Data Mining Density of States Spectra for Crystal Structure Classification: An Inverse Problem Approach

Scott R. Broderick[1], Hafid Aourag[2] and Krishna Rajan[1,*]

[1] *Department of Materials Science and Engineering, Iowa State University, Ames, IA, USA*

[2] *Department of Physics, University Abou Bekr Belkaid, Tlemcen, Algeria*

Received 19 February 2008; revised 26 September 2008; accepted 11 November 2008
DOI:10.1002/sam.10026
Published online 11 May 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: The ability to model the density of states has been a long-standing problem in condensed matter physics. The classical methods that have been used are based on a variety of approaches, ranging from maximum entropy methods to recursion methods involving high dimensional data. In this work, we classify the crystal structure of an alloy based on the electronic structure, the inverse process of first principles calculations which calculate the electronic structure from crystal structure-based inputs. Here the electronic structure is represented by the density of states, and the classification of crystal structures is achieved through data mining. In addition to classifying alloys by crystal structure, we can classify alloys based on the degree of tetragonality and stoichiometry using solely the density of states spectra. This work seeks to describe the relationship between crystal and electronic structure based on a quantitative interpretation of the density of states, while discussing how capturing the principal contributions of the density of states suggests future work in modeling the density of states using less computationally expensive data mining techniques. © 2009 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 1: 353–360, 2009

Keywords: principal component analysis (PCA); first principles calculations; density of states (DOS); design maps; electronic structure; metallic alloys

1. INTRODUCTION

If we know the irreducible representation of the crystal structure (i.e. equivalent positions) and the pair wise interactions associated with the bonding between the constituent elements, then we can calculate the statistical distribution of electronic states [i.e. density of states (DOS)] in the solid. From this knowledge, a library of equilibrium properties can be derived (e.g. elastic constants, thermal, and electronic properties)

As summarized by Finnis [1], it is well known that the main factors that govern the relative energies of different crystal structures are encoded within the sum of energies of the occupied states, the band energy or the bond energy. This focuses attention on the form of the DOS since

$$E^{\mathrm{band}} = \sum_n f_n \varepsilon_n = 2 \int_{-\infty}^{infty} f_{\mathrm{F}}(\varepsilon)\varepsilon DOS(\varepsilon)\mathrm{d}\varepsilon \qquad (1)$$

Correspondence to: Krishna Rajan (krajan@iastate.edu)

As noted by Finnis, Cyrot-Lackmann and others had long proposed the idea of modeling the DOS, that is to characterize the DOS in terms of its moments. The value of this lies in the fact that it is a real space expression, which can be evaluated without solving any Schrödinger equations. The idea of the moments approach is to calculate just the lowest few moments and with them to reconstruct somehow the DOS. The simplest and first approach was to use a Gaussian ansatz for the DOS, which of course limits you to first and second moments. The more general problem of recovering a function from knowledge of its moments has been of great study and one approach is the maximum entropy method, which is based on Bayesian probability theory and its virtue is that no information is introduced apart from the moments themselves which eliminates any spurious structure in the reconstructed DOS. A particularly efficient way of generating moments is the recursion method pioneered by Haydock *et al.* [2,3].

In this paper, we explore the feasibility of a different approach to accurately solve a problem that to the best of our knowledge has not been applied before, namely the use of data mining and statistical learning methods.

© 2009 Wiley Periodicals, Inc.

In our approach, we are not even taking the assumption that we need to explicitly characterize the DOS in terms of moments, but simply to see if we can extract a decomposition of the DOS spectra that can identify an orthogonality of variables that contribute to the spectrum. Apart from assessing the ability to use a singular value decomposition approach to seek clear classification of crystal chemistry, it is proposed that the eigenvector spectra associated with the major principal components (PC) can form the foundation of the next step in developing a 'soft modeling' approach to calculating the DOS spectra. This point is introduced here but will require future work and is not the main objective of this paper.

The question this paper addresses is if we have the DOS curve can we capture and identify the crystal structure from which it is derived without having to work through the complexity of Schrödinger's equation from an inverse problem perspective? In this paper we demonstrate for the first time, that by simply data mining the DOS spectra itself, we can in fact capture the principal contributions (ergo PC) that define the DOS, namely the crystal structure. The value of this work lies not only for its value of showing an interesting application of data mining to electronic structure calculations, but in fact has a far reaching impact.

A methodology to analyze the output of electronic structure calculations is needed due to the large amounts of data created by these calculations. Two limitations with these calculations are that their complexity requires substantial time to perform the calculation and the amount of output data makes it difficult to analyze the data quantitatively and efficiently. Data mining can be beneficial in both of these areas by providing a methodology to quantitatively analyze the data and classify the materials based on the output so that predictions of new systems can help reduce the number of future calculations necessary. This work addresses the issue of analyzing the data from electronic structure calculations by integrating data mining techniques with first principles calculations.

In this work, the electronic structure is represented by the DOS. The object of this work is to find the relationship between crystal structure and electronic structure. Crystal structure is based on the long-range organization of atoms, while electronic structure describes which energy bands that the electrons of a system occupy. Electronic structure and crystal structure have different properties associated with them, and finding a way to link electronic and crystal structure would allow a better understanding of how electronic properties are related with atomic properties.

This work consists of several stages. The first step was to perform first principles calculations to calculate DOS (see Section 2), the next step was to analyze the DOS using principal component analysis (PCA) (see Section 3), and the final stage was to analyze the results (see Section 4). Starting with inputs derived from the crystal structure, first principles calculations output the DOS. We have defined the inverse of this problem to find the crystal structure based solely on the DOS. Finding the relationship between electronic and crystal structures in a comprehensive manner has major implications for materials science, as does a methodology to analyze the DOS in a quantitative manner.

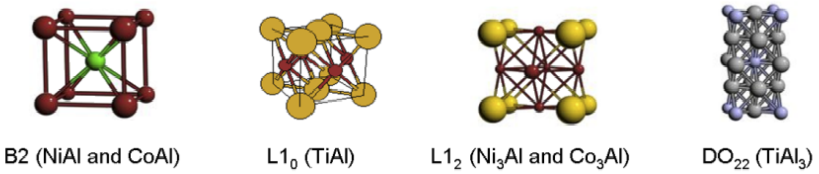
2. ELECTRONIC STRUCTURE CALCULATIONS

The electronic structures calculated in this work were created using WIEN2K [4], a first principles calculations program using the linear augmented plane wave method with local spin density approximation in density functional theory (DFT). First principles calculations are based on approximating solutions to Schrödinger's equation of a many-particle system, which serves as finding the statistical probability that an electron will have a specific energy. The solution to Schrödinger's equation contains all measurable properties of the system. However, finding an exact solution to Schrödinger's equation is nearly impossible for real problems. DFT overcomes this difficulty by approximating Schrödinger's equation with a different equation that can be solved. The process of calculating the electronic structure starting from crystal structure is shown in Fig. 1.

WIEN2K requires only the input of atom and equivalent atom position, lattice parameter and space group, as shown in the table of Fig. 1. Space group is identified by the crystal lattice and the symmetry operations possible for the atoms, therefore describing the crystal structure. Although crystal structure is input into the calculation, none of the inputs are suitable for a data mining analysis which seeks to integrate information on properties. These inputs are used to approximate a solution to Schrödinger's equation, from which, based on quantum mechanics, all properties of the system can then be solved. Further information on first principles calculations, solid state physics and quantum mechanics is provided in the references [5–11].

Numerous properties can be output from first principles calculations. The output this work focuses on is DOS, which contains information on materials properties [12–15]. DOS measures the number of electronic states per unit energy and is a factor of spacing between the number of electronic states at each energy level. An important value in the DOS is the Fermi energy (E_F), which is the maximum energy band occupied at temperature equal to zero K and has implications for conductivity. Various values at E_F are used to represent properties contained in the DOS. Also important is the cohesive energy (E_{Coh}), which is a measure of the energy change in an atom due to crystallinity instead of being a free atom. However, for most metals the E_{Coh}

(1) Crystal Structures



(2) Input Parameters

	Space Group	*Lattice Parameters*	*Equivalent Atom Position*
TiAl	P4/mmm (123)	a=4.001 c=4.071	Ti (0,.5,.5) Al (0,0,0) Al (.5,.5,0)
$TiAl_3$	$P6_3$/mmc (194)	a=3.854 c=8.584	Ti (0,0,0) Al (0,0,.5) Al (0,.5,.25)
CoAl	Pm-3m (221)	a=c=2.863	Co (.5,.5,.5) Al (0,0,0)
Co_3Al	Pm-3m (221)	a=c=3.655	Co (0,.5,.5) Al (0,0,0)
NiAl	Pm-3m (221)	a=c=2.848	Ni (.5,.5,.5) Al (0,0,0)
Ni_3Al	Pm-3m (221)	a=c=3.572	Ni (0,.5,.5) Al (0,0,0)

(3) First Principles Calculation/ DFT

Approximation to many particle Schrödinger equation

DFT: E related to electron density

$$DOS = \frac{1}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}$$

(4) Electronic Properties

DOS spectra, E_F, E_{Coh}, Charge Density, Elastic Constants, Modulus, optical properties

Fig. 1 The process for calculating electronic structure. From the alloys chosen, the crystal structure is used to create the input parameters for the calculation, which are space group (structure), equivalent atom positions, and lattice parameters (in angstroms). Using only these inputs, numerous electronic properties can be calculated. DFT calculates a solution approximating the many particle Schrödinger equation by relating the energy to the electron density. This work will consider only the DOS. This process creates electronic structure data from crystal structures.

is difficult to extract from the DOS. As a result, E_{Coh} and most other information taken from the DOS are qualitative and most of the DOS is not used.

Following the description in [16], the DOS can be thought of as a histogram describing the distribution of eigenvalues of the Hamiltonian matrix. These eigenvalues are the basis in calculating the electronic portion of total energy in some calculation methods. The energy eigenvalues consist of a wide range of energies, and the DOS is used as a probability function describing the spread of the eigenvalues. Equation 2 displays the equation describing DOS in an alternate form to that shown in Fig. 1, where i is the number of eigenvalues. This work looks to use all data in a quantitative analysis of DOS. Figure 2 displays the DOS of Ti, Al, TiAl, and $TiAl_3$, demonstrating how the combination of Ti and Al affect the DOS of the alloy due to the interactions of the valence electrons.

$$DOS = \sum_i \delta(E - E_i) \tag{2}$$

Additionally, the DOS shows the energy orbitals existing for a system because the energies where states are available would be expected to also be energies of available orbitals. As different structures have orbitals existing at different energies, it is anticipated that the DOS contains information on the structure of the system. Using this logic, by data mining the DOS we should be able to extract the structural information using solely the DOS. The difficulty in this extraction is selecting the proper descriptors. To overcome this difficulty, a multivariate analysis is being performed on the entire DOS, treating the DOS as spectral data.

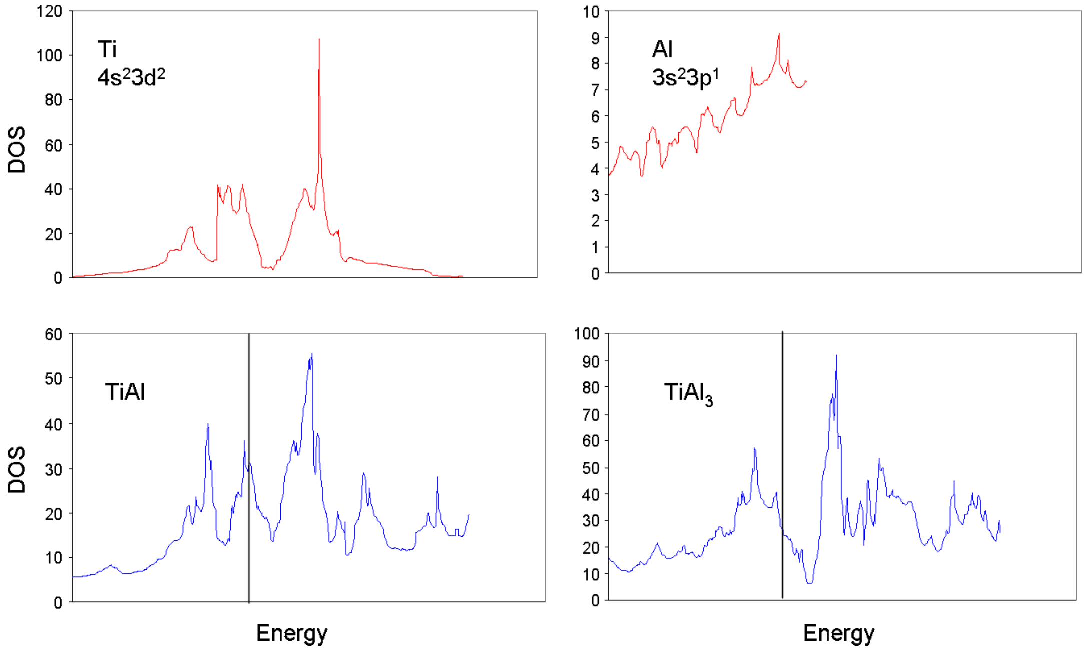


Fig. 2 Calculated total DOS of Ti, Al, TiAl, and TiAl3. Ti has s and d valence electrons with an unfilled d orbital, and Al has s and p valence electrons with an unfilled p orbital. The largest peak in TiAl corresponds with the peak in the Ti DOS which is due to the d orbital. $TiAl_3$ has the same peak due to the d orbital, but a broader peak than TiAl at lower energies due to the additional p-orbital states from having a stoichiometery with three Al instead of one Al. This figure demonstrates how the DOS of elements can be combined to form the DOS of an alloy, where bonding is represented by the overlap of orbitals. Here, only the total DOS of alloys and not partial DOS or DOS of elements is considered.

3. DATA ANALYSIS

PCA, described in [17–21], defines latent variables or PC that are defined to capture the maximum variance in a dataset and are comprised of a combination of properties to capture this variance. PCA is used in this work to quantitatively analyze the DOS to classify materials based on crystal structure. This poses the inverse objective of first principles calculations, and this inverse problem is shown in Fig. 3.

If the data is mean centered then the intersection of the PC will occur at the new origin in PC space, and the dataset will have a mean value of zero. The data also can be scaled to remove any effect due to changing units. In this work, the descriptors are all the same unit, and no scaling of each descriptor was necessary. However, the interest in this work initially is in the shape of the curves, as the shape of the DOS captures the information on the structure and some properties, and each curve was scaled so that the maximum DOS value of each curve is equal to unity.

PCA decomposes the initial data matrix **X** into a scores matrix **t** and the loadings matrix **p**. The equation PCA is based on is shown in Eq. 3, where **E** is the residual matrix.

$$\mathbf{X} = \mathbf{t} \cdot \mathbf{p}^T + \mathbf{E} \tag{3}$$

X contains all of the points of the DOS curves. As mentioned, the difficulty in extracting all of the data in DOS has been that choosing the right data from the DOS to analyze is difficult. To avoid this problem, we analyze every point in the DOS curve. Figure 4 shows the composition of **X, t,** and **p**.

Using Eq. 3, the scores and loadings matrices can be calculated. The scores matrix describes how a sample is related to other samples; in this case, the scores relates the DOS curve of all of the alloys included in **X**. The loadings matrix captures the features that most impact the scores matrix. From the loadings matrix, we can find the relationship between features in the DOS, and identify with which systems these features correspond. The value in each of the matrices is the respective PC values, which defines the alloy system or descriptor in PC space. In a scores or loadings plot, the proximity of points represents the degree of correlation, with the understanding that a difference in a PC capturing little variance is much less important than the same change in PC values for a latent variable capturing much more variation in the data. A scores plot will be presented and discussed in the next section.

Two PCs were included in the analysis and captured 91.49% of the variance. The rest of the information is not included in the analysis and is contained by the residual

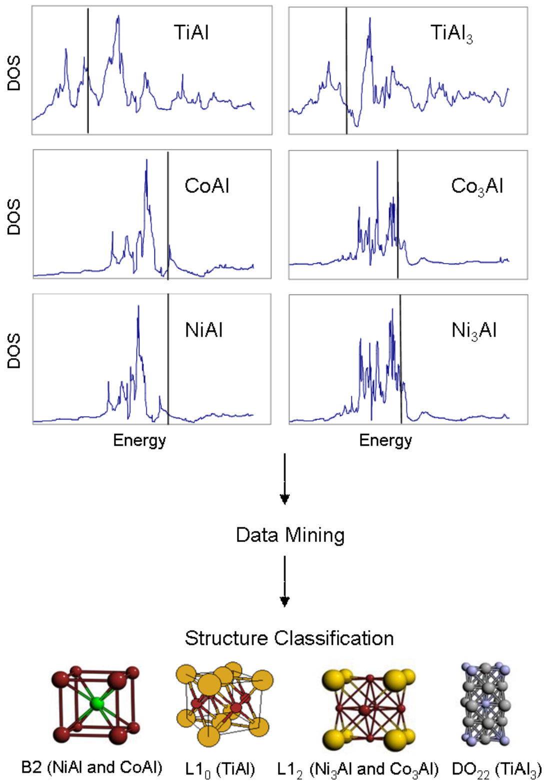


Fig. 3 The inverse problem. From Fig. 1, the process of first principles calculations was presented where from the crystal structure the electronic structure was calculated. In this work, data mining is used to determine the crystal structure from the electronic structure. The input parameters from Fig. 1 cannot be determined exactly, but the relative values can be determined.

matrix, **E** from Eq. 3, which consists of noise and background features. The minimum number of PCs employed should capture at least 90% of information with the last PC corresponding with the smallest eigenvalue that is still greater than unity. The amount of variance captured by each PC is defined by Eq. 4, where PCx refers to any PC between PC_1 and PC_n, and λ is the eigenvalues of the covariance matrix of **X**.

$$\%\text{variance of PC}x = \frac{\lambda_x}{\sum_{n=1}^{n} \lambda_n} \quad (4)$$

4. RESULTS AND DISCUSSION

By data mining the DOS with PCA, a structure classification based only on the DOS is created. The entire DOS curve provides an input for a data mining analysis that contains all properties and parameters at the electronic and atomic level. We will classify materials based on DOS, which additionally means that the material will be classified on structure and materials properties contained within the DOS.

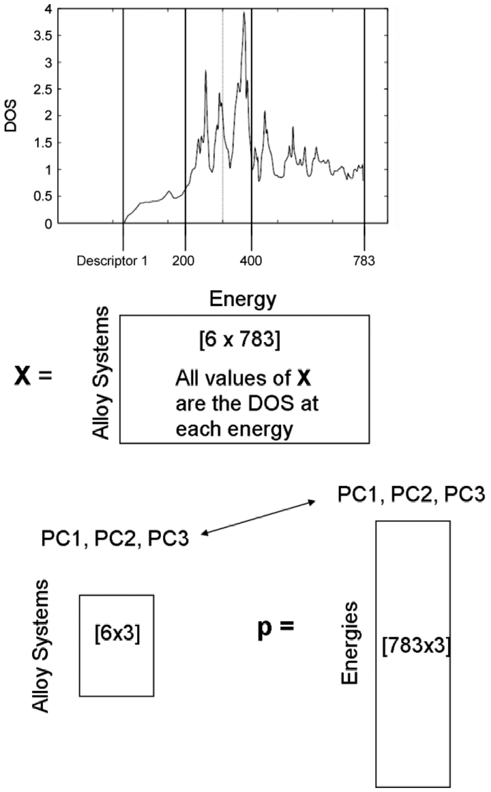


Fig. 4 The composition of the matrices from Eq. 3. X is the original data matrix, where the columns contain every data point in the DOS curve, the rows contain different alloy systems, and the value in the matrix is the DOS, or intensity, at the specified energy. The scores matrix, t, has the same rows as X, but the columns include the PCs included in the analysis. The loadings matrix, p, has rows of every energy in the DOS, and the columns are the same as the PCs in the scores matrix. The values in t and p are the PC values. As t and p have the same columns, the relationships between the alloys and energies to the PCs can be compared.

The systems calculated were transition metal aluminides: TiAl, $TiAl_3$, CoAl, Co_3Al, NiAl, and Ni_3Al. The input crystal structure for each calculation was TiAl ($L1_0$), $TiAl_3$ ($D0_{22}$), CoAl and NiAl (B2), and Co_3Al and Ni_3Al ($L1_2$). Transition metal aluminides are of interest because they constitute an exceptional class of structural materials for high temperature applications, with properties such as excellent high temperature strength, corrosion resistance, low density, and high melting point. However, most of these intermetallics exhibit brittle fracture and low ductility at ambient temperatures [22–24]. This work can be extended to ternary compounds to study the effect of alloying and site occupancy, as successful attempts have been made in improving the ductility with additional alloying of these compounds.

To perform this data mining exercise, DOS is treated as spectral data. Data mining has been used on spectral data previously [25–28]; however, in the case of those spectral sets the interpretation of the peaks was already developed and the alignment of the curves is obvious, while with DOS the meaning of changes in peaks is unclear and the alignment of data needs to be considered. The DOS was manually aligned by the tallest bonding peak, but other alignments such as focusing on the Fermi energy or using developed algorithms warrant consideration. This work provides a framework with which to better understand and interpret DOS curves and to use this information for materials design. The dataset is six samples versus 783 energy values, although some calculations can have thousands of energy values. Every point in the DOS curve is included into the dataset, and the impact of all energy points can be captured using PCA. Energy values that do not provide new information except for only background have a PC loadings value equal to the origin and thus have minimal impact in the developed model. Therefore, when using PCA every point can be included without creating a daunting problem of too much data.

Using PCA to analyze the data, 91.49% of the variance in the data is captured with two PCs. PCA is useful for analyzing spectral data because it can reduce large datasets to a minimal amount of dimensions; in this case, PCA reduces the data from 783 dimensions to two dimensions while maintaining 91.49% of the information. A result of this analysis is presented in the scores plot of Fig. 5. The important feature of this figure is that the aluminides are separated based on the structure of the compound, demonstrating that crystal structure information is contained in DOS. This analysis is particularly useful because it provides a *quantitative* analysis of the DOS, while most previous analyses were qualitative. This demonstrates the importance of integrating data mining with electronic structure calculations because this work makes a quantitative analysis of the entire DOS curve possible. This figure provides an analysis of the DOS resulting in crystal structure, the inverse operation of the first principles calculations.

PC1 and PC2 are comprised of a combination of the descriptors (energy levels) that contain the most variance in the data. The combination of energy levels which most differentiate the structures are redefined as PC1 and the combination of energy levels which capture the most difference in an orthogonal direction to PC1 are redefined as PC2. Using the loadings plot, the contribution of each descriptor to the PCA model can be identified. Knowing how the energy levels contribute to the PCA model can be used to identify which features of the DOS are most important and with what those features correlate. NiAl, Ni_3Al, CoAl, and Co_3Al all have negative PC1 values, and additionally all have the same space group, as shown in Fig. 1. The alloys with negative PC1 values have cubic structures (a=c) while the alloys with positive PC1 values are not cubic. PC2 is capturing the differences in stoichiometry. Positive PC2 structures have an AB stoichiometry and negative PC2 structures have an A_3B or AB_3 stoichiometry. From this interpretation, adding additional alloying elements will drive the system to a more negative PC2 value. PCA is able to differentiate structures, tetragonal degree, and stoichiometries based only on the DOS. This classification is again emphasized when examining the eigenvector spectrum from the loadings plot (Fig. 6). The loading plots for CoAl DOS, for instance, provide a set of orthogonalized spectrum of the decomposition of the DOS. While much of the work in the field has focused on developing by anstaz or other means (e.g. maximum entropy methods, recursion techniques), we are demonstrating for the first time the potential of using PCA methods to assess the contributions that can serve to model DOS.

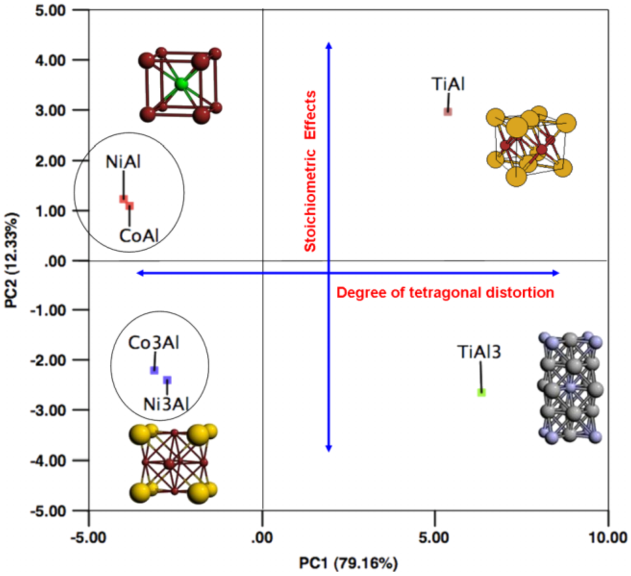


Fig. 5 PCA scores plot from data in matrix t of Fig. 3. PCA clusters points in the same structure, demonstrating that crystal structure is represented in the DOS. PC1 is capturing how tetragonal a structure is because the structures with negative PC1 values are cubic and the structures with positive PC1 are tetragonal. PC2 is capturing the differences in stoichiometery. The structures with positive PC2 have an AB stoichiometery and the structures with negative PC2 have stoichiometery of A_3B or AB_3. This analysis provides a framework with which to analyze a large number of DOS quantitatively.

As more systems are added to this work, we will explore how material properties appear in this diagram, creating a crystal structure-electronic structure-property map. The data has been *manually* aligned using the tallest bonding peak as the point of reference prior to performing the data

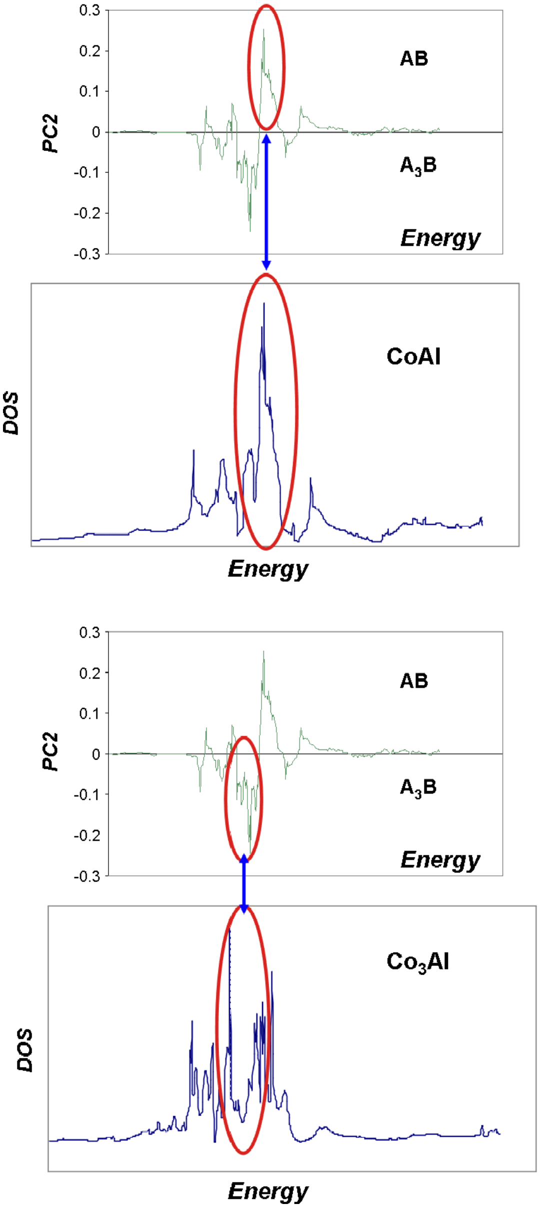


Fig. 6 Eigenvalue spectrum clearly showing the maximum (CoAl) and the minimum (Co_3Al) correspond to energies associated with the maximum in the density of states spectra for CoAl and Co_3Al, but of opposite weights in both eigenvectors. Hence we have identified in which part of the DOS curve the separation/classification according to stoichiometry/chemistry is most effectively captured with regards to the systems analyzed. As anticipated from the scores plot, similar results exist when identifying which part of the DOS curve captures stoichiometery for all the other aluminides analyzed, or when identifying parts of the DOS curve due to the tetragonal distortion (PC1).

mining analysis. Exploring how data alignment affects the extraction of properties will be a major focus as this work advances. This work will serve to better understand the DOS and improve its use as a design tool.

5. CONCLUSIONS

In this work, we are employing data mining tools to quantitatively analyze electronic structure calculations. Using PCA on the DOS of six transition metal aluminides we have demonstrated that the DOS is based on the crystal structure. The inverse approach of this analysis is demonstrated through the prediction of the crystal structure from the electronic structure, providing an inverse operation of the first principles calculations. The ability to model the DOS has been a long-standing problem in condensed matter physics. The classical methods that have been used are based on a variety of approaches, ranging from maximum entropy methods to the most successful being recursion methods proposed by Haydock *et al.* over 30 years ago. The aim of these modeling strategies is to avoid having to do a complete solution of the Schrödinger’s equation in order to model the DOS spectra. In this study, we discuss how data mining and statistical learning methods can provide a completely new way of addressing this problem in future work. Our paper here provides the first demonstration that data dimensionality reduction techniques such as PCA can indeed discriminate and identify the individual contributions of different physical attributes associated with crystal symmetry and crystal chemistry and their role in DOS spectra. Data mining is required to achieve these results because each electronic structure calculation results in large amounts of data, and a methodology is required so that the data can be analyzed efficiently. The PCA plot has axes that capture the most information based on a combination of all properties in the dataset, so that a linkage between features of the DOS and properties of an alloy can be formed. This work is a first step in creating a methodology to analyze electronic structure information, which can be extended to more systems and accelerate the structure prediction of new materials.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from the following program: National Science Foundation International Materials Institute program for Combinatorial Sciences and Materials Informatics Collaboratory (CoSMIC-IMI)—Grant no. DMR-0603644 (SB, HA and KR); AFOSR Grant no. FA95500610501 (KR) and the DARPA Center for Interfacial Engineering for MEMS (CIEMS)—Grant no. 1891874036790B (SB and KR).

REFERENCES

[1] M. Finnis, Interatomic Forces in Condensed Matter, Oxford, Oxford University Press, 2003.

[2] R. Haydock, V. Heine and M. J. Kelly, Electronic structure based on the local atomic environment for tight-binding bands, J Phys C: Solid State Phys 5 (1972), 2845–2858.

[3] R. Haydock, V. Heine and M. J. Kelly, Electronic structure based on the local atomic environment for tight-binding bands. II, J Phys C: Solid State Phys 8 (1975), 2591–2605.

[4] P. Blaha and K. Schwarz, Wien2K, Austria, Vienna University of Technology, 2002.

[5] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys Rev 136 (1964), B864.

[6] R. Terki, G. Bertrand, H. Aourag and C. Coddet, Structural and electronic properties of zirconia phases: A FP-LAPW investigations, Mater Sci Semicond Process 9 (2006), 1006–1013.

[7] G. Trambly de Laissardiere, D. Nguyen-Manh and D. Mayou, Electronic structure of complex Hume-Rothery phases and quasicrystals in transition metal aluminides, Prog Mater Sci 50 (2005), 679–788.

[8] D. J. Singh and L. Nordstrom, Planewaves, Pseudopotentials, and the LAPW Method (2nd ed.), New York, NY, Springer, 2006.

[9] W. Greiner, Quanum Mechanics: An Introduction (4th ed.), New York, NY, Springer, 2001.

[10] N. W. Ashcroft and N. D. Mermin, Solid State Physics, Philadelphia, PA, Brooks/Cole, 1976.

[11] C. Kittel, Introduction to Solid State Physics (8th ed.), Hoboken, NJ, John Wiley & Sons, 2005.

[12] L. Cheng-Bin, L. Ming-Kai, Y. Dong, L. Fu-Qing and F. Xiang-Jun, First principles study on the charge density and the bulk modulus of the transition metals and their carbides and nitrides, Chin Phys 14 (2005), 2287–2292.

[13] J. R. Alvarez and P. Rez, Calculation of electronic properties of boundaries in Ni3Al, Acta Mater 49 (2001), 795–802.

[14] W. Zhou, H. Wu and T. Yildirim, Electronic, dynamical, and thermal properties of ultra-incompressible superhard rhenium diboride: a combined first-principles and neutron scattering study, Phys Rev B: Condens Matter Mater Phys 76 (2007), 184113–184116.

[15] S. F. Matar and M. A. Subramanian, Calculated electronic properties of the mixed perovskite oxides: CaCu3T4O12 (T=Ti, Cr, Mn, Ru) within the DFT, Mater Lett, 58 (2004), 746–751.

[16] R. Phillips, Crystals, Defects and Microstructures: Modeling Across Scales, Cambridge, UK, Cambridge University Press, 2001.

[17] C. Suh, A. Rajagopalan, X. Li, K. Rajan, The application of principal component analysis to materials science data, Data Sci J, 1 (2002), 19–26.

[18] A. Daffertshofer, C. J. C. Lamoth, O. G. Meijer and P. J. Beek, PCA in studying coordination and variability: a tutorial, Clin Biomech, 19 (2004), 415–428.

[19] L. Ericksson, E. Johansson, N. Kettaneh-Wold and S. Wold, Multi- and Megavariate Data Analysis: Principles, Applications, Umea, Umetrics Ab, 2001.

[20] H. Berthiaux, V. Mosorov, L. Tomczak, C. Gatumel and J. F. Demeyre, Principal component analysis for characterising homogeneity in powder mixing using image processing techniques, Chem Eng Process 45 (2006), 397–403.

[21] Z. P. Chen, J. Morris, E. Martin, R. B. Hammond, X. Lai, C. Ma, E. Purba, K. J. Roberts and R. Bytheway, Enhancing the signal-to-noise ratio of X-ray diffraction profiles by smoothed principal component analysis, Anal Chem 77 (2005), 6563–6570.

[22] M. Jahnatek, M. Krajci and J. Hafner, Interatomic bonding, elastic properties, and ideal strength of transition metal aluminides: a case study for Al_3(V,Ti), Phys Rev B: Condens Matter Mater Phys, 71 (2005), 24101–24116.

[23] C. Jiang, Site preference of transition-metal elements in B2 NiAl: a comprehensive study, Acta Mater 55 (2007), 4799–4806.

[24] T. Li, J. W. Morris and D. C. Chrzan, Ideal tensile strength of B2 transition-metal aluminides, Phys Rev B: Condens Matter Mater Phys, 70 (2004), 054107.

[25] C. Suh, K. Rajan, B. M. Vogel, B. Narasimhan and S. K. Mallapragada, Informatics methods for combinatorial materials science, In Combinatorial Materials Science, S. K. Mallapragada, B. Narasimhan and M. D. Porter, eds. Hoboken, NJ, Wiley Interscience, 2007.

[26] B. K. Alsberg, M. K. Winson and D. B. Kell, Improving the interpretation of multivariate and rule induction models by using a peak parameter representation, Chemom Intell Lab Syst, 36 (1997), 95–109.

[27] M. E. Pate, M. K. Turner, N. F. Thornhill and N. J. Titchener-Hooker, Principal component analysis of nonlinear chromatography, Biotechnol Prog, 20 (2004), 215–222.

[28] F. Westad and H. Martens, Shift and intensity modeling in spectroscopy–general concept and applications, Chemom Intell Lab Syst, 45 (1999), 361–370.