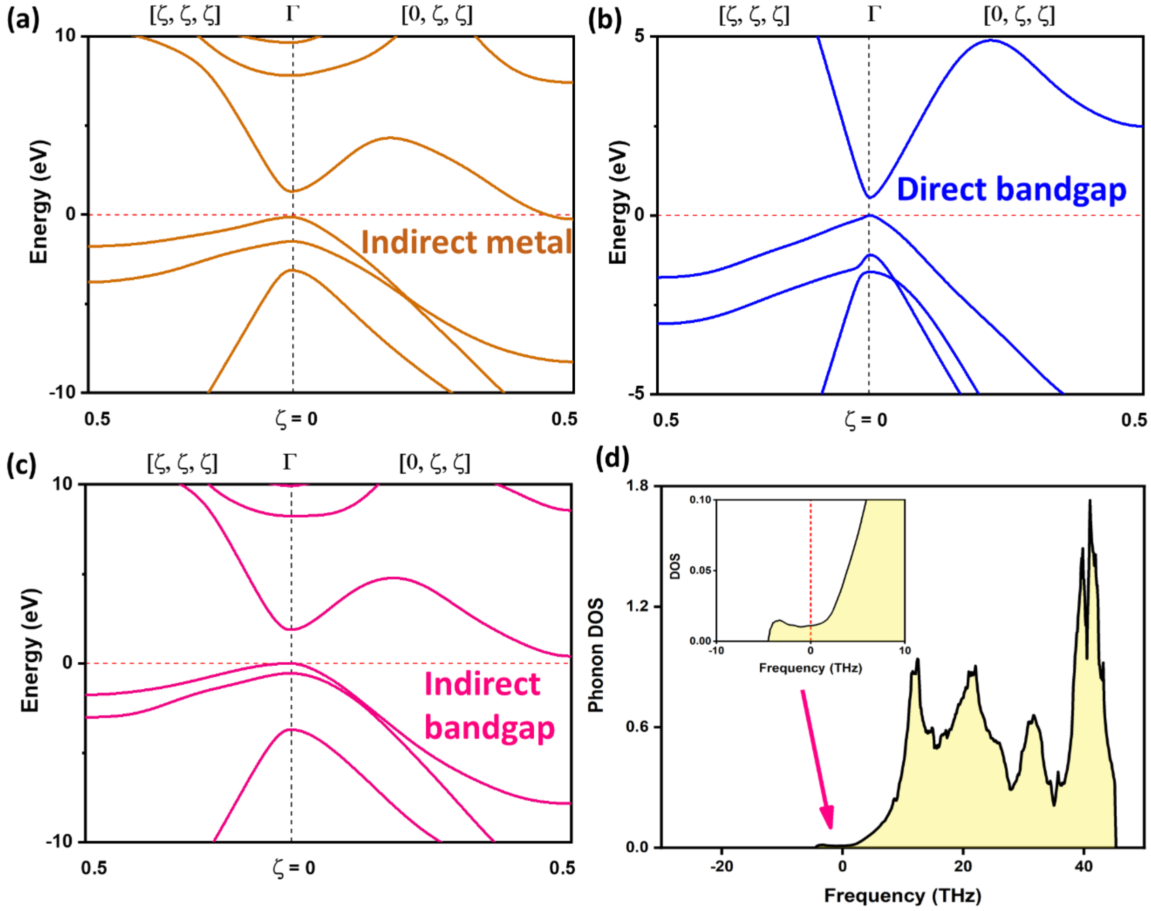


Figure 5.2 GW band structures and phonon stability of strained diamond. (a) Band structure of diamond strained within the "safe" metallization region resulting in an indirect metal. Strained diamond (b) with a direct bandgap (point d in Figure 5.1b), and (c) with an indirect bandgap (point e in Figure 5.1b). The strain region of phase transformation in diamond (usually associated with phonon instability) is shaded in gray color in Figure 5.1b. (d) A phonon DOS plot corresponding to point f in Figure 5.1b illustrates imaginary phonon frequencies (indicated by the magenta arrow) when structural instability occurs. A magnified view near-zero frequency is shown in the inset.

5.3. Joint machine learning-finite element modeling of bandgap

Experiments show that diamond nanoneedles exhibit ultra-large elastic bending before fracture [29]. Such deformation, resulting in local compressive strains larger than -10% and tensile strains in excess of 9%, is reversible upon release of the load. Here we apply simulations to determine bandgap modulation in bent diamond nanoneedles at maximum local strain levels that are known to be experimentally feasible (see Table 5.1). Figure 5.3a schematically illustrates the method whereby a diamond indenter tip pushes on a diamond nanoneedle to induce large deformation [29]. FEM is used to simulate the sideward bending moment of the diamond needle upon contact with the indenter tip and to account for nonlinear elasticity, the orientation of the cubic lattice with respect to the needle axis, the bending direction, and possible friction between the indenter tip and the needle.

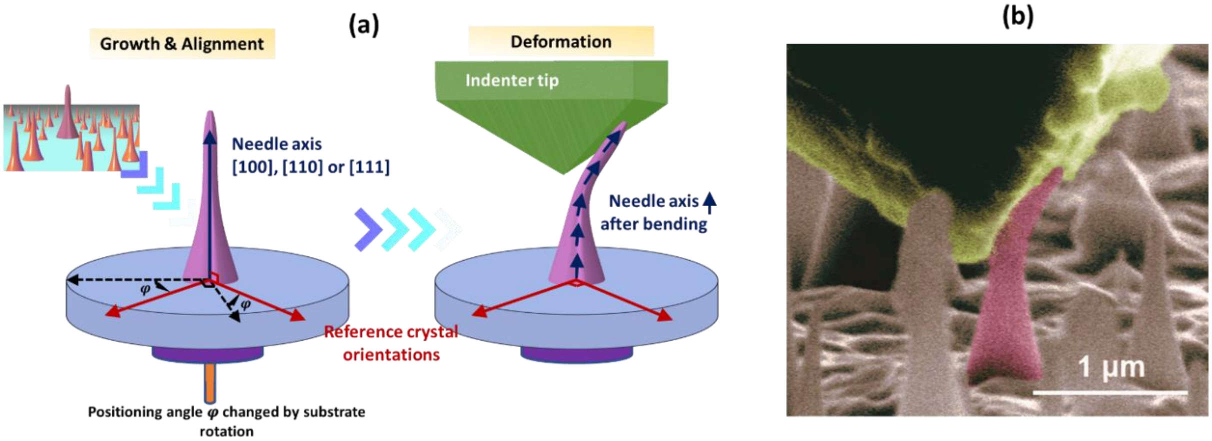


Figure 5.3 Experimental loading of diamond nanoneedles. (a) Schematic showing an as-grown, aligned, and bent nanoneedle. With the crystal orientation along the needle longitudinal direction known (blue arrows), the positioning angle φ defines the pre-selected crystal coordinates (black dotted arrows) versus the selected reference crystal coordinates (red arrows). φ is modulated by rotating the substrate in the alignment stage, introducing an additional degree of freedom and many more combinations of strain states in the bent needle. The needle is then bent when pushed by the side surface of a cube corner indenter tip, as described in Ref. [29]. Small blue arrows along the needle in the deformed configuration indicate the local crystallographic needle axis. (b) Bending of the diamond nanoneedle by diamond nanoindenter tip inside a scanning electron microscope. Micrograph is taken from Ref. [29].

Table 5.1 Limits for elastic strains and strain energy density from experiments [29,30,137] and calculations. The deformation of nanoneedle is limited by tension, i.e., failure first takes place on the tensile side of the nanoneedle. The strains listed below for the compressive strains in the nanoneedle correspond to the point at which failure first occurs on the tensile side of the nanoneedle. Since higher compressive strains can be achieved without failure in pure compression of diamond, the compressive strains listed below are lower bound estimates.

Crystallographic orientation of diamond nanoneedle	Experimental results		Theoretical limit for bending diamond nanoneedle (for experimental configuration)		Theoretical limit for “safe” metallization in general 6D strain space (this study)
	Tensile side	Compressive side	Tensile side	Compressive side	
<100>	13.4%	-14.0%	-	-	≤ 98.7 meV/Å^3
<110>	9.6%	-10.1%	12.1%	-14.5%	
<111>	8.8%	-9.9%	8.8%	-27.0%	

Figure 5.4a shows FEM results of local compressive and tensile strains of the deformed geometry of <110> diamond nanoneedle, with the maximum compressive and tensile strains of -10.8% and 9.6% respectively. The accuracy of FEM predictions is validated by direct comparison with experimentally measured indentation load plotted against displacement [29]. The corresponding predictions, from our simulations, of the distribution of bandgap are also plotted in Figure 5.4a. The onset of “safe” metallization appears in the severely strained compressive side of the nanoneedle at a local strain of -10.8%, as shown in Figure 5.3b. The propensity toward increasingly more metal-like behavior with increasing strain is independent

of friction between the indenter and the nanoneedle (see Section 5.6 Figure A5.3). The <110> nanoneedle can withstand up to 12.1% local tensile strain before incurring phonon instability on the tensile side, at a bandgap of 0.62 eV, as shown in Figure 5.4d. The maximum attainable local tensile strain of 9.6% on the tensile side of <110> single crystal natural diamond samples [30], as compared to theoretical predictions of higher values (Figure 5.4b and Table 5.1), could be attributed to the presence of dislocations and/or other surface-related defects [138–141]. The compressive side is more tolerant to deformation. The maximum attainable compressive strain could be on the order of -20% along a low-index orientation [137], suggesting that there is room for additional elastic deformation after achieving "safe" metallization in compression-dominated regions. Note that due to the zero-point motion effect [142] and the Varshni effect [143], for physical experiments performed at room temperature, the bandgap of diamond is expected to be even smaller than estimated here by 0.4-0.6 eV [144,145]. This understanding leads to the inference that safe metallization in diamond can occur at elastic strain levels somewhat smaller than indicated by our analysis, making it even more easily achievable than appears from the quantitative results plotted here (see Section 5.5 for details).

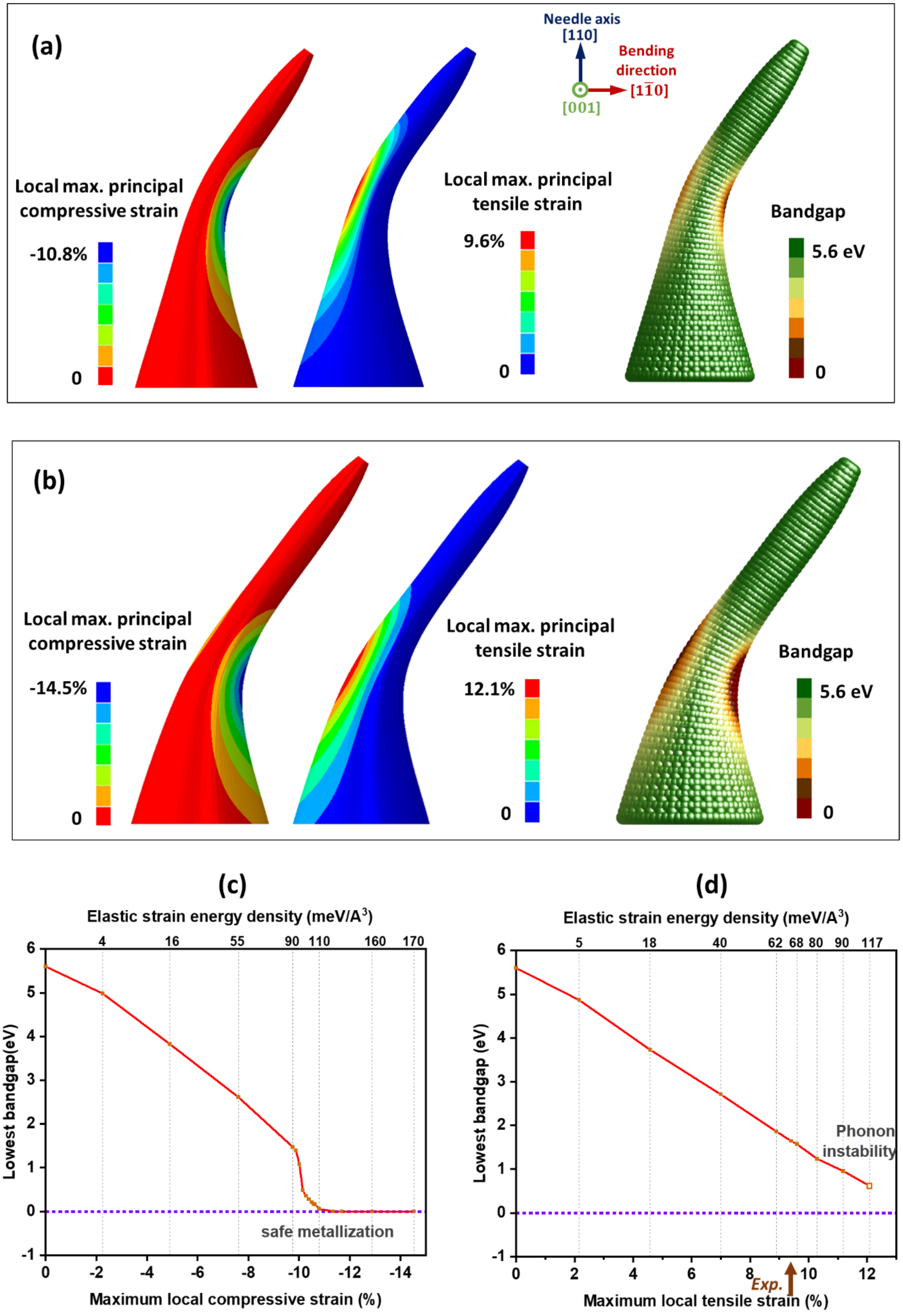

(a)
Needle axis
[110]
Bending direction
[1$\bar{1}$0]
[001]
Local max. principal compressive strain
-10.8%
0
Local max. principal tensile strain
9.6%
0
Bandgap
5.6 eV
0
(b)
Local max. principal compressive strain
-14.5%
0
Local max. principal tensile strain
12.1%
0
Bandgap
5.6 eV
0
(c)
Elastic strain energy density (meV/A^3)
4
16
55
90
110
160
170
Lowest bandgap(eV)
6
5
4
3
2
1
0
-1
0
-2
-4
-6
-8
-10
-12
-14
safe metallization
Maximum local compressive strain (%)
(d)
Elastic strain energy density (meV/A^3)
5
18
40
62
68
80
90
117
Lowest bandgap (eV)
6
5
4
3
2
1
0
-1
0
2
4
6
8
10
12
Phonon instability
Exp.
Maximum local tensile strain (%)

Figure 5.4 Metallization in diamond nanoneedles. (a) FEM predictions of the local compressive and tensile strain distributions and ML prediction of the distribution of bandgap for a diamond nanoneedle with its <110> crystallographic direction aligned with the needle axis. (b) FEM-ML predictions of the local maximum principal (compressive/tensile) strain and bandgap distributions for the <110> diamond nanoneedle deformed at the theoretically approachable maximum tensile strain of 12.1%. The FEM plots in (a) and (b) are contributed by Dr. Ming Dao. (c) Increasing magnitude of bending in the <110> nanoneedle causes a significant reduction in bandgap of diamond from 5.6 eV (zero strain) down to 0 eV for a maximum local compressive strain of -10.8% (the corresponding maximum local tensile strain on the tension side is 9.6%). (d) Local tensile strain beyond 12.1% results in fracture or graphitization on the tensile side of the nanoneedle according to our *ab initio* calculations, even when there are no pre-existing defects. See also Section 5.6 Figure A5.4 for the evolution of elastic strain energy density, bandgap and the corresponding band structure at the maximum compression site in the nanoneedle, showing the metallization process.

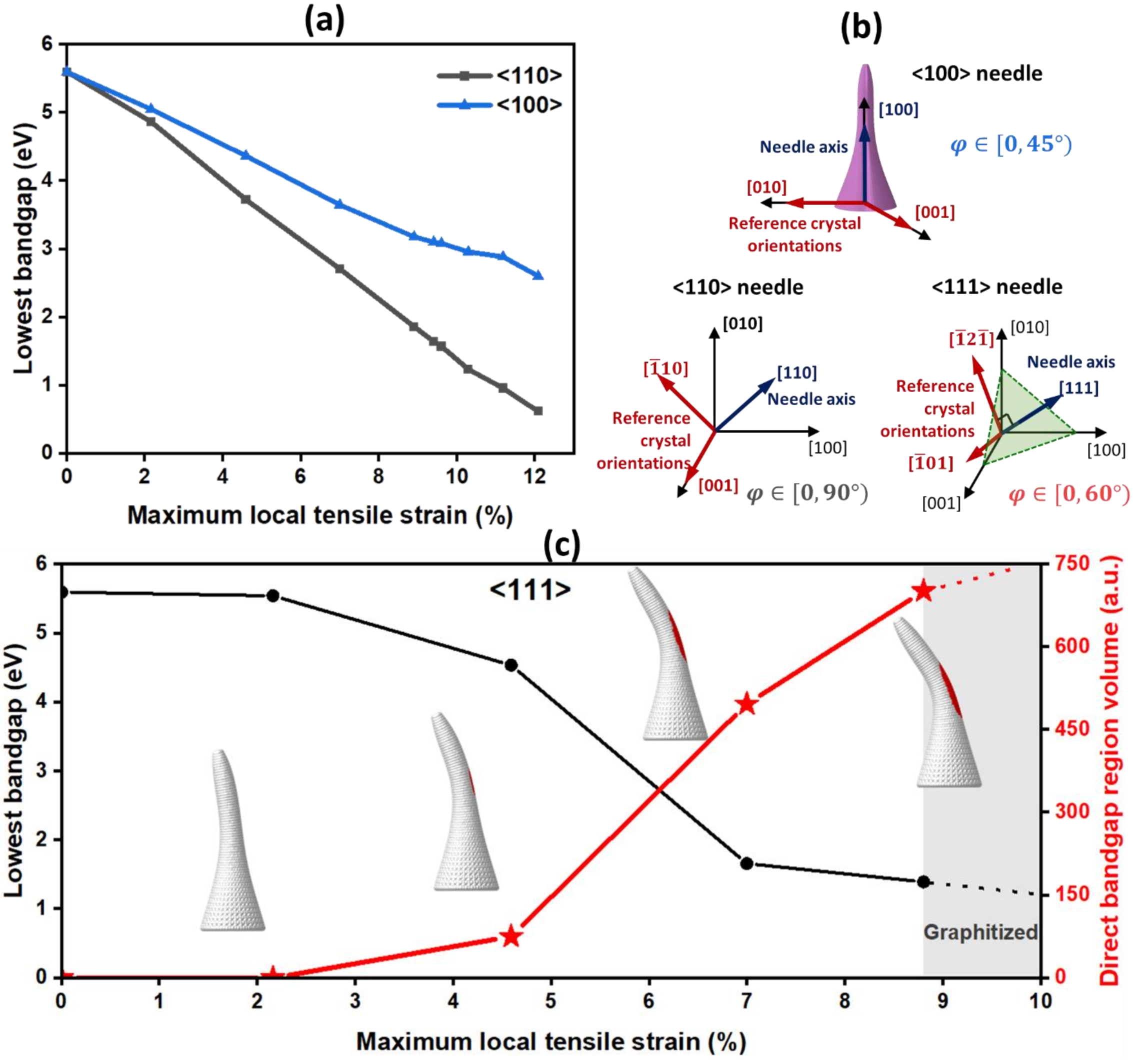


Figure 5.5 Orientation dependent bandgap changes and indirect-to-direct bandgap transitions. (a) Reduction of the lowest bandgap as a function of strain in nanoneedles of and orientations, respectively. (b) The definition of the reference crystal orientation for the three diamond nanoneedle families: [010]/[001] for a [100] needle, $[1\bar{1}0]/[001]$ for a [110] needle, and $[\bar{1}2\bar{1}]/[\bar{1}01]$ for a [111] needle. The green triangle indicates the (111) plane for the nanoneedle. (c) Reduction of the lowest bandgap (left axis) and development of direct bandgap region volume (right axis) in nanoneedles of orientation. These volumes are colored in red on the tensile side of the bent needles plotted next to the data points. Graphitization occurs in the nanoneedle right after 8.8% local tensile strain, as

indicated by the grey region. The direct bandgap region volume is expressed in terms of the number of FEM nodes that correspond to a direct bandgap.

Crystallographic orientation of the nanoneedle axis is another variable determining the extent of large deformation, and the resultant bandgap modulation. This orientation effect is illustrated in Figure 5.5a-b. Among the three types of nanoneedles studied, the <110> and <111>-oriented nanoneedles require relatively smaller tensile strains to reduce bandgap through straining, whereas the <100> orientation is the hardest orientation to reduce bandgap below 2 eV or approach metallization. This distinction can be attributed to the difference in flexibility to access all six components of the strain tensor expressed in the [100][010][001] coordinate frame. Despite the possibility of extremely large strain in a <100>-oriented nanoneedle, this orientation primarily facilitates normal strains (with the shear components ε_{23} , ε_{13} , and ε_{12} being relatively much smaller) and the resultant maximum bandgap reduction is limited before phonon instability is reached, causing fracture or phase transformation [9]. For deformation of the <110> and <111>-oriented needles, on the other hand, it is relatively easier to initiate both normal and shear strain components necessary for band structure engineering [146–149] and the resultant bandgap modulation. In the <111> oriented needles, these strain conditions further facilitate indirect-to-direct bandgap transitions in diamond. The spatial evolution of the “safe” direct bandgap regions in our nanoneedles can be found in Figure 5.5c. Bending direction is another geometrical factor. For a low-index oriented needle, we find bending direction has little influence on the maximum bandgap reduction in the bent needle.

Beyond the configurations considered here, more complex 3D loading geometries with holes and notches through topology optimization [125] and micro- and nano-machining of geometric features [150,151], can be designed without exposing the metallized zone to near-surface regions [152], further increasing possibilities for metallizing diamond. These methods for deep ESE are equally applicable to map the indirect-to-direct bandgap transition locations in diamond for the most general 6D straining case, as indicated in Figure 5.1a-b and Figure 5.2b. When a strained diamond is transformed into a direct bandgap semiconductor, even only locally at the site of maximum strain, it would exhibit a fundamental enhancement in its optical transitions around the adsorption edge compared to an undeformed diamond in its natural state. This transition arises from the absence of phonon involvement (momentum change of electron) in the adsorption or emission process. Since absorbance increases exponentially with thickness in a material, a light energy conversion device based on direct bandgap semiconductor with a high adsorption coefficient and rationally engineered bandgap value would require much less thickness to absorb the same amount of light with a variety of wavelengths, from the visible to the far-infrared. These considerations could pave the way for designing high-efficiency photo detectors and emitters from UV to the far-infrared on a single piece of diamond. As photons and excitons are the primary tools for quantum information processing, this extreme ability to mold diamond’s band structure will also be highly consequential for quantum sensing and quantum computing applications.

5.4. Loading geometry design

For future deep ESE practices in real-world devices, there are at least two more basic factors to consider: (i) The load-bearing structure should be easy to make into a device; (ii) Elastic strains should be relatively easy to be applied to the load-bearing structure. Therefore, it becomes immediately clear that standing nanoneedle structure, such as the ones in the previous section and those mentioned in Figure 1.1 of Section 1.1.1 is hard to be used as a device. But with the tools put forward in the previous section, we should be able to tackle (i) by coming up with geometries that are similar to that of a semiconductor device. But to tackle (ii), it turns out to be more challenging. The requirement for being “easy” in (ii) not only requires the straining pathway to be as energy-efficient as possible but also means we should try to use existing or experimentally feasible loading apparatus. Therefore, it becomes immediately clear that some of the peculiar strains, especially those with complex multi-dimensional strain states (Figure 3.1c-d) are not straightforward to achieve, and it may take years of significant efforts to be achieved in practical devices, despite Feynman’s prophecy to use “a hundred tiny hands” in the micro-world [26]. How to achieve a desirable FoM (again, say a 0-eV bandgap in diamond) in a familiar geometry by deep ESE with known experimental stressors at a reasonable amount of elastic strain energy cost?

To approach this four-fold challenge, we hereby come up with a relatively device-friendly diamond thin film design (most semiconductor electronic devices are based upon thin film technology), undergoing large biaxial straining in the horizontal directions and large uniaxial compression along the vertical direction. The FEM simulation result for this loading geometry is shown in Figure 5.6a, where a slab of diamond is compressed by vertically aligned spherical diamond indenters and at the same time experiencing in-plane equil-biaxial tension. The most highly strained region has a three-normal strain state of $\varepsilon_{11} = \varepsilon_{22} \approx -\frac{\varepsilon_{33}}{2}, \varepsilon_{23} = \varepsilon_{13} = \varepsilon_{12} \approx 0$. As shown in Figure 5.6b-c, this strain case gives a 0-eV bandgap in diamond. Separate GW band structure calculation (Figure 5.6d) and phonon stability check were conducted to confirm this case. It is also found that the onset of zero bandgap corresponds to $h \approx 98$ meV/Å^3, an energy cost that is smaller than what we proposed in the previous section.

Although only artificially incorporated in the study, we envision the large biaxial strain can be collected from two sources, namely substrate effect and external strain field. Previous theoretical [153] and experimental studies [154] have discussed the possibility of a 7.3% strain in diamond grown atop Si substrate with a 45° rotation. As for introducing additional strain, we point to the recent three-point bending design by Chi et al [45], which could be adapted to elevate the biaxial strain level and perform *in-situ* electrical properties measurement.

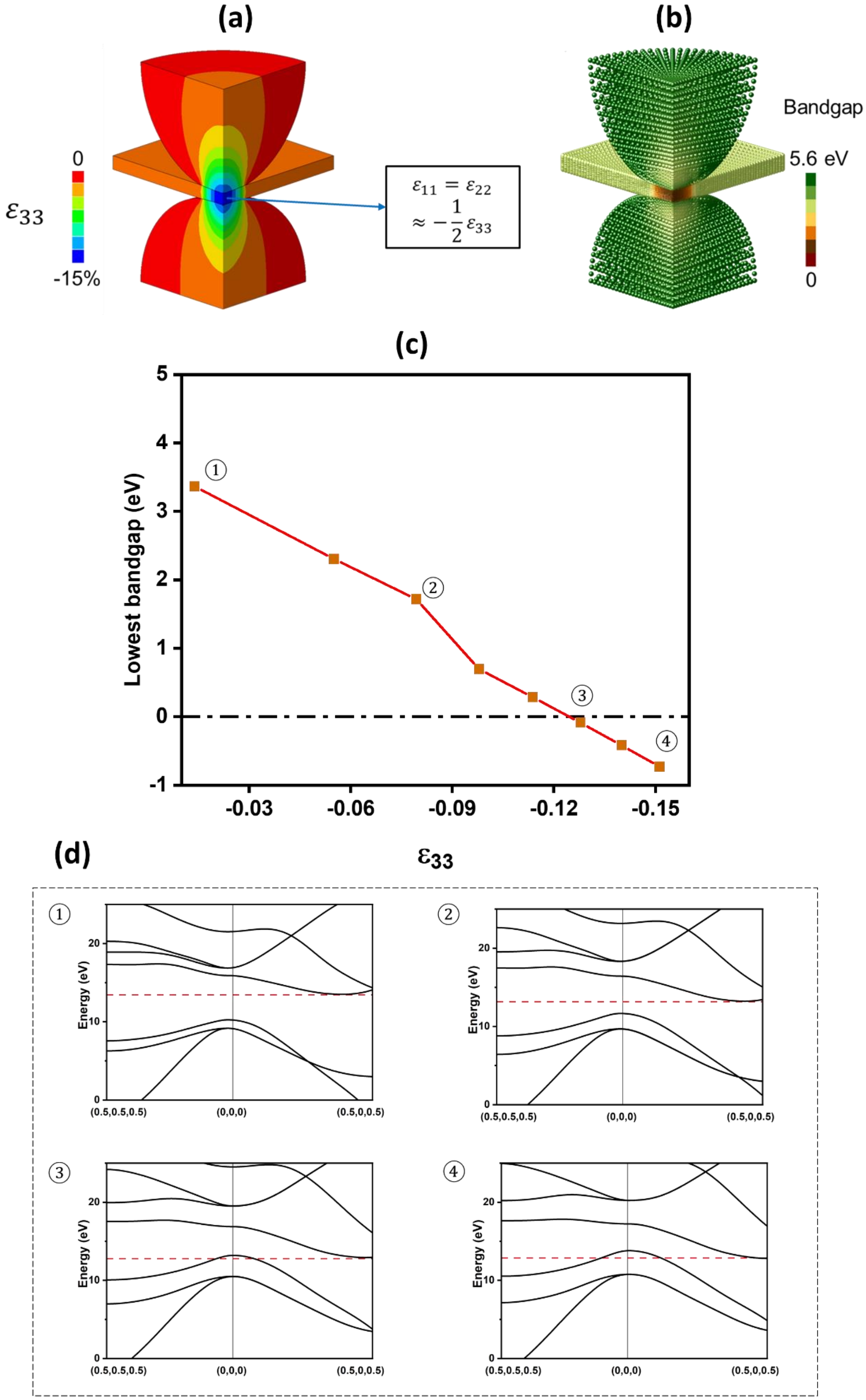

(a)
(b)
0
ε_{33}
-15%
$\varepsilon_{11} = \varepsilon_{22} \approx -\frac{1}{2}\varepsilon_{33}$
Bandgap
5.6 eV
0
(c)
Lowest bandgap (eV)
5
4
3
2
1
0
-1
-0.03
-0.06
-0.09
-0.12
-0.15
①
②
③
④
ε_{33}
(d)
①
②
③
④
Energy (eV)
0
10
20
(0.5,0.5,0.5)
(0,0,0)
(0.5,0,0.5)

Figure 5.6 Metallization of diamond in designed geometry. (a) FEM predictions of ε_{33} distributions in the loading geometry. The diamond thin film is compressed by two vertically aligned spherical diamond indenters on both sides (contributing to ε_{33}) and at the same time equil-biaxially stretched by $\varepsilon_{11} = \varepsilon_{22} \approx 7\%$. The plot is contributed by Dr. Ming Dao. (b) ML mapping of bandgap in the loading geometry. (c) Reduction of bandgap as increasing compression in the vertical direction. As strain state approaches $\varepsilon_{11} = \varepsilon_{22} \approx -\frac{\varepsilon_{33}}{2}$, the metallization of diamond is achieved. Not that the bandgap does not start to drop from 5.6 eV since we applied biaxial strain before the compressive loading. (d) The GW band structure plots of four cases labeled in (c). In this case the most deeply strained region in the diamond will turn into an indirect metal.

5.5. Theoretical details

First-principles calculations

In this section, the same VASP [108] was used for DFT calculations to predict the evolution of bandgap and band structure of diamond subjected to mechanical deformation. We invoked the generalized gradient approximation in the form of PBE exchange-correlation [106] functional and the projector augmented wave method (PAW) [107] in our DFT computation. A plane-wave basis set with an energy cutoff of 600 eV was adopted to expand the electronic wavefunctions. The Brillouin zone integration was conducted on a 13 × 13 × 13 Monkhorst-Pack [109] **k**-point mesh and atomic coordinates in all the structures were relaxed.

GW corrections were performed when bandgap evaluations were needed. It is known that an extremely accurate GW calculation would involve choosing “infinitely” large values for several interdependent parameters [155,156]. Given the situation that we need to construct a huge dataset of GW bandgaps for machine learning purposes and conduct many calculations for varied 6D strain cases, we hereby struck a balance between efficiency and effectiveness. Specifically, we chose the **q**-grid to be 6 × 6 × 6, the screened cutoff to be 600 eV, and the number of bands for both dielectric matrix calculation and Coulomb hole summation to be 600. In addition, beyond the single-shot G_0W_0 method, we allowed two to three iterations of Green’s function in our calculations to obtain accurate quasi-particle shifts. This partially self-consistent GW_0 calculation is known to yield results that are in agreement with available experimental measurement for semiconductor materials [134] and better than plain DFT calculations using hybrid functionals [133]. For undeformed diamond, our calculation indicates a +1.5 eV GW correction to the DFT-PBE bandgap, which matches values reported in recent literature [157]. For general 6D strain cases, this correction may vary (see Section 5.6 Figure A5.1 for an example).

We also acknowledge that, even at 0 K, due to the quantum zero-point motion, further corrections need to be made to the electronic levels of diamond. This renormalization of bandgap could be -0.6 eV to -0.4 eV for undeformed diamond [144,145]. We consider this correction value to be negative in other cases of our interest. According to the temperature-dependent “adiabatic Allen-Heine formula” [142,158], by setting $T = 0$ to zero-out the Bose-

Einstein occupancy factors, the zero-point renormalization of the band structure ($\Delta E_{n\mathbf{k}}^{\mathrm{ZP}}$) arising from the electron-phonon interaction could be expressed as:

$$\Delta E_{n\mathbf{k}}^{\mathrm{ZP}} \equiv \Delta E_{n\mathbf{k}}(T = 0) = \sum_{\nu} \int \frac{d\mathbf{q}}{\Omega_{\mathrm{BZ}}} \left[\sum_{n'} \frac{|g_{nn'\nu}(\mathbf{k}, \mathbf{q})|^2}{\varepsilon_{n\mathbf{k}} - \varepsilon_{n'\mathbf{k}+\mathbf{q}}} \right] + \Sigma_{n\mathbf{k}}^{\mathrm{DW}}, \tag{32s}$$

where ε_{nk} is the single-particle eigenvalue of an electron with crystal momentum $\mathbf{k}$ in the band n, the integral is over the Brillouin zone of volume Ω_{BZ}, the outermost summation is over all phonon branches ν, and the first-order electron-phonon matrix elements $g_{nn'\nu}(\mathbf{k}, \mathbf{q})$ describes the scattering from an initial state with wave vector $\mathbf{k}$ to a final state with wave vector $\mathbf{k} + \mathbf{q}$, with the emission or absorption of a phonon with crystal momentum $\mathbf{q}$ belonging to the phonon branch ν. The first term on the right-hand side is the Fan-Migdal self-energy term [159] and the $\Sigma_{n\mathbf{k}}^{\mathrm{DW}}$ term is the Debye-Waller (DW) self-energy term. Given the DW term are normally much smaller than the Fan-Migdal term (about 1:5 in diamond [144]), the deciding factors to the sign of $\Delta E_{n\mathbf{k}}^{\mathrm{ZP}}$ are the denominators $\varepsilon_{n\mathbf{k}} - \varepsilon_{n'\mathbf{k}+\mathbf{q}}$. The change of bandgap can be qualitatively evaluated by considering the relative shift of the VBM and CBM. For VBM, we can further assume the coupling primarily comes from scattering within the valence bands. Since no values of $\varepsilon_{n'\mathbf{k}+\mathbf{q}}$ in the valence bands can be larger than $\varepsilon_{n_{\mathrm{VBM}}\mathbf{k}}$, the denominators $\varepsilon_{n_{\mathrm{VBM}}\mathbf{k}} - \varepsilon_{n'\mathbf{k}+\mathbf{q}}$ would always be positive and the resultant $\Delta E_{n_{\mathrm{VBM}}\mathbf{k}}^{\mathrm{ZP}}$ would also be positive. Similarly, $\varepsilon_{n_{\mathrm{CBM}}\mathbf{k}} - \varepsilon_{n'\mathbf{k}+\mathbf{q}}$ at CBM and the resultant $\Delta E_{n_{\mathrm{CBM}}\mathbf{k}}^{\mathrm{ZP}}$ would always be negative. The upward shift of VBM and downward shift of CBM would, therefore, result in an overall reduction in the computed bandgap of diamond. Therefore, from this perspective, we provided a generally conservative estimation of the strain magnitude required for engineering the bandgap. The actual bandgap may be even smaller than we predicted at particular strain levels as in Figure 5.4, allowing metallization to be safely achieved more easily.

Diamond primitive cells were used for DFT and GW calculations. Supercells were not used in order to circumvent the problem caused by band folding when determining the direct/indirect nature of the bandgap. All band structures were plotted by VASP with a Wannier90 interface [160–162].

To identify the phonon instability boundaries, we performed phonon stability calculations for densely sampled strain points in 3D or 2D strain space. These calculations were primarily carried out using the VASP-Phonopy package [163]. 3 × 3 × 3 supercells were created, and phonon calculations were conducted with a 3 × 3 × 3 **k**-point mesh. Whenever accurate phonon stability check was needed for diamond primitive cell, DFPT [164] as implemented in Quantum ESPRESSO [165] was adopted, with a dense 11 × 11 × 11 **k**-grid and 6 × 6 × 6 **q**-grid.

Machine learning

The bandgap distribution in diamond nanoneedles deformed to different strains was computed using machine-learning algorithms. This is done by representing deformation as a strain tensor and using a NN to fit the strain states against respective bandgap values obtained accurately by first-principles calculations. The NN fitting is implemented within the TensorFlow framework, an end-to-end open-source machine learning platform released by Google [110]. The specific design, similar to our previous work [146], involves a feed-forward architecture with hidden layers capable of learning the variations of band structure and bandgap with respect to large mechanical deformation. In order to integrate both the PBE and GW datasets we prepared by first-principles calculations and to produce more consistent and accurate machine learning outcomes, the "data fusion" technique same as our work in Ref. [146] was used. It took the quantitative advantage of PBE and the qualitative advantage of GW by interpolating between them to achieve decent NN fitting results with only ~10^4 PBE and ~10^3 GW calculations, successfully alleviating the need for the otherwise impractical sub-million-level amount of computations.

Finite element modeling

The ABAQUS (Dassault Systèmes Simulia Corp., Providence, RI, USA) software package was employed to conduct FEM analyses on specimen models, which replicated the 3D geometry of the diamond nanoneedles. Both the nanoneedle and the cube corner indenter were treated as deformable solids with the same material properties. A sliding contact was specified between the tip of the nanoneedle and the top surface of the indenter. Geometric nonlinearity induced by large deformation was accounted for. Neo-Hookean nonlinear elasticity model was used to simulate large deformation. The equivalent small-strain Young's modulus is 1100 GPa and the Poisson's ratio 0.0725 [29]. Since friction makes a negligible change to the deformed shape, the friction coefficient between the nanoneedle and the indenter was taken to be 0.1.

5.6. Chapter appendix

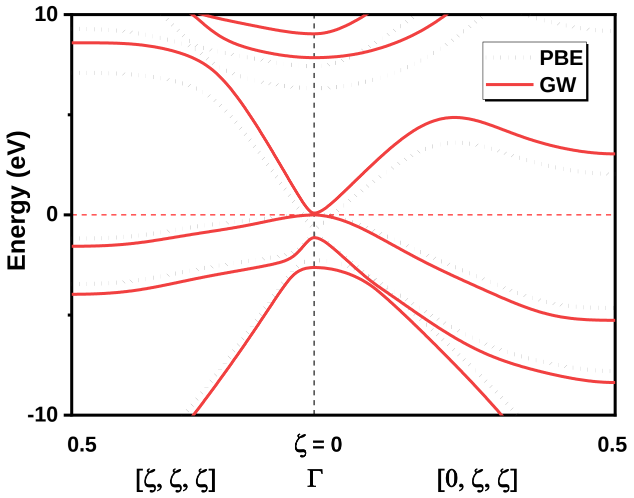


Figure A5.1. Deformed diamond band structures plotted in the scheme of DFT-PBE and GW. The 6D strain case is the same as in Figure 5.1d. There is about +0.68 eV GW correction in the DFT-PBE bandgap at this particular strain case.

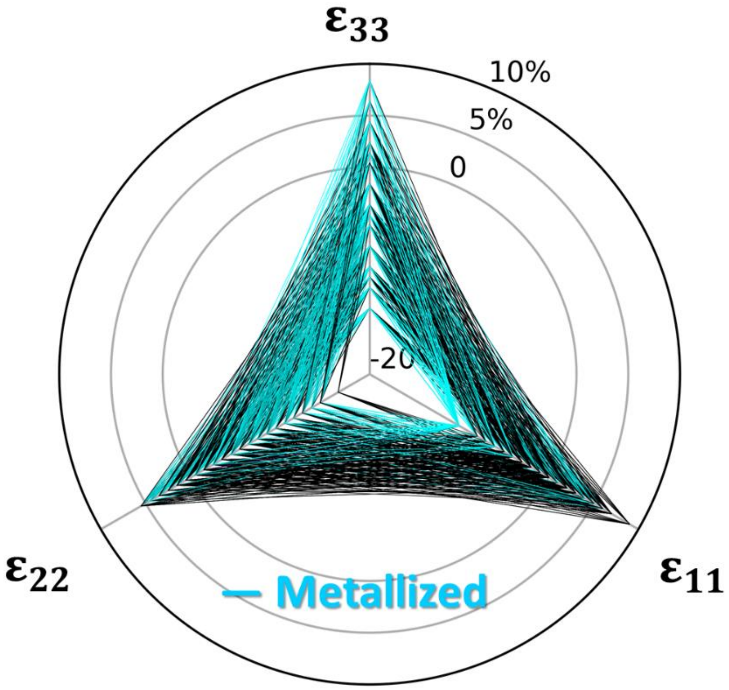


Figure A5.2. Spiderweb-plot illustrating the metallization strain cases (colored as cyan webs) in the 3D space of normal strains ε_{11}, ε_{22} and ε_{33} spanning −20% (i.e. compressive strain of 0.2) to +10% (i.e. tensile strain of 0.1), with shear components $\varepsilon_{23}, \varepsilon_{13}, \varepsilon_{12}$ all fixed to be constants as in Figure 5.1a. Strain components of the same magnitude belong to the same concentric circle in the plot.

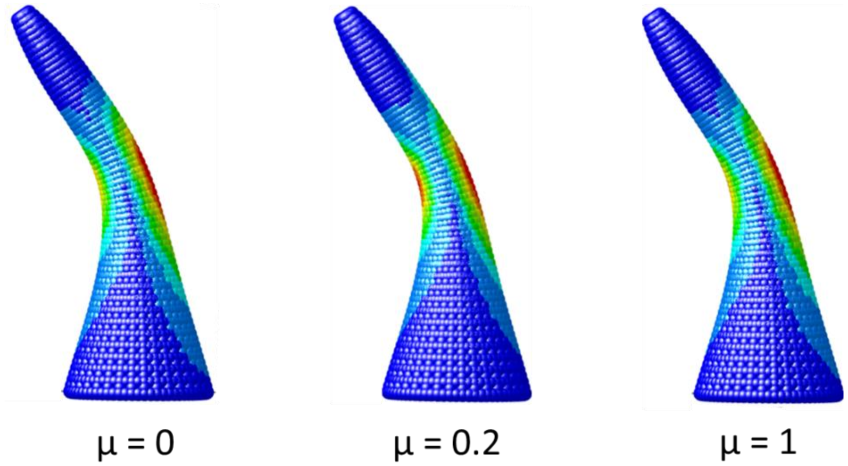


Figure A5.3 ML prediction of the bandgap distribution for the same <111> nanoneedle bent by the same amount and friction coefficient μ from 0 (perfectly smooth contact) to 1. The propensity of bandgap reduction during deformation is seen from our simulations to be independent of the level of friction between the indenter and the nanoneedle.

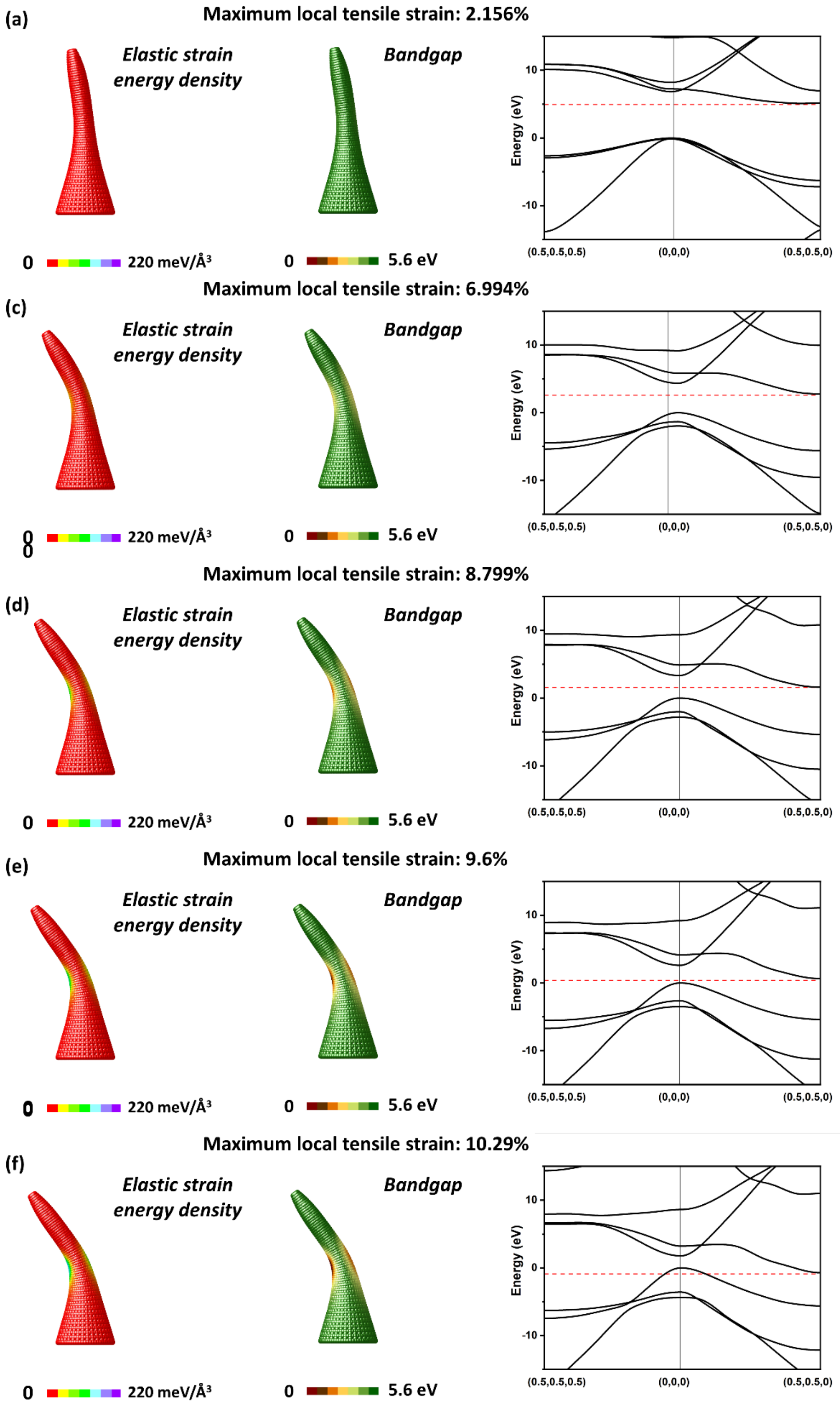


Figure A5.4. Evolution of elastic strain energy, bandgap, and the corresponding band structure at the maximum compression site in the nanoneedle, showing the metallization process.

Chapter 6. Engineering phonon and defect related properties by machine learning

6.1. Chapter introduction

As mentioned in Chapter 1, there are the idealistic limit ($\boldsymbol{\varepsilon}_{\text{ideal}}$) and realistic limit ($\boldsymbol{\varepsilon}_{\text{realistic}}$) to ESE. The former, also know as the ideal strain, is the upper bound of reachable strain in a perfect material at T = 0 K beyond which relaxation will find a way to set in, either through fracture, plasticity or phase transition. It should be noted that both the zero-temperature and defect-free requirements are too absolute to be true in a realistic experimental environment. The fact that every material must have a surface already makes the $\boldsymbol{\varepsilon}_{\text{ideal}}$ unattainable. In practice, the $\boldsymbol{\varepsilon}_{\text{realistic}}$ appears to be much more conservative and is a proper subset of $\boldsymbol{\varepsilon}_{\text{ideal}}$ in the 6D strain space, by considering temperature, microstructure, and defects that are present in the materials. Typically, the $\boldsymbol{\varepsilon}_{\text{ideal}}$ of deep ESE ought to be significantly larger than that of conventional ESE. It is no wonder that all the recent deep ESE experiments were conducted using nanostructured materials such as silicon and diamond, whose small sizes and single crystalline nature almost left no room for defects such as dislocations, grain/twin boundaries, etc. and made possible the approach of $\boldsymbol{\varepsilon}_{\text{ideal}}$.

Even though $\boldsymbol{\varepsilon}_{\text{realistic}}$ is more directly usable, it varies case by case and it is typically hard to be reproduced given slightly different experimental conditions even for the same material. With a limitation in computational tools and predictive capabilities of AI, in the first part of this chapter, the author focuses on mapping $\boldsymbol{\varepsilon}_{\text{ideal}}$ of a semiconductor in the 6D strain hyperspace, which is a simpler ML target and warrants numerical research.

Unwelcome as higher-dimensional defects are for deep ESE, 0-dimensional (point) defects in the format of dopants can sometimes be a good thing for enhancing functional electronic properties of materials. For example, to function as a semiconductor or conductor, a wide bandgap material may include defects with additional localized electronic states inside the bandgap of the material, but proximate to the edges of the CBM and/or the VBM. If the energy difference between the localized electronic states of the defect and either the CBM (n-doping) or the VBM (p-doping) is sufficiently small, then it is possible for the defects to be ionized by thermal fluctuation energy.

A material may comprise defects in suitable concentrations measurable by experimental methods such as X-ray photoelectron spectroscopy (XPS). Compared to the host element, the existence of these dopants is extremely rare, usually quantified by ppm, and normally has negligible detrimental effects on the mechanical strength of the material. Since their pros way outweigh the cons, dopants are often actively introduced and dispersed in the host material by means such as ion implantation and diffusion.

Defect doped material may be used in a variety of suitable semiconductor devices including, for example, photonic devices, optoelectronic devices, high-speed electronic devices, spintronic devices, photovoltaic devices, light-emitting devices (e.g., light-emitting diodes), and the like. In the second part of this chapter, the author applies ML to the study of point defect properties in semiconductors undergoing deep ESE. As before, the versatile ultra-wide bandgap material diamond is chosen to showcase the ML outcomes.

Some argumentation and figures/tables in this chapter are directly taken from the author's own manuscript in preparation as well as patent of Ref. [166]: *Elastic Strain Engineering of Defect Doped Materials,* (2020) International Publication Number WO/2020/076519.

6.2. Machine learning for phonon stability boundary in strain hyperspace

6.2.1. Phonon stability boundary in 3D and 6D hyperspaces

Chapter 3 and Chapter 4 offer an investigation upon how the electronic properties of different semiconductors change within a 6D strain hyperspace. The strain data points sampled for both silicon and diamond by the authors are conservative ones, i.e., all the strain cases sampled are phononic stable, but not all phononic stable strain cases are sampled. If we have enough computational resources to compute the phonon stability for all strain cases, what would be the aggregation of the phononic stable ones looks like in the strain hyperspace? Given the soft phonon criterion offers both necessary and sufficient conditions for crystal lattice instability, in other words, what would be the ideal limit for ESE? It is a must-ask question when practicing deep ESE.

For visualization purposes, we first trained ML models to showcase the stability boundaries in two 3D subspaces by fixing three of the six strain components (See Table 6.1 for accuracies report). We again rely on first-principles calculations for dataset curation. Phonon calculations were mainly conducted using the VASP-Phonopy package [163]. 2 × 2 × 2 supercells of 16 carbon atoms were created, and phonon calculations were conducted with a 3 × 3 × 3 **k**-point mesh. We also took full advantage of the known symmetries to further reduce the computations needed when collecting the strain data. Figure 6.1a and b illustrate the situation where only compressive and tensile normal strains are present ($\varepsilon_{23} = \varepsilon_{13} = \varepsilon_{12} = 0$). Figure 6.1c and d show the stability boundary for three-shear strain cases ($\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33} = 0$). Similar to the bandgap isosurfaces introduced in Chapter 3 and Chapter 4, the multifaceted nature of this boundary is attributed to the change of the onset of the soft phonon wave vector $\mathbf{q_c}$. In general, there are three types of $\mathbf{q_c}$. Following the same notation rules as introduced in Chapter 2, they can be denoted as:

- The Γ type: $\mathbf{q_c} = (0, 0, 0)$
- The 'Δ' type: $\mathbf{q_c} = (\xi, \xi, 0)$, $(\xi, 0, \xi)$ or $(0, \xi, \xi)$, where $0 < \xi < 0.5$
- The 'L' type: $\mathbf{q_c} = (0, 0, 0.5)$, $(0, 0.5, 0)$ or $(0.5, 0, 0)$

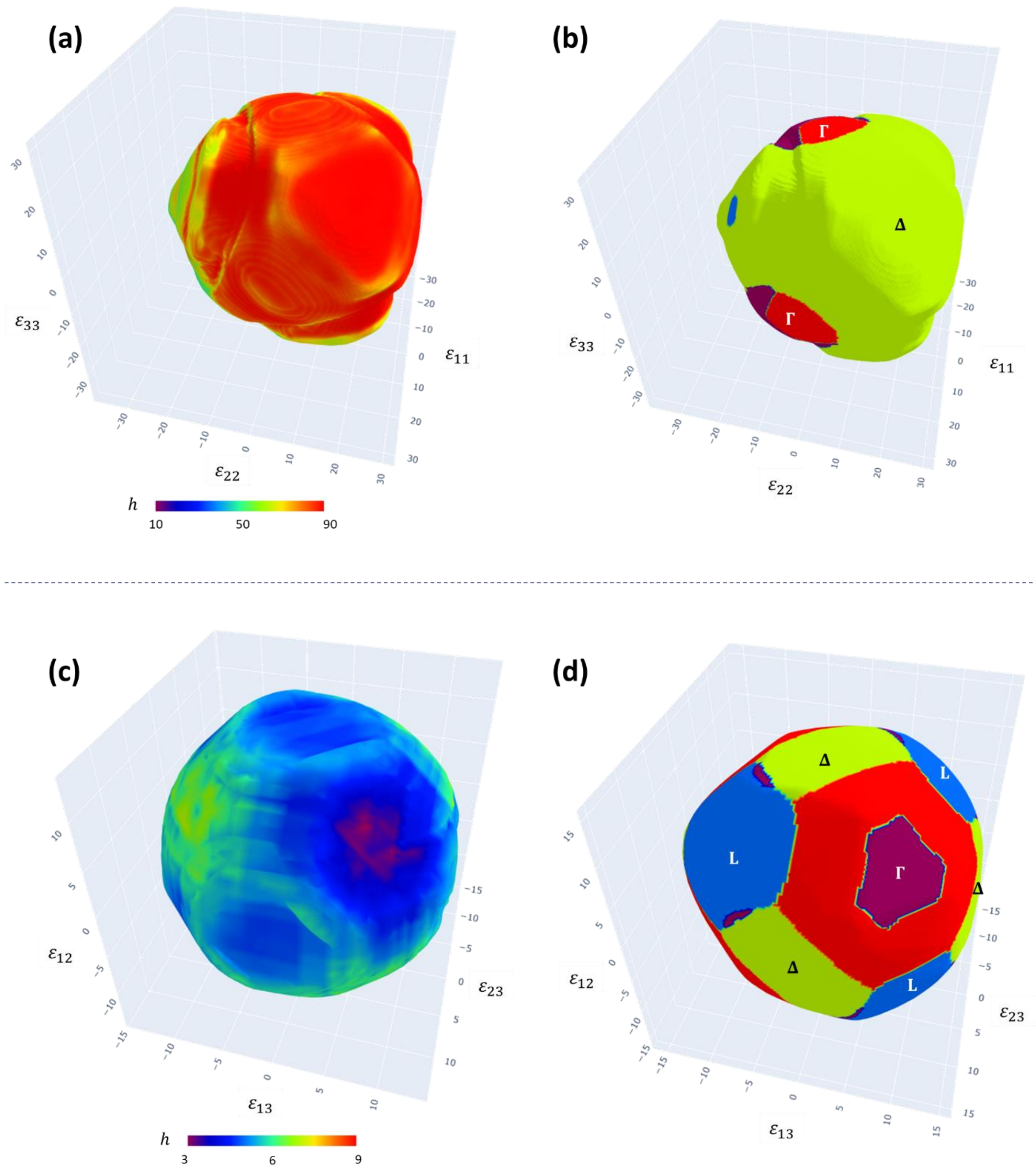


Figure 6.1 3D Phonon stability boundaries. The stability boundary for the three-normal strain ($\varepsilon_{11}\varepsilon_{22}\varepsilon_{33}$) space as colored by (a) the elastic strain energy density, h (meV/Å^3) and (b) onset of soft phonon wave vector $\mathbf{q_c}$. The

stability boundary for the three-shear strain ($\varepsilon_{23}\varepsilon_{13}\varepsilon_{12}$) space as colored by (c) h and (d) $\mathbf{q_c}$. All the strain axes are in terms of percentage. The rules for labeling the types of $\mathbf{q_c}$ on the facets of the stability boundaries in (b) and (d): red/purple for Γ/near-Γ type, blue for ‘L’ type and yellow for ‘Δ’ type.

We next move on to learn the phonon stability boundary in 6D and acquired the pair-plot visualization, as shown in Figure 6.2. This plot is made up of snapshots that characterize the shape of stability boundary in 2D subspace in 30 off-diagonal places and histograms in 6 diagonal places. Symmetries due to crystal deformation are obeyed when rendering this plot through ML models. One can think of the 6 subfigures in the top left corner of pairwise-normal-strain boundaries as origin-passing vertical or horizontal cuts of the 3D volume in Figure 6.2, and the 6 subfigures in the lower right corner of pairwise-shear-strain boundaries as origin-passing vertical or horizontal cuts of the 3D volume in Figure 6.2. The remaining 18 subfigures of mixed pairwise-normal/shear-strain boundaries take more interesting shapes that are subject to future detailed studies.

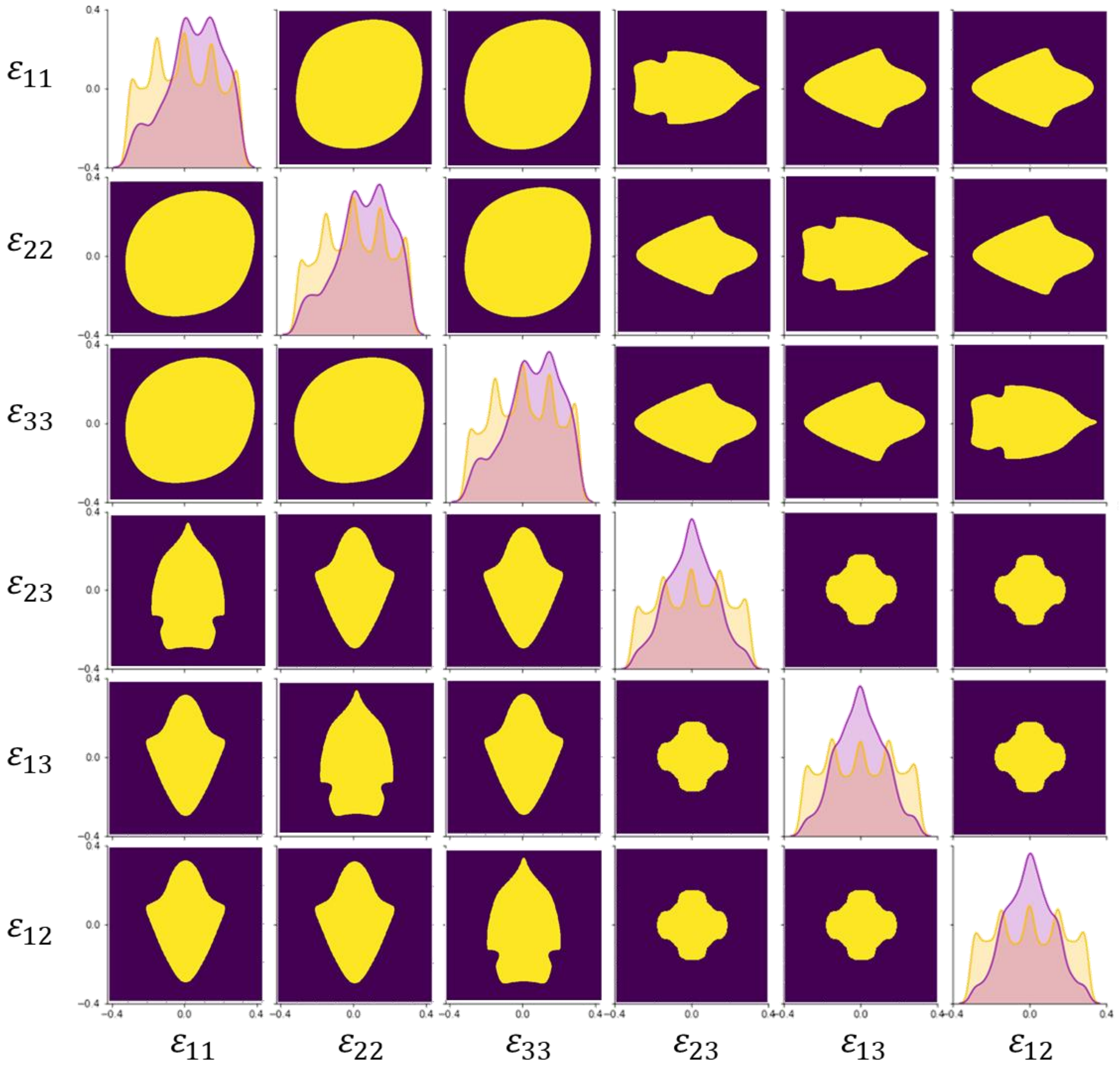


Figure 6.2 Pair-plot visualization of phonon stability boundary. The off-diagonal subfigures are 2D cuts of the 6D space with the other 4 strain components fixed at 0. For example, strains in the $\varepsilon_{11}\varepsilon_{12}$ subfigure in the lowest-left

corner have ($\varepsilon_{22} = \varepsilon_{33} = \varepsilon_{23} = \varepsilon_{13} = 0$). The diagonal subfigures are the histograms for all strain cases spanning -40% to 40%. Due to the symmetry in strain space, the 15 subfigures in the upper triangular region have a one-to-one correspondence with the other 15 subfigures in the lower triangular region at relative positions.

Table 6.1 Summary of ML accuracies for phonon stability, DOS, and band structure for the strain cases in the $\varepsilon_{11}\varepsilon_{22}\varepsilon_{33}$-space, the $\varepsilon_{23}\varepsilon_{13}\varepsilon_{12}$-space, and the general 6D hyperspace.

	Three-normal strain	Three-shear strain	General 6D strain
Phonon stability boundary	98%	97%	95%
Phonon DOS	MAE = 0.01	MAE = 0.01	MAE = 0.05
Phonon band structure	Max rel. error = 4.5%	Max rel. error = 5%	Max rel. error = 5%

6.2.2. Phonon band structure and density of states

Similar to electronic band structure, a crystal's phonon band structure $\omega_\nu(\mathbf{q}; \boldsymbol{\varepsilon})$ is a function of the 3D wave vector $\mathbf{q}$ and strain $\boldsymbol{\varepsilon}$, there are 9 dependent variables (10 with the integer phonon band index ν). Again, it is inadvisable to tabulate $\omega_\nu(\mathbf{q};\ \boldsymbol{\varepsilon})$ as billions of first-principles calculations may be required. Hereby, we adopted ML algorithms same as those introduced in Chapter 3 and Chapter 4 to fit the phonon dispersion, the results of which are shown in Table 6.1. We also studied the variation of phonon DOS, $g(\omega; \boldsymbol{\varepsilon})$, as a function of the 6D strain (Figure 6.3 and Table 6.1) and obtained decent results.

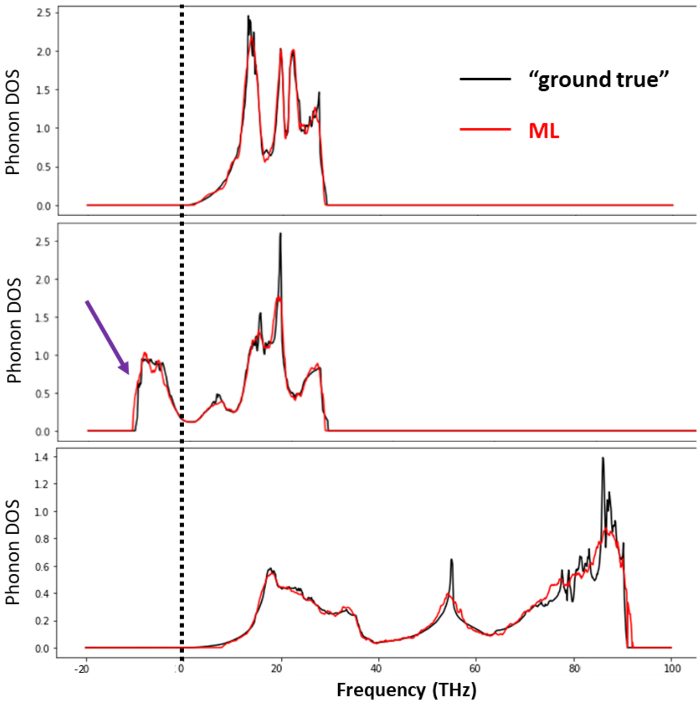


Figure 6.3 Comparison of the ML predicted phonon DOS and true phonon DOS at different strain states. The "ground truth" of phonon DOS here is obtained from first-principles calculations. The physical unit of the

horizontal axes is that of frequency (THz), and that of the vertical axes is the number of states per unit cell per THz. The occurrence of imaginary phonon frequencies in the second strain case is indicated by the purple arrow.

Finally, just as the derivative of the electronic band is related to group velocity, the derivative of phonon band ($\boldsymbol{\nabla}_{\mathbf{q}}\omega_{\nu}(\mathbf{q};\, \boldsymbol{\varepsilon})$) also contains important physical meaning, namely the speed of sound. Other important properties such as the lattice thermal conductivity or the Grüneisen parameter [167] can also be derived based on phonon calculation results. Similar to the Pareto optimization process introduced in Section 4.5.4, finding the best combination of thermal properties or even the best combination of thermal+electronic properties at various strain states warrants another good numerical research.

6.3. Machine learning for strain engineering defect ionization energy

Materials with a wide bandgap often do not have enough charge carriers (e.g., holes and/or electrons). For example, the implementation of diamond, an ultra-wide bandgap material, as a semiconducting or conducting material has conventionally been unsuccessful due to the difficulty in effectively doping the material with a defect capable of producing electrons (e.g., an n-type dopant). Furthermore, depending on the choice of defect, the defect ionization energy (E_{I}) can vary greatly within the same material. In some materials, the energy to ionize a defect in a deep dopant state is too large to be facilitated by room-temperature thermal fluctuations. As a result, a defect in a deep dopant state (such as substitutional nitrogen in diamond) typically does not contribute charge carriers to the conduction band and/or valence band, resulting in a material that is incapable of being used in a semiconducting or conducting device. Thus, there is a need for methods and systems to further tune the value of E_{I} for wide- and ultrawide-bandgap materials to facilitate their use in devices.

In view of the above, we have realized and appreciated that elastic strain can be used to control the doping level in a substitutional defect doped diamond. To find the strain cases which yields a transition of nitrogen substitutional dopant from a deep dopant state to a shallow dopant state (E_{I} no greater than several $k_{\mathrm{B}}T$) in diamond, we first have a study of its defect structure at an undeformed and several hydrostatically (non-deviatorically) compressed states. The defect phenomenon revealed by high-pressure physics [168] will give us useful insights into deep ESE later.

As shown in Figure 6.4a, the nitrogen point defect ($\mathrm{N_c}$) at equilibrium (undeformed) state is bound to four carbon atoms (C-atoms) and has tetrahedral symmetry. There exists, however, a spontaneous symmetry breaking of the tetrahedral symmetry to one of four equivalent low-symmetry variants. In each symmetry variant, the nitrogen atom breaks a bond with one of the four C-atoms it is bound to and forms shorter bonds with the other three C-atoms. As a result, the nitrogen point defect is in a deep donor state that is ~2.2 eV below the conduction band edge. As such, the nitrogen point defect is impossible to be ionized by room-temperature thermal fluctuations and therefore will not contribute charge carriers to the conduction band. The above phenomenon is backed by DFT calculations.

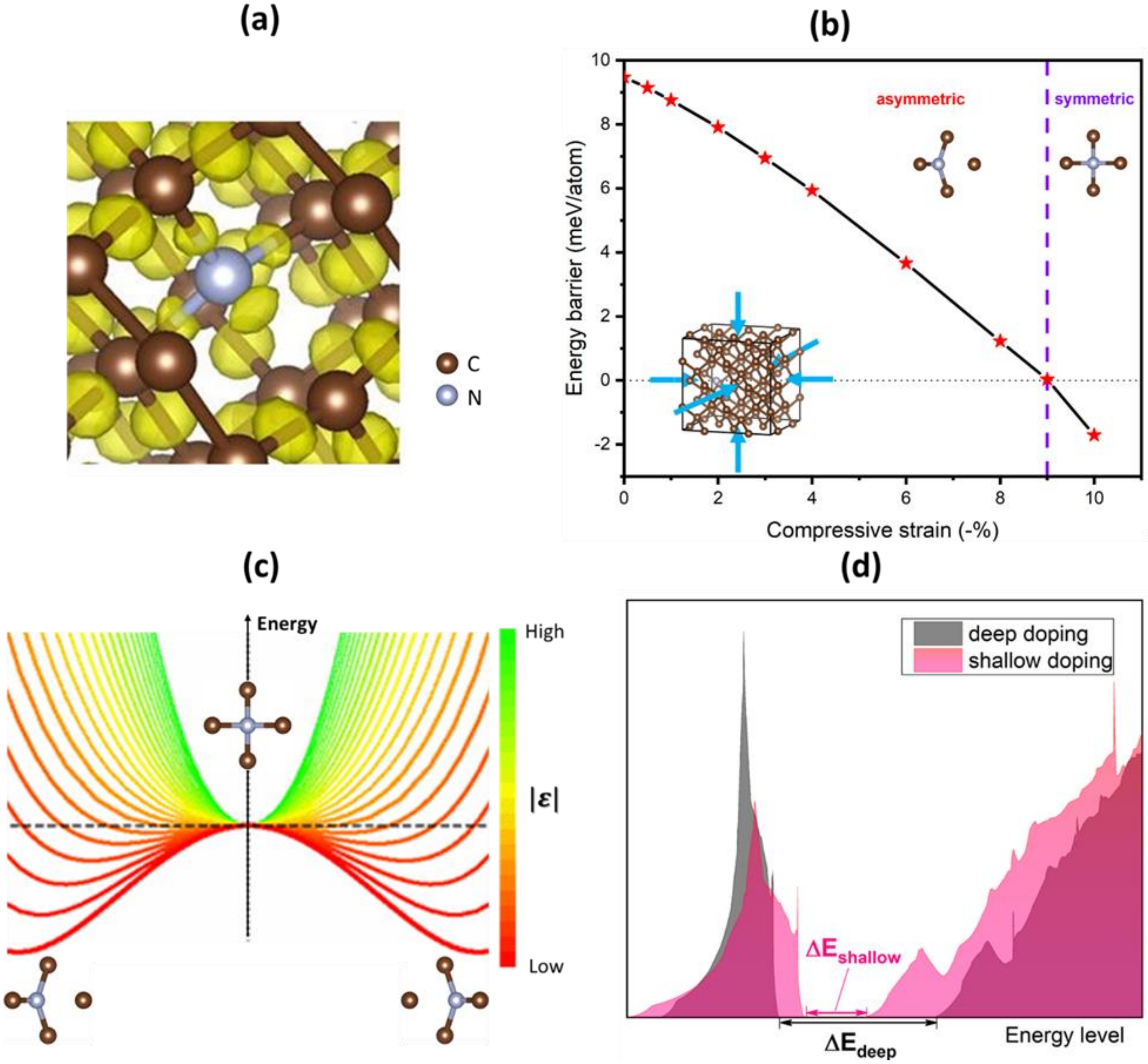


Figure 6.4 Atomic-level understanding of ESE in N-doped diamond. (a) Spontaneous tetrahedral symmetry breaking in diamond with nitrogen dopant, wherein one N-atom is bound to three out of the four C-atoms at the corners of the tetrahedron, forming a skewed defect structure. (b) Monotonic reduction of the energy barrier between skewed and centered N configuration upon straining. At the ground state, i.e., in an undeformed diamond, N tends to sway from the center of the sp^3 tetrahedral site. But with increasing hydrostatic compression, the N atom moves to the center of the tetrahedral site bounded by four other C atoms (symmetrical site). (c) Nudged Elastic Band calculation [169,170] provides another demonstration of this transition. The color bar represents the magnitude of compression. (d) DOS plot near CBM for undeformed and deformed structure. Note that the energy differences between CBM and defect level ($\Delta E_{\text{shallow}}$ and ΔE_{deep}) are not defect ionization energies. This is a schematic showing the realization of relatively shallower N dopant.

When the diamond is elastically compressed by as much as 10%, our DFT calculations reveal that the nitrogen defect structure will be structurally reconfigured. Figure 6.4b shows, as strain is applied to the nitrogen-doped diamond, the N-atom gradually moves to the center of the tetrahedral site bounded by four other C-atoms, until the crystal is symmetric. This process is further quantified in Figure 6.4c, which is a plot of the strain from asymmetrical to symmetrical orientation for nitrogen-doped diamond. The application of a 10% hydrostatic compressive strain allows the energy barrier to vanish, which indicates the symmetric, tetrahedral structure is energetically stable. Applying a compressive strain to a region of diamond comprising a nitrogen point defect can therefore provide a transition to a shallower n-type dopant state (Figure 6.4d).

To look for such transition in the 6D strain hyperspace, we followed the workflow similar to that in Chapter 3 and Chapter 4 to train a ML model that takes in a strain (within the $\boldsymbol{\varepsilon}_{\text{ideal}}$ boundary found in Section 6.2.1) and predicts the E_{I}. By plotting the resultant E_{I}'s against the elastic strain energy density (h), we can locate from Figure 6.5 a plethora of strain pathways towards shallow N$_c$ states with $E_{\text{I}} < k_{\text{B}}T$ that readily contribute delocalized electron carriers to the conduction band by thermal ionization. We can compare the donor level of N in elastically strained diamond with other substitutional defects such as Sb$_{\text{C}}$, B$_{\text{C}}$, P$_{\text{C}}$, and S$_{\text{C}}$.

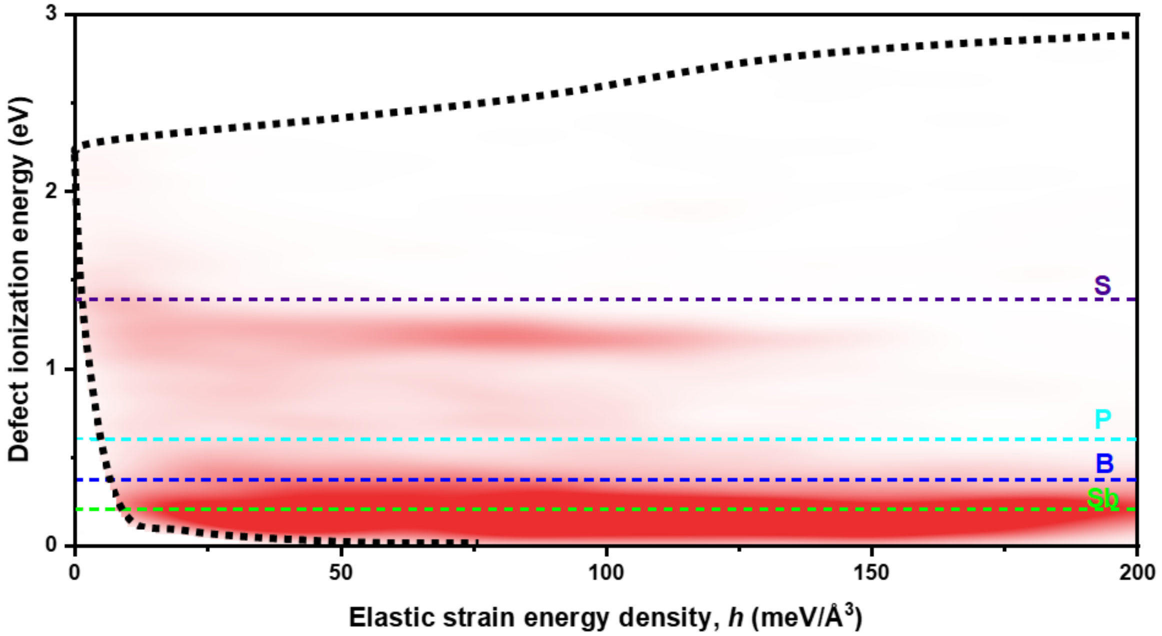


Figure 6.5 Density of states of donor ionization energy. Reachable E_{I} values for various h within the whole deformation space for diamond with N$_c$ centers. The red shading of the region reflects the distribution of available E_{I}. The bottom black dotted line delineating the lower boundary of the entire red envelope indicates the lowest energy penalty path for attaining a decreased E_{I} (including the realization of deep-to-shallow donor transition) in a general strain hyperspace. The upper bound of the reachable E_{I} is denoted by the black dotted line on the top.

Similar to the mathematical formulation of density-of-states of bandgap (Chapter 3) and density-of-states of effective mass (Chapter 4), a function that describes the resultant distribution of E_{I} arising from all possible elastically strained states in the $(h - \frac{dh}{2}, h + \frac{dh}{2})$

interval can be defined. Specifically, the cumulative density-of-states of defect ionization energy can be defined as

$$c(E_{\mathrm{I}}'; h') \equiv \int_{h(\boldsymbol{\varepsilon})<h'} d^6\boldsymbol{\varepsilon}\delta\big(E_{\mathrm{I}}' - E_{\mathrm{I}}(\boldsymbol{\varepsilon})\big) = \int d^6\boldsymbol{\varepsilon}\delta\big(E_{\mathrm{I}}' - E_{\mathrm{I}}(\boldsymbol{\varepsilon})\big)\Theta\big(h' - h(\boldsymbol{\varepsilon})\big), \quad (33)$$

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 ionization energy (g) at h' can then be defined by the derivative of $c(E_{\mathrm{I}}';\ h')$ with respect to h':

$$g(E_{\mathrm{I}}'; h') \equiv \frac{\partial c(E_{\mathrm{I}}'; h')}{\partial h'} = \int d^6\boldsymbol{\varepsilon}\delta\big(E_{\mathrm{I}}' - E_{\mathrm{I}}(\boldsymbol{\varepsilon})\big)\delta\big(h' - h(\boldsymbol{\varepsilon})\big). \quad (34)$$

In general, the methods described herein can be used to effectively n-dope and/or p-dope a material that was previously considered to be “undopable”, such that the material defects may transition from a deep dopant state to a previously inaccessible shallow dopant state upon the application of elastic strain. The ML methods of applying an elastic strain to alter the doping state of a defect doped material may be applied to any of a variety of suitable compositions. The defect doped material may comprise defect doped silicon, Ga_2O_3, GaN, BN, and/or any other appropriate material. Additionally, appropriate dopants may include, but are not limited to, nitrogen, boron, phosphorus, and/or combinations thereof. Besides semiconducting devices, the defect doped materials may be implemented in a memory device, due to the ability to dynamically toggle the defect doped material between these states rapidly and/or reversibly, akin to an “on-off” switch.

Chapter 7. Thesis summary and future works

7.1. What has been achieved in advancing deep elastic strain engineering

Deep ESE explores the full 6D space of admissible nonlinear elastic strain and its effects on physical properties. However, the complexity of controllably engineering materials properties by strains necessitates first-principles computations to first screen for a desirable figure-of-merit and then design an optimal straining pathway. In this work, to map the 6D strain space fully, we first combine ML and DFT/GW calculations to guide strain engineering whereby electronic properties could be designed. This method invokes deep NN algorithms and utilizes a limited amount of *ab initio* data for the training of a surrogate model, predicting various electronic properties within reasonable accuracy. On top of this, attempts have been made by us in developing a more versatile ML framework that adopts convolutional blocks, data fusion, and active learning to discover the indirect-to-direct bandgap transition and Mott transition in a material by scanning the entire strain space. Through this framework, we improve the state-of-the-art set by ourselves and achieve enhanced performance in every front, including more accurate bandgap and band structure prediction, band extrema detection, and effective mass calculation. Combining this method with experimentally validated FEM simulations, we predicted strain pathways that would reversibly transform an ultrawide-bandgap material such as diamond to a metalized state. Applying the model to phonon and defect related studies, we also visualized the phonon stability boundaries in 6D and predicted the deep-to-shallow donor transition in doped diamond.

7.2. Limitations and future works

Despite the capabilities introduced in previous chapters, several factors are limiting the current ML model. Firstly, we rely on excited-state calculations (GW) to give us the “ground truth” of electronic properties such as bandgap, CBM location, and effective mass. The only experimental value we can validate our GW calculation is that of the undeformed state. Therefore, we are very sure with the intrinsic bandgap of 1.1 eV for silicon and of 5.6 eV for diamond in their equilibrium state, but since there is no experimental bandgap measurement of silicon stretched in <111> direction by 10%, we are only left with GW results to trust when we practice deep ESE in a general 6D hyperspace.

Secondly, the current model does not consider the temperature effect. This is because the DFT data we collected is only for 0 K. This is another deviation from real-world finite temperature conditions. Thirdly, a successful deep ESE practice requires the ability to not only deforming the material to the uttermost, but also holding it there for an extended time period without relaxation of any kind. Time is also a factor that this ML model has not taken into consideration.

Also, strain rate is another factor neglected in the present study, and as a result, we are only at the extreme left side of the deformation map (see Figure 7.1). Lastly, the many other non-idealities in real-world material behavior can further complicate the practicality of the deep ESE model we discussed in this thesis. These limitations are the exact reasons we call for close collaboration with experimental colleagues, as discussed later in this chapter.

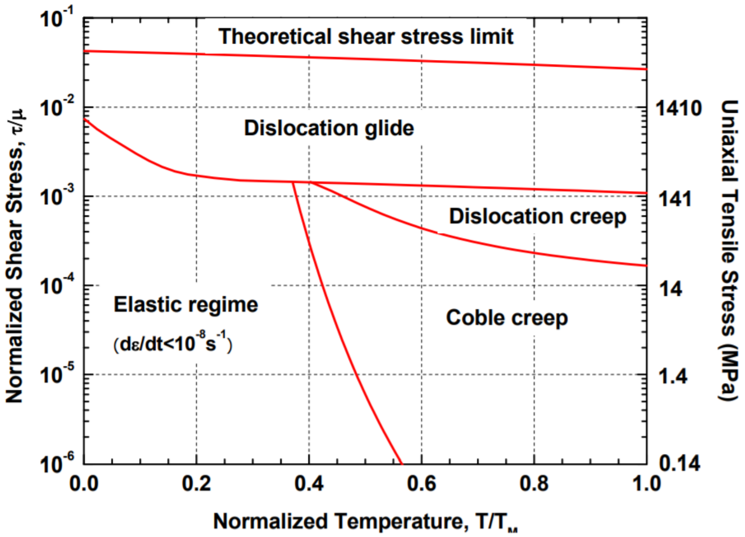


Figure 7.1 Deformation mechanism map for a particular alloy at a strain rate of 10^{-8} s^{-1}. The exact material itself is not important. The figure is taken from FIG. 8 of Ref. [171] just to show the concept.

7.2.1. Model extension

We only chose two materials, namely silicon and diamond, for which deep ESE is practiced, primarily because there exist near-$\boldsymbol{\varepsilon}_{\text{ideal}}$ experiments in these representative material systems. Important and versatile as they both are, there are many other semiconductors whose electronic properties deserve to be studied, as mentioned in Section 1.1.4. In addition, it is useful to use the model to learn properties that are a function of energy, such as the dielectric function $\epsilon_2(E;\, \boldsymbol{\varepsilon})$. As a comparison, the band structure tabulates E as a function of $\mathbf{k}$. Given that the phonon DOS has been learned in Section 6.2.2, it is straightforward to study strain dependent electronic DOS. The NN and CNN models follow what is shown in Figure 7.2a-b. Lastly, it is also desirable to extend the model to cover more external field effects. For example, as it is very common for devices to function inside an electric field **E** (consider the working environment of an FET), we can add **E** to the strain model and fit the piezoelectric tensor **e**(**E**; $\boldsymbol{\varepsilon}$), as shown in Figure 7.2b.

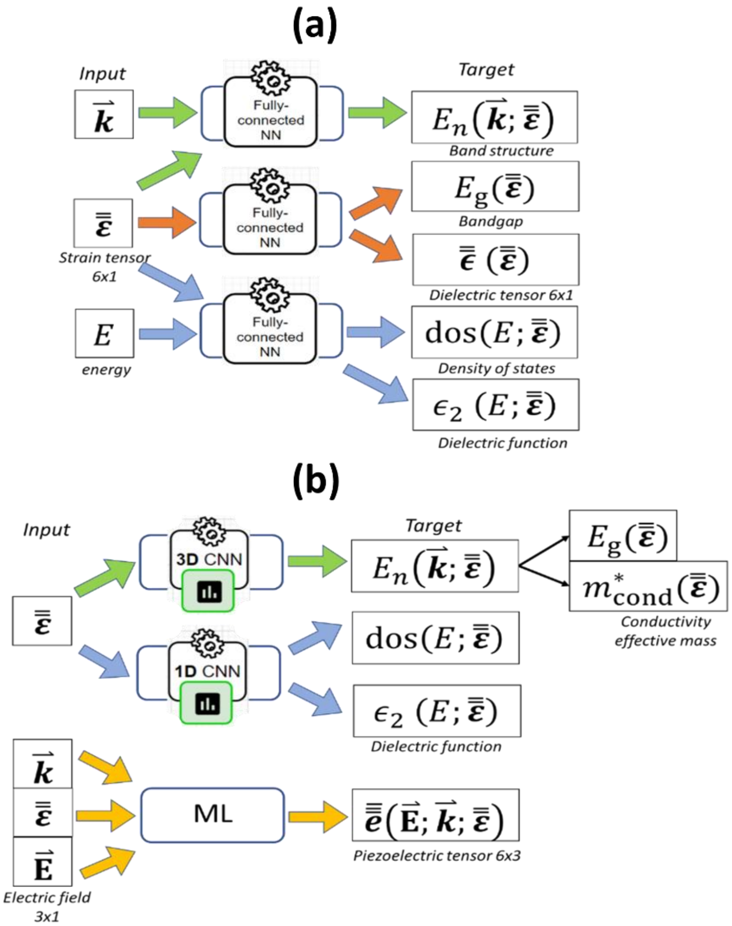

(a)
Input
$\vec{k}$
$\bar{\bar{\varepsilon}}$
Strain tensor 6x1
E
energy
Fully-connected NN
Fully-connected NN
Fully-connected NN
Target
$E_n(\vec{k};\bar{\bar{\varepsilon}})$
Band structure
$E_g(\bar{\bar{\varepsilon}})$
Bandgap
$\bar{\bar{\epsilon}}\ (\bar{\bar{\varepsilon}})$
Dielectric tensor 6x1
$\mathrm{dos}(E;\bar{\bar{\varepsilon}})$
Density of states
$\epsilon_2\ (E;\bar{\bar{\varepsilon}})$
Dielectric function
(b)
Input
$\bar{\bar{\varepsilon}}$
3D CNN
1D CNN
Target
$E_n(\vec{k};\bar{\bar{\varepsilon}})$
$E_g(\bar{\bar{\varepsilon}})$
$m^*_{\mathrm{cond}}(\bar{\bar{\varepsilon}})$
Conductivity effective mass
$\mathrm{dos}(E;\bar{\bar{\varepsilon}})$
$\epsilon_2\ (E;\bar{\bar{\varepsilon}})$
Dielectric function
$\vec{k}$
$\bar{\bar{\varepsilon}}$
$\vec{\mathbf{E}}$
Electric field 3x1
ML
$\bar{\bar{e}}(\vec{\mathbf{E}};\vec{k};\bar{\bar{\varepsilon}})$
Piezoelectric tensor 6x3

Figure 7.2 ML for different targets with (a) NN and (b) CNN. If the target is a function of a vector, such as a band structure $E(\mathbf{k})$, then 3D convolution is used. On the other hand, if the target is a function of scalar, such as DOS(E), then the 1D convolution is used.

7.2.2. Device-level design and simulation

Sections 5.3 introduce our work on coupling strain mapping using FEM with bandgap distribution mapping using ML. In Section 5.4 we further came up with a loading design that looks more friendly for device designers and engineers. However, in a metal-oxide-semiconductor field-effect transistor (MOSFET), people measure I-V characteristics and study macroscopic electronic properties or FoMs such as the subthreshold swing, largest current driven by the transistor, dynamic power factor, and intrinsic delay time [172]. These as well as many other properties are suggested in the International Technology Roadmap for Semiconductors [173].

Thus, it is a natural next step to go beyond fundamental electronic properties such as bandgap and try to incorporate the strain model into Technology Computer Aided Design (TCAD) simulations to understand the I-V curves of a device undergoing deep ESE. This integration would be a meaningful contribution to the academic field of device simulation where models exist only for moderately strained channel materials [174–178]. It is noted that in real semiconductor devices, there are other issues to be dealt with, including the contact resistance and imperfections aforementioned, and experimental verification is eventually needed.

7.2.3. Experimental works and collaborations

Carrying out *in-situ* NEMS loading experiments inside a TEM with built-in electron energy loss spectroscopy (EELS) is another obvious next step to show the impact of the strategies introduced in this thesis to improve functional properties through deep ESE of semiconductors. It is known [179–181] that EELS is reliable for assessing the bandgap value (including surface plasmon mapping) as well as indirect-to-direct bandgap transition in diamond. We have started the experiments with our collaborators (Figure 7.3a-b), and we plan to conduct line scans across the tensile to the compressive sides of a bent diamond needle to map bandgap changes, as demonstrated in the red arrow in Figure 7.3b. In addition, a colleague from the author's group designed an *in-situ* mechanical loader actuated by heat (see Figure 7.3c for its SEM micrograph). It can offer ultralarge uniaxial tension and compression to free-standing semiconductor samples and perform *in-situ* electrical measurement (Figure 7.3d-e), making possible experimental verification of future TCAD simulation results that are obtained from Section 7.2.2.

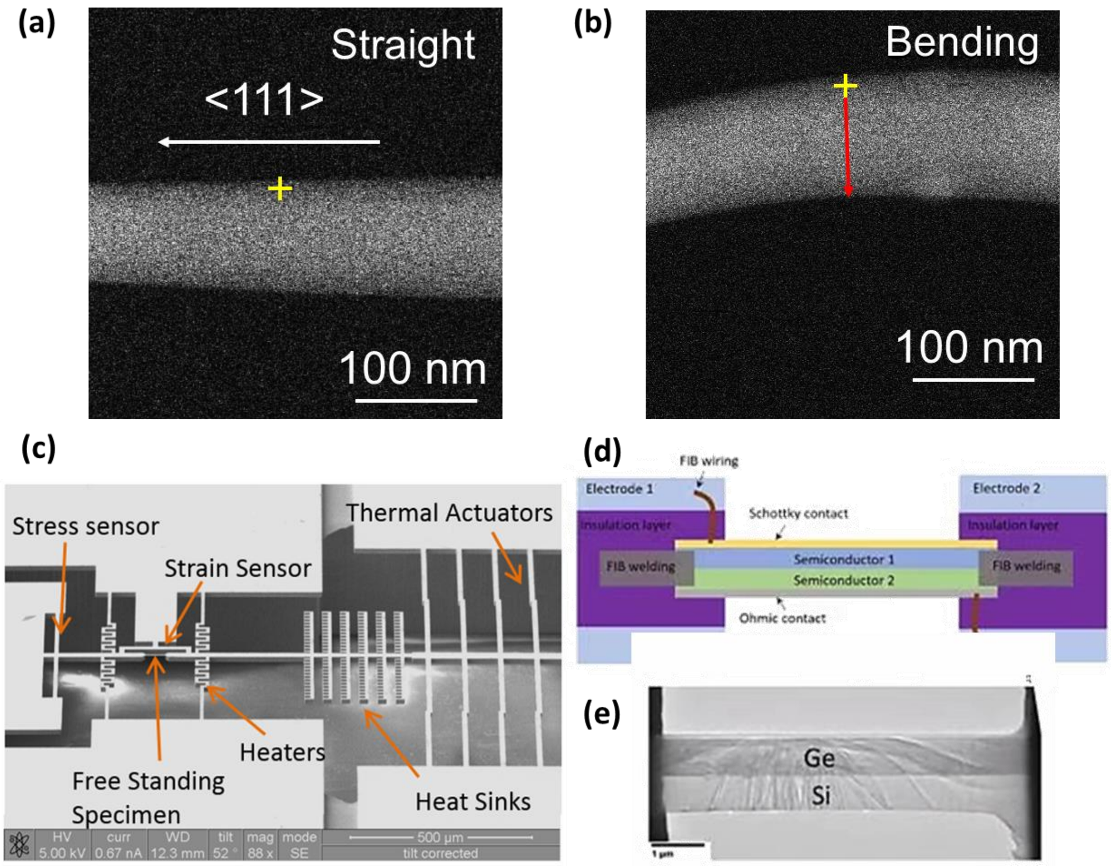


Figure 7.3 Deep ESE experiments. (a) An undeformed <111> nanoneedle. (b) Bent diamond nanoneedle ready to be scanned *in-situ* by EELS. The yellow marks in (a) and (b) indicate the same place on the nanoneedle. TEM

images of (a) and (b) are credited to the author’s colleagues from Zhejiang University and Nanyang Technological University. (c) An *in-situ* thermo-mechanical loading device. (d-e) A microchannel of Si|Ge subject to tension by the thermo-mechanical loader. Images in (c-d) are used with the permission of Dr. Baoming Wang.

Indentation and anviling (compression under extreme pressures) coupled with *in-situ* photoluminescence [182–184] or cathodoluminescence [16] spectroscopy as well as electrical resistivity measurement [185] further add to the toolbox for characterization of mechanically-induced properties in semiconductors. In Section 6.3, the author predicted defect ionization energy change under ultrahigh strains, and an associated high-pressure physics experiment (Figure 7.4) is currently underway to confirm the theoretical findings.

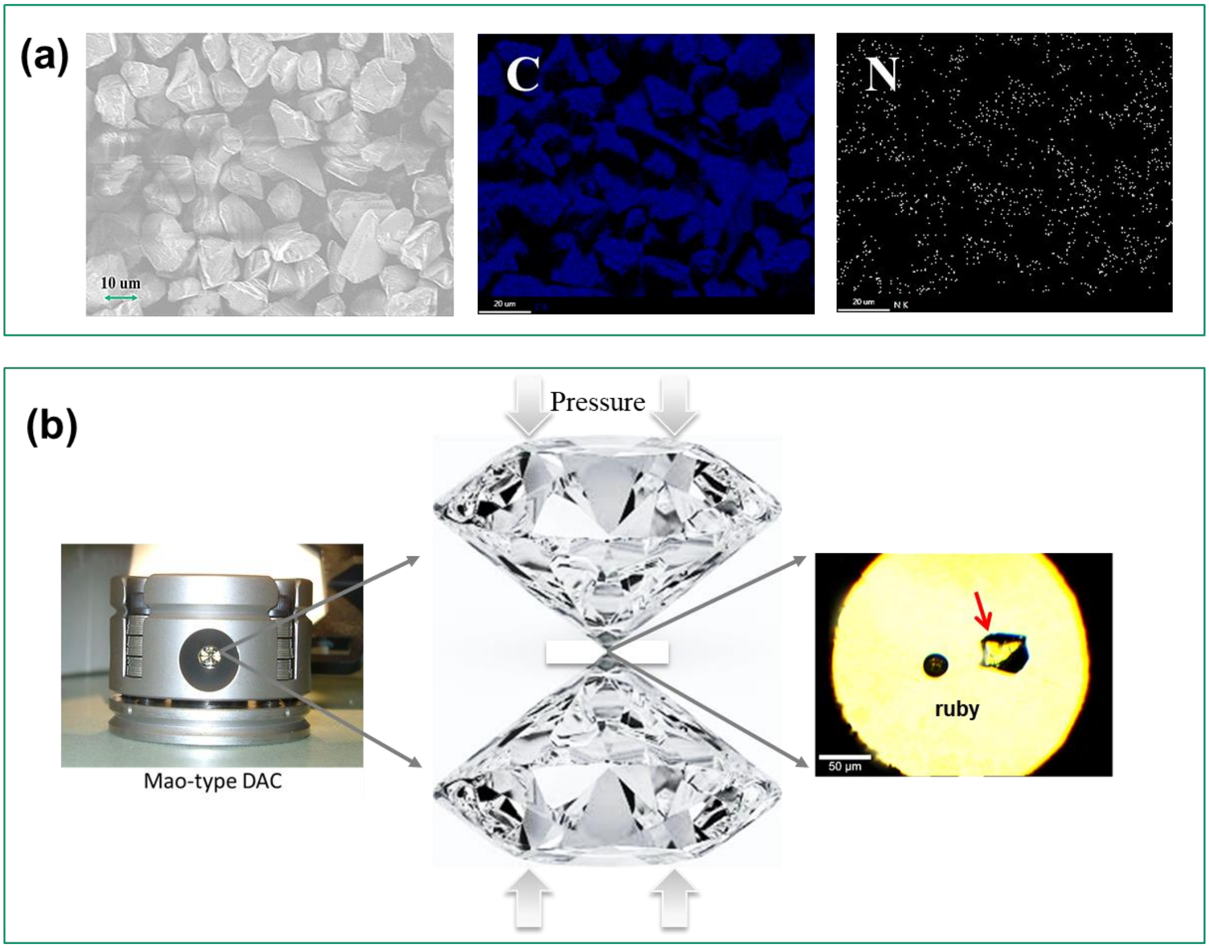


Figure 7.4 Diamond anvil cell (DAC) experiment. (a) SEM micrograph of N-doped diamond micro-particles and energy-dispersive X-ray (EDX) spectroscopy analysis of carbon and nitrogen element. (b) From left to right: Mao-type DAC used in the experiment, schematic of the diamond compressors, and the optical micrograph of a compressed N-doped diamond particle which is ready to be measured by *in-situ* photoluminescence spectroscopy. The red arrow indicates the diamond particle of concern and a ruby next to it is used for pressure calibration.

These and other experiments are still undergoing. As the reader has noticed, such experiments are beyond the scope of the present work, and they are being pursued at this time in collaboration with several different research teams.

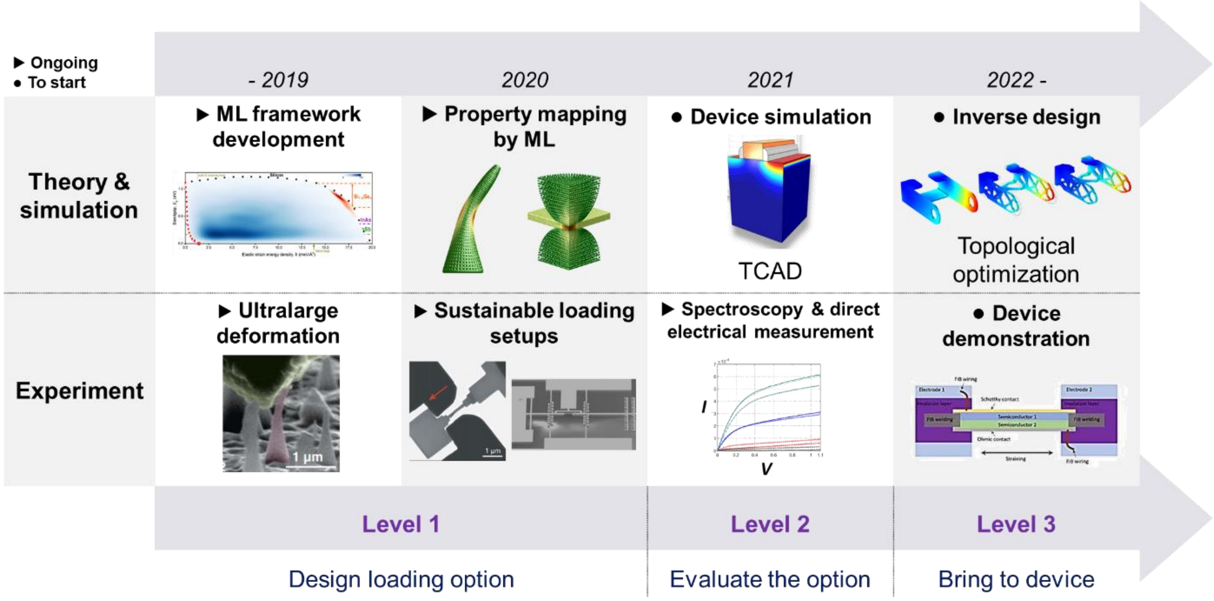


Figure 7.5 The development roadmap for the introduction of deep elastic strains in experimental devices.

As a closing remark, side-by-side simulation and experimental efforts outlined in Figure 7.5 are needed for achieving deep ESE in devices. In Level 1, with ML framework developed for predicting electronic properties of experimentally feasible loading geometries, the research goals are then redefined towards the development of sustainable and up-scalable deep ESE options and apply the as-developed ML model to evaluate materials electronic behavior under deformation. In Level 2, device simulations together with direct electrical measurement (such as retrieving the I-V curves) would be carried out. Non-contact spectroscopy is also involved in this phase to study the deformation mechanism and evaluate the selected loading option. Knowledge acquired from this stage will be helpful in experimental device demonstration in Level 3. Also, the concept of inverse design can be brought into the simulation works at this level. Specifically, given that $\boldsymbol{\varepsilon}$ dictates the material property and that many loading options become accessible, one may first articulate a needed property or FoM (a given bandgap, donor level, Baliga's FoM, etc.), and then adaptively look for a loading geometry through topological optimization that yields the desirable $\boldsymbol{\varepsilon}$ corresponding to this particular FoM. Deep ESE can only be successful if all the device challenges listed in Figure 7.5 are properly resolved. However, in the near term, devices with embedded deep elastic strains will be mainly for simple electronic and optical applications whose material and integration requirements are less stringent compared to the needs for CMOS, for example. The route to practicing deep ESE in future devices is via sustained research and development to meet the requirements of semiconductor device commercialization.

References

[1] S. S. Brenner, Growth and Properties of “Whiskers”: Further Research Is Needed to Show Why Crystal Filaments Are Many Times as Strong as Large Crystals, Science **128**, 569 (1958).

[2] S. S. Brenner, Tensile Strength of Whiskers, J. Appl. Phys. **27**, 1484 (1956).

[3] C. Herring and J. K. Galt, Elastic and Plastic Properties of Very Small Metal Specimens, Phys. Rev. **85**, 1060 (1952).

[4] G. L. Pearson, W. T. Read, and W. L. Feldmann, Deformation and Fracture of Small Silicon Crystals, Acta Metall. **5**, 181 (1957).

[5] Dynamical Theory of Crystal Lattices (Oxford University Press, Oxford, New York, 1998).

[6] J. Li, K. J. Van Vliet, T. Zhu, S. Yip, and S. Suresh, Atomistic Mechanisms Governing Elastic Limit and Incipient Plasticity in Crystals, Nature **418**, 307 (2002).

[7] K. J. Van Vliet, J. Li, T. Zhu, S. Yip, and S. Suresh, Quantifying the Early Stages of Plasticity through Nanoscale Experiments and Simulations, Phys. Rev. B **67**, 104105 (2003).

[8] R. E. Miller and D. Rodney, On the Nonlocal Nature of Dislocation Nucleation during Nanoindentation, J. Mech. Phys. Solids **56**, 1203 (2008).

[9] X. Liu, J. Gu, Y. Shen, and J. Li, Crystal Metamorphosis at Stress Extremes: How Soft Phonons Turn into Lattice Defects, NPG Asia Mater. **8**, 10 (2016).

[10] Y.-T. Chi, M. Youssef, L. Sun, K. J. Van Vliet, and B. Yildiz, Accessible Switching of Electronic Defect Type in SrTiO3 via Biaxial Strain, Phys. Rev. Mater. **2**, 055801 (2018).

[11] Yen-Ting Chi et al., Prep. (n.d.).

[12] Z. Huang, Z. Q. Liu, M. Yang, S. W. Zeng, A. Annadi, W. M. Lü, X. L. Tan, P. F. Chen, L. Sun, X. Renshaw Wang, Y. L. Zhao, C. J. Li, J. Zhou, K. Han, W. B. Wu, Y. P. Feng, J. M. D. Coey, T. Venkatesan, and Ariando, Biaxial Strain-Induced Transport Property Changes in Atomically Tailored $SrTiO_3$-Based Systems, Phys. Rev. B **90**, 125156 (2014).

[13] A. Cavallaro, M. Burriel, J. Roqueta, A. Apostolidis, A. Bernardi, A. Tarancón, R. Srinivasan, S. N. Cook, H. L. Fraser, J. A. Kilner, D. W. McComb, and J. Santiso, Electronic Nature of the Enhanced Conductivity in YSZ-STO Multilayers Deposited by PLD, Solid State Ion. **181**, 592 (2010).

[14] B. Yildiz, “Stretching” the Energy Landscape of Oxides—Effects on Electrocatalysis and Diffusion, MRS Bull. **39**, 147 (2014).

[15] D. G. Schlom, L.-Q. Chen, C. J. Fennie, V. Gopalan, D. A. Muller, X. Pan, R. Ramesh, and R. Uecker, Elastic Strain Engineering of Ferroic Oxides, MRS Bull. **39**, 118 (2014).

[16] X. Fu, C. Su, Q. Fu, X. Zhu, R. Zhu, C. Liu, Z. Liao, J. Xu, W. Guo, J. Feng, J. Li, and D. Yu, Tailoring Exciton Dynamics by Elastic Strain-Gradient in Semiconductors, Adv. Mater. **26**, 2572 (2014).

[17] J. Cao, E. Ertekin, V. Srinivasan, W. Fan, S. Huang, H. Zheng, J. W. L. Yim, D. R. Khanal, D. F. Ogletree, J. C. Grossman, and J. Wu, Strain Engineering and One-Dimensional Organization of Metal–Insulator Domains in Single-Crystal Vanadium Dioxide Beams, Nat. Nanotechnol. **4**, 11 (2009).

[18] M. J. Süess, R. Geiger, R. A. Minamisawa, G. Schiefler, J. Frigerio, D. Chrastina, G. Isella, R. Spolenak, J. Faist, and H. Sigg, Analysis of Enhanced Light Emission from Highly Strained Germanium Microbridges, Nat. Photonics **7**, 466 (2013).

[19] J. R. Jain, A. Hryciw, T. M. Baer, D. A. B. Miller, M. L. Brongersma, and R. T. Howe, A Micromachining-Based Technology for Enhancing Germanium Light Emission via Tensile Strain, Nat. Photonics **6**, 6 (2012).

[20] B. H. Lee, A. Mocuta, S. Bedell, H. Chen, D. Sadana, K. Rim, P. O'Neil, R. Mo, K. Chan, C. Cabral, C. Lavoie, D. Mocuta, A. Chakravarti, R. M. Mitchell, J. Mezzapelle, F. Jamin, M. Sendelbach, H. Kermel, M. Gribelyuk, A. Domenicucci, K. A. Jenkins, S. Narasimha, S. H. Ku, M. Ieong, I. Y. Yang, E. Leobandung, P. Agnello, W. Haensch, and J. Welser, Performance Enhancement on Sub-70 Nm Strained Silicon SOI MOSFETs on Ultra-Thin Thermally Mixed Strained Silicon/SiGe on Insulator (TM-SGOI) Substrate with Raised S/D, in Digest. International Electron Devices Meeting, (2002), pp. 946–948.

[21] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson, and M. Bohr, A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors, in IEEE International Electron Devices Meeting 2003 (2003), p. 11.6.1-11.6.3.

[22] S. Narasimha, P. Chang, C. Ortolland, D. Fried, E. Engbrecht, K. Nummy, P. Parries, T. Ando, M. Aquilino, N. Arnold, R. Bolam, J. Cai, M. Chudzik, B. Cipriany, G. Costrini, M. Dai, J. Dechene, C. DeWan, B. Engel, M. Gribelyuk, D. Guo, G. Han, N. Habib, J. Holt, D. Ioannou, B. Jagannathan, D. Jaeger, J. Johnson, W. Kong, J. Koshy, R. Krishnan, A. Kumar, M. Kumar, J. Lee, X. Li, C.-H. Lin, B. Linder, S. Lucarini, N. Lustig, P. McLaughlin, K. Onishi, V. Ontalus, R. Robison, C. Sheraw, M. Stoker, A. Thomas, G. Wang, R. Wise, L. Zhuang, G. Freeman, J. Gill, E. Maciejewski, R. Malik, J. Norum, and P. Agnello, 22nm High-Performance SOI Technology Featuring Dual-Embedded Stressors, Epi-Plate High-K Deep-Trench Embedded DRAM and Self-Aligned Via 15LM BEOL, in 2012 International Electron Devices Meeting (2012), p. 3.3.1-3.3.4.

[23] A. Domenicucci, S. Bedell, R. Roy, D. K. Sadana, and A. Mocuta, Use of Moire Fringe Patterns to Map Relaxation in SiGe on Insulator Structures Fabricated on SIMOX Substrates, in Microscopy of Semiconducting Materials, edited by A. G. Cullis and J. L. Hutchison (Springer, Berlin, Heidelberg, 2005), pp. 89–92.

[24] Layout-Dependent Strain Optimization for p-Channel Trigate Transistors | IEEE Journals & Magazine | IEEE Xplore, https://ieeexplore-ieee-org.libproxy.mit.edu/abstract/document/6068239.

[25] A. Nainani, S. Gupta, V. Moroz, M. Choi, Y. Kim, Y. Cho, J. Gelatos, T. Mandekar, A. Brand, E.-X. Ping, M. C. Abraham, and K. Schuegraf, Is Strain Engineering Scalable in FinFET Era?: Teaching the Old Dog Some New Tricks, in 2012 International Electron Devices Meeting (2012), p. 18.3.1-18.3.4.

[26] R. P. Feynman, There's Plenty of Room at the Bottom, Eng. Sci. **23**, 22 (1960).

[27] J. Li, Z. Shan, and E. Ma, Elastic Strain Engineering for Unprecedented Materials Properties, MRS Bull. **39**, 108 (2014).

[28] Z. Shi, E. Tsymbalov, M. Dao, S. Suresh, A. Shapeev, and J. Li, Deep Elastic Strain Engineering of Bandgap through Machine Learning, Proc. Natl. Acad. Sci. **116**, 4117 (2019).

[29] A. Banerjee, D. Bernoulli, H. Zhang, M.-F. Yuen, J. Liu, J. Dong, F. Ding, J. Lu, M. Dao, W. Zhang, Y. Lu, and S. Suresh, Ultralarge Elastic Deformation of Nanoscale Diamond, Science **360**, 300 (2018).

[30] A. Nie, Y. Bu, P. Li, Y. Zhang, T. Jin, J. Liu, Z. Su, Y. Wang, J. He, Z. Liu, H. Wang, Y. Tian, and W. Yang, Approaching Diamond's Theoretical Elasticity and Strength Limits, Nat. Commun. **10**, 1 (2019).

[31] C. Dang, J.-P. Chou, B. Dai, C.-T. Chou, Y. Yang, R. Fan, W. Lin, F. Meng, A. Hu, J. Zhu, J. Han, A. M. Minor, J. Li, and Y. Lu, Achieving Large Uniform Tensile Elasticity in Microfabricated Diamond, Science **371**, 76 (2021).

[32] H. Zhang, J. Tersoff, S. Xu, H. Chen, Q. Zhang, K. Zhang, Y. Yang, C.-S. Lee, K.-N. Tu, J. Li, and Y. Lu, Approaching the Ideal Elastic Strain Limit in Silicon Nanowires, Sci. Adv. **2**, e1501382 (2016).

[33] J. Y. Tsao, S. Chowdhury, M. A. Hollis, D. Jena, N. M. Johnson, K. A. Jones, R. J. Kaplar, S. Rajan, C. G. V. de Walle, E. Bellotti, C. L. Chua, R. Collazo, M. E. Coltrin, J. A. Cooper, K. R. Evans, S. Graham, T. A. Grotjohn, E. R. Heller, M. Higashiwaki, M. S. Islam, P. W. Juodawlkis, M. A. Khan, A. D. Koehler, J. H. Leach, U. K. Mishra, R. J. Nemanich, R. C. N. Pilawa-Podgurski, J. B. Shealy, Z. Sitar, M. J. Tadjer, A. F. Witulski, M. Wraback, and J. A. Simmons, Ultrawide-Bandgap Semiconductors: Research Opportunities and Challenges, Adv. Electron. Mater. **4**, 1600501 (2018).

[34] M. J. Hÿtch and A. M. Minor, Observing and Measuring Strain in Nanostructures and Devices with Transmission Electron Microscopy, MRS Bull. **39**, 138 (2014).

[35] J. J. Gilman, Electronic Basis of the Strength of Materials, 1 edition (Cambridge University Press, Cambridge, 2008).

[36] R. G. Parr and Y. Weitao, Density-Functional Theory of Atoms and Molecules (Oxford University Press, 1989).

[37] N. F. MOTT, Metal-Insulator Transition, Rev. Mod. Phys. **40**, 677 (1968).

[38] J. J. Gilman, Chemical Reactions at Detonation Fronts in Solids, Philos. Mag. B **71**, 1057 (1995).

[39] K. Yang, W. Setyawan, S. Wang, M. Buongiorno Nardelli, and S. Curtarolo, A Search Model for Topological Insulators with High-Throughput Robustness Descriptors, Nat. Mater. **11**, 7 (2012).

[40] F. Liu, P. Ming, and J. Li, Ab Initio Calculation of Ideal Strength and Phonon Instability of Graphene under Tension, Phys. Rev. B **76**, 064120 (2007).

[41] L. S. Pan and D. R. Kania, editors , Diamond: Electronic Properties and Applications (Springer US, 1995).

[42] C. J. H. Wort and R. S. Balmer, Diamond as an Electronic Material, Mater. Today **11**, 22 (2008).

[43] J. H. Haeni, P. Irvin, W. Chang, R. Uecker, P. Reiche, Y. L. Li, S. Choudhury, W. Tian, M. E. Hawley, B. Craigo, A. K. Tagantsev, X. Q. Pan, S. K. Streiffer, L. Q. Chen, S. W. Kirchoefer, J. Levy, and D. G. Schlom, Room-Temperature Ferroelectricity in Strained SrTiO3, Nature **430**, 758 (2004).

[44] C. J. Fennie and K. M. Rabe, Magnetic and Electric Phase Control in Epitaxial EuTiO3 from First Principles, Phys. Rev. Lett. **97**, 267602 (2006).

[45] MIT.nano, Yen-Ting Chi—External Field Effects on Defects in Functional Oxides: Experiments and Simulations (2020).

[46] C. Lee, X. Wei, J. W. Kysar, and J. Hone, Measurement of the Elastic Properties and Intrinsic Strength of Monolayer Graphene, Science **321**, 385 (2008).

[47] S. Bertolazzi, J. Brivio, and A. Kis, Stretching and Breaking of Ultrathin MoS2, ACS Nano **5**, 9703 (2011).

[48] L. Cheng, C. Zhang, and Y. Liu, Why Two-Dimensional Semiconductors Generally Have Low Electron Mobility, Phys. Rev. Lett. **125**, 177701 (2020).

[49] F. Aryasetiawan and O. Gunnarsson, The GW Method, Rep. Prog. Phys. **61**, 237 (1998).

[50] J. Venderley, V. Khemani, and E.-A. Kim, Machine Learning Out-of-Equilibrium Phases of Matter, ArXiv171100020 Cond-Mat (2017).

[51] G. Carleo and M. Troyer, Solving the Quantum Many-Body Problem with Artificial Neural Networks, Science **355**, 602 (2017).

[52] G. Hautier, C. Fischer, A. Jain, T. Mueller, and G. Ceder, Finding Nature’s Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory, Chem. Mater. **22**, 3762 (2010).

[53] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Predicting Crystal Structures with Data Mining of Quantum Calculations, Phys. Rev. Lett. **91**, 135503 (2003).

[54] J. Behler, Constructing High-Dimensional Neural Network Potentials: A Tutorial Review, Int. J. Quantum Chem. **115**, 1032 (2015).

[55] N. Artrith, T. Morawietz, and J. Behler, High-Dimensional Neural-Network Potentials for Multicomponent Systems: Applications to Zinc Oxide, Phys. Rev. B **83**, 153101 (2011).

[56] J. Behler, First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems, Angew. Chem. Int. Ed. **56**, 12828 (2017).

[57] L. Lu, M. Dao, P. Kumar, U. Ramamurty, G. E. Karniadakis, and S. Suresh, Extraction of Mechanical Properties of Materials through Deep Learning from Instrumented Indentation, Proc. Natl. Acad. Sci. **117**, 7052 (2020).

[58] W. Ben Chaabene, M. Flah, and M. L. Nehdi, Machine Learning Prediction of Mechanical Properties of Concrete: Critical Review, Constr. Build. Mater. **260**, 119889 (2020).

[59] M. S. Scheurer and R.-J. Slager, Unsupervised Machine Learning and Band Topology, Phys. Rev. Lett. **124**, 226401 (2020).

[60] M. Nuñez, Exploring Materials Band Structure Space with Unsupervised Machine Learning, Comput. Mater. Sci. **158**, 117 (2019).

[61] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman, Machine Learning Bandgaps of Double Perovskites, Sci. Rep. **6**, 19375 (2016).

[62] V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, and K. S. Kim, Machine Learning for Predicting the Band Gaps of ABX3 Perovskites from Elemental Properties, J. Phys. Chem. C **124**, 8905 (2020).

[63] G. Pilania, J. E. Gubernatis, and T. Lookman, Multi-Fidelity Machine Learning Models for Accurate Bandgap Predictions of Solids, Comput. Mater. Sci. **129**, 156 (2017).

[64] 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).

[65] T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, and J. C. Grossman, Graph Dynamical Networks for Unsupervised Learning of Atomic Scale Dynamics in Materials, Nat. Commun. **10**, 1 (2019).

[66] Y. Zhuo, A. Mansouri Tehrani, and J. Brgoch, Predicting the Band Gaps of Inorganic Solids by Machine Learning, J. Phys. Chem. Lett. **9**, 1668 (2018).

[67] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, Prediction Model of Band Gap for Inorganic Compounds by Combination of Density Functional Theory Calculations and Machine Learning Techniques, Phys. Rev. B **93**, 115104 (2016).

[68] S. Chaube, P. Khullar, S. Goverapet Srinivasan, and B. Rai, A Statistical Learning Framework for Accelerated Bandgap Prediction of Inorganic Compounds, J. Electron. Mater. **49**, 752 (2020).

[69] K. Choudhary and F. Tavazza, Convergence and Machine Learning Predictions of Monkhorst-Pack k-Points and Plane-Wave Cut-off in High-Throughput DFT Calculations, Comput. Mater. Sci. **161**, (2019).

[70] Y. Dong, C. Wu, C. Zhang, Y. Liu, J. Cheng, and J. Lin, Bandgap Prediction by Deep Learning in Configurationally Hybridized Graphene and Boron Nitride, Npj Comput. Mater. **5**, 26 (2019).

[71] J. Wang, J. Li, S. Yip, S. Phillpot, and D. Wolf, Mechanical Instabilities of Homogeneous Crystals, Phys. Rev. B **52**, 12627 (1995).

[72] W. Kohn and L. J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects, Phys. Rev. **140**, A1133 (1965).

[73] J. P. Perdew, W. Yang, K. Burke, Z. Yang, E. K. U. Gross, M. Scheffler, G. E. Scuseria, T. M. Henderson, I. Y. Zhang, A. Ruzsinszky, H. Peng, J. Sun, E. Trushin, and A. Görling, Understanding Band Gaps of Solids in Generalized Kohn–Sham Theory, Proc. Natl. Acad. Sci. **114**, 2801 (2017).

[74] Á. Morales-García, R. Valero, and F. Illas, An Empirical, yet Practical Way To Predict the Band Gap in Solids by Using Density Functional Band Structure Calculations, J. Phys. Chem. C **121**, 18862 (2017).

[75] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, Concrete Problems in AI Safety, ArXiv160606565 Cs (2016).

[76] A. Shapeev, K. Gubaev, E. Tsymbalov, and E. Podryabinkin, Active Learning and Uncertainty Estimation, in Machine Learning Meets Quantum Physics, edited by K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller (Springer International Publishing, Cham, 2020), pp. 309–329.

[77] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, Active Learning with Statistical Models, J. Artif. Intell. Res. **4**, 129 (1996).

[78] C. M. Bishop, Bayesian Neural Networks, J. Braz. Comput. Soc. **4**, (1997).

[79] Y. Gal and Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, ArXiv150602142 Cs Stat (2016).

[80] Z. Lu and J. Bongard, Exploiting Multiple Classifier Types with Active Learning, in Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation (ACM, New York, NY, USA, 2009), pp. 1905–1906.

[81] D. W. Opitz and J. W. Shavlik, Generating Accurate and Diverse Members of a Neural-Network Ensemble, in Proceedings of the 8th International Conference on Neural Information Processing Systems (MIT Press, Cambridge, MA, USA, 1995), pp. 535–541.

[82] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, Less Is More: Sampling Chemical Space with Active Learning, J. Chem. Phys. **148**, 241733 (2018).

[83] K. Neklyudov, D. Molchanov, A. Ashukha, and D. Vetrov, Structured Bayesian Pruning via Log-Normal Multiplicative Noise, ArXiv170507283 Stat (2017).

[84] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors, ArXiv12070580 Cs (2012).

[85] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).

[86] F. E. White, Data Fusion Lexicon (PN, 1991).

[87] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. **98**, 146401 (2007).

[88] V. L. Deringer and G. Csányi, Machine Learning Based Interatomic Potential for Amorphous Carbon, Phys. Rev. B **95**, 094203 (2017).

[89] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, Neural Network Models of Potential Energy Surfaces, J. Chem. Phys. **103**, 4129 (1995).

[90] S. Saha, A Comprehensive Guide to Convolutional Neural Networks — the ELI5 Way, https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[91] S. W. Bedell, A. Khakifirooz, and D. K. Sadana, Strain Scaling for CMOS, MRS Bull. **39**, 131 (2014).

[92] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach, J. Chem. Theory Comput. **11**, 2087 (2015).

[93] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, Multisensor Data Fusion: A Review of the State-of-the-Art, Inf. Fusion **14**, 28 (2013).

[94] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, AFLOWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations, Comput. Mater. Sci. **58**, 227 (2012).

[95] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation, APL Mater. **1**, 011002 (2013).

[96] A. M. Saxe, J. L. McClelland, and S. Ganguli, Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks, ArXiv13126120 Cond-Mat Q-Bio Stat (2013).

[97] S. J. Pan and Q. Yang, A Survey on Transfer Learning, IEEE Trans. Knowl. Data Eng. **22**, 1345 (2010).

[98] B. J. Baliga, Semiconductors for High-voltage, Vertical Channel Field-effect Transistors, J. Appl. Phys. **53**, 1759 (1982).

[99] J. Bardeen and W. Shockley, Deformation Potentials and Mobilities in Non-Polar Crystals, Phys. Rev. **80**, 72 (1950).

[100] J. L. Corkill and M. L. Cohen, Band Gaps in Some Group-IV Materials: A Theoretical Analysis, Phys. Rev. B **47**, 10304 (1993).

[101] T. Inaoka, T. Furukawa, R. Toma, and S. Yanagisawa, Tensile-Strain Effect of Inducing the Indirect-to-Direct Band-Gap Transition and Reducing the Band-Gap Energy of Ge, J. Appl. Phys. **118**, 105704 (2015).

[102] null People, Indirect Band Gap of Coherently Strained GexSi1-x Bulk Alloys on <001> Silicon Substrates, Phys. Rev. B Condens. Matter **32**, 1405 (1985).

[103] F. Capasso, Compositionally Graded Semiconductors and Their Device Applications, Annu. Rev. Mater. Sci. **16**, 263 (1986).

[104] W. Shockley and H. J. Queisser, Detailed Balance Limit of Efficiency of P-n Junction Solar Cells, J. Appl. Phys. **32**, 510 (1961).

[105] D. P. Jenkins, Calculations on the Band Structure of Silicon, Proc. Phys. Soc. Sect. A **69**, 548 (1956).

[106] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. **77**, 3865 (1996).

[107] P. E. Blöchl, Projector Augmented-Wave Method, Phys. Rev. B **50**, 17953 (1994).

[108] G. Kresse and J. Furthmüller, Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set, Comput. Mater. Sci. **6**, 15 (1996).

[109] H. J. Monkhorst and J. D. Pack, Special Points for Brillouin-Zone Integrations, Phys. Rev. B **13**, 5188 (1976).

[110] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, TensorFlow: A System for Large-Scale Machine Learning, in Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (USENIX Association, Berkeley, CA, USA, 2016), pp. 265–283.

[111] D. P. Kingma and L. J. Ba, Adam: A Method for Stochastic Optimization, (2015).

[112] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-Learn: Machine Learning in Python, J Mach Learn Res **12**, 2825 (2011).

[113] L. Breiman, Random Forests, Mach. Learn. **45**, 5 (2001).

[114] J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat. **29**, 1189 (2001).

[115] Z. Shi, M. Dao, E. Tsymbalov, A. Shapeev, J. Li, and S. Suresh, Metallization of Diamond, Proc. Natl. Acad. Sci. **117**, 24634 (2020).

[116] E. Tsymbalov, Z. Shi, M. Dao, S. Suresh, J. Li, A. Shapeeva, Machine Learning for Deep Elastic Strain Engineering of Semiconductor Electronic Band Structure and Effective Mass, Npj Comput. Mater. (2021).

[117] D. Ciregan, U. Meier, and J. Schmidhuber, Multi-Column Deep Neural Networks for Image Classification, in 2012 IEEE Conference on Computer Vision and Pattern Recognition (2012), pp. 3642–3649.

[118] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, DeepInsight: A Methodology to Transform a Non-Image Data to an Image for Convolution Neural Network Architecture, Sci. Rep. **9**, 1 (2019).

[119] D. P. Kingma and M. Welling, An Introduction to Variational Autoencoders (Now Publishers, 2019).

[120] Y. Gal and Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (JMLR.org, New York, NY, USA, 2016), pp. 1050–1059.

[121] E. Tsymbalov, S. Makarychev, A. Shapeev, and M. Panov, Deeper Connections between Neural Networks and Gaussian Processes Speed-up Active Learning, Proc. Twenty-Eighth Int. Jt. Conf. Artif. Intell. 3599 (2019).

[122] H. Lian, A. N. Christiansen, D. A. Tortorelli, O. Sigmund, and N. Aage, Combined Shape and Topology Optimization for Minimization of Maximal von Mises Stress, Struct. Multidiscip. Optim. **55**, 1541 (2017).

[123] A. N. Christiansen, J. A. Bærentzen, M. Nobel-Jørgensen, N. Aage, and O. Sigmund, Combined Shape and Topology Optimization of 3D Structures, Comput. Graph. **46**, 25 (2015).

[124] A. N. Christiansen, M. Nobel-Jørgensen, N. Aage, O. Sigmund, and J. A. Bærentzen, Topology Optimization Using an Explicit Interface Representation, Struct. Multidiscip. Optim. **49**, 387 (2014).

[125] N. Aage, E. Andreassen, B. S. Lazarov, and O. Sigmund, Giga-Voxel Computational Morphogenesis for Structural Design, Nature **550**, 84 (2017).

[126] Y. G. Gogotsi, A. Kailer, and K. G. Nickel, Transformation of Diamond to Graphite, Nature **401**, 663 (1999).

[127] F. Nava, C. Canali, C. Jacoboni, L. Reggiani, and S. F. Kozlov, Electron Effective Masses and Lattice Scattering in Natural Diamond, Solid State Commun. **33**, 475 (1980).

[128] M. Willatzen, M. Cardona, and N. E. Christensen, Linear Muffin-Tin-Orbital and K·p Calculations of Effective Masses and Band Structure of Semiconducting Diamond, Phys. Rev. B **50**, 18054 (1994).

[129] H. Löfås, A. Grigoriev, J. Isberg, and R. Ahuja, Effective Masses and Electronic Structure of Diamond Including Electron Correlation Effects in First Principles Calculations Using the GW-Approximation, AIP Adv. **1**, 032139 (2011).

[130] B. J. V. Zeghbroeck, Principles of Semiconductor Devices (Bart Van Zeghbroeck, 2011).

[131] M. D. McKay, R. J. Beckman, and W. J. Conover, A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics **21**, 239 (1979).

[132] B. Regan, A. Aghajamali, J. Froech, T. T. Tran, J. Scott, J. Bishop, I. Suarez-Martinez, Y. Liu, J. M. Cairney, N. A. Marks, M. Toth, and I. Aharonovich, Plastic Deformation of Single-Crystal Diamond Nanopillars, Adv. Mater. **9**, 1906458 (n.d.).

[133] J. Kaczkowski, Electronic Structure of Some Wurtzite Semiconductors: Hybrid Functionals vs. Ab Initio Many Body Calculations, Acta Phys. Pol. A **121**, 1142 (2012).

[134] M. S. Hybertsen and S. G. Louie, Electron Correlation in Semiconductors and Insulators: Band Gaps and Quasiparticle Energies, Phys. Rev. B **34**, 5390 (1986).

[135] O. H. Nielsen, Optical Phonons and Elasticity of Diamond at Megabar Stresses, Phys. Rev. B **34**, 5808 (1986).

[136] X. Liu, J. Gu, Y. Shen, J. Li, and C. Chen, Lattice Dynamical Finite-Element Method, Acta Mater. **58**, 510 (2010).

[137] A. Nie, Y. Bu, J. Huang, Y. Shao, Y. Zhang, W. Hu, J. Liu, Y. Wang, B. Xu, Z. Liu, H. Wang, W. Yang, and Y. Tian, Direct Observation of Room-Temperature Dislocation Plasticity in Diamond, Matter **2**, 1222 (2020).

[138] X. Li, Y. Wei, L. Lu, K. Lu, and H. Gao, Dislocation Nucleation Governed Softening and Maximum Strength in Nano-Twinned Metals, Nature **464**, 7290 (2010).

[139] J. Xiao, H. Yang, X. Wu, F. Younus, P. Li, B. Wen, X. Zhang, Y. Wang, and Y. Tian, Dislocation Behaviors in Nanotwinned Diamond, Sci. Adv. **4**, eaat8195 (2018).

[140] J. R. Greer and W. D. Nix, Nanoscale Gold Pillars Strengthened through Dislocation Starvation, Phys. Rev. B **73**, 245410 (2006).

[141] C. Chisholm, H. Bei, M. B. Lowry, J. Oh, S. A. Syed Asif, O. L. Warren, Z. W. Shan, E. P. George, and A. M. Minor, Dislocation Starvation and Exhaustion Hardening in Mo Alloy Nanofibers, Acta Mater. **60**, 2258 (2012).

[142] P. B. Allen and V. Heine, Theory of the Temperature Dependence of Electronic Band Structures, J. Phys. C Solid State Phys. **9**, 2305 (1976).

[143] Y. P. Varshni, Temperature Dependence of the Energy Gap in Semiconductors, Physica **34**, 149 (1967).

[144] F. Giustino, S. G. Louie, and M. L. Cohen, Electron-Phonon Renormalization of the Direct Band Gap of Diamond, Phys. Rev. Lett. **105**, 265501 (2010).

[145] S. Poncé, G. Antonius, P. Boulanger, E. Cannuccia, A. Marini, M. Côté, and X. Gonze, Verification of First-Principles Codes: Comparison of Total Energies, Phonon Frequencies, Electron–Phonon Coupling and Zero-Point Motion Correction to the Gap between ABINIT and QE/Yambo, Comput. Mater. Sci. **83**, 341 (2014).

[146] Z. Shi, E. Tsymbalov, M. Dao, S. Suresh, A. Shapeev, and J. Li, Deep Elastic Strain Engineering of Bandgap through Machine Learning, Proc. Natl. Acad. Sci. **116**, 4117 (2019).

[147] G. L. Bir and G. E. Pikus, Symmetry and Strain-Induced Effects in Semiconductors (Wiley, 1974).

[148] I. Yu. Sahalianov, T. M. Radchenko, V. A. Tatarenko, G. Cuniberti, and Y. I. Prylutskyy, Straintronics in Graphene: Extra Large Electronic Band Gap Induced by Tensile and Shear Strains, J. Appl. Phys. **126**, 054302 (2019).

[149] J. C. Hensel, H. Hasegawa, and M. Nakayama, Cyclotron Resonance in Uniaxially Stressed Silicon. II. Nature of the Covalent Bond, Phys. Rev. **138**, A225 (1965).

[150] D.-D. Cui and L.-C. Zhang, Nano-Machining of Materials: Understanding the Process through Molecular Dynamics Simulation, Adv. Manuf. **5**, 20 (2017).

[151] G. M. Robinson and M. J. Jackson, A Review of Micro and Nanomachining from a Materials Perspective, J. Mater. Process. Technol. **167**, 316 (2005).

[152] P. Hess, Predictive Modeling of Intrinsic Strengths for Several Groups of Chemically Related Monolayers by a Reference Model, Phys. Chem. Chem. Phys. **20**, 7604 (2018).

[153] W. S. Verwoerd, Diamond Epitaxy on a Si(001) Substrate: A Comparison of Structural Models, Surf. Sci. **304**, 24 (1994).

[154] Q. Chen, 45-Degree Rotated Epitaxial Nucleation of Diamond on Silicon Using Chemical Vapor Deposition, ArXivcond-Mat9708124 (1997).

[155] B. D. Malone and M. L. Cohen, Quasiparticle Semiconductor Band Structures Including Spin–Orbit Interactions, J. Phys. Condens. Matter **25**, 105503 (2013).

[156] B.-C. Shih, Y. Xue, P. Zhang, M. L. Cohen, and S. G. Louie, Quasiparticle Band Gap of ZnO: High Accuracy from the Conventional G0W0 Approach, Phys. Rev. Lett. **105**, 146401 (2010).

[157] K. Ramakrishna and J. Vorberger, Ab Initio Dielectric Response Function of Diamond and Other Relevant High Pressure Phases of Carbon, J. Phys. Condens. Matter **32**, 095401 (2019).

[158] F. Giustino, Electron-Phonon Interactions from First Principles, Rev. Mod. Phys. **89**, 015003 (2017).

[159] H. Y. Fan, Temperature Dependence of the Energy Gap in Semiconductors, Phys. Rev. **82**, 900 (1951).

[160] A. A. Mostofi, J. R. Yates, G. Pizzi, Y.-S. Lee, I. Souza, D. Vanderbilt, and N. Marzari, An Updated Version of Wannier90: A Tool for Obtaining Maximally-Localised Wannier Functions, Comput. Phys. Commun. **185**, 2309 (2014).

[161] I. Souza, N. Marzari, and D. Vanderbilt, Maximally Localized Wannier Functions for Entangled Energy Bands, Phys. Rev. B **65**, 035109 (2001).

[162] N. Marzari and D. Vanderbilt, Maximally Localized Generalized Wannier Functions for Composite Energy Bands, Phys. Rev. B - Condens. Matter Mater. Phys. **56**, 12847 (1997).

[163] A. Togo and I. Tanaka, First Principles Phonon Calculations in Materials Science, Scr. Mater. **108**, 1 (2015).

[164] X. Gonze, Perturbation Expansion of Variational Principles at Arbitrary Order, Phys. Rev. A **52**, 1086 (1995).

[165] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials, J. Phys. Condens. Matter **21**, 395502 (2009).

[166] M. Dao, J. Li, Z. Shi, and S. Suresh, Elastic Strain Engineering of Defect Doped Materials, (16 April 2020).

[167] E. Grüneisen, Theorie des festen Zustandes einatomiger Elemente, Ann. Phys. **344**, 257 (1912).

[168] Y. Ma, M. Eremets, A. R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A. O. Lyakhov, M. Valle, and V. Prakapenka, Transparent Dense Sodium, Nature **458**, 182 (2009).

[169] G. Henkelman and H. Jónsson, Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points, J. Chem. Phys. **113**, 9978 (2000).

[170] G. Henkelman, B. P. Uberuaga, and H. Jónsson, A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths, J. Chem. Phys. **113**, 9901 (2000).

[171] M. Li and S. J. Zinkle, Deformation Mechanism Maps of Unirradiated and Irradiated V-4Cr-4Ti, J. ASTM Int. **2**, 1 (2005).

[172] G. Pizzi, M. Gibertini, E. Dib, N. Marzari, G. Iannaccone, and G. Fiori, Performance of Arsenene and Antimonene Double-Gate MOSFETs from First Principles, Nat. Commun. **7**, 12585 (2016).

[173] P. Gargini, The International Technology Roadmap for Semiconductors (ITRS): "Past, Present and Future," in GaAs IC Symposium. IEEE Gallium Arsenide Integrated Circuits Symposium. 22nd Annual Technical Digest 2000. (Cat. No.00CH37084) (2000), pp. 3–5.

[174] L. Smith, TCAD Modeling of Strain-Engineered MOSFETs, MRS Online Proc. Libr. OPL **913**, (2006).

[175] R. Tiwari, N. Parihar, K. Thakor, H. Y. Wong, S. Motzny, M. Choi, V. Moroz, and S. Mahapatra, A 3-D TCAD Framework for NBTI, Part-II: Impact of Mechanical Strain, Quantum Effects, and FinFET Dimension Scaling, IEEE Trans. Electron Devices **66**, 2093 (2019).

[176] G. Wang, J. Luo, J. Liu, T. Yang, Y. Xu, J. Li, H. Yin, J. Yan, H. Zhu, C. Zhao, T. Ye, and H. H. Radamson, PMOSFETs Featuring ALD W Filling Metal Using SiH4 and B2H6 Precursors in 22 Nm Node CMOS Technology, Nanoscale Res. Lett. **12**, 306 (2017).

[177] S. Carapezzi, S. Reggiani, E. Gnani, and A. Gnudi, Electron Mobility of Strained InGaAs Long-Channel MOSFETs: From Scattering Rates to TCAD Model, Solid-State Electron. **172**, 107902 (2020).

[178] S.-T. Chang, W.-C. Wang, C.-C. Lee, and J. Huang, A TCAD Simulation Study of Impact of Strain Engineering on Nanoscale Strained Si NMOSFETs with a Silicon–Carbon Alloy Stressor, Thin Solid Films **518**, 1595 (2009).

[179] S. Korneychuk, G. Guzzinati, and J. Verbeeck, Measurement of the Indirect Band Gap of Diamond with EELS in STEM, Phys. Status Solidi A **215**, 1800318 (2018).

[180] C. S. Granerød, W. Zhan, and Ø. Prytz, Automated Approaches for Band Gap Mapping in STEM-EELS, Ultramicroscopy **184**, 39 (2018).

[181] Ph. Redlich, F. Banhart, Y. Lyutovich, and P. M. Ajayan, EELS Study of the Irradiation-Induced Compression of Carbon Onions and Their Transformation to Diamond, Carbon **36**, 561 (1998).

[182] B. Li, C. Ji, W. Yang, J. Wang, K. Yang, R. Xu, W. Liu, Z. Cai, J. Chen, and H. Mao, Diamond Anvil Cell Behavior up to 4 Mbar, Proc. Natl. Acad. Sci. 201721425 (2018).

[183] L. Dubrovinsky, N. Dubrovinskaia, V. B. Prakapenka, and A. M. Abakumov, Implementation of Micro-Ball Nanodiamond Anvils for High-Pressure Studies above 6 Mbar, Nat. Commun. **3**, 1163 (2012).

[184] T. Yin, Y. Fang, W. K. Chong, K. T. Ming, S. Jiang, X. Li, J.-L. Kuo, J. Fang, T. C. Sum, T. J. White, J. Yan, and Z. X. Shen, High-Pressure-Induced Comminution and Recrystallization of CH3NH3PbBr3 Nanocrystals as Large Thin Nanoplates, Adv. Mater. **30**, 1705017 (2018).

[185] L. Lu, Y. Shen, X. Chen, L. Qian, and K. Lu, Ultrahigh Strength and High Electrical Conductivity in Copper, Science **304**, 422 (2004).