

Jury Member Report – Doctor of Philosophy thesis.


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

Name of the Reviewer: Sergey Levchenko

I confirm the absence of any conflict of interest	Signature:  Date: 19.09.2020
---	---

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

- Brief evaluation of the thesis quality and overall structure of the dissertation.

Scientifically, the thesis is of the highest quality. It is multidisciplinary in nature, since it includes development of machine learning methods and data analysis approaches, and also their application to physical problems. It is clear that the candidate has built a strong and unique expertise, and is a capable independent researcher. The structure of the thesis, however, is not ideal. The text is filled with literature review, but it is often not clear outright what the relation between the reviewed literature and the thesis subject is. Another problem is the quality of English. The text can be understood, but there it is full of grammatical mistakes. These issues must be corrected, but I do not see any serious obstacles to addressing them.

- The relevance of the topic of dissertation work to its actual content

The topic does not actually reflect the main content of the thesis. In fact, the thesis title does not do justice to the main contribution of the author: method development. While only application to strain

engineering is mentioned in the title, a large portion of the thesis is devoted to the development of machine-learning methodology, and a few other applications are discussed prominently.

- The relevance of the methods used in the dissertation

Method development and tailoring to particular applications is the main part of the work, so the methods are very relevant for the thesis.

- The scientific significance of the results obtained and their compliance with the international level and current state of the art

I assess the scientific significance of the results as very high. The study has not only addressed important practical challenges, but has also gone beyond state-of-the-art in development of machine-learning approaches to materials science problems. The high scientific significance of obtained results is reflected in the high level of the journal (PNAS) where an important part of the work was published. The research presented in the thesis significantly advances the field of machine-learning in materials science.

- The relevance of the obtained results to applications (if applicable)

The thesis contains results of great practical importance. The main application is elastic strain engineering that is gaining more and more attention as means of tuning materials properties due to development of advanced approaches to materials fabrication. In addition, although not reflected in the thesis title, other practical applications are discussed, including fluid flow in a wellbore for drilling applications, airplane flight scheduling, and others.

- The quality of publications

The main publication is of very high quality. It is published in the Proceedings of the National Academy of Sciences, a high-level peer-reviewed journal. In addition, there are two publications in conference proceedings, one other publication submitted to a conference proceedings, and one manuscript in preparation.

The summary of issues to be addressed before/during the thesis defense

The comments listed below are meant to be addressed directly by modifying the thesis text, unless stated otherwise or the comment is unclear and needs further discussion.

Correct grammar in the whole thesis; some comments below are related to grammar, but there are many more mistakes

"material science" -> "materials science", "material design" -> "materials design", "material properties" -> "materials properties"

"heppening" -> "happening"

"and the experimental part starts from the methodology description, given in Chapter 4" - "experimental" here is misleading

"diamond cubic, also known as face-centered cubic (FCC)" - diamond cubic and FCC are not the same; diamond cubic denotes two atoms in a certain relative position, repeated according to FCC Bravais lattice

"Points of k-space (referred to as k-points) are used to describe electromagnetic properties within the ideal crystal." - "electromagnetic" usually means something else (electromagnetic fields); I suggest to replace with "electronic"

"this introduction will follow the gentle steps from" - "gentle" does not seem to be the right word here

"A complete representation of ψ should also include an electron spin, yet we omit it due to clarity of presentation and the fact that spin does not affect our calculations." - "for the sake of clarity"; also, the part "spin does not affect our calculations" should be discussed more; spin affects many properties, why is it not important in your work?

"much less rapidly" -> "much slower"

"The second approximation we take into account is a one-electron approximation, or Hartree product, which treats the wave function ψ 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

"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 ψ_i ; clarify what simplification you imply here

"It turns out that the equation above may be described not in terms of the electronic wave function ψ 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)

"referred to as exchange-correlational (XC)" - there is no term "exchange-correlational", it should be "exchange-correlation functional"

"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

"we also use the GW0 correction" - GW0 or G0W0? also specify on top of which reference

"We can finally describe in general terms the self-consistent procedure of solving the equation (2.7)" - you do not solve eq. 2.7, you solve eq. 2.8

"1. Define an initial trial electron density $n(r)$ " - actually, you need initial trial ψ_i

"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.

"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

"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)?

"As the optimization algorithms are iterative" - explain what you mean by iterative here

"In our models, we would rely on the weight regularization and dropout in the case of the neural networks. For other algorithms, we would use the standard means of regularization as well." - I would remove both "would".

"In some cases, his technique makes it possible to explicitly express the variance of the model output " - "this", not "his"

"It is worth noting that finding suitable prior weighting distributions of scales is especially relevant in the context of this approach" - did you want to say "important" instead of "relevant"?

"Another direction is the incorporation of the GP-like elements" - check if you have defined abbreviation GP

"Another direction is the incorporation of the GP-like elements into the NN structure" - explain why this is good/needed

"However, most of the tasks of the so-called "physics of processes" are devoted to regression. This approach is attractive..." - clarify what "This approach" refers to.

"Dropout-based uncertainty quantification" - explain briefly the general idea of the dropout method

check if UE abbreviation was defined; and then either use it everywhere (except maybe in abstract and conclusions) after it is defined, or do not use it at all

"In practice, there is a number of scenarios for which we cannot be sure that there is a better solution" -
> "In practice, there are a number of scenarios when this may not be possible"

"one could just simply recalculate " - remove either "just" or "simply"

"covariation matrix" -> "covariance matrix"

"one could simply run more epochs on the updated data set" - explain what is "epoch" in this context for non-specialists

"This concludes the general introduction to the machine learning models, which would be used in this work" - "would be" -> "are"

FIGURE 3.1: show the dates at the top plot

"A more technical description of the strained Si technology may be found in ..." - "can be found in" ("may" here implies that you are not sure if they can be found there or not); correct similar cases in the whole text

"As for the diamond crystal, an ultra-wide band gap of 5.5 eV in the unstrained state connected with superior properties in terms of durability and melting temperature makes it quite challenging to bend." - strange sentence; how a wide band gap can make diamond challenging to bend? why are you talking about bending now?

"or experimentally via transmission electron microscope" - "microscopy"

"A short and shallow overview" -> "A short overview"

choose one way of writing words/phrases (e.g., band gap or bandgap, but not both)

"describes a direct Ge on Si sheet provided by 5.7% uniaxial strain (epitaxy)" - seems like a word is missing after "direct"

"One of the large problems with the early development is a small band gap value (0.6 eV), which turns into the zero band gap for a simple PBE calculation." - this is not a good place to mention PBE problems

"In Even et al. (2014), strain effects for hybrid organic perovskites are studied via DFT" - this should not be listed within "Aluminium and other metal alloys" section

Section "3.2.1 ML-simulation taxonomy" - describe more specifically how this is related to your work

"First of all, it is QMX (QM7, QM9, etc.) and GDB-X (GDB-7, GDB-9, GDB-11) databases" - some of these (QMX) are datasets, not databases; clarify this part

"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?

"that describes the system evolution in time and is usually related to the interaction of chemical molecules" - "chemical molecules" is poor English, should be corrected

"Two important steps that often occur in a typical calculation" -> "Two important types of multi-step ab initio calculations"

"A classic approach requires the calculation of pairwise distances between pairs of atoms, and angles between the triples of atoms, within a certain cutoff radius of each atom." - why are you stopping at triples of atoms? what about quadruples, and so on?

3.2.2 Selected works in ML-assisted simulation - re-write to make clear and focus on how this is related to your work

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

"is a handy approximation describing the first-principles mentioned in Section 2.2" - "describing the first-principles" is poor English, please improve

"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

"Hypothesis set is an underlying model class we use for training: NNs and CNNs, with a particular specimen described in details in Sections 5.1.1 and 5.1.2, correspondingly." - "specimen" does not fit here, use another word (variants?)

"well-bore" -> "wellbore"

"discoveries presented in this work would simply be incomprehensible for plain DFT calculations" - "incomprehensible" is not a good word here, use another one

"This is a standard preparation step, as GW calculations always require a one-electron basis set." - not for self-consistent GW

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

"We have sampled > 10000 strains within the following limits:" - specify strain units (I guess it is in %) in the equations

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

"The details on the estimation of the translational parameter" -> "The details on the estimation of the line translation parameter in eq. 4.6"

"The percent of unstable strains as a function of the translation parameter" -> "The percent of unstable strains as a function of the line translation parameter"

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

"Figure 4.3 shows that for most of the bands, the dependence is really close to linear, with an excited bands ($n \geq n_{CB}$) showing a stronger relationship." - explain how you define excited bands

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

"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"?

"As we had seen in a previous section, " - tell in which section

"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)

Figure 4.6: explain the figure better; what are the plots at the end of each row?

"For a given k -grid invariant property (e.g., E_g)" - E_g is not k -grid invariant; modify or clarify

"The mean absolute deviance" - deviation

"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?

"Another example is a possible prediction of a free electronic mass tensor" - I think you did not want to write "free"

"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

"To this extend" -> "To this extent"

"bandgap-narrow" -> "narrow-bandgap"

"and a more advanced and compatible algorithm" - "compatible" or "competitive"? or did you mean "suitable"? compatible usually implies "compatible with ..."

"Therefore, the E^{PBE}_g consistent with the query strain case is learned using exclusively e and E^{PBE}_g as input, as illustrated in Figure 5.1." - I guess you meant " E^{GW}_g ... is learned using exclusively e and E^{PBE}_g as input."

"an image and E_n denoting the "color-scale"" - I guess it should be E_n (n is subscript)

"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?

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?

"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

FIGURE 5.4: explain in the caption the meaning of n and b

"In the first part, we trained our model on the large dataset (~ 35,000 samples) " - explain what is a sample (different strains?)

"Within this approach, one can sample some number T of i.i.d. realizations" - what is "i.i.d."?

"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.

"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

"10'000" -> "10000" or "10,000", choose one and use consistently

"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

Table 6.2, 6.3, 6.4, 6.5: report maximum absolute error as well; or even better make a plot with error distribution

"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

"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

Figure 6.3: in the top panel there are 4 curves but six labels

"However, even for the unpleasant case shown in Figure 6.4" - "However, even for..."

"train set" -> "training set"

"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

Figure 6.22: explain what RND means

"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

"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

Fig. 7.1: explain better the details in the caption; what is marked by the red filled circles? similar comment on fig. 7.4

Fig. 7.2: can you show the zero-bandgap state on the figure?

describe how strain can be realized in practical applications

"a solar cell based on direct bandgap Si with high adsorption coefficient" - "absorption coefficient"

"3D shear strains can also give rise to direct bandgap in diamond where CBM is at the point" - "... where CBM is at the Gamma point."

"The following Section 7.2 is dedicated to deeper analysis" -> "Section 7.2 is dedicated to deeper analysis"

"Figure 7.6 requires 8,000,000 (200³)" - the power in 200³ is not formatted correctly

Figure 7.5 should be better explained; what do the white circles denote? what do the numbers mean? I do not see a "red web"

Figure 7.9: explain the meaning of legends (a, b, c, ...)

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-to-direct transition"

"were made in co-authorship with the Zhe Shi from Massachusetts Institute of Technology" - no articles before a name (Zhe Shi)

"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)

Provisional Recommendation

I recommend that the candidate should defend the thesis by means of a formal thesis defense

I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report

The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense