Accelerating Materials Discovery Based on Generalized Low-
dimensional Conformation Performance Relationships

Yuanyuan Qu1, Xiangying Meng1*, Zixi Jia2, Xiaofeng Liu3, Dongyan Liu1, Shanshan Li1 and Fang Bian1

1College of Sciences; 2Faculty of Robot Science and Engineering; 3Information Construction and Network Security Office; 3College of Information Science and Engineering, Northeastern University, Architecture Building, Room 115, No.3-11, Wenhua Road, Heing District, Shenyang, P. R. China, 110819
Email: x_y_meng@mail.neu.edu.cn

Abstract. The construction of conformation-performance relationships (CPRs) is the core issue of efficient material design. Different from traditional CPRs originating from experience summarizations or rational model cognitions, a strategy of generalized low-dimensional conformation-performance relationships (GLD-CPRs) for rapid material discovery has been established in this work by the aid of material information technology, providing us with a new perspective to material gene (MG) and its function on performance. From a pure computing perspective, the strategy has been concreted by a case study of semiconductor bandgap engineering. Based on more than 20,000 items of self-produced semiconductor bandgap and attributes data, a set of CPRs for bandgap regulation towards arbitrary semiconductor of the same kind has be generated using GLD-CPRs strategy. By comparing with the reported studies, the reliability of the strategy is confirmed. Finally, future improvements needed by GLD-CPRs are addressed.

1. Introduction
One principal goal of materials science is to understand conformation-performance relationships (CPRs) of material systems such that specific performances may be achieved via material engineering. For centuries, we have been accustomed to build traditional CPRs based on experience summarizations or rational model cognitions. Now, it is feasible to construct CPRs for producing materials and precisely designing their interiors to address the performance issues with the boom of high-throughput [1] and data processing technologies [2]. In general, a specific substance can be defined by its structure S, composition C, and ingredient I, and CPRs for a given performance is a 3D point-to-point function P(S, C, I) in the conformation-performance space. Since there has been much interest in accelerating the discovery of materials, it is necessary to fix one or two variables in P(S, C, I) to construct point-to-face or point-to-line low-dimensional CPRs (LD-CPRs) to narrow the searching space and thus speed the progress of materials screening.

Although not explicitly proposed, the concept of LD-CPRs has been implied in many successful data-processing-based materials designs. For example, Xue [3] et al determined the optimal composition in Ni_{50-x-y},Ti_{50},Cu_{x},Fe_{y},Pd_{z} low thermal hysteresis shape memory alloys based on the LD-CPR of thermal to composition ΔT(C)|S=R,B19,B19';I=Ni,Ti,Cu,Fe,Pd. Lin [4] et al obtained carbon-capture materials in 2.6 million zeolite-like structures of silica with the LD-CPR of parasitic energy to structure E(S)|I=Si,O; C=1:2. Castelli [5] et al predicted new photoelectrochemical cells with improved light absorption in ABO_{2}N perovskite oxy-nitrides by mapping the change of bandgap (BG)
with respect to A/B site cations BG(IA,B)|S=Perovskite; C=1:1. Different from traditional CPRs pursuing universality, data-processing-based LD-CPRs can be tailored for the specific material system and directly produces materials with merits or demerits for performance screening, making it more attractive in materials discovery.

Another cause for the rise of material data processing technology is the emergence of advanced experimental instruments. With this process, more and more sophisticated intrinsic attributes, namely material genes (MGs), are expected to involve in the material design to affect performances, while the traditional CPRs based on experiences or models are becoming less and less good at describing the role of personalized MGs. The character of multiple attributes of performance leads to a multivariate function of performance to attributes as \( P(x_1, \ldots, x_i) \), where \( x_i \)-set is a collection of various intrinsic attributes that may affect the material performance. Following the LD-CPRs strategy, the generalized low-dimensional CPRs (GLD-CPRs) need to be established to accelerate the process of material oriented design.

To fulfil the goal of GLD-CPRs, three key links, i.e. correlation analysis, weight evaluation and data mining, are necessary to guarantee the rationality and reliability of the strategy. Firstly, it is of great significance to evaluate the correlation between attributes to avoid overlapping representation of material genetic information. We can list arbitrary attributes as original MGs, and then construct nearly independent derived MGs through correlation analysis. Secondly, a weight evaluation mechanism should be established to assess the contribution of MGs to performance. The weight of genes on performance (WGP) is used to judge whether attributes are major or minor MGs. Thirdly, a mature machine learning algorithm ought to be adopted to produce a reliability regression model. The final GLD-CPRs between major MGs and performance in a specific material system will be excavated out based on the regression model, which serves to be material design atlases to guide the subsequent experiments.

The energy bandgap of semiconductor is a fundamental performance restricting light harvesting, conversion and transport technologies, such as photovoltaics [6] and photocatalysts [7]. In this work, the function, efficiency and reliability of GLD-CPRs strategy are embodied in the bandgap design of rutile-TiO\(_2\) binary semiconductors, with a target of screening effective visible light responsive semiconductors. More than twenty thousand items, covering multi-dimension attributes and bandgap, are produced by ourselves and used to train a mature reliable regression model based on data-preprocessing and random forest algorithm [8]. By evaluating WGP, the major MGs that fundamentally influence bandgap are picked out to build GLD-CPRs to disclose new facts and guide bandgap engineering. Our strategy is universal to the target optimization design of materials, also will arouse extensive interest and thinking in other research fields.

2. Details of Algorithms
The implementation of GLD-CPRs is schematically descripted in figure 1. In our own work, we have favoured the analysis of computational data and the validation of predictions through both reported experiments and calculations. Four modules including interface to high-throughput computing database, data pre-processing, regression model training and GLD-CPRs construction are integrated in the routine. In this work, data of attributes and bandgap are uniformly generated from workflows in the high-throughput computing platform AiiD\(_A\) [9]. An interface to the platform has been developed in our group to automatically download computational data and then push it to the subsequent data preprocessing procedure. To guarantee the reliability of regression model from data layer, data-prerecession is executed to convert data set to training set. By the aid of random forest machine learning algorithm, a mature reliable regression model for bandgap predicting is built. Finally, various constraints for accelerating target materials discovery are manifested through GLD-CPRs. As a result, more than twenty thousand cleaned bandgaps are prepared for the subsequent data training.

3. Results and Discussion
A mature and reliable regression model for bandgap predicting has been established through data processing technique and machine learning algorithm. Then the model is utilized to assess WGP and
yield GLD-CPRs. By taking bandgap designs of rutile-TiO₂ binary as a case study, we illustrate the function of the strategy.

Figure 1. Implement of GLD-CPRs strategy in bandgap design

The big data from AiiDA high-throughput computing platform is divided into two parts: original attributes (MGs) and performance (bandgap). Bandgap cleaning process produces a set of non-zero averaged bandgap. After data pre-processing, a clean training set is obtained and input into the random forest machine learning algorithm for building a reliable regression model. By evaluating the weight of MGs on performance (WGP), major MGs are picked out to establish corresponding GLD-CPRs to suggest new experiments, with the triple goals of model improvement, mechanism recognition and materials discovery.

3.1. General Evaluation of WGP for Binary Semiconductors

In figure 2, there is a clear division between the first five MGs and the rest MGs. The first five MGs, each weights twice the average weight and total weights greater than 60%, are considered to be major MGs affecting bandgap profoundly. The rest 18-dimension MGs, their respective weight is close to or largely less than the average weight are thus regarded as minor MGs affecting bandgap weakly. As expected in the section of PCC analysis, we found that the weights of some derived MGs have suppressed that of original MGs.

We are more concern about the feedback of WGP and its impact on our physicochemical cognition of materials. First of all, bandgap is found to be a localized electronic property and most sensitive to the nearest interaction distance between cations and anions. This finding is consistent with the understanding of the origin of electronic energy bands in solids, because the electronic bands are
formed by the overlap of atomic orbitals and the extent of overlap is determined by the interatomic distance.

In the second place, the deference between ionization energy and affinity energy of ingredients, namely $IP_i - EA_j$ or $IP_j - EA_i$, is found to be the major MGs significantly affecting the bandgap. Here WGP evaluation provides a basis for the widely used model of approximating the bandgap of an infinite periodic solid with the atomic energy level difference of a finite ingredient cluster. For the finite system, the first ionization potential ($IP$) and electron affinity energy ($EA$) of ingredients respectively defines their highest occupied atomic orbitals (HOAO) and the lowest unoccupied atomic orbitals (LUAO). WGP informs us that the real bandgap $E_g$ of binary clusters can be approximated by $E_{ga} = IP_i - EA_j$, where subscript $i$ and $j$ represent different kinds of ingredients.

![Figure 2. WGP histogram of bandgaps for binary semiconductors](image)

The weights of total 23-dimention MGs on bandgap are sorted according to the size of their contributions. The blue column represents the weight of original MGs and the red one is the weight of derived MGs.

More significantly, WGP shows that bandgap is largely related to the valence electron structure of cation and electronegativity of anion, which suggests an effective way to regulate the bandgap by controlling ingredients. Interestingly, although being the same physical properties, valence electron number of anions and electronegativity of cations are minor-MGs that affect bandgap weakly. Therefore, the influence of ions on the bandgap will vary with their roles, that is, the same element will have different effects on the bandgap with different cationic or anionic identities.

Through above data analysis, the validity and importance of WGP evaluation are verified by checking with the existing knowledge. As long as reasonable MGs and reliable regression model are constructed, WGP can provide us with strong supports for the existing axioms, profound understanding of phenomena and new research topics and directions.

### 3.2. Construction of Specific GLD-CPRs
Since major-MGs have been distinguished from minor-MGs by WGP evaluation, it is critical to reveal the quantitative relationships between bandgap and major MGs by GLD-CPRs. In our strategy, this process needs to be realized by means of dimensionality reduction, that is, the relation of $P(x_i | x_{jai} = \text{constant})$ need to be clearly expressed based on the bandgap regression model, in which the multivariate function of performance versus attributes $P(x_i, ..., x_n)$ has been implicated.

Valence electron structure of cation and electronegativity of anion are two major MGs affecting bandgap profoundly according to WGS evaluation. By the aid of GLD-CPRs between bandgap and the
two quantities, bandgaps of rutile-TiX₂ and rutile-XO₂ (X represents the specie of cations or anions) are presented in figure 3. Based on the results, we can predict the effect of substitutional doping in rutile-TiO₂.

With the aim of reducing bandgap, there are much more rooms for cation-doping than that for anion doping. For cationic doping, the GLD-CPR suggests that rare-earth elements (from Tb to Lu), d-filled transitional metals (Cu, Zn, Ag, Cd) and post-transitional metals (Ga, Ge, As, In, Sn, Sb) are promising dopments for reducing the bandgap of rutile-TiO₂, while alkali, alkaline-earth and 3d- or 4d- unfilled transition dopments have a fat chance in reducing the bandgap of rutile-TiO₂. As for anionic doping, the GLD-CPRs prompt that B, C, N, P and S are potential dopments for the bandgap reduction of rutile-TiO₂, but doping halogens (F, Cl, Br, I) in rutile-TiO₂ is not a good choice. These enlightenments from GLD-CPRs are basically coincident with the reported facts [10], indicating that the strategy is generally reasonable and convincing. Further experimental verification and deep-seated explorations should be aroused by these predictions.

![Figure 3. Bandgap variation induced by ingredient change in rutile structure](image)

The bandgap of intrinsic rutile-TiO₂ is set to zero. The bandgap spectrums of XO₂ (blue curve) and TiX₂ (green curve) in the same lattice prototype as rutile structure are predicted based on GLD-CPRs, which provides a guidance for the bandgap regulation of rutile by means of element doping.

4. Future Perspectives
There is no doubt that the prospects of material informatics are promising, but at present the method is far from mature. Aiming at discovering materials with expected performances rapidly and correctly, a complete procedure, including data-generation, data-management and data-utilization, should be established. In these aspects, the reliability of attributes construction, data-cleaning and classification, regression model training and CPRs strategy is directly related to the effectiveness and practicability of material informatics method. In the GLD-CPRs frame, there are two main causes that lead to the uncertainty of the strategy and need to be further refined.

Firstly, attributes (MGs) are not orthogonal to each other as independent variables, so adiabatic approximation is introduced into the model of GLD-CPRs. Adiabatic approximation is reasonable for those weakly correlated MGs, but for strongly correlated MGs, orthogonalization scheme need to be developed to obtain as complete a set of attributes as possible. During the process, the physical analysis of those new emerging derivative genes is challenging but worth looking forward to. In this regard, one of the main tasks of GLD-CPRs strategy is to approach a complete set of MGs, and each element of the set has a clear physical meaning towards specific performance requirements.

Secondly, the reliability of GLD-CPRs relies on the accuracy of regression model, and machine-learning models are only as good as the data used to construct them. Therefore, the model, algorithm...
and parameter adjustment of machine learning need to be protected at data layer. In addition, the current material design based on machine learning method adopts a “forward” strategy, which means that the performances of unknown materials are predicted by those known materials data, and then the expected target materials are selected by exhaustive method. In the future, it is necessary to further develop “reverse” machine learning technology, that is, by setting the performance requirements, target material systems that meet the criteria are autonomously generated by machine learning method itself. This will be a challenging but significant work.

5. Summary
Data processing technology has shown its great effectiveness in the design of structural and functional materials. At present, the complementary advantages between material data technology and traditional material rational design have emerged. However, material data technology now exists as a “tools” rather than a “field” of material science research. The main reason lies in that it has not formulated a standard data research strategy from the perspective of material design to raise specific scientific issues.

In this work, the strategy of GLD-CPRs for material design has been established based on material informatics method. In the framework of the strategy, performance is a multivariate function of material genes, and constraints imposed on material system are concreted by GLD-CPRs that are used to narrow the searching space and thus speed the materials discovery. With the help of machine learning algorithm, the proposed strategy highlights the construction of GLD-CPRs, adds new depth and width to our understanding of material gene on performance, thus makes materials design more fast, reasonable, and traceable. This strategy offers a universal solution to the efficient material oriented design, as required by the material genome engineering. Facing the challenge of GLD-CPRs, it is hopeful to broaden the research field of materials science.

6. Acknowledgments
This work is financially supported by the National Key Research and Development Program of China 2016YFB0701100, the Fundamental Research Funds for the Central Universities of China (N160208001) and National Natural Science Foundation of China (No. 51525101).

7. References
[1] Curtarolo S, Hart G, Nardelli M, Mingo N, Sanvito S and Levy O 2013 Nat. Mater. 12 191.
[2] Nosengo N 2016 Nature 533 22.
[3] Xue D, Balachandran P, Hogden J, Theiler J, Xue D and Lookman T 2016 Nat. Commun. 7, 11241.
[4] Lin L, Berger A, Martin R, Kim J, Swisher J, Jariwala K, Rycroft C, Bhown A, Deem M, Haranczyk M and Smit B 2012 Nat. Mater. 11, 633.
[5] Castelli I, Olsen T, Datta S, Landis D, Dahl S, Thylesen K and Jacobsen K 2012 Energy Environ. Sci. 5, 5814.
[6] Yu P, Wu J, Liu S, Xiong J, Jagadish C and Wang Z 2016 Nano Today 11 704.
[7] Fujishima A and Honda K 1972 Nature 238 37.
[8] Ho T 1998 IEEE T. Pattern Anal. 20 832.
[9] Pizzi G, Cepellotti A, Sabatini R, Marzari N and Kozinsky B 2016 Comp. Mate. Sci. 111, 218.
[10] Stengl V, Bakardjieva S and Murafa N 2009 Mater. Chem. Phys. 114 217.