Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Evgenii Tsymbalov

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Machine Learning for Elastic Strain Engineering

Supervisor: Associate Professor Alexander Shapeev

The thesis document includes the following changes in answer to the external review process.

Dear Reviewers,

I would like to thank you for the useful comments which led to further improvements in the thesis text, as well as for highlighting possible future research directions. Please find the responses per each reviewer (sorted in the alphabetical order) below.

Prof. Nikolai Brilliantov

Q1: The study addresses an ideal crystal without any defects. This is correct for the idealized case of zero temperature, T=0. Each real crystal however has inevitably equilibrium point defects for non-zero temperature and generally, non-equilibrium defects like dislocations. Hence, a question arises, how the presented results would change, if these important properties of real crystals were taken into account? I do not expect quantitative estimates, but just a qualitative answer – whether the observed decrease of the bandgap persists? Will it shift for larger or smaller strain?
R1: To the best of my knowledge, the defects may introduce the “defect bands” (or intermediate bands) to the electronic bandstructure, which corresponds to the discrete eigenvalues of the Hamiltonian (in contrast to the continuous spectrum, which is related to the bands). Since the bandgap is defined as the difference between the conduction band minimum and the valence band maximum, technically, if these bands are not changed, then the bandgap value remains the same. However, the energy required to excite an electron to become a conduction electron decreases.

Q2: What would be the qualitative change of the results for non-zero temperature?
R2: A short answer is that for non-zero temperature the bandgap appears to be even smaller. This means that the metallization of both silicon and diamond is generally expected to happen earlier, for smaller strain values and elastic strain energy density values (*h* in the text).

To address the question fully, please allow me to cite a paragraph from our recent paper "Metallization in diamond":

"Recent experimental discovery (1) has established that monocrystalline and polycrystalline diamond nanoneedles (diameter ~ 300 nanometers) can be deformed reversibly to local elastic tensile strains higher than 9% and 3.5%, respectively, at room temperature. Note that due to the zero-point motion effect (17) and the Varshni effect (18), 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 (19, 20). 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 provided in the text."

(1) A. Banerjee, et al., Ultralarge elastic deformation of nanoscale diamond. Science 360, 300–302 (2018).
(17) P. B. Allen, V. Heine, Theory of the temperature dependence of electronic band structures. J. Phys. C: Solid State Phys. 9, 2305–2312 (1976).
(18) Y. P. Varshni, Temperature dependence of the energy gap in semiconductors. Physica 34, 149–154 (1967).
(19). F. Giustino, S. G. Louie, M. L. Cohen, Electron-Phonon Renormalization of the Direct Band Gap of Diamond. Phys. Rev. Lett. 105, 265501 (2010).
(20) S. Poncé, et al., 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. Computational Materials Science 83, 341–348 (2014).

Q3: Is it possible to extend the research for the multi-component crystals or 2D materials?
R3: Yes, it is possible to do it directly. It would be computationally easier to do for 2D materials on both *ab initio* simulations side and machine learning side. We focused on the 3D case as it seemed less covered by the researchers in general and posed a harder problem to tackle.

As for multicomponent materials, the machine learning part of my research, including the model and developed methodology, would be the same. A problem may arise from the increased time of *ab initio* calculations: we have tried to make a similar work on GaAs, but the accurate G0W0 correction procedure is too expensive for the general deformation case. It may be possible, however, if one is willing to trade some accuracy by e.g. reducing the ***k***-mesh size or using other functionals like HSE06, which are less computationally expensive. This part is left for future research.

Q4: What is the role of active learning in the developed methodology?
R4: It is used to increase the accuracy of the model by providing a smarter sampling procedure compared to the random sampling. A standard approach for the accuracy increase in surrogate models is to add more data, and active learning helps to choose the sampling points wisely by addressing the model uncertainty. This leads to smaller errors (compared to the random sampling) after the same number of samples; it is demonstrated in Figure 6.1, where random sampling compared to the two active learning approaches.

Q5: What are the limitations of the machine learning model proposed? May it benefit more from the multiple non-connected sources of data?

R5: One of the main shortcomings is the need for a fixed ***k***-mesh size, which does not allow to combine various calculations provided they have different settings. Another is that the phonon instability is not taken into account within the model, so model users must be aware of the strain range within which the model can operate.

Different sources may be combined, provided they have the same or compatible input (strain values) and output (electronic bandstructure) data format.

Q6: I suggest the author to comment more on the limitations of the proposed approach predictions, including surface effects, DFT-related errors, etc.
R6: Thank you for the suggestion. I added the following text to the end of Section 4.2.1:
“We want to note that even this setup and methods may be far from the truth, according to the various studies. For instance, Giustino et al. (2010) suggests that zero-point renormalization of the bandgap for the diamond crystal may be as large as 0.6 eV. We also do not consider the Varshni effect (Varshni, 1967), which suggests that the bandgap for the non-zero temperature is, in fact, smaller. Another significant limitation is connected with the possible defects of the crystal and surface effects, which may alter the result (see Nie et al. (2019)); to address these challenges, one needs to consider larger supercell in the calculation, including vacuum or defects. Last but not least, is a DFT convergence error: we did not use large parameters of energy cutoff or dense k-mesh in order to find a balance between calculation time and accuracy. A thorough discussion on the accuracy of ab initio approaches is out of scope for this work; however, ML machinery developed here can be applied to more rigorous data as well.”

Q7: In the section devoted to the DFT more theory is needed. Why did not the author mention the Slater determinant, which would naturally explain the exchange interactions? Otherwise it is not clear what is the physical meaning of these interactions.
R7: Thank you for the suggestion. Based on your and other reviewers’ comments, I rewrote and improved the DFT section to avoid possible confusion.

Q8: In Figs. 6.21 and 6.22 the values of about 1kcal/mole are shown. Is this a large or small error? This should be discussed explicitly in the text or in the figure captions.
R8: Thank you for the suggestion. This error is considered to be small: in the related DTNN paper (a), authors claim their model “an accurate approach”, discussing the 1 kcal/mol error on this particular dataset. I added the corresponding note to the text:
“For example, to reach the RMSE of 1 kcal/mol <u>(a state of the art error reported in Schutt et al. (2017))</u> starting from the SchNet...”.

(a) Schütt, Kristof T., et al. "Quantum-chemical insights from deep tensor neural networks." *Nature communications* 8.1 (2017): 1-8.

Q9: I do not think that it is a good idea to discuss the hydraulic simulator in the main text – this is rather far from the claimed topic of the thesis. I suggest briefly mention this approach in the main text and move the according material to the Appendix.
R9: Thank you for the suggestion. I have moved it into the separate Appendix.

Q10: While the author introduces the density-of-states plots for the band gap as a figure of merit, it would be interesting to see similar plots for other quantities considered in the work, like the effective mass tensor components.
R10: Thank you for this valuable suggestion. The following text was added (which will also be included in the upcoming paper on the diamond straining):

“We also data-mined the 6D strain space to study the conduction related properties and the elastic strain energy density against ε. Here, we adopted our ML model to acquire the many-to-many relation between conductivity effective mass for the conduction electron $m^*_{cond}(\varepsilon)$ and $h(\varepsilon)$, as shown in Figure 6a. The values of scalar m^*_{cond} are obtained by averaging individual longitudinal and transverse effective masses, as in Van Zeghbroeck (2010):

$$m^{\star}_{cond} = \frac{3}{\frac{1}{m_{11}} + \frac{1}{m_{22}} + \frac{1}{m_{33}}}.$$

The purple shading in Figure 6a 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. In principle, by using our model, we can accurately predict any components of the m^* tensor and their arithmetic averages for every **k**-point at every strain level.

In the design of photovoltaic cells and scintillators, it is desirable to adopt a semiconductor material with a direct bandgap and small effective mass to allow for a combining high light yield and conductivity. When ESE is used to modulate the bandgap and effective mass together, a lower elastic strain energy density is often preferable than a higher energy density for reaching the same property design. However, in our case of material properties optimization, the best solution that simultaneously minimizes all objectives (namely E_g, m^*_{cond}, and h) does not exist. Instead, we found out Pareto-efficient solutions that cannot be better off (decreased) in any of the three values without worsening off (increasing) at least one of the other two values. As shown in Figure 6b, the 3D Pareto front of minimized E_g, m^*_{cond}, and h indicates a trade-off must be made in simultaneously having a small bandgap and conductivity effective mass, where h could increase up to more than 120 meV/Å . One cannot achieve, for example, a near-zero bandgap and m*cond < 0.25me without paying a considerable price in h by deforming diamond, as indicated by the “infeasible region” in Figure 6b. Also, one can usually find higher h values that correspond to the same (E_g, m^*_{cond}) combination. The strain cases with such h values are in the “feasible region” in Figure 6b. In addition, if one would like to access to all possible combinations of (E_g, m^*_{cond}) achieved by straining diamond and to find the lowest elastic strain energy density (h_min) for each combination, Figure 6c could be a blueprint for this purpose. Note that it is not a 2D projection of the 3D Pareto front of Figure 6b where only minimized Eg and are present.”

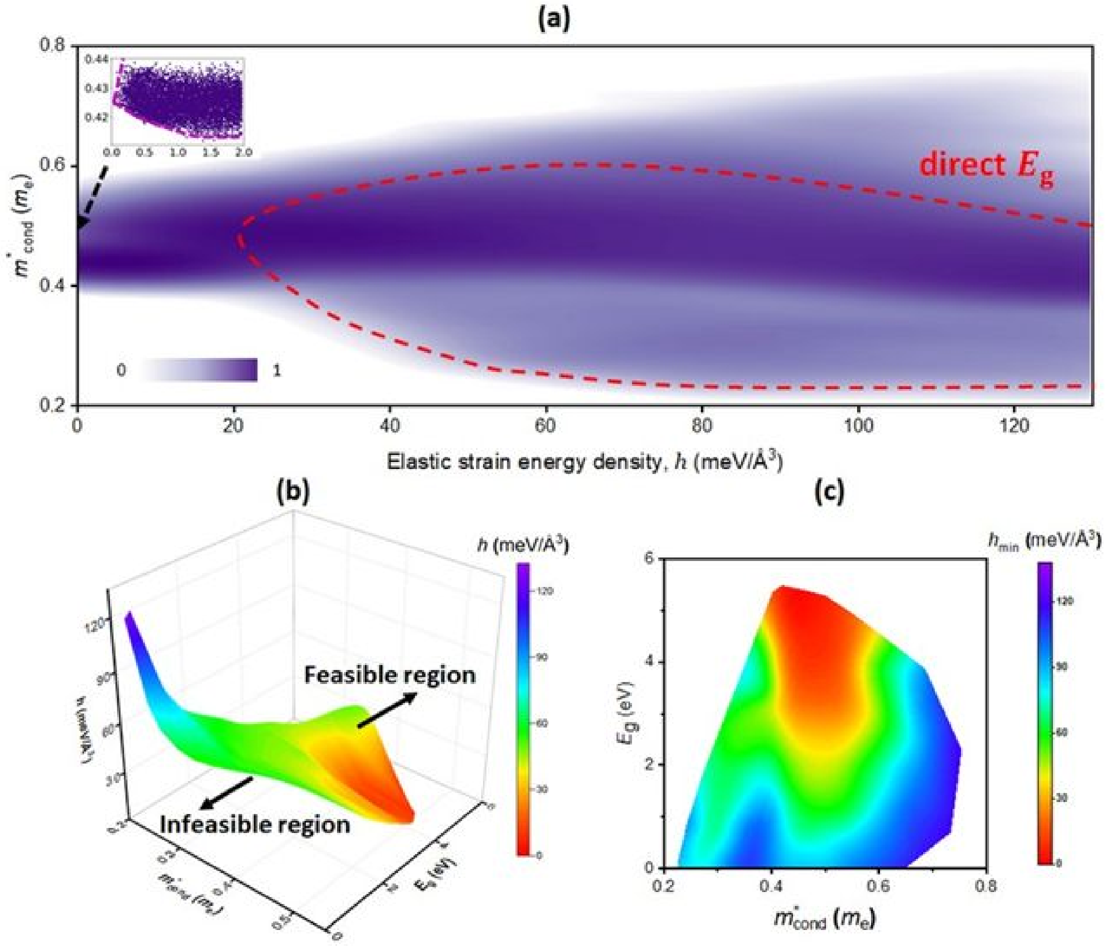

FIGURE 6: ML of electron effective mass. **(a)** Distribution and density of states of conductivity effective mass. Strain region where direct bandgap may appear is bound by the red dashed line. Inset is the zoomed-in plot near h = 0 of the m^*_{cond} distribution. **(b)** Pareto front for minimizing m^*_{cond}, bandgap, and h. The color contours indicate different elastic strain energy densities h. Points within the Pareto front are feasible while those beyond the Pareto front (under the colored surface) are infeasible. **(c)** Color contour plot of the lowest elastic strain energy density (hmin) required for achieving any combinations of bandgap and m^*_{cond}. (c) contains more (E_g, m^*_{cond}) points and is not a 2D projection of (b).

Q11: In page 126 the author makes a strange statement that the simulation results are "…more accurate, compared to the experimental values". How could it be? This sounds awkward, please reword.
R11: Thank you for the suggestion, I rewrote this part. It meant to be : "provides answers closer to experimental values"

Q12: There are also a few typos and minor mistakes in the text, especially in the last chapters of the work. I would advise drawing the author's attention to the following pages: 20, 39, 83, 91, 93, 110, 115, as well as captions to the figures in the ML experiments chapter.
Q12: Thank you for mentioning the particular places in the thesis text that indeed contain typos and mistakes; I've corrected them.

Dr. Bohayra Mortazavi

Q13: Could this model be used to predict other bandgap-related properties of the materials, such as thermal conductivity, optical absorption, or phononic properties?
R13: This particular trained model may help in obtaining some values, e.g., the bandgap, but in general, new data is needed. It is easy to train the model on new data of the same format; for instance, we are now working on the phonon bandstructure for the diamond crystal. For the thermal conductivity \kappa, we can obtain the necessary data with e.g. phonopy or phono3py Python packages provided the VASP OUTCAR files. As for the optical absorption, it is a slightly different problem as the output format is not of the bandstructure type but an imaginary part of the dielectric tensor, which may be obtained after additional calculations. I suppose that it may be fitted by using the pre-trained layers of the convolutional neural network proposed in the thesis, but it seems easier to me to treat it as a separate problem and use, e.g., simple fully-connected NN to fit it.

Q14: Enhancement of carrier mobilities in diamond and silicon by straining could be considered as the further extension of this work. In this case, the pattern of valance and conductance bands should be considered for the engineering. More discussions should be included with this regard.
R14: This is definitely an interesting research direction to consider, as our model may be directly used to estimate the effect of strain on valence and conduction bands. I have added the following text to the future work section:
“As for the possible extensions in terms of the values to tune and predict, we would like to mention the phonon bandstructure fitting, deeper exploration of the calculated properties (e.g., exploring the hole conductivity and enhancement of carrier mobilities, which involves tuning more Hessians like the one for the effective mass tensor). Another desirable application is the coupling of the proposed models with the FEM simulators, which was tackled in Section 7.3 yet can be used for the broader range of material science tasks.”

Q15: I also suggest to include more discussion on the effect of type of atomic lattices on the strain engineering, because by changing the lattice type the symmetry also changes.
R15: This is true, yet symmetries vanish quickly with deformations in the general case. The change of atomic lattice may result in some other strain-bandstructure symmetries, like the one in Eq. (4.12), to disappear, yet this will not affect the rest of the research much. One needs to carefully choose settings of ab-initio to include symmetric points in ***k***-mesh; the rest of the methodology and the model itself remain unchanged and thus ready for use.

Q16: Please also comment on the extension of this approach for the case of 2D materials, I think that would be highly appealing and computationally more feasible.
R16: This approach may be extended to include 2D materials, please see R3 (Response #3) for the full answer.

Q17: Personally, I would suggest moving the details of numerical experiments to the appendices.
R17: Thank you for the suggestion. I have moved a few parts to the Appendix, including the hydraulic simulator case and image classification parts.

Q18: One promising aspect is to couple the developed approach with machine learning interatomic potentials in order to accelerate the computations, please add more discussions about such a possibility.
R18: At the beginning of the research, MLIP-based approach was considered as the primary; we then moved to the simpler methods (fully-connected NN), which grew into advanced ones (convolutional NN). One of the options is indeed to use MLIP as an additional driver for the NN: to use it between the complex NN and *ab initio* calculations, as an intermediate model that is powerful enough to provide the NN with the approximate answers instead of VASP. I have included the following sentence in the text:
“One of the options is indeed to use MLIP (Podryabinkin et al., 2019) as an additional driver for the NN: to use it between the complex NN and *ab initio* calculations, as an intermediate model that is powerful enough to provide the NN with the approximate answers instead of VASP.”

Q19: A significant part of the content is related to the uncertainty estimation for the neural networks, yet it is used merely for active learning in the case of strain engineering task. I would appreciate if the author could elaborate more on the precise mechanism on uncertainty estimation and active learning for the diamond crystal and indicate whether there other uses of the produced uncertainty estimates, e.g., confidence intervals for the predicted values of band gap.
R19: The dropout-based approach to the uncertainty estimation (and related NN+GP approach as well) actually produces uncertainty estimates for each component of the bandstructure tensor. In the research, I average these estimates to produce a single value for each strain, and this value is then used to rank the strain values from “less” to “more” difficult. The most “difficult” ones (the strains with the largest average uncertainty) are then passed to the VASP, so the NN may train on them later.

It is indeed possible to consider these uncertainty estimates from the perspective of confidence intervals for each value: one can derive, e.g., the standard deviation of both conduction band minima and valence band maxima and then sum them to provide a standard deviation for the bandgap value. However, these estimates need to be properly calibrated first (for instance, on a separate validation set), and this direction is set for the future work.

Prof. Sergey Levchenko

Q20: Correct grammar in the whole thesis; some comments below are related to grammar, but there are many more mistakes.
R20: I would like to thank you for noticing and reporting numerous typos and mistakes within the thesis text (which are not listed here). I have corrected them as well as many other typos I noticed after your suggestions.

Q21: "spin does not affect our calculations" should be discussed more; spin affects many properties, why is it not important in your work?
R21: We have tested various strains in VASP and came to the conclusion that our system is spin-degenerate (i.e., both up and down spins correspond to the same eigenvalue). Nevertheless, the methodology and the proposed machine learning model do not change in case of systems that are affected by spin (e.g., Fe) -- this will result in more electronic bands to fit.

The following footnote was added: “This was tested by double-checking selected results with an *ab initio* calculations that account for the spin.”

Q22: "The second approximation we take into account is a one-electron approximation, or Hartree product, which treats the wave function psi as a product of individual electron wave functions" - this wave function does not obey the permutation rule for fermionic wave functions (it should change sign upon permutation of any pair of particles), and therefore Pauli exclusion principle; such a crude approximation is not used for a long time now
R22: Thank you for pointing out this mistake. I removed this part from the text.

Q23: "The next step to further simplification is to consider a density of electrons at a particular position in space" - this is not a simplification, since you still need to know psi_i; clarify what simplification you imply here.
R23: I rewrote this part as follows: “The next step is to consider …” to avoid ambiguity.

Q24: "It turns out that the equation above may be described not in terms of the electronic wave function y but in terms of the electron density n(r), which significantly reduces the number of unknowns to 3" - clarify which "equation above"; in general, this is a strange logic; first you should introduce Hohenberg-Kohn theorems, and then explain why they are so powerful (because n(r) depends only on three variables)
R24: Thank you for pointing this out. Based on this and another suggestion (Q7) I rewrote this part of the thesis to provide a better introduction to both Schrodinger equation and DFT.

Q25: "since the whole DFT approach (for any level of XC functional) is known for the poor performance within the semiconductors properties estimation." - this is wrong; DFT is in principle exact; even in approximate DFT hybrid functionals work quite well for semiconductors
R25: Thank you for pointing out this mistake. I rewrote this part (see R26 for the full correction text).

Q26: we also use the GW0 correction" - GW0 or G0W0? also specify on top of which reference
R26: For obtaining the training data, we used G0W0 on the top of PBE-PAW to speed up the process. However, we did a double-check for the most important deformations we discovered, which confirmed our G0W0-based findings. The corresponding correction was introduced to the text:
“In this work, we will use the Perdew–Burke–Ernzerhof (PBE) (Perdew et al., 1996) functional, which is a standard choice for solid-state calculations. This functional is known for the poor performance within the semiconductors properties estimation. Hence, on the top of it, we also use the G0W0 correction (Shishkin and Kresse, 2006), which accounts better for excited states”

Q27: "1. Define an initial trial electron density n(r)" - actually, you need initial trial psi_i
R27: Thank you for pointing out this detail. In this explanation, I was following the Sholland Steckel (2011)’s book on the DFT; however, I’ve found out that VASP indeed starts with the trial psi_i. The corresponding footnote was added to the text:

“Actually, in most ab initio packages, including VASP, this procedure starts with the trial ψ_i. After that, the electron density is calculated (step 3), and the procedure continues. The text remains unchanged since this part cites Sholl and Steckel (2011).”

Q28: "metals, which have a zero band gap with the conduction band being partially filled due to overlapping with the Fermi level" - "overlapping with the Fermi level" is not a reason, but a consequence.
R28: This text was changed to: **“**metals, which have a zero bandgap with the bands being filled up to the Fermi level”.

Q29: "Our ab initio calculations are restricted to the case of zero temperature (0 K); in the general case, the band gap also depends on the temperature." - you also did not include zero-point renormalization of the band gap; you've mentioned it later, but also mention it here
R29: Thank you for the suggestion. I added a mention at this particular place:
“Our *ab initio* calculations are restricted to the case of zero temperature (0 K); in the general case, the bandgap also depends on the temperature (Varshni, 1967), and the zero-point renormalization (Giustino et al., 2010) should also be taken into account.”.

Q30: "Another property that could be obtained from the electronic band structure is the Hessian of the conduction band, evaluated at the conduction band minima, called the free electron effective mass tensor" - why only conduction band minima, what about valence band maxima (hole conductivity)?
R30: The main purpose of considering the effective mass tensor was to demonstrate the accuracy of our model. The hole conductivity may be studied as well, I’ve added the corresponding text:
“We would like to note that this tensor can also be calculated for the valence band at the valence band maxima, accessing the hole conductivity. However, in this work, we focus on the m*.”, and to the future studies section, see also R14.

Q31: "Another direction is the incorporation of the GP-like elements into the NN structure" - explain why this is good/needed
R31: The *ad hoc* dropout-based uncertainty estimate often demonstrates poor performance compared to the GP-empowered approaches. Within the GP framework, an uncertainty estimate is natural and more accurate, due to strong theoretical guarantees. Unfortunately, straightforward use of GP scales poorly to large data sets. One of the solutions, which recently gained popularity in the research community, is the incorporation of GP-like elements into the NN, where NN often serves as a powerful trainable nonlinear transformation from the input data to the intermediate representation, which is then fed to the GP.

I have added the following explanations to the text: “.. into the NN structure, which may provide theoretical guarantees on the uncertainty estimates while preserving the power of NN, see ...”

Q32: "Dropout-based uncertainty quantification" - explain briefly the general idea of the dropout method
R32: Thank you for pointing this out. I’ve added the following description into the text:
“The main idea of dropout is to omit ("drop out") part of the activations of the hidden layer (while preserving the sample mean) during the training time. This allowed state of the art models to

reach better accuracy at the cost of increased training time. Nowadays, most of the modern deep learning architectures use dropout.“

Q33: Section "3.2.1 ML-simulation taxonomy" - describe more specifically how this is related to your work
R33: This section serves as a general overview of how ML and simulations may benefit from each other, on the higher methodology level, while particular details are described in Section 5.1. I've added the following text to clarify the relation:
“This section aims to answer the general methodological question: in what ways machine learning models and simulators can interact and benefit from each other? This allows our approach to find a place among similar approaches used in material science and surrogate modeling.“

Q34: "which does not provide any properties yet may be suitable for the large unsupervised exploration in future." - clarify this; if there are no properties provided, what can be learned then?
R34: Unsupervised machine learning studies the data without labels, and there are three particular applications I can think of:

1) learning the lower-dimensional representations, which may be more suitable for specialized models later, for the small pieces of data that may eventually obtain labels;
2) clusterization, which splits the data into the groups based on some similarity metric;
3) anomaly detection, which may point out particular data samples that differ dramatically from the rest of the data.

I added the corresponding clarification to the text:
“may be suitable for the large unsupervised exploration in future by, e.g., learning the lower-dimensional representations, or grouping molecules into the clusters.”

Q35: 3.2.2 Selected works in ML-assisted simulation - re-write to make clear and focus on how this is related to your work
R35: An intended target auditory for the thesis are computer scientists and graduate students that work on problems related to coupling machine learning algorithms with simulators. This is the reason why a large part of the thesis is devoted to the methodology description and uncertainty estimation, with a number of examples not related to the elastic strain engineering at all. For this reason, I have included this section to give a very brief overview of the problems and typical solutions in this field, providing readers with the necessary information. In the process of constructing the methodology and model for the elastic strain engineering -- which is the main problem I address in my work -- I found inspiration in the various solutions within the field, solutions that the computer science community is not likely to be familiar with. I want to share this part of my experience with the readers, in a hope that this may help them in future studies. I agree that this shifts the focus from the main problem in the thesis text, so I added a brief introduction to the beginning of this section:
“This section can be considered as a short overview of main data sources, problems, models, and challenges in the field of machine learning applications to chemoinformatics and materials science, as many of these models influenced the methodology described in this work.“

Q36: "We begin the methodology chapter with the high-level formulation on how exactly the machine learning model for ESE and the corresponding first principles calculations affect and complement each other" - this should have been discussed in the previous sections

R36: Previous chapters were mostly aimed at providing the readers with the necessary background in order to understand the methodology: DFT calculations, ML introduction, both in a Chapter 2, and a high-level overview in methodology used in ML-simulation coupling and related problems in Chapter 3. Starting from Chapter 4, readers are provided with the particular details on how the elastic strain engineering task is solved in this work. Some hints and details, however, appear in the previous chapters and contain a reference to the fine details in this and further chapters.

Q37: "Simulation result is ultimately an electronic band structure, represented as a rank-4 tensor, or a set of these calculations" - describe the tensor representation of band structure or give a reference

R37: I have provided a short description with a reference:
“Three out of four dimensions represent the k-space coordinates, and the last dimension denotes the band number, see Section 5.1.2 for details.”

Q38: "that will provide an essential road map for deep ESE, as will be demonstrated in Section 7.1" - explain what deep ESE is

R38: I was referring to the deep neural network model for the elastic strain engineering. The word “deep” was removed to avoid possible confusion.

Q39: FIGURE 4.1: explain criteria for the classification "stable/unstable" (how the "blue" and "red" classes are defined)

Q40: Thank you for pointing this out. I added a short description to both figure caption and text mentioned this stability comes from phonon calculations.

Q41: Eq. 4.8 - this is certainly not true for all k=(a,b,c), a,b,c in Z; clarify

R41: Thank you for pointing this out. The correct description, related to the Bloch theorem, includes the reciprocal space lattice vectors; the text was corrected.

Q42: "Figure 4.3 shows that for most of the bands, the dependence is really close to linear, with an excited bands (n ≥ nCB) showing a stronger relationship." - explain how you define excited bands

R42: I was referring to the conduction band and the bands above (n ≥ nCB); the text is changed to “ with the conduction and upper bands (n ≥ nCB) showing a stronger relationship.”

Q43: FIGURE 4.5: specify units on the axes; this should be done for all figures where units are applicable

R43: Thank you for pointing this out. The units here are in eV, the figure was updated.

Q44: "One possible reason for this is that both exchange-correlational functionals used for the diamond and the GW approximation do operate in an extreme regime." - this must be clarified; were different functionals used for DFT and as reference for GW?? what do you mean by "extreme regime"?

R44: No, we used the same functional. By “extreme regime” I mean that there is a large strain applied to the crystal lattice, and the PBE-PAW calculation step (and GW step as well) takes more time to converge compared to the small strain cases. Calculations at very large strains are more likely to fail. Therefore, we say that the PBE bandgap drop from 4 eV to the negative values may be an extreme situation for the underlying simulation. I agree that this may confuse the readers as such details are not provided in the thesis text and removed this sentence for clarity.

Q45: "This plot indicates that the direct band gap usually requires the shear (off-diagonal) strain components e_xy, e_yz, e_xz to be far from zero." - clarify what you mean by "far from zero" (the gap or the probability to find a direct gap)
R45: This sentence was rephrased for clarity: “This plot indicates that the typical requirement for a direct bandgap to appear involves the strain cases with the shear (off-diagonal) components being far from zero: **|e_xy|, |e_yz|, |e_xz| >> 0**”.

Q46: Figure 4.6: explain the figure better; what are the plots ar the end of each row?
R46: This description was added to the caption:
“At the end of each row, density plots, or smoothed histograms, are provided for the strain components.”

Q47: "For a given k-grid invariant property (e.g., E_g)" - E_g is not k-grid invariant; modify or clarify
R47: This text was modified to: “For the bandgap as an electronic bandstructure property, …”

Q48: "For these, there is a complicated one-to-one correspondence between the dispersion energies in k-points of the corresponding electronic band structures we found empirically" - can you explain these symmetries?
Q48: We tried to obtain theoretical explanations for these symmetries, yet the complexity arises from the usage of the specific strain tensor, presented in Section 2.1.3. We believe that rigorous proof exists, and this is left for the future work.

Q49: "Second, obtaining a full band structure at once offers a more comprehensive description of what is going on in terms of effects caused by the band structure change. " - you can also mention that the full band structure calculation has a negligible cost on top of specifically band gap determination
R49: Thank you for the suggestion. The corresponding text was added:
“... band structure change, and is required for a proper bandgap estimation”

Q50: "energy bands evolve piecewise-smoothly with changes in k, and the information within the energy dispersion" - clarify how you treat band crossings; how do you identify a smooth band in that case?
R50: Band crossing usually happens at the boundary of the Brouillon zone. We treat it the way the VASP treats it -- ordering the bands for each ***k***-point according to the energy dispersion value. The band changes smoothly as a function of ***k***, and the piecewise-smoothness arises from this ordering approach.

Q51: FIGURE 5.2: Usually bands are calculated on a much denser set of k-points along the lines, while you represent the bands on a 3D mesh in the figure; how do you actually represent bands?
R51: In order to obtain the canonical electronic bandstructure image -- the “spaghetti plot” -- one usually chooses a special regime on the top of existing calculations, specifying the ***k***-path, along which dispersion energies are calculated (the tight-binding model is another option to consider); this is also used to produce such images as Figure 7.10. Within the NN model, we do not actually represent the bands fully but the energy dispersion relations at the certain ***k***-points. These values are stored within the rank-4 tensor, please refer to R37 for the additional information on the bandstructure representation.

Q52: "to describe the energy dispersion near the Fermi level of diamond" - you should explain how you define the Fermi level for a system with a gap; it is not well defined in fact, usually it is placed at the middle of the gap
R52: We merely use the VASP output settings to define the Fermi level (although this is not used in the thesis directly). In general, this sentence was related to the fact that in order to predict the bandgap we may fit the valence and conduction bands only. In fact, we also fit the neighbouring bands (one band below the valence band and one band above the conduction band) since data analysis indicates that they correlate with their neighbours -- this is used within the NN convolutional structure.

To avoid ambiguity, this sentence was rewritten as “...to describe the energy dispersion near the bandgap...”.

Q53: FIGURE 5.4: explain in the caption the meaning of n and b
R53: n refers to the electronic band number, and **b** is a batch size (NN calculations are done in a parallel manner, so **b** data samples are predicted at once). I have added these explanations to the caption.

Q54: "In the first part, we trained our model on the large dataset (~ 35,000 samples) " - explain what is a sample (different strains?)
R54: Yes, 35 000 samples means 35 000 different strains; I’ve added the corresponding correction to the text to avoid ambiguity.

Q55: "Within this approach, one can sample some number T of i.i.d. realizations" - what is "i.i.d."?
R55: i.i.d. refers to “independent and identically distributed”. I’ve added the corresponding explanation to the text to avoid ambiguity.

Q56: "We propose to overcome the difficulties mentioned above by considering the full approximate posterior distribution." - it should have been clearly explained BEFORE the long description of various approaches, that the commonly used approaches have drawbacks, and YOU suggest an approach that addresses at least some of these drawbacks (if not all). In other words, you should better motivate the long review parts of the thesis.
R56: I agree that Section 5.2, which is devoted to the description of various dropout-based methods for the uncertainty estimation, steps off the general line of the thesis (the elastic strain engineering). One possible way was to incorporate it into the thesis text is to split this section into the smaller chunks, which contain a literature review, description of known approaches, and developed methods, and put it into corresponding chapters. However, I do believe that this may puzzle the reader way more than putting this part into a single section; therefore, I decided to concentrate the methodology description for the uncertainty estimation in a few pages, so interested readers can delve deep into details of the dropout-based methodology, and the readers more interested in the elastic strain engineering can skip it. Taking your consideration into account, I added a remark to the end of the section related to the description of the baseline approach:
“We would like to note that the basic dropout-based UE and the corresponding active learning procedure have several drawbacks, including biased estimation, greedy sampling algorithm, and loose theoretical guarantees, which will be discussed in detail in the next sections. In order to

address these drawbacks, two approaches are proposed: GP-based enhancing of a trained NN, described in Section 5.2.3, and dropout mask diversification, introduced in the next section."

Q57: "Effective mass estimation. The partial derivatives used in the effective mass estimation (2.10) were approximated using central differences with the mesh size proportional to the k-grid internal distance" - clarify if you calculate an effective mass tensor or just a scalar along a chosen direction in k-space
R57: I calculate components of the effective mass tensor, but only diagonal values of the tensor are reported; the following sentence was added to the text to avoid ambiguity: "All the other derivatives are approximated in the same way to achieve the effective mass tensor approximation."

Q58: "BE-PAW data could be sampled on a much larger scale since it is 100-1000 times cheaper in terms of computational time of first-principles calculations." - clarify cheaper than what
R58: Than G0W0 data. The sentence was corrected to avoid ambiguity.

Q59: Table 6.2, 6.3, 6.4, 6.5: report maximum absolute error as well; or even better make a plot with error distribution
R59: Thank you for this valuable suggestion. I included the error distribution plot for the selected values into the text:

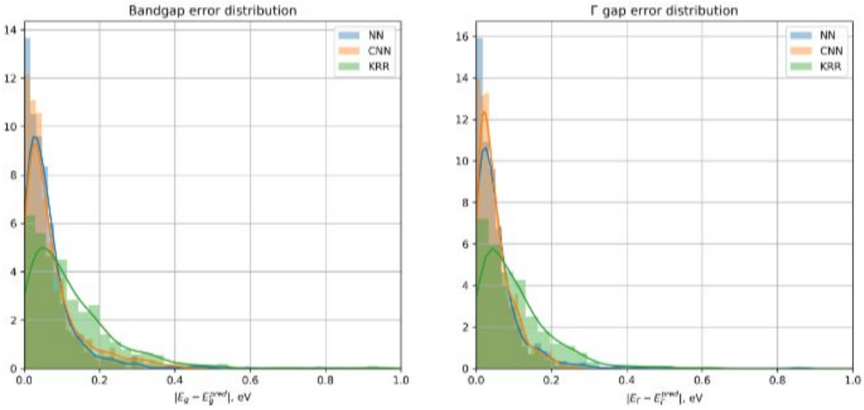

FIGURE 1: Error distribution for different algorithms for the bandgap and Γ gap prediction tasks. CNN and NN demonstrate similar performance while KRR shows the long tail error distribution.

Q60: "It is a quantity used to model the behavior of a free electron with that mass" - this reads as a tautology, and in fact does not add anything to the discussion; remove or replace with something more meaningful
R60: While this may be obvious for the reader with a physical background, I believe that the reader without such background (e.g., coming from the field of computer science) may benefit from this short explanation.

Q61: "it reveals not only the shape of an energy band but also the curvature of it" - sounds strange, curvature is part of shape, clarify what you mean by shape
R61: I rewrote this sentence to avoid ambiguity: "... it reveals not only the shape of an energy band but also provides more detailed information of energy dispersion.".

Q62: "In this section, we aim to show the applicability of the proposed methods to the classification tasks, computer vision problems in particular." - clarify how this is related to the topic of the thesis (elastic strain engineering)?
Section 6.4 - clarify how this is related to the topic of the thesis
"The rest of the section is dedicated to the general active learning and uncertainty estimation for neural networks." - summarizing some comments above, this part looks disconnected from the topic of the thesis
R62: These parts are not related to the elastic strain engineering but to uncertainty estimation for the neural networks in general. These experiments are shown to test the proposed approaches to the uncertainty estimation in other problems, and, in my opinion, can help other researchers as well. In order to cut this and other sections not related to ESE, I moved some parts to Appendices, including hydraulic simulator and image classification experiments.

Q63: Figure 6.22: explain what RND means
R63: RND refers to the random sampling procedure. I added an explanation to the caption.

Q64: "This section represents the quintessence of this work – namely, results and insights discovered by the use of high-throughput machine learning models." - it is a problem that "quintessence of this work" comes close to the end of the thesis
R64: I deliberately moved the most impressive results related to the discoveries in elastic strain engineering to the end of the thesis, in order to shift the attention from the physical part to the machine learning methodology.

Q65: describe how strain can be realized in practical applications
R65: While this is definitely a weak point of the work not to consider the ways to realize the found deformations, to the best of my knowledge, this is still an unsolved problem for the general case. The closest example is obtained by the bending of nanoneedles in (a) and is reported in Section 7.3.

(a) Banerjee, Amit, et al. "Ultralarge elastic deformation of nanoscale diamond." *Science* 360.6386 (2018): 300-302.

Q66: Figure 7.10 in caption: "(A and B) represents the ‘D-L’ transition and (Band C) shows the indirect-to-direct transition." - clarify the difference between A and B more; it does not look like B "shows the indirect-todirect Transition"
R66: Thank you for pointing this out. These figures are aimed to demonstrate the “journey” of the surface of the bandgap isosurface shown at Figure 7.6: it starts at the \Delta_3 face (A), then moves to the L_1 face (B) and ends at the \Gamma “tip” (C). The B -> C part shows the indirect-to-direct transition. I added the explanation text into the caption.

Q67: "8.3 Author’s contribution" - I think some of this info should be reflected in other parts of the text (e.g., when you write "we use..." it should always mean you, and other contributions should be mentioned separately)
R67: Thank you for the suggestion. I added references to my collaborators wherever appropriate.

Dr. Justin Smith

Q68: The initial random data set seemed very large compared the data generated in the very few active learning iterations. This made the active learning work seem to be tackled on, rather than an ultimate effort to produce the best model with minimal data. Perhaps a justification on why one wouldn't start with a much smaller initial data set, the active learn to error convergence, is warranted. Could it be possible to build the model with much less PBE data? This would be important information for future studies wanting to use these methods for other materials.
R68: I agree that the current application of the active learning methods could be considered as a fine-tuning of the existing model rather than the full active learning pipeline. Two interconnected reasons for that is the current model not being optimal in terms of the number of parameters to train, and the need for the large fully-connected layers in order for dropout-based uncertainty estimation to work. However, I believe that the proposed methodology may be of greater use in more data-intensive problems, and experiments with other tasks (e.g., image classification) indicate that this might be the case.

The model may be built with less PBE data (e.g., using 6,000 samples instead of 32,000) yet this kind of data is cheap, and we used all the data produced in the exploration (data analysis) stage. I added a footnote to reflect this information: "We used all the data calculated on the exploratory analysis part of the research; in practice, the model may be pre-trained on a much smaller amount of data (~ 5 000 strains)"

Q69: Perhaps I overlooked it, but I could not find how the transfer learning data (GW) was selected. Was this data simply a random subsample of the larger PBE data?
R69: The data was selected "uniformly randomly" with the Latin Hypercube Sampling. It was not a subsample of PBE data since the latter is pretty cheap to sample. I added the clarification to the text.

Q70: Ensemble-based UE is said to "increase training time". This is not necessarily true since models in the ensemble can be trained (and evaluated) in an embarrassingly parallel manner. It does, however, increase the overall computational resources required.
R70: I rewrote this sentence to refer to the computational resources instead of the training time.

Q71: Are there any major disadvantages to using the dropout-based UE method? Dropout tends to require a very large number of model parameters compared to non-dropout models, does this slow the model training and evaluation down significantly? How much larger did you need to make the dropout-based models compared to a normal model?
R71: The major disadvantages for the dropout-based UE are related to the low quality of the corresponding uncertainty estimates; this problem was partially addressed by the diverse masks and NNGP introduced in this work. Dropout is widely used across various architectures in ML: from CNNs and RNNs up to Transformer models, and the main point of using dropout-based UE is that dropout is already incorporated into the NN model in most of the cases.

Dropout-based training indeed slows down the convergence for the large dropout rates but often results in a better generalization error, as was shown in the original paper (a); I used dropout as a universal regularization tool. If one may come up with a small yet powerful NN model (and by small I mean having less than 128-256 neurons in the hidden layers), there is no need to artificially incorporate the dropout; other options, like Gaussian noise injection, can be used instead.

As for the evaluation time, one usually just turns off the dropout during the inference stage, so the evaluation time remains unchanged. Theoretically, the right way is to consider an average prediction between the multiple stochastic passes, yet in practice, predictions barely differ; considered models are not an exception in this regard.

(a) Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15.1 (2014): 1929-1958.

Q72: How (if possible) could these methods be extended to have a single model predict the bandstructure on multiple crystal phases and/or elements? This could make an interesting discussion in the future work section.
R72: Personally, I do believe that the paradigm of having separate surrogate models for each problem (element or compound) is far more powerful and accurate than the approaches aimed at universal approximation. I believe that the current methods may be extended to work on multiple crystal phases provided enough data, yet will advise on having separate models instead.
These models, however, may share some of their parts (e.g., weights on the earlier layers) or have multiple “heads” to train in a multi-task fashion. These ideas are reflected in the future work section.

Dr. Alexey Zaytsev

Q73: Can the author provide an evidence that considered GP models provide results of reasonable quality for both their mean prediction and uncertainty estimates? How to take advantage of the structure of multidimensional output while constructing GP model?
R73: Considered GP models are not used for the mean prediction; to obtain the predictions, we use an underlying NN model. As a consequence, GP part is trained separately, at the moment the user needs uncertainty estimates for, e.g., an active learning step. Moreover, uncertainty estimates are not used for the confidence intervals of the predicted values; this direction is set for the future work, see R19 for the related answer.

Inner relations within the multidimensional output may be addressed directly by the GP in the correlation matrix or with the choice of an appropriate kernel. However, I believe that a better, pure GP model may be derived for this particular task. If one is willing to consider the ***k***-points as an input to the model, this may result in a dramatic decrease of the output data size (from 2048D to 4D), making the modern GP methods applicable. The corresponding discussion is mentioned in the future work section.

Q74: It would be interesting to compare other uncertainty estimation approaches in the scope of the main problem of the work to provide practical recipes on how to construct and use surrogate models in this area.
R74: Thank you for the suggestion. This direction is indeed interesting yet left for the future work.

Q75: The author doesn’t use specific multifidelity approaches to construct surrogate models, while there is a number of GP regression models designed to work with multifidelity data, e.g., cokriging and derivative works in this area.

R75: In the thesis, GP was used as a simple extension to the proposed model, and indeed did not benefit from the rich heritage of GP-related works (including ones connected to the multifidelity data handling), this is an interesting direction for the future work. However, I do believe that these approaches are more suitable for a more compact GP-based model, as the one mentioned in R73.

Q76: Why the author use Bayesian Neural Networks? Now GP model is based on mean prediction by NN model and don't require multiple runs of NN with different dropout masks.

R76: There is a number of approaches for coupling NN and GP models mentioned in the text. However, they require an extra training along the NN, and may significantly decrease the training time. I used Bayesian Neural Networks mainly because of the preference of the dropout as a regularization tool, see R71 for further discussion.