Computational Materials Science 44 (2009) 1042–1049

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Data mining and accelerated electronic structure theory as a tool in the search for new functional materials

C. Ortiz [1], O. Eriksson[2], M. Klintenberg [*,3]

Department of Physics and Materials Science, Uppsala University, Box 530, SE-751 21 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 29 April 2008
Received in revised form 2 July 2008
Accepted 15 July 2008
Available online 2 September 2008

PACS:
29.40.Wk
71.20.Mq
71.15.−m

Keywords:
Electronic structure theory
Data mining
Scintillating materials

ABSTRACT

A highly accelerated electronic structure implementation and data mining algorithms have been combined with structural data from the inorganic crystal structure database to generate materials properties for about 22,000 inorganic compounds. It is shown how data mining algorithms employed on the database can identify new functional materials with desired materials properties, resulting in a prediction of 136 novel materials with potential for use as detector materials for ionizing radiation. The methodology behind the automatized *ab initio* approach is presented, results are tabulated and a version of the complete database is made available at the internet web site <http://gurka.fysik.uu.se/ESP/> (Ref. [1]).

Sensors, solar cells, advanced batteries, and magnetic strips in credit cards are examples of functional materials present in every-day life. One important task for the research in materials science is the continuous improvement and discovery of new such advanced materials. *Ab initio* electronic structure calculations as a tool for predicting materials properties have steadily increased in use over the years [2] and play today an important role due to the relatively inexpensive and versatile guidance it offers. There are currently some 8000 studies published annually with this method.

Electronic structure theory applied in materials research is typically done in a fashion where a calculation follows, or accompanies, an experimental result. Knowledge on an atomistic level is thus gained which can help in understanding the experimental results [3]. In some rarer cases the theoretical calculations predict a materials property which subsequently may be realized experimentally. An example of the latter is the newly proposed tetragonally distorted FeCo alloy with exceptional out of plane magnetic anisotropy [5,4]. These materials simulations are done in a one-by-one mode, where one calculation accompanies one experiment. However, with an increasing demand of an accelerated speed in finding or predicting new materials this may not be the most efficient approach. Alternatives to this methodology have indeed been discussed, for instance, numerical algorithms which obey evolutionary principles borrowed from biology have been applied to find structural data of compounds and alloys [6], and in a somewhat similar study where formation probabilities derived from correlations mined for in experimental data were used to guide *ab initio* calculations for unknown structures [7].

In this article a method for automatizing the generation of a new database with electronic structure results for a large number (22,000) of inorganic compounds is presented. The necessary structural information for the *ab initio* calculations is extracted from the inorganic crystal structure database (ICSD) [8]. The electronic structure results are generated within the local density approximation (LDA) of density functional theory (DFT) in combination with a highly accurate full potential linear muffin-tin orbital (FP-LMTO) method [9].

We describe how data mining algorithms can be applied to the database when searching for any particular class of potentially new advanced materials. This may involve semiconductors with tailored band gaps for solar cells and light harvesting, i.e. materials with desired optical properties, and magnetic compounds for energy conversion or magnetocaloric applications. To demonstrate in detail the power of data mining algorithms with automatic electronic structure calculations we also give a full account of the search and identification of 136 novel compounds with potential

* Corresponding author.
E-mail addresses: olle.eriksson@fysik.uu.se (O. Eriksson), Mattias.Klintenberg@fysik.uu.se (M. Klintenberg).
[1] Partly affiliated at Lawrence Berkeley National Laboratory, Berkeley, California during years 2002–2005.
[2] Tel.: +46 18 4713625; fax: +46 18 4713524.
[3] Partly affiliated at LBNL during years 1998–2005.

0927-0256/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.commatsci.2008.07.016

use as radiation detector materials. In addition, we prove a high success rate of the algorithm by an un-biased identification of several known, successful cerium activated materials.

Radiation detection systems are generally used in areas such as biomedical imaging, nuclear security, nuclear non-proliferation and treaty verification, and in industry. The limiting factors for the performance of these systems are found within the detector material, and improvements are desired in properties like energy resolution for isotope identification in nuclear security, temporal resolution in biomedical imaging, or simply effective detection in small sized systems.

Standard simulations involve to a large extent manual work, where input data files must be generated. We have circumvented this time consuming step by a fully automatic method, where all data files are computer generated after certain materials specific criteria. In addition we have designed a software system which, once all necessary input files have been generated, carries out the simulation automatically (optimization of parameters, truncation criteria, decision-making, etc.). It should also be noted that the algorithm employed has to undertake several steps of learning, in order to decide on a proper set of parameters. For instance, it was found that the definition of the base geometry and basis set used in these calculations [9] needed to be updated after several trial calculations, in order to ensure accuracy of the electronic structure and total energy of the material. Hence all steps of the simulations of this work have been made by artificial intelligence and high performance computation.

The algorithm which executes first principles calculations, with general rules for the computational details as described in Appendix 1, have been applied to some 22,000 compounds from the ICSD database, and the results of these calculations are available on the web site of Ref. [1]. There are a number of electronic structure databases available on the web, e.g. [10–13], however all contain at least two orders of magnitude smaller number of entries than published here, and in addition these studies have been done with different focus. The web-based databases all complement each other however. It should be noted that the crystallographic data from the ICSD originates from different experimental settings, with small variations, giving rise to slightly different electronic structure results and that several entries in ICSD must be disregarded because at least one site has non-trivial occupancy. The control files for our calculations are available upon request.

The crystallographic data needed to construct the control files involves information about the cell geometry, bravais lattice, and the coordinates for each atom and space group. This information is available from the ICSD database [8] in, for example, the CIF (crystallographic information file) format. With access to the CIF files the coordinates can then be unfolded and transformed to the minimal bravais lattice[14]. Our approach can make use of any electronic structure method and can be applied to any compilation of structural geometries, even hypothetical ones which have not yet been identified.

For each entry in Ref. [1] the electronic structure results are presented as figures illustrating band structure, density of states (DOS), partial DOS, and charge density contour maps; furthermore, properties like density, total energy, Fermi energy and band gap (if available) are also listed. Note that the density is calculated using the experimental lattice parameters. Optimizing the lattice parameters, i.e. calculating the bulkmodulus, as well as performing spin polarized calculations will be subjects of future work.

We now proceed with a detailed example on how mining algorithms on the electronic structure information in Ref. [1] may be used for identifying novel scintillator materials. The general philosophy of the mining algorithm is to compare specific electronic structure related properties of a larger set of compounds (i.e. the data in Ref. [1]) to a peer group, which is known to have desired properties connected to a certain functionality of the material. We focus here on suitable candidates for nuclear radiation detector materials, and have chosen two sub-groups of materials: (1) cerium activated scintillating materials, and (2) activated semiconductor materials, e.g. Ga doped ZnO (ZnO:Ga) [15]. In identifying principles of data mining for these materials, we consider experimental information regarding characteristic electronic properties for known cerium materials, that show 5d to 4f luminescence, as well as known semiconductor materials which have been found to have encouraging materials properties. The data mining results in 136 candidate materials proposed for further investigation.

A first desirable property for the materials of interest for this study is high detection probability in small sized units, which is associated to the number of available electrons per unit volume. A high density and high atomic number (Z) are therefore desired [16]. Moreover, a short attenuation length is needed and it is also advantageous that photons scatter mainly through the photoelectric channel. These two properties can be characterized with the photoelectric attenuation length (PAL). PAL is the ratio between the calculated attenuation length ($\lambda = \mathrm{FW}/(\rho \cdot [\sigma_{\mathrm{pe}} + \sigma_{\mathrm{C}}])$ of the incoming radiation in the material and the calculated fraction between the photoelectric (σ_{pe}) to Compton scattering (σ_{C}) cross-sections (or rather, the ratio $\sigma_{\mathrm{pe}}/[\sigma_{\mathrm{pe}} + \sigma_{\mathrm{C}}]$) at some energy, e.g. 511 keV, which is the energy scale relevant for positron emission tomography (PET)[16,17]. FW is the formula weight and ρ is the calculated density of the material. The atomic masses and photoelectric- and Compton scattering cross-sections are measured, element specific entities and are listed in Appendix 1. PAL summarizes attenuation length and the efficiency of the photoelectric scattering channel and the lower the PAL value is, the higher is the chance that an incoming γ-ray is absorbed in the material after a short distance by the photoelectric effect, which makes the material more relevant to our study.

We show in Fig. 1 the distribution and cumulative sums of materials densities and the PALs. A high density requirement is imposed as $\rho > 6.5$ g/cm^3, and we find 4602 materials satisfying this criterion. As an upper limit for the PAL, the value of a well-known detector material is used, i.e. Tl doped NaI with $\mathrm{PAL}_{\mathrm{NaI:Tl}} = 17$ cm. This limit is satisfied for about 87% of the materials also satisfying the high density condition and the selection is reduced to 3983 entries.

We are now left with some four thousand compounds which according to the density and PAL criteria might be suitable detector materials. The synthesis and testing of thousands of candidate compounds is clearly not realistic, nor is it feasible with standard, manually controlled computational methods to go through such a large body of potential materials, to try to identify successful compounds. Hence, more efficient methods are needed, where efficient algorithms for calculating the electronic structure must be combined with data mining techniques. Using a peer group of materials (discussed below) the data mining algorithms can learn selection rules related to electronic structure properties as given by the band structure or the density of states. We will here use the LDA bandgap (E_{g}), the width of the highest valence band (vbw), the width of the lowest conduction band (cbw), the width of the highest occupying electron in the valence band (dEe), and the width of the lowest available state in the conduction band (dEh) to further narrow down the list of candidate materials. The definitions of these properties are shown in Fig. 2. It should be noted that the parameter vbw measures how delocalized the highest energy band of the valence states is. In the same way cbw measures the degree of delocalization of the lowest band amongst the conduction band states. The remaining two parameters, dEe and dEh, give related information, and stand in direct proportion to the effective mass of the highest electron state and the lowest hole state, respectively. As a matter of fact our mining algorithm could have made use of effec-