Data-mining study of the electronic charge density of semiconductors

A. Sari[1, 2]

[1]Division of Study and Prediction of Materials, Research Unit for Materials and Renewable Energies, EPM-URMER, University of Tlemcen, Algeria.

[2]University Center of Maghnia Tlemcen, Algeria.

***Corresponding author:** sari.aouatef@yahoo.fr

Received 1 June 2018, Received in final form 26 July 2018, Accepted 26 July 2018

Abstract

An unusual approach is proposed to extract knowledge and information from electronic charge density (ECD). Instead of the classical model witch use topological way to identify the nature of the bond; Data-mining technique are employed in this work. Initially, ab-initio method is used to investigate calculations of electronic charge densities (ECD), thus the statistical technique is considered to illustrate correlations existing between semiconductors. This work represents a remarkable approach to modeling the electronic properties of a material which may be used to identify new promising semiconductors and is one of the few efforts utilizing data-mining at an electronic level.

Keywords: *Electronic charge density (ECD), Semiconductors, FP-LAPW, Data-mining, Principal Component Analyses (PCA)s*

1. Introduction

The electronic charge density (ECD) is a topological way to identify the nature of the bond, it contains all the information, features and all the properties of a given material. It has a very significant role in analysis of the bonds and seems of capital importance for the study and the interpretation of the properties of materials. In DFT [1,2], it is the second observable after the total energy and is given by:

$$\rho_n(r) = \sum_k e\left|\Psi_{n,k}(r)\right|^2 \qquad (1)$$

Where $\left|\Psi_{n,k}(r)\right|$ is the wave function and *e* is the electronic charge. The summation covers all states of the Brillouin zone for a given band *n*.

The ECD has a meaning only in the area located between two atoms. Technically the segment between these atoms or the line of bonding is discretized on 128 points. Each point contains value of the charge density. The various values of the ECD between two atoms related to the interatomic distance are represented by a curve or profile which gives a visualization of the distribution of electrons. For each structure the profile is drawn in a specific direction, depending on the atomic environment. Examples of profiles are illustrated in **Fig. 1**. A considerable amount of experimental and theoretical work has been done during several years ago on the structural, mechanical and optical properties on tetrahedral coordinated semiconductors [3-6]. However until now, methods of interpretation of ECD are qualitative and are based on approximations.

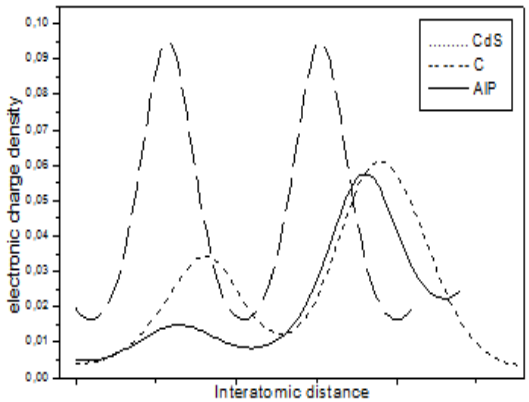


Fig. 1. *Electronic charge densities of AlP, CdS and C semiconductors*

The classical methods of analysis ECD found on the ionicity. It is the single quantitative amount but is evaluated by different models (Pauling [7], Phillips [8], Harrison [9]...). Although the basic principle of the density functional theory (DFT) [1, 2] and thus of any ab-initio calculations is that the ECD contains large information about an investigated system. It would be logical to expect that the knowledge of ECD can provide access to all

other physical and chemical properties; it would suffice to have an analysis tool that will be direct and effective. It is time now to find an efficient process to analysis the ECD and to provide chemical knowledge using quantitative investigation.

Therefore, we propose a distinct approach to extract knowledge and information from ECD. The idea is to use statistic tools to analyze data, since this method has such evidence in the field of medicine, biology… In this paper, we focus our study on A^NB^{8-N} semiconductors for the reason that their properties are already known. This is what allows verifying the accuracy and the power of the method used. An important fact is that the data analysis should present the results in a manner consistent with the knowledge, philosophy and theory of chemistry.

In the context of data analysis, data-mining techniques are now well established as effective and fast tools. They can handle and examine a large amount of data so that trends and correlations become apparent. In the other hand, the number of properties required to explain one system can be reduced to a minimum. Since, the design of a material requires essentially the definition of its properties; the various parameters (nature of bonding, structure of the cell occupancy of space, electronic structure, etc...) provide the conceptual basis for the study of materials plausible. In this aim, the flood of data that generates calculations of semiconductors' charge densities is exploited to construct the database. Thus, after calculations of ECD, each obtained profile is treated as a distinct individual which will belong to the database used as input for further investigation.

2. Computational details

(a). Calculation method

The calculations of electronic charge densities are performed by using the self-consistent first principles full-potential linearized augmented plane wave (FP-LAPW). This method, implemented in the Wien code [10] within the density functional theory formalism, is among the most accurate methods in the calculation of the electronic structure of solids. The Generalized Gradient Approximation (GGA) of Perdew et al [11] was taken to include the exchange–correlation energy to the total energy. The self consistant cycle that allows the resolution of Kohn and Sham equations gives access primarily to the wave function of the electronic system and to its ground state energy.

(b). Principal component analysis:

In the second step, we used a descriptive data-mining technique via Principal Component Analysis (PCA) which is a powerful tool for the analysis of a database. Considering data sets in the form of N×K matrices, the number of K (matrix columns) is large when exceeding, say, 100 000 or 1 000 000; while in 1970,this number was large when exceeding 20. In the same time, the number of observations N (matrix rows), has increased from around 10 000 to more than 500 000 [12]. PCA projects the spatial data onto a set of principal components and maps the data on a dimensionally reduced space, while preserving the best information conveyed [13,14]. Technically, the covariance matrix is calculated, the eigenvalues are referred to as the scores, and the eigenvectors are called loadings. The scores plot classifies the samples, while the loadings plot offer information on the relationships between the properties [15, 16]. In our study, the loadings plot does not give significant information since the relationship between various points of charge density has not a physical interpretation. The first principal component (PC1) is the eigenvector that corresponds with the largest eigenvalues of the covariance matrix, it captures the most information. The method used provides a strategy for utilizing the richness of information for exploiting and summarizing data, classification, and identification relationships between samples. PCA is unsupervised method based on the rich of information in multivariate data; multi-dimensional data measured on a set of similar samples, compounds, cases, process points. The discussion on the mathematical background of PCA can be found in different sources [17-19].

3. Results

The aim of this section is to understand the relationship between different semiconductors of type A^NB^{8-N} and identify trends and clustering. The initial matrix of database covers 22 semiconductors and 128 values of ECD. Owing to the PCA, this matrix is reduced to 2-dimensions while capturing more than 72% of the variance in the data.

Fig. 2 shows the scores plots of the studied semiconductors. The materials are scattered over a two-dimensional graph (PC1, PC2). PC1 captures 51, 92% of disaccord in the data set whereas PC2 captures 20.58%.

By first look, it appears five important clusters. Cluster 1 includes III-V compounds that crystallize in zinc-blend structure (AlP, AlAs, AlSb, GaP, GaAs, GaSb, InP, InAs, InSb). Cluster 2 contains II-VI compounds which crystallize in zinc-blend structure (ZnS, ZnSe, ZnTe, CdTe, HgTe). Cluster 3 includes all compounds based on boron (BN, BP, BAs, BSb). Cluster 4 encloses wurtzite compounds of II-VI and III-V semiconductors (CdSe, CdS, InN,

AlN, GaN). Finally, in cluster 5 we find diamond compounds (Ge, Si, C).

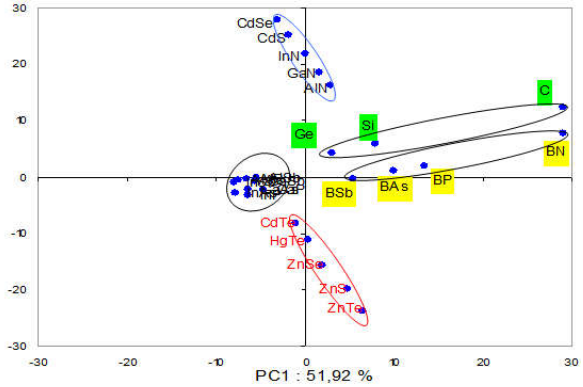


Fig. 2. *The scores plot of PCA with clusters.*

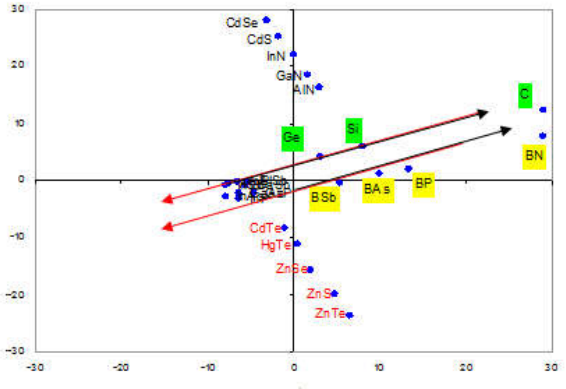


Fig. 3. *The increased gap energy (black arrow) and the decrease interatomic distance (red arrow) for borides and IV-IV semiconductors.*

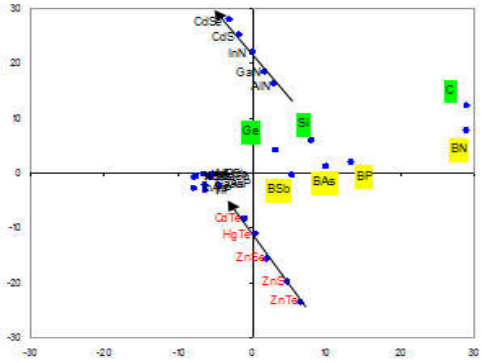


Fig. 4. *The increased Phillips ionicity for wurtzite compounds and zinc-blend II-VI compounds.*

Borides of cluster 3, even they crystallize in zinc-blend structure, and they don't belong to the cluster 1. This can be due to the fact that the small size of boron and the absence of *p* electrons in its core ($1s^2 2s^2 2p^1$) offer to borides special properties [20, 21]. Also, the small size of Nitride ($1s^2 2s^2 2p^3$), can explain that BN is localized in the extremity of PC1 far from other compounds of cluster 3, more the compound is far from the origin more the information conveyed is significant [17-19].

Table. 1.*Interatomic distance and energy gaps of borides and IV-IV semiconductors.*

Material	structure	Interatomic distance(Å)	Energy Gap (e.v)
BN	Zinc-blend	1,4926	6 [22]
BP	Zinc-blend	1,9710	2 [23]
BAs	Zinc-blend	2,0685	1.25 [24]
BSb	Zinc-blend	2,2812	0 .62 [25]
C	Diamond	1,5416	5.4 [6]
Si	Diamond	2,3687	1.17 [26]
Ge	Diamond	3,9880	0.74 [26]

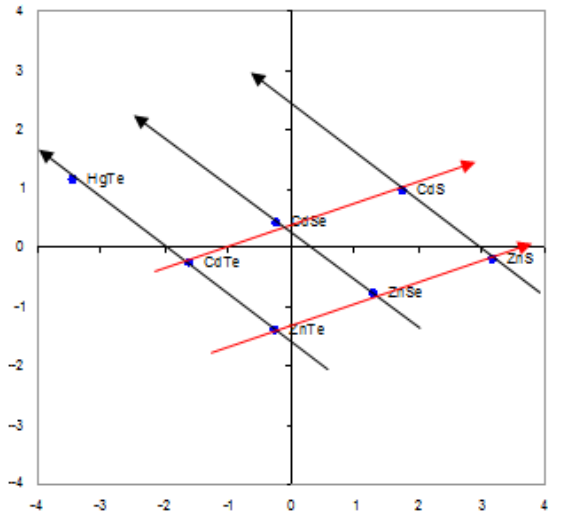
Table. 2. *structure and Phillips ionicity of semiconductors clusters 2 and 4* [27].

Compounds	structure	Phillips ionicity
ZnTe	Zinc-blend	0.609
ZnS	Zinc-blend	0.623
ZnSe	Zinc-blend	0.630
HgTe	Zinc-blend	0.650
CdTe	Zinc-blend	0.717
AlN	wurtzite	0.449
GaN	wurtzite	0.500
InN	wurtzite	0.578
CdS	wurtzite	0.685
CdSe	wurtzite	0.699

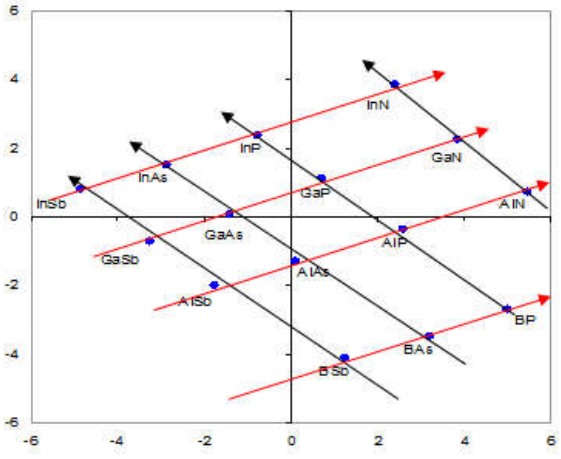
Table 3: *Bulk modulus of III-V and II-VI compounds* [27]

Compounds	*Bulk modulus (GP)*
BSb	*100*
BAs	*137*
BP	*162*
AlSb	*66*
AlAs	*77*
AlP	*86*
AlN	*212*
GaSb	*59*
GaAs	*65*
GaP	*76*
GaN	*173*
InSb	*55*
InAs	*61*
InP	*63*
InN	*130*
ZnTe	*51*
ZnSe	*62.4*
ZnS	*77.1*
CdTe	*42.4*
CdSe	*53.1*
CdS	*61.6*
HgTe	*42.3*

One can see from **Fig. 3** that the positioning of these materials along the PC1 axis (BSb, BAs, BP and BN) is consistent with an ascending order of the value of the gap energy following the direction of the black arrow. It is also consistent with a decreasing order of the interatomic distance following the direction of the red arrow. The same observations can be made for germanium, silicon and diamond carbon. The gap energy increases from 0.74 and 1.17 to 5.4 (black arrow) and the interatomic distance decreases from 3.988 and 2.3687 to 1.5416 (red arrow) for Ge, Si and C respectively. In **table 1**, gap energy and interatomic distance of borides and diamond compounds are mentioned. Hence, the increasing direction of PC1 indicates the increasing values of the energy gap for these materials, and the decreasing values of the interatomic distance.



(a)



(b)

Fig. 5. *The scores plot of (a) II-VI and (b) III-V semiconductors.*

A particular attention to the compounds of cluster 2 and 4, in **Fig. 4**, reveals that their positioning along PC2 is according the growing value of the Phillips ionicity. In **table 2**, Phillips ionicity values of cluster 2 and 4 compounds are exposed. Compounds of cluster 2 and 4 are in an ascending order of the Phillips ionicity since this one enhances from 0.609, 0.623, 0.630, 0.650 to 0.717 for ZnTe, ZnS, ZnSe, HgTe and CdTe respectively and from 0.449, 0.500, 0.578, 0.685 to 0.699 for AlN, GaN, InN, CdS and CdSe respectively.

In order to identify trends between II-VI semiconductors and between III-V ones, we performed others PCA analysis. The first one includes these II-VI compounds. The scores plot is displayed in **Fig. 5. a**. The second PCA analysis is done for III-V compounds, and the scores plot is displayed in **Fig. 5.b**. One can see from **Fig.5.a**, that II-VI compounds are spread according to a decreasing anion size (Te, Se and S) along the red arrow, and an increasing cation size (Zn, Cd and Hg) along the black arrow.

The same result is obtained from **Fig. 5.b**. since III-V semiconductors are spread according to a decreasing anion size (Sb, As, P and N) following the direction of the red arrow, and an increasing cation size (B, Al, Ga and In) following the direction of the black arrow. A similar interpretation is done in the paper [28]. In term of mechanical properties, III-V and II-VI compounds are in a growing direction of the bulk modulus along red arrow. Bulk modulus of II-VI and III-V semiconductors are given in **table 3.**

4. Conclusions

This work is among the few investigations in field of data-mining material's science. This paper illustrates the effectiveness of clustering and scattering of compounds into distinct groups. It demonstrate also, the correlation between alignment along sharp lines with some properties of these materials (atomic size, bulk modulus…) particularly those related to their chemical bonding. Thus, based on the ECD, materials are classified according to their energy gaps, crystal structures, interatomic distances and their ionicities. In other words, the PCA confirms that ECD is an important and fundamental concept which contains information about all these properties. The PCA allows us to reveal this information hidden at first sight and inaccessible when treating ECD in the usual way: plotting profiles and contours to identify qualitatively the type of chemical bonds. The use statistic approaches produces results comparable to experimental measurements. It and shows also how

computers can be utilized to exploit information and so to determinate what is necessary and useful. This work is a statistic methodology for analyzing the information contained in the ECD which can be developed to other systems and can speed up the prediction of new materials.

References

[1]. P. Hohenberg, W. Kohn, Phys. Rev. 136 (1964) B864.
[2]. W. Kohn, L. J. Sham, Phys. Rev. 140 (1965) A1133.
[3]. M .S. Omar, Mater. Res. Bull. 42 (2007) 961.
[4]. A .E. Merad, M .B. Kanoun, G. Merad, J. Cibert, H. Aourag, Mater. Chem. Phys. 92 (2005) 333.
[5]. S .Q. Wang, H.Q. Ye, J. Phys.: Condens. Matter 17 (2005) 475.
[6]. S. Zh. Karazhanov, P. Ravindran, A. Kjekshus, H. Fjellvag, B. G. Sevensson, Phys. Rev. B 75 (2007) 55104.
[7]. J. C. Phillips, Rev. Mod. Phys. 42 (1970) 317.
[8]. J. C. Phillips (1973) "Bonds and Bands in Semiconductors", First Edition, Academic, New York, USA.
[9]. W. A. Harrison, Phys. Rev. B 8 (1973) 4487.
[10]. P. Blaha, K. Schwarz, G. K. H. Madsen, D. Kvasnicka, J. Luitz: Wien2k. Institut fur Physikalische und Theoretische Chemie. Getreidemarkt 9/156. A-1060 Wien/Austria. ISBN 3-9501031-1-2. Release 2014.
[11]. S. Cotennier (2004) "Density functional Theory and the family of (L) APW-methods: a step by step introduction" (institute voor Kern- en stralingsfysica, K.U.Leuven, Belgium). ISBN 90-807215-1-4.
[12]. N. Kettaneh, A. Berglund, S. Wold, Computational statistics & data Analysis 48, 2005.
[13]. J. R. Nowers, S. R. Broderick, K. Rajan, B. Narasimhan, Macro-molecular Rapid Communications 28 (2007) 972.
[14]. S. C. Sieg, C. suh, T. Schmidt, M. Stukowski, K., Rajan, W. F. Maier, QSAR & Combinatorial Science 24 (2005) 114.
[15]. C. Suh, K. Rajan, Appl. Surf. Sci. 223 (2005) 148.
[16]. C. K. R. Suh, B. M. Vogel, B. Narasimhan, S. K. Mallapragada (Combinatorial Materials Science, Eds: S. K. Mallapragada, B. Narasimhan, and M. D. Porter, Hoboken, NJ John Wiley-interscience, 2007.
[17]. L. Ericksson, E. Johansson, N. Kettaneh-wold and S. Wold, Multi- and Megavariate Data Analysis: Principles, Applications, Umetrics Ab, Umea, 2001.
[18]. M. E. Pate, M. K. Turner, N. F. Thornhill, N. J. Titchener-Hooker, Biotechnol. Progress 20 (2004) 215.
[19]. C. Suh, A. Rajagopalan, X. Li, and K. Rajan, Data Science Journal 1 (2002) 19, 2002.
[20]. B. Bouhafs, H. Aourag, M. Certier, J. Phys. condens. Matter 12 (2000) 5655.
[21]. A. Zaoui, S. Kacimi, A. Yakoubi, B. Abbar, B. Bouhafs, Physica B 367 (2005) 195.
[22]. V. A. Fomichev, M. A. Rumsh, J. Chem. Phys. 48 (1968) 555.
[23]. B. Paulus, P. Fulde, H. Stoll, Phys. Rev. B 54 (1996) 2556.
[24]. R. M. Wentzcovitch, M. L. Cohen, J. Phys. C: Solid State Phys. 19 (1986) 6791.
[25]. O. Madclung, M. Schulz, H. Weiss (1982) "Semiconductors: physics of Group IV elements and III-V compounds. In", Landolt-Bornstein New Series, vol. III/17a , Springer (eds.), Berlin.
[26]. C. Kittel (1996) "Introduction to solid state physics", Seventh Edition, John Wiley & Sons, Inc, UK.
[27]. S. Adachi (2005) Properties of Group IV, III-V and II-VI Semiconductors, John Wiley & Sons Ltd, UK.
[28]. H. Zenasni, H. Aourag, S. R. Broderick, and K. Rajan, J. Phys. Status Solidi B 247 (2009) 115. .

