Research Networking Programmes

Science Meeting – Scientific Report

The scientific report (WORD or PDF file - maximum of seven A4 pages) should be submitted online within two months of the event. It will be published on the ESF website.

***Proposal Title*: White nights of materials science: From physics and chemistry to data analysis, and back**

Application Reference N°: 5247

1) Summary (up to one page)

In recent years, the rapid development of computer hardware and electronic-structure software has made first-principles materials design not only feasible, but also a powerful and cost-effective alternative to experimental screening approaches. It has also opened up new possibilities and with them new challenges, some of which were addressed at this conference.

On the first-principles modelling side, the challenges are the identification and production of the key common and application-specific data, computational-data reliability and accuracy, and the development of multiscale approaches with control of error propagation across different scales. On the data-analysis side, the challenge is the development of a comprehensive materials data structure and analysis tools that are capable of handling large amounts of data. Usually, the specifics of a particular area of computational research, and data storage and analysis issues are considered as separate. Computational materials scientists usually assume that their data is not of the quantity that merits the attention from a database expert, or detailed analysis, and indeed rarely consider data that are not central to a specific project. Database specialists often abstract data from its origin. As a result, there are still only few scientists who are experts in both advanced electronic-structure theory and database structure/analysis, while communication between experts in these different fields is often difficult because

they “speak different languages". Overcoming these barriers has the potential to change how research is done worldwide, and was one of the incentives of this conference.

It was discussed during the conference how computational data that are often discarded (for example, method validation studies, wave functions, information on the potential-energy surface obtained during atomic relaxation, molecular dynamics, or a genetic-algorithm run) could be stored and used to determine hidden correlations and find descriptors that facilitate the exploration of a vast compositional space across multiple time and length scales. The suitable data structures were suggested and thoroughly discussed. Consensus was achieved on the need of having hierarchical data structures in which scalar and higher dimensional objects are stored efficiently (*e. g*., by means of the hdf5 standard), with the use of *dictionaries* which allow for trans-code comparisons. Novel data analysis tools were proposed that could reveal what data are still missing, providing a rational motivation for new computational studies. In particular, the tools are based on the so-called NoSQL architectures, *i. e*., beyond the tabular relations of the traditional relational data-bases. The following problems with present approaches to solving the above challenges were discussed: (i) Typically only a limited amount of data is stored; thereby valuable information is lost, which greatly hinders the discovery and understanding of correlations between properties of different materials. (ii) The existing databases do not allow for complex queries of the data, and are not designed to store very large amounts of data. (iii) The first important steps towards data analysis in the materials sector are so far limited to only a few examples and these initiatives are decoupled from large databases. The constructive extraction of hidden information from the enormous amount of data that can be generated is an essential aspect that has so far been largely ignored.

The conference brought together researchers who are experts in electronic-structure theory or database structure/analysis, and who are interested in both fields. We created an environment where communication between people with different expertise and backgrounds was strongly encouraged, which helped to understand and overcome barriers on all sides. Advances in electronic-structure method development, multiscale modelling, materials data analysis, and machine learning were discussed in close connection with technologically important applications, such as heterogeneous catalysis, thermoelectric materials, optoelectronics, hybrid organic-inorganic interfaces, and others. The goals of the workshop (to identify both common and application-specific features of a comprehensive materials data structure, specific areas where data analysis would be especially helpful, and the availability or absence ("white spots") of data) were successfully reached due to the very high level and relevance of the expertise of the invited and contributing speakers.

2) Description of the scientific content of and discussions at the event (up to four pages)

The conference was opened by two researches who are leading experts in materials database analysis, and who have demonstrated successful applications of their analysis methods to material’s design. One of the presented applications was designing new polymer materials for capacitor-based energy storage. The goal was to design materials with higher dielectric constant and larger band gap (as a measure of the intrinsic breakdown voltage). The study was a combination of high-through density-functional theory calculations and experimental synthesis and testing of the most promising candidates. Due to the nature of the material class (organic polymers), the presenter emphasized the importance of multiscale description, in particular 3D-morphology

prediction. A series of new organo-metallic materials was predicted theoretically and tested experimentally, and was found to surpass the state-of-the-art materials (biaxially-oriented polypropylenes) in terms of the higher dielectric constant. The discussion concentrated mainly on the possible role of defects. Also, it was pointed out that finding physically-meaningful descriptors would allow for a more efficient exploration of the vast configurational space.

The second speaker presented his novel approach to data analysis, and demonstrated their performance by applying to the design of new multiferroic materials and establishing reversible electronic-structure/geometric-structure relationships. The main focus of his approach is the dimensionality reduction and validation of the corresponding methods. The strategy is to combine all available relevant methods of data analysis into a general procedure that is self-validated by construction. The presented argued that the standard view of the dimensionality reduction as a functional relationship needs to be replaced by a set of design rules presented by a graph. The nodes and connections in the graph are established by machine learning methods. The discussion focused mainly on the issues of physical meaning and causality of the developed design rules. Also, the necessity for developing a common language has been once again clearly emphasised.

The next contribution discussed a combination of quantum-mechanical (QM) and machine-learning (ML) models. A detailed analysis of machine-learning-based interpolation of quantum-mechanical reference data was presented. Both successes and challenges were discussed. The successful applications include learning the electron kinetic energy from the electron density, learning molecular properties from the atomic structure, and learning the error between different levels of theory. The following challenges were discussed: compound representation for QM/ML models, learning atoms-in-compounds and non-scalar properties (such as forces and multipole moments). The debate after the presentation mainly focused on whether physically-sounding models should be preferred to “general” ML regression models such as Gaussian kernel-ridge regression. Even though in general the participants agreed that “dedicated” statistical models should be developed for addressing the physical problems, at this point it remained open how to possibly follow in practice this route, except for the somewhat simple requirement of imposing physical symmetries to the learning model, rather than expecting a model to find such regularities.

In the next series of talks, applications of *ab initio* methods (in particular density-functional theory) to real materials and conditions were discussed. A method for predicting the thermal conductivity of materials from first principles based on *ab initio* molecular dynamics and Green-Kubo relations was presented. The advantages of the method are the absence of the empirical parameters and the applicability to realistic system (unit-cell) sizes. The performance of the method was demonstrated by application to Y-stabilized zirconia. A common line across the talks in the series was establishing structure-property correlations for technologically important materials classes, which emphasised the complexity of the problem and the importance of creating a database that can reflect this complexity. In particular, defects in binary clathrates, properties of skutterudites, and surface energies of GaN were discussed. A first-principles methodology for discovering structures with desirable properties, and its application to TiO_2 clusters with low ionization potential or high electron affinity, was presented. The methodology is based on a genetic algorithm. Two studies directly demonstrated how atomic properties of a material (atomic and electronic structure of adsorbed molecules or defects at a surface) can be affected by properties of materials at larger scales, such as charge-carrier conductivity and macroscopic polarizability. An important point raised by the presentations was the sensitivity of the theoretical

predictions to the level of theory employed, which emphasized once again the importance of developing data validation strategies.

Validation strategies for *ab initio* methods were discussed in the next series of talks. The discussion focused on identifying explicit and implicit approximations in density-functional theory (DFT) methods, and developing more accurate approaches with minimum penalty of increasing the computational cost. The approximations were identified and improved by comparing approximate DFT with many-body perturbation theory. In particular, a set of exchange-correlation kernels for the time-dependent DFT were evaluated based on *GW* and Bethe-Salpeter equations. Also, an all-electron exact-exchange implementation using a basis set that is close to completeness was presented. Long-range first-principles van der Waals interaction correction to density functionals including a collective medium response was reported. Originally formulated as a post-processing correction, it was then extended to allow for self-consistency of the Kohn-Sham orbitals, thus permitting a validation of the one-shot approximation. Two presenters in the same series of talks reported efficient all-electron implementations of the stress tensor and density-functional perturbation theory. Both implementations can employ hybrid functionals, which allows for validation of the standard LDA and GGA functionals for modeling pressure-dependence of the electronic structure, and lattice dynamics and electron-phonon coupling.

The discussion of the challenges of *ab initio* modelling of real materials and processes was continued by a series of three talks. The first one emphasised the importance and power of symmetry for predicting structures and electronic properties of transition-metal-oxide nanotubes. The second described the behaviour of water under extreme conditions, *i. e*., at large undercooling, where a debated high density liquid-low density liquid (first order) phase transition is expected to take place, and at high pressures. The challenges of modelling matter under such conditions were discussed. In particular, the theoretical approach is computationally feasible only through semi-empirical force-fields, and the obvious limits and intriguing advantages (*i. e*., possibility to tune the force-fields for a sensitivity analysis, in order to understand what contributions are actually important) where discussed after the talk. The last talk in this series addressed the true multiscale nature of the active material phase in heterogeneous catalysis. A method allowing for a first-principles modelling of mass and heat transfer at the macroscale was presented, based on a combination of the first-principles kinetic Monte Carlo simulations and computational fluid dynamics. Relationships “input material-resulting material as a function of environmental parameters” and “input material-material function” were discussed.

The next series of talks described some of the current and planned materials databases. The first talk presented the “Automated Interactive Infrastructure and Database for Atomistic simulations”, which is in its beta version. This infrastructure implements several useful ideas for the management of large amount of data in materials science, such as the full record of the data provenance and the possibility of automatizing workflows (*i. e*., “completing” existing calculations with missing pieces of information or with similar materials by constructing the suitable input files and launch the calculation).
In the next talk, an under-construction materials database was presented, where the focus is on the “normalization” of the data, *i. e*., the creation of *dictionaries* that allow for comparing data obtained with different codes.

3) Assessment of the results and impact of the event on the future directions of the field (up to two pages)

The results of the conference clearly demonstrate the necessity to further develop a common “language” for the experts in database analysis and materials science. Moreover, the data structure itself requires an intensive dialogue of (until now rare) specialists in materials databases, with the active participation of (other) physicists and chemists. Although important successes in the development of materials databases and database analysis tools were reported (*e. g*., discovery of new polymer materials with favourable dielectric properties, design of novel multiferroic materials, development of new electronic structure methods using machine learning), also numerous challenges were identified and discussed (*e. g*., establishing the causality of descriptors and extracting the physical meaning from the machine-learning models, the hidden complexity of widely used concepts such as “material”, “descriptor”, and “validation”, control over approximations and error propagation, multiscale nature of a “material”, database flexibility and accessibility).

One of the important results of the conference is the realization of the fact that comprehensive materials data analysis tools cannot be developed without taking into account the specifics and complexity of the data. In particular, the issues of data validation and the causality of the correlations derived based on this data are much more important in materials science than in any other area where machine-learning methods have been successful so far. These issues are a direct consequence of the complexity of both obtaining and structuring the materials data, leading to the relative scarcity of the reliable data. In particular, this complexity reveals itself in the multitude of factors that can affect materials properties across orders of magnitude of time- and length scales. For example, concentration and distribution of point (atomic scale) defects at an oxide surface can be strongly affected by the dopants that are hundreds of nanometers away from the surface; or surface structure, stoichiometry, and chemistry at the atomic level can be determined by the limitations of mass and heat transport at macroscopic scales. In view of the data complexity, the existing data analysis tools require further development.

Another important result of the conference is the identification of the classes of materials and applications that would greatly benefit from a comprehensive data storage and analysis. The following materials and applications were discussed: thermoelectric materials, inorganic-organic interfaces, thermal-barrier coatings, oxide clusters, oxide semiconductors, semiconductor surfaces, photo-active materials, heterogeneous catalysts, matter under extreme conditions, biomolecules, polymers for energy storage, and ferroelectric and multiferroic materials.

In summary, the conference succeeded in initiating and promoting the dialogue between world-leading experts in the fields of materials databases, data analysis, and materials science. The discussions during the conference emphasized the importance of maintaining the dialogue, also in the future. Both successes and challenges of statistical learning in materials science were discussed. The complexity of materials data was demonstrated by many examples of practical importance. This complexity prompts the development of a flexible data structure and novel data analysis tools, including validation of data and analysis of the causality (physical relevance) of the found correlations (descriptors). Thus, the conference prepared the basis for developing a comprehensive materials data structure, suitable for an efficient design of novel functional materials. This provides an opportunity to create a worldwide materials database that will eventually absorb data and experience of all existing materials databases.

4) Annexes 4a) and 4b): Programme of the meeting and full list of speakers and participants

Annex 4a: Programme of the meeting

Monday, June 16, 2014 (near Assembly Hall)		
18:30 - 20:00	Dinner (Pushkin Hall)	
20:00 - 20:10	Introduction	
Session "Material science at the time of Big Data: Successes and challenges"		
20:10 - 20:50	Ramamurthy Ramprasad	"Materials Discovery in the Midst of the Materials Genome Initiative: The Example of Polymer Dielectrics Design"
20:50 - 21:30	Krishna Rajan	"Mapping the Topology of Big Data: Discovering Pathways for Materials Discovery"
Tuesday, June 17, 2014 (near Assembly Hall)		
7:45 - 8:45	Breakfast (Pushkin Hall)	
8:15 - 8:45	Registration	
8:45 - 9:15	Poster parade	
Session "Material science at the time of Big Data: Successes and challenges" (continued)		
9:15 - 9:55	Matthias Rupp	"Quantum Mechanics / Machine Learning Models. Recent Successes and Challenges"
Session "*Ab initio* modelling of real materials and processes" (part I)		
9:55 - 10:20	Christian Carbogno	"High Temperature Thermal Conductivity from First Principles"
10:20 - 10:45	Amrita Bhattacharya	"Role of Vacancies in Binary Clathrates"
10:45 - 11:05	Coffee break	

11:05 - 11:30	Susmita Basak	"First-Principles Investigation of Skutterudites"
11:30 - 11:55	Saswata Bhattacharya	"Unraveling Structure-Property Relationships: Property-Based*Ab Initio* Cascade Genetic Algorithm Applied to $(TiO_2)_n$ Clusters"
11:55 - 12:20	Oliver Hofmann	"Interface Dipole Formation and Charge (De)localization at Inorganic/Organic Interfaces"
12:20 - 13:50	Lunch (Pushkin Hall)	
13:50 - 14:15	Hong Li	"Surface Energies and Equilibrium Crystal Shapes in GaN"
14:15 - 14:40	Sergey Levchenko	"Defect-Defect and Adsorbate-Adsorbate Interactions at Realistic Conditions"
14:40 - 15:00	Break	
15:00 - 22:00	Conference outing and dinner (gather at 15:00 at the entrance to the "new hotel")	
Wednesday, June 18, 2014 (Theater Hall)		
8:00 - 9:00	Breakfast (Pushkin Hall)	
Session "Achieving greater accuracy: Advanced electronic structure methods"		
9:00 - 9:25	Claudia Draxl	"White Paper of Theoretical Spectroscopy: From DFT to Many-Body Perturbation Theory and Back"
9:25 - 9:50	Angel Rubio	"Hybrid-Organic Photovoltaic Devices from First-Principles Simulations"
9:50 - 10:15	Heiko Appel	"Photon Energy Density and Correlated Photon-Electron Wavefunctions in Quantum Electrodynamics"
10:15 - 10:40	Johannes Flick	"Optimized Effective Potential Approach for Correlated Light-Matter Interactions in Cavity QEDFT"
10:40 - 11:00	Coffee break	

11:00 - 11:25	Victor Ruiz Lopez	"The Role of the Collective Substrate Response in the van der Waals Interactions of Organic/Inorganic Systems"
11:25 - 11:50	Caterina Cocchi	"Optical Spectra from Molecules to Solids: Insights from Many-Body Perturbation Theory"
11:50 - 12:15	Santiago Rigamonti	"TDDFT for Extended Systems: Evaluation of Exchange-Correlation Kernels"
12:15 - 12:40	Björn Bieniek	"Ultra-thin ZnO on Metal Substrates"
12:40 - 14:10	Lunch (Pushkin Hall)	
14:10 - 14:35	Franz Knuth	"Pressure-Dependent Electronic Structure from First Principles"
14:35 - 15:00	Honghui Shang	"Density-Functional Perturbation Theory for Lattice Dynamics with Numeric Atom-Centered Orbitals"
15:00 - 15:25	Ute Werner	“Hybdrid Functionals and Exact Exchange: An All-Electron Implementation”
15:25 - 15:50	Nicola Ferri	"Electronic Properties of Surfaces and Interfaces with Self-Consistent van der Waals Density Functional"
15:50 - 16:15	Igor Ying Zhang	"Toward Simple Orbital-Dependent Density Functionals for Molecular Dissociation"
16:15 - 16:35	Coffee break	
Session "*Ab initio* modelling of real materials and processes" (part II)		
16:35 - 17:15	Robert Evarestov	"Symmetry and *Ab initio* Modeling of Nanotubes Based on Binary and Ternary Transition Metal Oxides"
17:15 - 17:55	Roberto Car	"Water at Extreme Conditions: From Deep Undercooling to Ultrahigh Pressure"
17:55 - 18:20	Karsten Reuter	"Identification of the Active Phase of Heterogeneous Catalysts through *In Situ* Reaction Product Imaging"
18:20 - 21:00	Dinner (Pushkin Hall)	

22:45 - 2:00	City tour (gather at 22:45 at the entrance to the "new hotel")	
Thursday, June 19, 2014 (Theater Hall)		
9:00 - 11:00	Breakfast (Pushkin Hall)	
11:00 - 11:25	Hanna Krauter	"How to Apply for an ERC Starting Grant"
Session "Managing large amounts of data: Structure and sharing"		
11:25 - 12:05	Nicola Marzari	"Phase Stability of Ferroelectric Perovskites: A Case Study in Data and Workflow Management"
12:05 - 12:30	Pasquale Pavone	"The exciting way towards novel materials discovery"
12:30 - 14:00	Lunch (Pushkin Hall)	
14:00 - 14:35	Fawzi Mohamed and Evgeny Blokhin	"Towards Database-driven Novel Materials Discovery"
14:35 - 15:15	Georgy Zegrya (Ioffe institute)	"Electronic archive: *New Semiconductor Materials. Characteristics and Properties*"
15:15 - 19:00	Visit to the Ioffe Institute	
19:00 - 21:00	Dinner (Pushkin Hall)	
Friday, June 20, 2014 (near Assembly Hall)		
8:00 - 9:00	Breakfast (Pushkin Hall)	
Session "(Bio)molecules in motion: Statistical mechanics"		
9:00 - 9:25	Adriana Supady	"Sampling the Conformational Space for Accurate Property Predictions"

9:25 - 9:50	Mariana Rossi	"Structure and Dynamics of Biomolecules and their Interactions: Challenges for First-Principles Methods and Nuclear Quantum Effects"
Session "Towards discovery of new materials: The quest for the optimal descriptors"		
9:50 - 10:15	Jan Vybiral	"Sparsity and Kernel Methods in Machine Learning"
10:15 - 10:40	Luca Ghiringhelli	"Big Data of Materials Science -- Critical Role of the Descriptor"
10:40 - 10:45	Concluding remarks	
10:45	Conference ends	

Posters

1	Wael Chibani	"Self-Consistent Dynamical Embedding in Real Space"
2	Tanja Dimitrov	"Exact Hohenberg-Kohn-Functional for a Lattice Model"
3	Arvid Ihrig	"RI-LVL: Efficient and Accurate Four-Center Integral Evaluation"
4	Nora Salas-Illanes	"Electronic Structure of Hybrid Materials by Means of Self-Consistent *GW*"
5	Rene Jestädt	"Real-Time Evolution of Coupled Maxwell-Schrödinger Systems
6	Sebastian Kokott	"Towards Understanding Charge-Carrier Conductivity in Oxides: Intrinsic and Extrinsic Point Defects in MgO"
7	Archana Manoharan	"*Ab Initio* Studies of X-Ray Absorption Spectra of Kesterite Materials"
8	Mateusz Marianski	"Dissecting Structures and Interactions in Polypeptide Chains"
9	Aliaksei Mazheika	"Theoretical Study of CO_2 Methanation on Ni/MgO and Ru/Ni/MgO solid Solutions"
10	Pablo García	"High-Performance Electronic Structure Calculations: Optimization of the Evaluation of

	Risueño	the Hartree Potential"
11	Markus Schneider	"Histidine-Cation Interaction and Microsolvation from First Principles"
12	Xunhua Zhao	"Sub-Monolayer Water Adsorption on Alkaline Earth Metal Oxide Surfaces: A First-Principles Cascade Genetic Algorithm Study"

Annex 4b: Full list of speakers and participants

Heiko	Appel	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Carsten	Baldauf	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Susmita	Basak	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Saswata	Bhattacharya	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Amrita	Bhattacharya	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Bjoern	Bieniek	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Evgeny	Blokhin	Humboldt-Universität zu Berlin
Dirk	Broßke	Humboldt-Universität zu Berlin
Roberto	Car	Princeton University
Christian	Carbogno	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Wael	Chibani	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Caterina	Cocchi	Humboldt-Universität zu Berlin
Tanja	Dimitrov	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Claudia	Draxl	HU Berlin
Robert	Evarestov	Saint Petersburg State University
Anne	Eyb-Reuter	Technische Universität München
Nicola	Ferri	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Johannes	Flick	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Pablo	Garcia Risueno	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Anna Rita	Gentile	Humboldt-Universität zu Berlin
Luca	Ghiringhelli	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Vivekanand	Gobre	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Karsten	Hannewald	Humboldt-Universität zu Berlin
Oliver	Hofmann	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Arvid Conrad	Ihrig	Fritz Haber Institut der Max Planck Gesellschaft
René	Jestädt	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Franz	Kntuh	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Sebastian	Kokott	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Hanna	Krauter	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Gitta	Kutyniok	Technische Universität Berlin
Birgit	La Monaco	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Sergey	Levchenko	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Hong	Li	Humboldt-Universität zu Berlin
Archana	Manoharan	Humboldt-Universität zu Berlin
Mateusz	Marianski	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Nicola	Marzari	École polytechnique fédérale de Lausanne
Aliaksei	Mazheika	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Fawzi	Mohamed	Qt-Digia
Julia	Pach	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Pasquale	Pavone	Humboldt Universität zu Berlin
Krishna	Rajan	Iowa State University
Ramamurthy	Ramprasad	University of Connecticut
Karsten	Reuter	Technische Universität München
Santiago	Rigamonti	Humboldt-Universität zu Berlin
Mariana	Rossi	University of Oxford

Angel	Rubio	European Theoretical Spectroscopy Facility
Victor G.	Ruiz Lopez	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Matthias	Rupp	University of Basel
Nora	Salas Illanes	Humboldt-Universität zu Berlin
Matthias	Scheffler	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Markus	Schneider	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Honghui	Shang	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Adriana	Supady	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Jan	Vybiral	Technische Universität Berlin
Ute	Werner	Humboldt-Universität zu Berlin
Igor Ying	Zhang	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Xunhua	Zhao	Fritz-Haber-Institut der Max-Planck-Gesellschaft
Du	Zheng	Humboldt-Universität zu Berlin