Ab initio Materials Genomics: High-Throughput Study of 50'000 Single-Phase Materials

Evgeny Blokhin,[a,b] Pierre Villars[c] and Roberto Dovesi[d]

[a]*Tilde Materials Informatics, Straßmannstraße 25, 10249, Berlin – GERMANY*

[b]*Materials Platform for Data Science, Sepapaja 6, 15551, Tallinn – ESTONIA*

[c]*Material Phases Data System, Unterschwanden 6, 6354, Vitznau – SWITZERLAND*

[d]*Universita di Torino, via Giuria 5, 10125, Turin – ITALY*

e-mail: eb@tilde.pro

Analogously to the Human Genome Project, the today's efforts in materials informatics tackle the so called Materials Genome. The artificial intelligence techniques, such as deep learning and logic reasoning, had already enabled the powerful innovations like voice assistants, self-driving cars, face recognition outperforming humans *etc.* The key feature is a bio-inspired algorithm, able to learn and reason on a very complex subject (*e.g.* solid state physics). The aim of this work is to apply such techniques for materials data, to develop the machine-learning approach for the design of the new materials, and to search for the new patterns and hidden dependencies. However the main requirement is a giant amount of training data, which must be of very high quality.

The CRYSTAL *ab initio* engine is proposed for the data generation. With the fine-tuned basis sets, it provides higher quality of predictions in shorter terms, than the other *ab initio* engines.[1] The first phase of this project is the high-throughput study of 50'000 single-phase materials, planned to take 1.5 years. The starting point for the automated simulations is the experimental materials database PAULING FILE, in its online implementation called Materials Platform for Data Science.[2] The data were collected manually from the hundreds of thousands of publications in materials science (1891—2017) and critically evaluated by an international team of editors. Now this database contains the machine-readable data for more than 50'000 distinct phases of materials. These data back up such commercial products as SPRINGER Materials® and MedeA Materials Design.® Although all the results of this project, including raw data, as well as the created software tools, will be published as the open-access online interactive encyclopedia for educational purposes, allowing reproduction and enhancement.

1. R. Evarestov, E. Blokhin, D. Gryaznov, E. Kotomin, J. Maier, *Phys. Rev. B*. **2011**, 83, 134108

2. E. Blokhin, P. Villars, Handbook of Materials Modeling, 2nd ed. by S. Yip, **2018**, In Section “Materials Informatics”