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: Sergey V. Levchenko

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Signature:**  **Date: 28-08-2019**

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 by Konstantin Gubaev on the topic “Machine-learning interatomic potentials for multicomponent alloys” is of very high quality. The topic is timely and important from both scientific and application point of view. The topic exactly reflects the content of the thesis. The thesis is clearly written in very good English, and is clearly structured. It includes seven sections. First, a comprehensive review of the existing approaches in the field of interatomic potentials is presented. The review covers most used approaches, from empirical force fields to fully quantum-mechanical methods and machine learning. For the latter, all existing state-of-the-art methods are mentioned.

The general methodology of machine-learning potentials is discussed in a separate section. All possible caveats and challenges are discussed in great detail and illustrated by insightful figures, which is very important for further development of the methodology. The main methodologies of the thesis, moment tensor potentials and active learning, are described very clearly, which demonstrates deep understanding of the subject by the candidate. The personal contribution of the candidate to the development of the methodologies is also clearly explained, and is significant. Since the methods development is an important part of the thesis, the methods are obviously relevant for the work.

In the next section, applications of the newly developed approaches to important problems in materials science are presented. The performance of the methodology is impressively demonstrated by a prediction of several previously unknown ternary metal alloys. This is a significant achievement considering how much effort has been put into the search for new alloy materials of the considered compositions for various applications worldwide. This also clearly shows the relevance of the obtained results to practical applications, including heterogeneous catalysis. The numeric results and achieved efficiency also demonstrate the advantage of the developed methodology over the state-of-the-art methods in the field.

The details on the implementation of the methods and on how to use the program are clearly written in the next section. Then, a potential impact of the work on materials science is discussed, in particular new possibilities for multiscale modelling and materials design. The discussion is objective and demonstrates nicely the importance of the work. The suggested future developments are also very interesting and important for the field in general.

The two publications listed in the thesis are published in international peer-reviewed journals Computational Materials Science (impact factor 2.6) and The Journal of Chemical Physics (impact factor 3.0). In both publications Konstantin is the first author. The high quality of publications and the importance of the work are demonstrated by a very impressive citation statistics: this year’s publication is cited already 15 times, while the last-year’s one is cited 22 times (according to google scholar).

The summary of remaining minor issues is given below:

General comments:

1) Please clarify how magnetism was treated in Co-Nb-V and Al-Ni-Ti alloys.

2) Please report distribution of errors for your application examples, including maximum errors, not only mean errors.

3) Mention tight-binding and other quantum-mechanical force fields.

4) There is no clear motivation for developing "yet another" approach (MTP) machine-learning potential. Please motivate clearly why your approach is better than the approaches currently developed by other groups.

5) Please motivate better why one needs to study the considered alloy systems.

Minor comments:

Eqs 10-11: there is an "a" on top of H; Eq. 11 index k should be j

Eq. 11 and others: discuss spin

Eq. 12 is messed up

"There are approaches which search the solution of (11) in a form of a linear combination of atomic orbitals (Hartree-Fock, post-Hartree-Fock methods), which provide great accuracy but scale as O(N^4) with the number of atoms, which limits their applicability to the systems containing not more than a few dozens of atoms." - There are linear-scaling methods, also even without additional approximations thousands of atoms can be calculated nowadays with HF.

First equation in section 2.3: n(r) should be in the square brackets instead of (r)

Check units in front of the Hartree term in all equations.

"sufficient for qualitative analysis only" -> "sufficient for qualitative analysis only at best"

"Suppose there is a large number, N, of configuration whose structure " -> "...configurationS..."

"The model has a number of free parameters ... of total number m" -> "The model has a number m of free parameters ... "

"it can result it bad convergence" -> "it can result in bad convergence"

"continuous values ranging from -inf to +inf" - insert proper infinity notation

"However, in practice it is unsolvable in principle" - clarify that for some choices of F(\theta) it is solvable (namely linear; maybe some others?)

"extend the opportunities of MD/MC approaches" - "MC" was not defined in the text

"they account interaction of each atom" -> "...account for interaction..."

"where the contributions of the long range Coulomb forces is essential" - "...are essential"

"While simulation the atoms will displace" - "During simulation..."

"not to include correlated configurations" -> "in order to exclude correlated configurations"

"In this sampling approach it is considered that even with some simple (therefore computationally fast) interatomic interaction model" - I guess you mean "In this sampling approach it is assumed that even with some simple (therefore computationally fast) interatomic interaction model"

"Their typical values are: Ce = 1; Cf = 10-2 A2; Cs = 10-3 A6 ..." - explain where these typical values come from

Figure 6: it is unclear what t's are

"The symbol x stands for the outer product of vectors, thus in (25) rij x ... x rij is the tensor of rank nu" - bold "r"?

"As follows from, (26) a number of parameters" - redundant comma

"The proof [76] holds for a single-component case, but can be easily extended to a multicomponent case." - this statement is a bit in the air; if this was not published, some proof should be given in the thesis

"After the calculation of ab initio energies, forces, and stresses of the selected configurations are added to the training set." -> "After the calculation, ab initio energies, forces, and stresses for the selected configurations are added to the training set."

Seems that Figure 9 is not referenced in the text

Figures are referenced sometimes as "Fig. xxx", and sometimes as "Figure xxx", it is better to consistently use only one of these ways

"Even despite its favorable accuracy/efficiency trade-off" - part of the sentence occupies the page margin; this could be avoided by adding spaces in "accuracy / efficiency"

"Surrogate models such as the cluster expansion and standard machine learning approaches do not have the broad applicability and exceptional accuracy of the moment tensor potentials-based [76] approach we demonstrate here." - can you prove this?

"Though such a shallow ground state is typically not significant beyond academic interest" - clarify your criterion for shallow/non-shallow ground state

"Cu-rich ground states are believed to have an effect on the experimental" - I think it would be better to write "Cu-rich phases are believed..."

Figure 11: why not show the two panels on one plot?

"large amount of molecules" -> "large number of molecules" (amount does not fit here)

"The other problem we address with our algorithm is the issue of the socalled outliers. ..." - The logic of this paragraph seems strange. You say that other approaches cannot describe atypical structures, but your approach can do it AFTER you add these structures to the training set. But this may be also true for the other approaches. Besides, are the outliers remain outliers after you add them? I guess you wanted to say that you can IDENTIFY a structure as an outlier before you actually calculate it?

"expose our machine-learning model" - "expose" seems the wrong word here

Figure 18: I suggest to add to the caption a brief explanation of active 1 and active 2

"We argue that the latter can be useful in those applications where the region of interest in the chemical space is fixed a priori." - This is somewhat confusing. Will active learning be useless when the chemical space is NOT fixed a priori?

"and these expenses scale cubically with the number of atoms" - not necessarily, there are linear-scaling implementations

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*