Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Konstantin Gubaev

PhD Program: Materials Science and Engineering

Title of Thesis:

Machine-learning interatomic potentials for multicomponent alloys

Supervisor: Prof. Alexander Shapeev

Date of Thesis Defense: 02 October 2019

Name of the Reviewer:

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Signature:** **Date: 30-08-2019**

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

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

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

The thesis is written with good english. It is structured properly to understand the work done by Konstantin Gubaev. After reading the thesis I got an understanding that the work was focused on method development and approach extension in the field of machine learning in materials science. The topic is in the international research front and therefore it is highly suitable for PhD studies. Konstantin Gubaev has extended the approach of machine learning interatomic potentials (MLIP) for multicomponent systems. He suggest an active learning algorithm to calculate molecular properties and to predict phase diagram of alloys. As it is summarized in section 2.5 and 6 of the thesis, simulations of realistic sized systems (millions of atoms) with quantum mechanical accuracy have strong impact on developing next generation electronic, industrial products or drugs for medical applications. The included work is published in two papers in highly rank international materials science journals. The obtained results have high relevance for research community which is underlined by the number of citations. Paper 1 is published in year 2018 and has already 14 citations while paper 2 published online in late 2018 count 8 citations.

The thesis starts with a basic (and gentle) introduction to the methodology, including density functional theory and the machine learning approach how to train interatomic potentials. After this part a section is devoted to the detailed description of the developed multicomponent moment tensor potential (sections 3.4 and 3.5) and active learning approach (sections 3.6). The application of the method, which is the published two papers is summarized in section 4.

In general, I am pleased to read the thesis and suggest public defense after the candidate answers my questions, comments listed below. The two most relevant points are that I am missing a state-of-the-art kind introduction of other techniques based on applications - see Q1) below. The other one is the choice of the presentation order of the results, papers – see Q5).

Q1) I suggest to elaborate/extend the last paragraph of section 2.4 as a state-of-the-art and put the MLIP approach into a context of other methods in crystal structure prediction with application and limitations.

Q2) In section 3.1 you introduce the "non-conventional" terminology of (un)labeled dataset. You use training errors, validation error, holdout set and validation test (pages 23, 29). However, you also use the "standard" out-of-sample errors term on page 46, for example. It generates a certain type of confusion. I suggest to introduce and use only one version.

Q3-a) On page 33 you write that validation can be done with predicting "high-level" quantities. Fitting the model with quantum mechanical total energies and then predict derivatives of energy, such as elastic constants, phonons can be inaccurate. Is there any argument/motivation to underline your statement?

Q3-b) On page 46 you write that Ref.[40] gives a method in active learning to reduce the out-of-sample (validation) errors. Would not it be appropriate to describe this method in section 3.2 together with sampling?

Q3-c) There is also a typo on page 33 "... is running and MD and checking ...".

Q4) The moment tensor descriptor works in the combined radial and angular motion of freedoms - nicely shown in Figure 7. Is it possible the extend the ordinary 3D space with spins space and define a descriptor for magnetic materials.

Q5) Paper 1 introduces a generalization of the original single component MTP approach to multicomponent systems. This extension is used in paper 2. Confusingly, the papers are

presented in opposite order - section 4.1 and 4.2. Why?

Q6) Paper 1 discuss the question of locality and presents a figure, Fig.4. Could you explain the meaning of non-locality, because both Figure 21 and the equation (29) in the thesis suggest rather the incompleteness of the models MTM1 and MTM2 ,instead of non-locality. If both MTM1 and MTM2 were complete then they would give the same result, or?

Q7) The MaxVol concept as a measure of extrapolation is introduced in Ref.[60]. Fig.3 in Ref. [60] underlines its usage for bulk Li. In this thesis you extend the approach of Ref.[60] to multicomponent materials. Can one use Fig.3 with force error to validate the MaxVol approach for multicomponent systems, why?

Q8) In your paper 2 you extend the active learning approach for searching new alloys. You mention on page 55 that "In this work we extend the approach [61] ...". In Ref.[61] the authors use an evolutionary algorithm while you use randomly generated configurations. Do you also suggest to use your algorithm (paper 2 section 2.3 or 3.6.3 in the thesis) using random configurations instead of the evolutionary method? You include configurations with 12 or less atoms. What has motivated this choice? You utilized the algorithm from Ref.[35]. Why do not you explain this algorithm in the introduction part of your thesis (section 3.1), for completeness? In general, selection of configuration (active, random) is an important subject of the thesis, it might deserve more state-of-the-art description – see Q1).

Q9) Noisy data in section 3.1 would mean inaccurate quantum mechanical calculations (DFT). Can you explain how to chose the DFT calculational parameters for the best "accurate prediction"/"computational demand" ratio? Do you have any strategy? Can you consider to add some description (a table of parameters) to the thesis?

Q10) A. van der Walle has developed a tensorial cluster expansion technique [Nat. Mater. 7, 455 (2008)]. What is your opinion, is it possible to generalized the moment tensor descriptors to anisotropic quantities? See my question 4) about magnetic materials.

Q11) The thesis refers to the relaxation of the atomic configurations using the available MTP model. What type of relaxation approach has been used in the two projects? Section 5 refers to LAMMPS implementation. Please clarify this in the thesis.

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*