

Chapter 1 : Virtual Screening and QSAR Formulations for Crystal Chemistry - [PDF Document]

Data Mining and Multivariate Analysis in Materials Science www.nxgvision.com, A. Rajagopalan and www.nxgvision.com: in. *Molten Salts - Fundamentals to Applications* ed. M. Gaune-Escard Kluwer Academic p ().

The analysis is applied to the prediction of bulk modulus based on a combinatorial analysis of crystal-chemistry descriptors used in ab-initio calculations. However, the key to minimizing the search process even in combinatorial ex- periments is to identify the key combinations required to achieve the desired functionality in the class of materials being studied. In previous work, we proposed the concept of virtual combinatorial experiments [1]. In this type of materials selection and design strategy we showed how one may design combinatorial libraries a priori by inte- grating data-mining techniques with physically robust mul- tivariate data. This involves a process of strategically se- lecting appropriate physical parameters that can be ana- lyzed in a multivariate manner. The analysis can lead to the identification and development of new materials chemistries that show promise for the desired functionali- ty. In this paper we extend our previous work to develop explicit quantitative relationships that identify the relative contributions of different data descriptors, and the result- ing relationship between all these descriptors as a linear combination, to the final property i. The foundation of our study is a combinatorial analysis of first-principles-derived data on crystal chemistry and bulk modulus. The engineering motivation behind such a study is that the use of combinatorial- and informatics- based methods can help in rapid screening and identifica- tion of new materials chemistries for hard materials. As noted by McMillan [2], the search for alternative super- hard materials, some of which might even be harder than diamond, has stimulated high-pressure research for over half a century. The approaches have been to develop com- putational predictions either from first principles or through experimental means that take advantage of new advances in instrumentation. Despite this progress, our knowledge base on possible new materials chemistries is still limited to a relative small number of systems. In this paper we will describe work that is part of a larger effort in our group to use informatics techniques to accelerate this screening process. While there is a major concern in the combinatorial ex- perimentation community for integrating data-mining tools into high-throughput experimentation data, we are developing a parallel strategy of developing a computa- tional informatics infrastructure. The question that we can address via QSAR formulation is: To experimentally synthesize and characterize a vast array of possible compo- sitions is, of course, not realistic. Similarly, despite the ad- vances and sophistication of first-principles predictions, it is equally prohibitive to computationally predict properties for a large combinatorial array of crystal chemistries. This database contains a vast array of property in- formation that is relevant to issues of interest here such as thermal expansion, compressibility and modulus. We wish to explore the individual correlations of the specific variables i. The Partial Least Squares PLS regression method is particularly appropriate for QSAR formulations as it is used to predict properties of properties based on variables even some which may have only indirect im- pact which collectively relate to these properties. In addi- tion, model parameters in PLS can be more accurately cal- culated with increasing number of relevant variables and observations [3]. PLS also has the advantage over multiple linear regressions for handling of collinearity and missing data [3]. The mathematical formulation of this method is one that calcu- lates a projection which captures the variance in the pre- dictor variables X and the correlation between it and the predicted variables Y. Detailed geometrical or mathematical descriptions can be found in the literature [4]. All descriptors were auto-scaled because they have different ranges and scales. Auto-scaled data matrix has zero mean and unit variance. Cross-valida- tion is used to choose optimum number of LVs of the cali- bration model. In this process, several model parameters were calculated: Several variables are shown in Table 1. A series of studies have systematically explored up to 39 of the total possible of 49 compounds [5 â€” 7]. While one may, in principle, repeat these complex first-principles calculations, the issue we wish to address in this paper is whether one can develop QSAR to be able to predict properties without arduously repeating first-principle calculations and fitting the energy â€” volume curve to an equa- tion of state. In this paper, we describe how we have used an informatics approach to gain further insight into this theoretically derived data set, and hence

provide a means of rapidly enhancing critical data for structure-property relationships in materials. Besides the descriptors described by Ching et al. Since the Pauling electronegativity scale is not good for bulk materials [8], we chose Martynov-Batsanov electronegativity, which is fully based on quantum mechanical bases [9]. As in a previous paper [1], the parameterization of electronegativity for ternary compounds is based on a linear weighting model that was originally proposed by Villars and Hulliger [10]. KGaA, Weinheim Figure 1. Combinatorial selection of stoichiometries used in this study based on the work described by Ching et al. However, since lattice constants l_c correlate with cation radii or bond length $B \propto \sqrt[3]{N} \sqrt{BL} \propto \sqrt[3]{N}$, a column of lattice constants was excluded. Bond order was also not used, because it is a measure of the bond strength and we used bond length. The theoretical foundations of these individual parameters are described elsewhere [5-7] and will not be repeated here; suffice it to say that the quantitative assessment of these variables is based on a first-principle methods. PLS can be used to identify outliers. With 16 compounds whose bulk moduli are shown in Table 1, preliminary PLS was tested to detect outliers in the dataset. Therefore $c\text{-ZrTi}_2\text{N}_4$ was identified as an outlier and this was removed from the data. In order to assess the value of which and how many LVs may be appropriate to avoid overfitting the data, we tested different models based on the selection of LVs the results of which are summarized in Table 2. List of single and double nitrides which are used in this paper, taken from Ching et. Statistical parameters of each model. Rajan test sets for model I are shown in Table 3. Four LVs were chosen and explained. Calculated model parameters were as follows: After using LOO as an internal validation, the test set was tested as an external validation. Figure 2 shows predicted versus ab-initio-derived bulk modulus values in the model with the resulting QSAR formulation: Mulliken effective charge for N ion. Calculated model parameters for model II are shown in Table 2. A very low value for Q_2 is shown in Table 2. Since the balance of R^2 and Q^2 represents a desirable model [4], we chose model I as our final model. The PLS regression equation provides us with the QSAR for a linear combination of LVs, which, in turns, provides a means of computationally engineering the modulus of spinel nitrides. It should be noted that the apparent discrepancies of the outer points do not necessarily detract from the statistical analysis but are actually of more value in identifying the limitations of the computational parameters associated with these compounds. KGaA, Weinheim Table 3. Predicted versus ab-initio-derived bulk modulus for PLS model I. This is consistent with theoretical studies [5, 8] that show it is the effective charge parameter which helps to define the degree of charge transfer and the level of covalency associated with the specific site occupancy of a given species. Ab-initio calculations of this effective charge can be then be used as a major screening parameter in identifying promising crystal chemistries for promoting the modulus. Hence, using PLS to develop a QSAR formulation combined with an interpretation of the physics governing these materials can indeed be valuable. Our predictions fit well with systems of similar electronic structure and allow us to clearly identify outliers based on these quantum mechanical calculations. As noted in the beginning of this paper, we are using PLS to detect outliers as well. Using the above QSAR formulation of model I, we predicted the bulk moduli of a new external test set. For six samples, we compared PLS-derived bulk moduli with those of ab-initio calculations Figure 4. It should be noted that comparison to theoretical calculations are our primary source for validation as experimental studies in this field are extremely limited. Predicted versus ab-initio-derived bulk modulus for six external test sets using PLS model I. The ab-initio-derived values used are from the literature [11]. In this manner we can rapidly develop predictions of vast new arrays of chemistries. Based on these predictions we can now seriously and effectively accelerate materials design by focusing on promising candidate chemistries. Those selected can then be subjected to further analysis via experimentation and computational methods to validate crystal-structure-level properties. B , 61, B , 63, B , 54,

LIST OF CONTRIBUTORS A.K. ADYA www.nxgvision.com R.T. CARLIN www.nxgvision.com www.nxgvision.comND
M. GAUNE-ESCARD www.nxgvision.comEV www.nxgvision.com University of Abertay Dundee, School of Molecular of
Life Sciences.

Seeing Millions of Atomsâ€. Sinnott â€” National Science Foundation - [http:](http://) Sastry, Metallurgical Transactions 64, Wallach; Journal of Crystal Growth 49 Vander Sande, Journal of Materials Science 17, Beddoes, Journal of Materials Science 17 Rajan, Metallurgical Transactions 14A Wallace, Metallurgical Transactions, 15A Terada, Scripta Metallurgica 18 Hewitt, in Advances in Fracture Research, eds. Rajan Spine 9 Rajan Acta Metallurgica 32 Rajan Metallurgical Transactions 16A Subramanian and MA Imam, p. Sewell Journal of Metals 38, 30 Sewell Journal of Metals 38 34 Rajan, Philips Electron Optics Publ. Group, Mahwah, NJ, p. Denhoff , Journal of Applied Physics 62 Denhoff, J-M Baribeau, D. Rajan Journal of Crystal Growth, 81, Moore, Journal of Applied Physics, 62, Denhoff, Solid State Communications, 61, Rajan Canadian Journal of Physics, 65, Rajan Thin Solid Films, , Rajan in Dislocations and Interfaces in Semiconductors, p. Gundlier, Elsevier Science Publ. Status and Prospects, p. Rajan MRS Proc Wright an d K. Rajan Journal of Metals, 41 28 Rajan, Applied Physics Letters, 54 Rajan Metallurgical Transactions A 21 Webb, Applied Physics Letters, 57 Rajan Applied Physics Letters, 57 Rajan; Journal of Electronic Materials, 19 Rajan Journal of Electronic Materials, 19 Rajan in Superconductivity and Applications, eds. Rajan in MRS proc. Rajan Physica C, Rajan in MRS proc, vol. Rajan Journal of Electronic Materials, 20 C Corelli and K. Rajan Journal of Applied Physics, 70 Rajan Journal of Crystal Growth Rajan in Materials Developments in Microelectronic Packaging: Performance and Reliability, eds. Performance and Reliability, ed. Rajan Journal of Applied Physics, 71 Rajan Materials Science Forum, Rajan Journal of Electronic Materials, 21 Rajan in MRS Proc. Tripathi, Vedam Books Intl. Rajan Journal of Electronic Materials, 22 Rajan Journal of Crystal Growth, , Norberg Physica D, 60, Rajan Journal of Applied Physics, 73 Rajan in Modeling of Coarsening and Grain Growth, pp. Rajanin Advanced Composites Fabrication, eds. Frear, van Nostrand and Reinhold, NY Precious Metals Institute Conf. Rajan Journal of Electronic Materials, 23 Topological Events in 2 dimensional grain growth: Rajan in Processing of Long Lengths of Superconductors, eds. Rajan Journal of Metals, 52 Rajan Journal of Materials Science, 29 Rajan Materials Science and Eng. B, 13 2 A 26 Rajan Chemical Engineering Communications: Festschrift Issue for Prof. Gill, Rajan Journal of Engineering Failure Analysis 2 Rathore, The Electrochemical Society, pp. Rajan Journal of Engineering Failure Analysis, 3 Petkie in Polycrystalline Thin Films, p. Higher Reliability Through Processin g , ed. Rajan in Beam Processing of Materials, eds. European Ceramic Society 17 Rajan in Mathematics of Microstructural Evolution, pp. Rajan Thin Solid Films Electronic Materials, 26 Smith Memorial Symposium, pp. A Visualization Tool for Fuzzy Clustering. Rajan in Microstructure Evolution: Characterization and Modeling, pp. Marsh, TMS ,Warrendale Characterization and Modeling pp. Marsh, TMS, Warrendale E Glicksman and K. Rajan Acta Materialia 46 Rajan in MRS Proceedings Rajan Materials Science and Engineering A Key Engineering Materials Philosophical Magazine 79 European Ceramic Society 19 Finding, understanding and using information about our physical world - DOE Panel report: Journal of Metals Molten Salts â€” Fundamentals to Applications ed.

Chapter 3 : Changwon Suh | Aspuru-Guzik Group

K. Rajan, A. Rajagopalan, C. Suh (3rd author) (Symposium Proceedings) The Application of Support Vector Machines to the Identification of Materials Attributes.

In engineering design, we are constantly faced with the need to describe the behavior of complex engineered systems for which there is no closed-form solution. There is rarely a single multiscale theory or experiment that can meaningfully and accurately capture such information primarily due to the inherently multivariate nature of the variables influencing materials behavior. Seeking structure-property relationships is an accepted paradigm in materials science, yet these relationships are often not linear, and so the challenge is to seek patterns among multiple length and time scales. In this paper, we present two separate but complementary examples of addressing the issue of high-dimensional data in materials science in the spirit of the intellectual focus of this new journal. The first example uses principal component analysis and the second example uses statistical analysis coupled to dimensional analysis. Statistical Analysis and Data Mining 1: The modeling efforts can be once and engineering revolves around understanding the divided into two main types: The enormous com- strategies involving advanced discretization, parallel algo- plexity in studying materials can be traced to the uncom- rithms, and a software architecture for distributed com- monly large number of variables involved in the relation puting systems. Among these approaches are atomistic between any two of these features. Compounding this chal- models and ab initio calculations, thermodynamic model- lenge, we find that it is impossible in practice to uncover all ing, phase field simulation, and finite element modeling variables, and the theoretical and experimental limitations at a microstructural level. Soft modeling was first intro- of traditional approaches sometimes fail to uncover even duced by the life sciences and organic chemistry commu- the dominant variables. When important variables have not nity, and it relates to statistically based, model-independent approaches. Among these approaches are the uses of regres- Correspondence to: Krishna Rajan krajan iastate. The designed to study the structure-properties aspect of the innovative aspect of these techniques is that the statisti- materials paradigm and the other designed to study the cal approaches employed are enhanced by including the processing-properties aspect. Other techniques introduced basic physics of the problem; for example, requiring that in this special issue address relationships involving the other the predictions made have meaningful units. As noted by Searls [6], understanding the relative roles One of the earliest soft modeling efforts to address the of the different attributes governing systems behavior is challenge of excessive variables between materials prop- the foundation for developing models Fig. Materials erties was the one done by Ashby, who showed that by design is a process that helps us determine the optimal merging phenomenological relationships in materials prop- combinations of material chemistry, processing routes, and erties with discrete data on specific materials characteristics, processing parameters to meet specific performance require- one can begin to develop patterns of classification of materi- ments robustly such as mechanical properties and corrosion als behavior [4]. The visualization of multivariate data was resistance. As an example, one such map an effect can be the result of many different causes. The second approach described in Section 3 materials behavior we approach it from a broader per- applies a set of computational strategies to represent pro- spective. By exploring all types of data, such as crystallo- censing and properties of data at a system level within a graphic, electronic, and mechanical data over a wide range unit-consistent framework, which also allows for a reliable of materials, that may have varying degrees of influence pruning of secondary effects. Materials Informatics Tools Fig. Associated with entities are attributes squares , which comprise features or properties such as molecular mass. Attributes take on particular values, and each entity can then correspond to a table in a database, so that the model specifies a schema for that database. Increasing connectivity can simply result in a proliferation of data, but at the level of classes of entities in underlying data models, additional connections increase the complexity of those models and resulting database designs from [6]. Please refer to the online version for color legends. PCA is a projection technique that can be used to han- The new axes, which are linear combinations of the original dle multivariate data that consists of many interrelated variables, are the PCs. By reducing the information dimen- The corresponding

eigenvalue-eigenvector pairs of Σ . PCA finds uncorrelated axes that create hyperplanes. Then, the i -th PC is given by \mathbf{v}_i . For bivariate datasets can be visualized through hyperplanes spanned in multidimensional space. Each PC is a suitable linear combination of all the original descriptors. The M -dimensional PC space has retained trace of the covariance matrix can be expressed as: $\sum_{i=1}^k \lambda_i \mathbf{v}_i \mathbf{v}_i^T$. PCA Methodology Then the total population variance due to the k -th PC is λ_k . Following the treatment of Johnson and Wichern [8], we can describe this mathematically as follows. When the variables have different ranges and are measured on different scales as is the case with most materials problems, they are standardized. The i -th PC is given by \mathbf{v}_i from these patterns. Thus, three PCs contain Geometrical explanations of PCA seen right plot of Fig. All the data points refer to different compounds and their spatial position indicates how they relate to each other when defined simultaneously by all the latent variables or parameters.

2. Application of PCA to Superconducting Behavior. A fundamental challenge in materials science is the discovery of new superconducting materials. One prominent example is the discovery of high temperature superconductors. Since the discovery of ceramic superconductors, our work, which includes the data since then, demonstrates the broader impact of the valency clustering criterion in superconductors. While the PC axes themselves in this case do not have a discrete physical meaning, other directions in the PC projections can. We will later show examples of how the PC axes can correlate directly to trends in a physical parameter. The loading plot Fig. Every as it is near the origin $(0, 0)$ position of the loading point is an attribute and their spatial correlations suggest the strength of the relative correlations between attributes on the component plot. However, parameters such as pseudopotential radii and ionization energy play an important but lesser role. The loading plot yielded: Note that for ionization energy and pseudopotential radii in terms of their PC3, the weighting coefficients for valency and electronegativity are the same 0 . Also the weighting coefficients for cohesive energy are very small 0 . The strong effect of valency on the linear pattern of clustering in the scoring plot is consistent with its large distance rather than transition temperature. The highest transition from the origin on the loading plot. In the above figure, compounds are the cupric oxides marked in light orange. DA is a standard tool used by many engineering and science disciplines. In essence, DA reduces the number of parameters in a problem by considering their units. For example, many fluid mechanics problems involve four parameters viscosity, density, length, and velocity, which, by using this technique, can be reduced to a single dimensionless parameter the Reynolds number. Thus, DA can significantly reduce the number of experiments necessary to characterize a system. While the application of DA to relatively simple systems is well understood, when a problem involves many parameters, DA typically yields too many dimensionless groups and so do not enhance understanding or intuitive interpretations. This problem of excess number of dimensionless groups is more often the rule rather than the exception shows the trend in transition temperature across the linear clusters. The peak corresponds to those with the highest recorded when modeling structure-properties relationships in materials. By comparison to the earlier scoring plot, and using different visualization scheme, we can now capture a more complete perspective of trends in this multivariate dataset. It is interesting to note that the more recent discovery of MgB_2 as a high temperature superconductor actually shows up in this plot small

use of DA in these fields. We will use the materi- indicating how this multidimensional analysis appears to capture als informatics approach to select the most representative the critical physics governing high Tc materials. Please refer to dimensionless groups in order to reduce the dimensionality the online version for color legends. Previous efforts to use DA to reduce the number of that dataset. This reduction in dimensionality now offers us adjustable parameters in regressions were pioneered by Li better opportunities to: The Artificial Intelligence community has also pro- fraction of the total set of dimensions. The materials informatics approach that we are pursuing combines elements of DA and elements of regressions and 3. First, the dimensionless Here we present an algorithm that combines a linear groups employed are generated by an algorithm instead of regression model of the experimental data with physical being postulated a priori ; second, it does not require integer considerations of the process; namely, the units of the exponents in the scaling laws; third, it allows for datasets in resulting model match the units of the dependent vari- which variables change value simultaneously; and fourth, it able. We look for the power law model that minimizes explicitly searches for the simplest predictive formulation the prediction error only among models that have the cor- using a heuristic formulation. DA generates results in the rect units. The output of the algorithm is a physically form of power laws. Power laws yield estimates in the form meaningful and simple power law, representing the pro- of a function of the problem parameters raised to constant cess and a set of dimensionless groups ordered by their exponents. For example, if L is a characteristic value of relevance to the problem. The user input in selecting the length in the x direction, La is a power law, while x a is simple model, and the ability to correct it further using not. Materials Informatics Tools Power laws are ubiquitous in engineering and science and in the variables e. The dimensionless are especially appealing to materials informatics because groups have the expression they can provide estimations for a whole family of systems. This way, outliers can be readily identified indicating either errors or physical phenomena If we approximate function f as a power law of the that had been disregarded but were relevant for that outlier. SLAW The constrained linear regression is performed in the is an algorithm designed to generate power laws from logarithmic space statistical data. The first assumption is that the target quantity can logarithmic scale. As discussed earlier, this is Considering p experimental observations of the phys- generally a good hypothesis. We denote the The second assumption is that the optimal dimensionality p observations of the target magnitude Y by y1 , A third assumption is that the exponents of the x1j ,. This reduces the effect of experimental independent identically distributed IID random variables. The estimate for the coefficients in model that minimizes If Y is the target magnitude that we want to model, DA the residual sum of squares is the solution to the system states that it can be represented exactly as: Each point represents a different ceramic to metal joint. In this problem, the dependent variable Y. In this case, residual sum of squares while satisfying the units constraint:

Chapter 4 : Rajan, Krishna - Materials Design and Innovation - University at Buffalo

K. Rajan and A. Rajagopalan "Informatics Based Optimization of Crystallographic Descriptors for Framework Structures", Combinatorial and High Throughput Discovery and Optimization of Catalysts and Materials, ed. W. Maier and R.A. Potyrailo (Boca Raton, FL: CRC Press,).

All of these technologies are linked by the general characteristics of molten salts that can function as solvents, have good heat-transfer characteristics, function like a fluid, can attain very high temperatures, can conduct electricity, and also may have chemical catalytic properties. The Janz molten salt database is the most comprehensive compilation of property data about molten salts available today and is widely used for both fundamental and applied purposes. These static data can be transformed by informatics and data mining tools into a dynamic dataset for analysis of the properties of the materials and for making predictions. While this approach has been successful in the chemical and biochemical sciences in searching for and establishing structure-property relationships, it is not widely used in the materials science community. Because the design of the original molten salt database was not oriented toward this informatics goal, it was essential to evaluate this dataset in terms of data mining standards. Two techniques were used—a projection principal components analysis PCA and a predictive method partial least squares PLS—in conjunction with fundamental knowledge acquired from the long-term practice of molten salt chemistry. By using PCA, the information contained thermodynamical, or physical—that may be used to de- in multiple dimensions was compressed into a more com- scribe any given chemistry Table I. Hence, the analysis pact space helping to identify patterns and outliers among of pre-existing empirical and theoretical data as well as the the data points and to study the connections between the virtual design of new materials is a multivariate problem. By using PLS, one physical property density Statistical analysis tools have to be used to find the unex- was predicted from other physical properties. Solving this class of materials problems requires an acceptable precision. The present work, while it made use of the same numer- This approach was applied for the first time to molten ical techniques, was concerned with systems for which salts[1] using the large data sets reported by Janz in several data were not available in the Janz database. Indeed, systematic thermodynamic investigations of alent conductance, and specific conductance, to character- molten salt systems showed that both the relative ionic ize the large number of compounds and their properties potential and the difference of electronegativity between chosen from the Janz molten salt database. The database cation and anion permitted categorization of melts, espe- was originally designed as a static compilation of materials cially those including ions with different charges. Two data analytical techniques were employed. University, Ames, IA The use of easier in this compressed space. PLS techniques is well established in many fields such as psychology, chemometrics, process control, biology, and economics. Descriptors for the Molten Salts Reference 1. The prediction was performed for the following com- pounds, named as the test set: The method was applied successfully both to divalent and trivalent rare earth halides. This test set is given in Table III. Figures 1 and 2 show predicted against experimental values of DH_{form} and DG_{form} , respectively, for the training set. All numbers in Figures 1 and 2 correspond to the compounds in Table II. The large R^2 values Shortly after these calculations, in our laboratory, we performed an experimental thermodynamic investigation on two compounds, $CeBr_3$ and $GdBr_3$. They are in good agreement with those ob- tained using the predictive model! Later on, the same procedure was applied on the entire temperature range from the melting temperature up to Fig. Some experimental measurements improved our prediction results and our dataset. The results show that carefully selected parameters and descriptors sometimes it may even include some general attributes such as atomic number or equivalent weight are most important for a modern data-mining base to be used for analysis and virtual materials design. Volume 2, Section 2. Alloys Compounds, , vol. Very satisfactory results were obtained, as indicated by vol. DG_{form} , and are presented in Figures 3 and 4. As an example, in Table IV are pp. Principal Component Analysis, Springer-Verlag, Berlin, given the equations that illustrate the temperature depend- London A, , We have provided examples of statistical approaches of vol. By using PCA, we manage multiple physical We also predict two important thermodynamic

