JAIST Repository
https://dspace.jaist.ac.jp/

Title	Data mining for materials design: A computational study of single molecule magnet
Author(s)	Dam, Hieu Chi; Pham, Tien Lam; Ho, Tu Bao; Nguyen, Anh Tuan; Nguyen, Viet Cuong
Citation	The Journal of Chemical Physics, 140(4): 044101
Issue Date	2014-01-23
Type	Journal Article
Text version	author
URL	http://hdl.handle.net/10119/12154
Rights	This is the author's version of the work. Copyright 2014 American Institute of Physics. This article may be downloaded for personal use only. Any other use requires prior permission of the author and the American Institute of Physics. The following article appeared in Hieu Chi Dam, Tien Lam Pham, Tu Bao Ho, Anh Tuan Nguyen and Viet Cuong Nguyen, The Journal of Chemical Physics, 140(4), 044101 (2014) and may be found at http://dx.doi.org/10.1063/1.4862156
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Data Mining for Materials Design: A Computational Study of Single Molecule Magnet

Hieu Chi Dam[1,2], Tien Lam Pham[1], and Tu Bao Ho[1]
[1] *Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*

Anh Tuan Nguyen[2]
[2] *Faculty of Physics, Vietnam National University, 334 Nguyen Trai, Hanoi, Vietnam*

Viet Cuong Nguyen[3]
[3] *HPC Systems, Inc., 3-9-15 Kaigan, Minato-ku, Tokyo 108-0022, Japan*
(Dated: December 31, 2013)

We develop a method that combines data mining and first principles calculation to guide the designing of distorted cubane $Mn^{4+}Mn_3^{3+}$ single molecule magnets. The essential idea of the method is a process consisting of sparse regressions and cross-validation for analyzing calculated data of the materials. The method allows us to demonstrate that the exchange coupling between Mn^{4+} and Mn^{3+} ions can be predicted from the electronegativities of constituent ligands and the structural features of the molecule by a linear regression model with high accuracy. The relations between the structural features and magnetic properties of the materials are quantitatively and consistently evaluated and presented by a graph. We also discuss the properties of the materials and guide the material design basing on the obtained resutls.

PACS numbers: 31.15.E-, 75.50.Xx, 02.70.-c

I. INTRODUCTION

Quantum calculation plays a very important role in the process of materials design nowadays. For a material with a given hypothesized structural model, the electronic structure, as well as many other physical properties can be predicted by solving the Schrödinger equation. Conventionally, the ground state's potential energy of a material is calculated using atomic positions in the hypothesized structure model. By optimizing the ground state's potential energy, the optimal structure can be derived. The features of an optimal structure model of materials, as well as its derived physical properties, results in a series of optimizing processes, and in addition has strong multivariate correlations. The task of materials design is to make these correlations clear and to determine a strategy to modify the materials to obtain desired properties. However, such correlations are usually hidden and difficult to uncover or predict by experiments or experience. As a consequence, the design process is currently performed through time-consuming and repetitive experimentation and characterization loops, and to shorten the design process is clearly a big target in materials science. In an effort to improve on existing techniques, we propose a first principle calculation-based data mining method and demonstrate its potential for a set of computationally designed single molecular magnets with distorted cubane $Mn^{4+}Mn_3^{3+}$ core (Mn_4 SMMs).

Data mining is a broad discipline that aims to develop and use methods for extracting meaningful information and knowledge from large data sets. To the field of computational materials science, data mining methods have recently been used with successes, for example, in solving Fokker-Planck stochastic differential equations [1], in predicting crystal structure and discovering new materials [2, 3], in parametrizing interatomic force fields for fixed chemical composition [4, 5], and in predicting molecular atomization energies [6, 7] by merging data mining with quantum calculations. Motivated by using data mining to solve data-intensive problems in materials science, we develop a method to quantitatively model a family of materials by graph, using their quantum calculated data. The key idea of our method is to use advanced statistical mining algorithms, in particular multiple linear regression with LASSO regularized least-squares [8, 9] to solve the *sparse* approximation problem on the space of structural and physical properties of materials. We use cross-validation [10] to consistently and quantitatively evaluate the conditional relations of each feature on to *all* the other features in terms of prediction. Based on the obtained relations, a graph representing relations between all properties of materials can be constructed. Furthermore, we propose a graph optimization method to have better visual representation and easier inferences on the controlling features of the materials. The obtained graph is not only significant for the comprehension of the physics relating to the materials, but also valuable for the guidance of effective material design.

The main contribution of this work includes: (1) a quantitative and rational solution to the modeling of the structural and physical properties of the distorted cubane $Mn^{4+}Mn_3^{3+}$ SMMs; (2) a first principles calculation-based data mining approach that can be applied to accelerate the understanding and designing of materials.

II. MATERIAL SYSTEM

In this paper, we focus on single-molecule magnets (SMMs) which are recently being extensively studied due

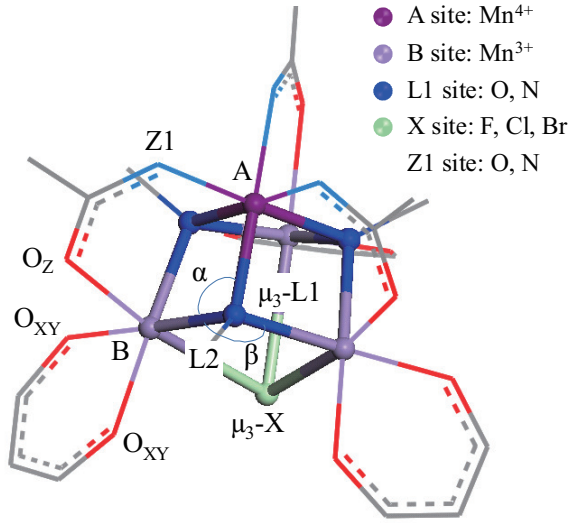

FIG. 1: Schematic geometric structure of $[Mn^{4+}Mn_3^{3+}(\mu_3\text{-}L)_3^{2-}(\mu_3\text{-}X)^{-}Z_3^{-}(CH(CHO)_2)_3^{-}]$ molecules, with L = L1L2, Z = $(CH_3COZ1)_3Z2$, Z1$_3$-Z2 = O_3 or N_3-$(CCH_2)_3CCH_3$. Color code: Mn^{4+} (violet), Mn^{3+} (purple), L1 (blue), X (light green), Z1 (light blue), C (grey). H atoms and Z2 group are removed for clarity.

to their potential technological applications in molecular spintronics [11–16]. SMMs can function as magnets and display slow magnetic relaxation below their blocking temperature (T_B). The magnetic behavior of SMMs results from a high ground-state spin combined with a large and negative Ising type of magnetoanisotropy, as measured by the axial zero-field splitting parameter [17–19].

SMM consists of magnetic atoms connected and surrounded by ligands, and the challenge of researching SMM consists in tailoring magnetic properties by specific modifications of the molecular units. The current record of the T_B of SMMs is only several degrees Kelvin, which can be attributed to weak intra-molecular exchange couplings between magnetics metal ions [16]. The design and synthesis of SMMs with higher T_B that are large enough for practical use, are big challenges for chemists and physicists. In the framework of computational materials design, the SMM with distorted cubane $Mn^{4+}Mn_3^{3+}$ core is one of the most attractive SMM systems because their interesting geometric structure and important magnetic quantities can be well estimated by first-principles calculations [14, 15].

In this paper, we construct and calculate a database of structural and physical properties of 114 distorted cubane $Mn^{4+}Mn_3^{3+}$ SMMs with full structural optimization by first-principles calculations (Fig. 1). A data mining method is applied to the calculated data to explore the relation between structural and physical properties of the SMMs. We quantitatively model the structural and physical properties of the SMM by a graph that allows us to infer and to guide the molecular design process (Fig. 2).

1. Construct molecular structural models of SMMs and carry out first principles calculation to optimize the molecular structures.
2. Calculate structural, chemical, and physical property features using the optimized molecular structures. Use these features to represent all the constructed molecules in a feature space.
3. Take each feature as a response feature and predict it by a regression analysis using the other features.
4. Evaluate quantitatively the impact of each feature on the prediction accuracy of the regression analysis of the other features.
5. Build a directed graph with features as nodes and their impacts on other features as edges to represent the whole picture of the relation between features.
6. Simplify the obtained graph by removing unnecessary features for specific materials design purposes.

FIG. 2: Framework of first principle calculation based-data mining to model the physical properties of SMMs.

III. METHODOLOGY

A. Data generation

1. *Molecular structure construction*

New distorted cubane $Mn^{4+}Mn_3^{3+}$ SMMs have been designed by rational variations in the μ_3-O, μ_3-Cl, and O_2CMe of the synthesized distorted cubane $Mn^{4+}Mn_3^{3+}(\mu_3\text{-}O^{2-})_3(\mu_3\text{-}Cl^{-})(O_2CMe)_3^{-}(dbm)_3^{-}$ (here after Mn_4-dbm) molecules [20–24].

In Mn_4-dbm molecules, the μ_3-O atoms form Mn^{4+}-$(\mu_3\text{-}O^{2-})$-Mn^{3+} exchange pathways between the Mn^{4+} and Mn^{3+} ions. Therefore, substituting μ_3-O with other ligands will be an effective way to tailor the geometric structure of exchange pathways between the Mn^{4+} and Mn^{3+} ions, as well as the exchange coupling between them.

To preserve the distorted cubane geometry of the core of $Mn^{4+}Mn_3^{3+}$ molecules and the formal charges of Mn ions, ligands substituted for the core μ_3-O ligand should satisfy the following conditions: (i) To have the valence of 2; (ii) The ionic radius of these ligands must be not so different from that of O^{2-} ion. From these remarks, nitrogen-based ligands, NR (R = a radical), must be the best candidates. Moreover, through variation in the R group, the local electronic structure as well as electronegativity at the N site can be controlled. As a consequence, the Mn-N bond lengths and the Mn^{4+}-N-Mn^{3+} angles (α), as well as delocalization of d_{z^2} electrons from the Mn^{3+} sites to the Mn^{4+} site and the exchange coupling between them (J_{AB}) are expected to be tailored. In addition, through variations in the core μ_3-Cl ligand and the O_2CMe ligands, the local electronic structures at Mn sites are also changed. Therefore, combining variations

in μ_3-O, μ_3-Cl, and O_2CMe ligands is expected to be an effective way to seek new superior $Mn^{4+}Mn_3^{3+}$ SMMs with strong J_{AB}, as well as to reveal magneto-structural correlations of $Mn^{4+}Mn_3^{3+}$ SMMs. By combining variations in μ_3-O, μ_3-Cl, and O_2CMe ligands, 114 new $Mn^{4+}Mn_3^{3+}$ molecules have been designed. For a better computational cost, the dbm groups are substituted with $CH(CHO)_2$ groups, which shows no structural and magnetic properties change after the subtitution [25, 26]. The designed molecules have a general chemical formula $[Mn^{4+}Mn_3^{3+}(\mu_3\text{-}L^{2-})_3(\mu_3\text{-}X^-)Z_3^-(CH(CHO)_2)_3^-]$ (hereafter Mn_4L_3XZ) with L = O, NH, NCH_3, NCH_2-CH_3, NCH=CH_2, NC≡CH, NC_6H_5, $NSiH_3$, NSiH=CH_2, $NGeH_2$-GeH_3, NCH=SiH_2, NSiH=SiH_2, $NSiH_2$-CH_3, NCH_2-SiH_3, $NGeH_2$-CH_3, NCH_2-GeH_3, $NSiH_2$-$GeCH_3$, $NGeH_2$-SiH_3, or $NSiH_2$-SiH_3; X = F, Cl, or Br; and Z_3 = $(O_2\text{-}CMe)_3$ or $MeC(CH_2\text{-}NOCMe)_3$. Details of the constructed SMMs can be found elsewhere [12–15, 25, 26].

2. *Molecular structure optimization*

The constructed molecular structures were optimized by using the same computational method as in our previous paper[25, 26]. All calculations have been performed at the density-functional theory (DFT) level [27] by using $DMol^3$ code with the double numerical basis sets plus polarization functional (DNP) [28, 29]. For the exchange correlation terms, the revised generalized gradient approximation (GGA) RPBE functional was used [30]. All electron relativistic was used to describe the interaction between the core and valence electrons [31]. The real space global cutoff radius was set to be 4.7 Å for all atoms. The spin unrestricted DFT was used to obtain all results presented in this study. Since the experimental results reported so far indicate the colinearity of the magnetic properties of the materials, all the DFT calculations are carried out within a collinear magnetic framework [22, 32, 33]. The atomic charge and magnetic moment were obtained by using the Mulliken population analysis [34]. For better accuracy, the octupole expansion scheme is adopted for resolving the charge density and Coulombic potential, and a fine grid is chosen for numerical integration. The charge density is converged to 1×10^{-6} a.u. in the self-consistent calculation. In the optimization process, the energy, energy gradient, and atomic displacement are converged to 1×10^{-5}, 1×10^{-4}, and 1×10^{-3} a.u., respectively. In order to determine the ground-state atomic structure of each $Mn^{4+}Mn_3^{3+}$ SMM, we carried out total energy calculations with full geometry optimization, allowing the relaxation of all atoms in molecules.

3. *Data representation*

One of the most important ingredients for data mining is the choice of an appropriate data representation that reflects prior knowledge of the application domain, i.e., a model of the underlying physics. For representing structural and physical properties of each distorted cubane $Mn^{4+}Mn_3^{3+}$ SMMs, we use a combination of 17 features. We divide all the features into four groups. The first group pertains to the features for describing the electronic properties of the constituent ligands, including (1) electron negativity of X (χ_X), (2) electron negativity of $L1$ (χ_{L1}), (3) electron negativity of $Z1$ (χ_{Z1}) [35, 36], (4) electron affinity of L (E_L^{EA}) [37]. The selection of these features comes from the physical consideration that the local electronic structures, as well as electron negativities at ligand sites, will determine the d orbital splitting at Mn ion sites. Furthermore, since we intentionally vary ligand groups, these electronic features are just considered as explanatory features in the following analysis process.

To have a good approximation of the physical properties of SMMs, it is natural to introduce intermediate features. From the domain knowledge, we know that information on molecular structure, such as bond length, bond angle, and structure of octahedral sites, is very valuable in relation to understanding the physics of molecular materials with transition metal. Therefore, we design the second group with structural features which represent the core structure and the structures of the octahedral fields at A and B sites. The features for the core structures are: (5) the distance between the A site and B site (d_{AB}), (6) the distance between B sites (d_{BB}), (7) the distance between the A site and $L1$ site (d_{AL1}), (8) the distance between the B site and $L1$ site (d_{BL1}), (9) the angle $\angle AL1B$ (α), and (10) the angle $\angle BL1B$ (β). The features for the structures of octahedral fields at A and B sites are (11) the distance between the A site and $Z1$ (d_{AZ1}), (12) the distance between the B site and O_{xy} ($d_{BO_{xy}}$), and (13) the distance between the B site and O_z (d_{BO_z}). These features are calculated from the optimized molecular structure and considered as structural intermediate features.

The third group of features includes (14) the magnetic moment of Mn^{4+} ion at site A (m_A) and (15) the magnetic moment of Mn^{3+} ions at site B (m_B). These two features are magnetic intermediate features. The last group includes targeting magnetic properties, which are (16) exchange coupling between Mn^{4+} and Mn^{3+} ions at sites A and B (J_{AB}/k_B), and (17) exchange coupling between Mn^{3+} ions at sites B (J_{BB}/k_B). The magnetic moments of the Mn ions are calculated by the Mulliken method. The exchange coupling parameters of the molecules are calculated by using the total energy difference method. Details of the calculation method are described elsewhere [25, 26, 38]. It should be noted that the features in the first group are the only features that can be obtained at a very low cost, without first principles calculations.

B. Data analysis

1. Parallel Regression

We perform a parallel regression process on the calculated data. With each feature, we perform a regression in which the feature we are focusing on is considered as a response variable, and the other features are considered as explanatory variables. The response variable is expressed as a linear combination of selected explanatory variables (from all availables) that have the lowest prediction risk. The main purpose of this regression is to extract a set of features that are sensitive in predicting the value of the feature we are focusing on. Commonly, regression methods use the least-squares approach. However, for the sparse data with ill condition, it is often the case that a bias-variance tradeoff must be considered to minimize the prediction risk. For this purpose, in the regression process, the LASSO regularized least-squares has been applied [8, 9].

In a standard regression analysis, we solve a least-squares problem, that minimizes

$$\frac{1}{m}\sum_{i=1}^{m}(y_i^{predict}-y_i^{obs})^2,$$

where m is the total number of samples in the data set; $y_i^{predict}$ and y_i^{obs} are the predicted and the measured values, respectively. The predicted values $y_i^{predict}$ are calculated from the linear regression function:

$$y_i^{predict}=\sum_{j=1}^{n}\beta^j x_i^j+\beta^0,$$

where n is the total number of variables considered in the regression model; x_i^j represents the value of the explanatory variable j for the sample i, and β^j are the sought coefficients corresponding to explanatory variable j, which determines how the explanatory variables are (optimally) combined to yield the result $y_{predict}$. In LASSO regularized least-squares regression [8], we minimize the penalized training error with ℓ_1-norm of regression coefficients:

$$\frac{1}{m}\sum_{i}(y_i^{predict}-y_i^{obs})^2+\lambda\sum_{j=1}^{n}\left|\beta^j\right|.$$

To estimate the prediction risk, we do not use the training error $\frac{1}{m}\sum_{i\in training}(y_i^{predict}-y_i^{obs})^2$, since it is biased. Instead we use leave-one-out cross-validation. In this validation, one sample (ith sample) is removed and the remaining $m-1$ samples are used for training the regression model. The removed sample (ith sample) is used to test and calculate the test error $(y_{i-left}^{predict}-y_{i-left}^{obs})^2$. The process is repeated m times for every sample, so that every sample has a chance to be the removed once. Finally we take the average of the test errors:

$$\hat{R}(\lambda)=\frac{1}{m}\sum_{i}(y_{i-left}^{predict}-y_{i-left}^{obs})^2,$$

where the sum is taken over all the $m folds$ in the cross-validation. We use it as a measure for the prediction risk, and the value of λ will be tuned to minimize this prediction risk. The explanatory variables of which the corresponding coefficients β^j are non-zero, are considered as sensitive explanatory variables to the response variable in the regression. By using the LASSO, we can assess the relation between the features we used for the data representation.

To evaluate quantitatively the relation between a specific sensitive explanatory variable x_j and the response variable, we carry out again the procedure of regression and prediction risk estimation by a leave-one-out cross-validation, using all but one (x_j) sensitive explanatory variables. The prediction risk $\hat{R}_j$ obtained from this procedure reflects quantitatively how the prediction of the response variable is impaired by removing the concerning variable x_j. In the case of weak correlation between explanatory variable x_j and the response variable, the prediction risk must not change much and $\hat{R}_j \simeq \hat{R}_{opt}$. On the other hand, if the explanatory variable x_j has a strong relation with the response variable, the removal of x_j from the set of sensitive explanatory variables for the regression will impair the model for prediction, and therefore, dramatically increase the prediction risk and $\hat{R}_j \gg \hat{R}_{opt}$. Another consideration is that if the score s_{total} [39] of a regression for all samples using all the sensitive explanatory variables is low, the linear relation between every explanatory variable and the response variable must be poor. Therefore, we normalize the prediction risk $\hat{R}_j$ with considering the total score s_{total} by:

$$I_j = s_{total}\times\frac{\hat{R}_j}{\sum_i \hat{R}_i},$$

and use these values to quantitatively evaluate the relative impact of a sensitive explanatory variable to the response variable. The I_j can take a value between 0 and 1, and the sum of all I_j is s_{total}. The I_j with a larger value indicates the higher impact of the explanatory variable j to the response variable. The impacts of the other non-sensitive variables to the response variable are set to 0. This procedure is repeated for every feature and we can obtain the relations (in terms of sensitivity for prediction) between every pair of features. It should be noted that the difference in prediction risk is estimated in the context that all the other sensitive explanatory variables are used in the regression model. Therefore, the obtained relative impact of a sensitive explanatory variable on the response variable should be different from simple correlations between two variables. In other words, the relation between each pair of features is evaluated with the consideration of all the other relations.

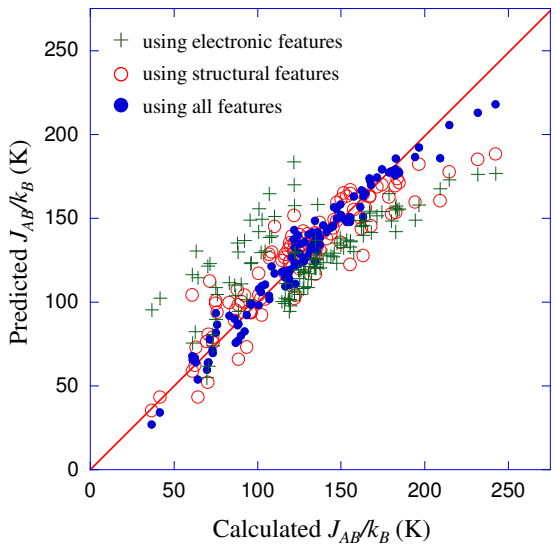

FIG. 3: Calculated (by DFT) and predicted (by data mining) exchange couplings J_{AB}/k_B for 114 distorted cubane $Mn^{4+}Mn_3^{3+}$ single molecular magnets. The green crosses represent the results of a linear regression using electronic features. The red circles represent the results of a linear regression using structural features α, d_{AB}, and d_{BB}. The blue solid circles represent the results of a linear regression using electronic features and structural features together. The red line represents the ideal correlation between calculated and predicted results.

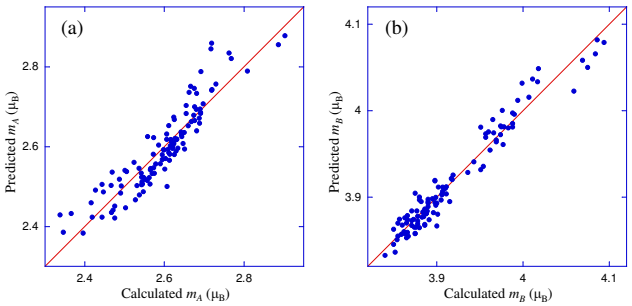

FIG. 4: Calculated (by DFT) and predicted (by data mining) magnet moments of Mn^{4+} ion at site A and Mn^{3+} ion at sites B (m_A and m_B) for 114 distorted cubane $Mn^{4+}Mn_3^{3+}$ single molecular magnets. The red line represents the ideal correlation between calculated and predicted results.

2. *Modeling relations between features by graph*

From the obtained relations we can build a directed graph in which nodes are features and edges are the relations between features, thus representing the whole picture of the relations between the features. Directions of edges are from response variables to explanatory variables in the regression. For the purpose of materials design, we added weights to the edges with the values of the obtained relative impacts of the sensitive explanatory variable on the response variable. Further, the edges are assigned with colors (red and blue) to differentiate the respective positive and negative correlations between variables which can be extracted from the corresponding coefficients in the linear regression models.

The relation between features can be asymmetric, therefore there may be two edges with vice versa direction and different weights (the relative impact I_j) between two nodes. It should be noted that Bayesian network is another choice for modeling the relations between features by a graphical model. However, automatical learning of a graph structure from data for a Bayesian network is an extremely heavy task. In contrasts, with this method a structure together with parameters of the network can be automatically derived from data at the same time with a parallelism [40].

We repeat the following steps to simplify the obtained graph: (1) remove all independent features that are not sensitive to any other features; (2) remove all intermediate features that are not sensitive to any other features; (3) remove an intermediate feature that can be predicted perfectly (regression score $\simeq 1$) by using the other features that are not sensitive to targeting magnetic properties features; (4) then recreate the graph using the remaining features. Steps (1) and (2), remove features that do not make sense in the prediction of the targeting magnetic properties. Step (3) removes unnecessary intermediate features. Features are removed one by one, and step (4) preserves the consistency of the outcome graph.

IV. RESULTS AND DISCUSSIONS

A. Magnetic property prediction

We first examine whether the exchange coupling J_{AB}/k_B can be directly predicted from electronic properties (features (1) - (4)) of the constituent ligands. Only a rough linear regression with an average relative error of more than 25% ($R < 0.6$) is obtained for the exchange coupling J_{AB}/k_B by using χ_X, χ_{L1}, χ_{Z1}, and E_L^{EA} as explanatory variables. This result indicates that it is hard to observe a simple linear correlation between the magnetic properties and the electronic properties of the constituent ligands for the SMMs. However, it should be noted that this result does not mean that the exchange coupling J_{AB}/k_B of the SMMs has no correlation with the electronic properties of the constituent ligands. It will be a great interest if these correlations appear when we take the other features into account.

Next, the relation between the exchange coupling J_{AB}/k_B and the geometrical structures of SMMs are studied. A linear regression using structural features (features (5) - (13)) is performed. It is found that the exchange coupling J_{AB}/k_B can be predicted quite well by a linear model using α, d_{AB}, and d_{BB} with an average relative error of 11% ($R = 0.9$). This result implies

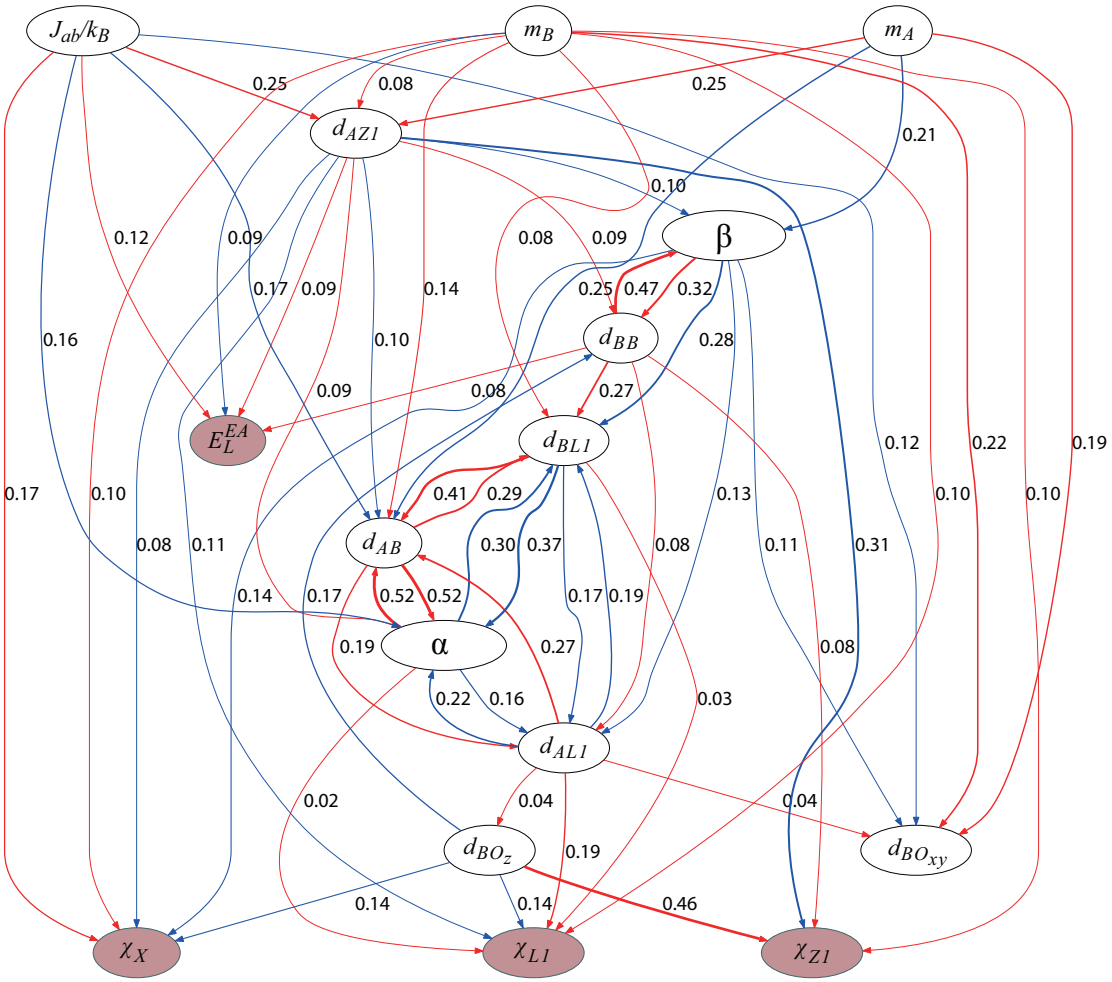

FIG. 5: The graph represents all relations between the features. Brown nodes and white nodes indicate independent and dependent features, respectively. Red edges and blue edges indicate positive and negative correlation, respectively. The arrows are from response variables to explanatory variables. The edges are plot with pen-widths in proportion to the values of the corresponding relations.

that the geometrical structure of the distorted cubane $Mn^{4+}Mn_3^{3+}$ core is the determinant factor for the magnetic properties of the SMMs. The prediction accuracy of the regression is dramatically improved when we take together the electronic properties of ligands into account. With a linear model using α, d_{AB}, d_{AZ1}, $d_{BO_{xy}}$, χ_X, and E_L^{EA}, the exchange coupling J_{AB}/k_B of SMMs can be predicted accurately with an average relative error of less than 5% ($R = 0.98$) (Fig. 3).

From this result, it is obvious that the electronic properties of the constituent ligands strongly correlate with the geometrical structure factors, and all of these features cooperatively contribute to the determination of the exchange coupling J_{AB}/k_B. Furthermore, it is interesting that the features representing the structures of octahedral fields at the A and B sites (d_{AZ1} and $d_{BO_{xy}}$) become strongly sensitive in the prediction of J_{AB}/k_B when the electronic features are considered. This result implicitly shows the relations between d_{AZ1}, $d_{BO_{xy}}$, and the electronegativities of constituent ligands which are well known in the ligand field theory with the effect of d orbital splitting [41].

Similar analyses are done for the other magnetic properties. The obtained results show that exchange coupling J_{BB}/k_B can not be predicted by a linear regression model using the features. This result can be explained by the facts that the exchange coupling J_{BB}/k_B is derived from a complicated formula of the total energies of three magnetic states of SMMs including the antiferromagnetic state, the ferromagnetic state, and the mix state (in which the Mn ion at the A site is ferromagnetically coupled to a Mn ion at the B site, and both of them are antiferromagnetically coupled to the other two Mn ions at the B site)[38]. The constituent ligands (especially ligand L) involved in both the magnetic interaction between Mn ions at the A and B sites, and the magnetic interaction between Mn ions at the the B sites. Further, the value of the exchange coupling J_{BB}/k_B is one order smaller than that of the exchange coupling J_{AB}/k_B. The design for new features that are more informative

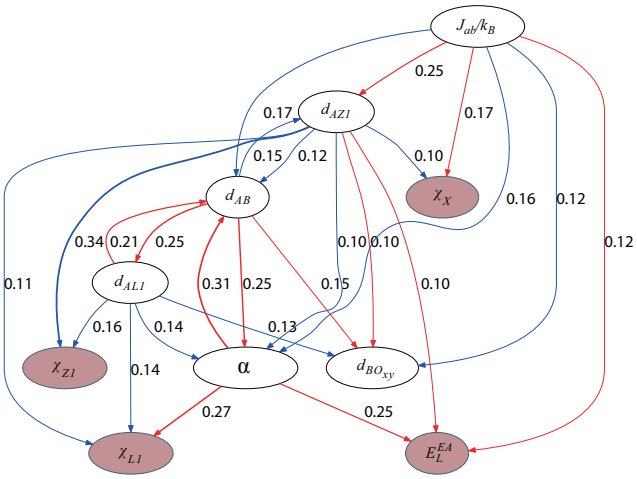

FIG. 6: The simplified graph represents the relations between selected features. Brown nodes and white nodes indicate independent and dependent features, respectively. Red edges and blue edges indicate positive and negative correlation, respectively. The arrows are from response variables to explanatory variables. The edges are plotted with pen-widths in proportion to the values of the corresponding relations.

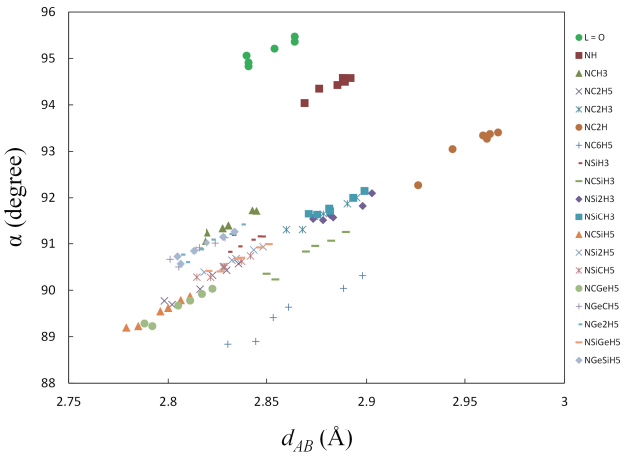

FIG. 7: The correlation between α and d_{AB} of $Mn^{4+}Mn_3^{3+}$ SMMs.

to estimate the two magnetic interactions is promising to improve the predictive power of the method on the exchange coupling J_{BB}/k_B.

The magnetic moment m_A of the Mn^{4+} ion at the A site can be fairly predicted by a linear regression model using four features: β, d_{AB}, d_{AZ1}, and $d_{BO_{xy}}$ with an average relative error of 1.3% ($R = 0.91$) (Fig. 4a). On the other hand, the magnetic moment m_B of Mn^{3+} ions at sites B can be accurately predicted by a linear regression model using d_{AB}, d_{AZ1}, d_{BL1}, $d_{BO_{xy}}$, and all the four electronic features with an average relative error of 0.33% ($R = 0.96$) as shown in figure 4b.

B. Correlations between features of the SMMs and a molecular design strategy

Figure 5 shows the graph built from the obtained relations between all the features. It is clearly seen that the obtained graph appears with two groups of structural features, in which features are strongly correlated to each other: the group of features α, d_{AB}, d_{AL1}, and d_{BL1}, and the group of features d_{BB} and β. The values of d_{AB} positively correlate with the values of all the three features α, d_{AL1}, and d_{BL1}. The values of d_{BB} positively correlate with the values of β in the same manner. These correlations can be qualitatively estimated from the rigid geometrical structure of the distorted cubane $Mn^{4+}Mn_3^{3+}$ cores of the SMMs.

We carry out the above mentioned graph simplification process. The features d_{BB}, d_{BL1}, and β are removed since they can be predicted well by using the other features. The features m_A, m_B, and d_{BO_z} are also removed since they are not sensitive to targeting magnetic properties features. The relations between the remaining features are recalculated and summarized in the simplified graph as shown in Figure 6.

Interestingly, it is clearly seen that the distance $d_{BO_{xy}}$ is sensitive to the exchange coupling J_{AB}/k_B, but can not be predicted by a linear regression model using the electron negativities of the constituent ligands. Further investigation for seeking the features that are sensitive to $d_{BO_{xy}}$ is promising.

To have a better understanding about the correlations between features, we plot all the constructed SMMs in a 2D plane using the distance d_{AB} and angle α as axes (Fig. 7). The structures of SMMs with $L1 = O$ have larger angle α within a range of 94° to 95.5°. For the SMMs with $L1 = N$, the angle α is within a broad range of 89° to 93.5°. For the SMMs with the same L, the α linearly varies with the distance d_{AB}, and this correlation can be understood by considering the magnetic interaction between Mn ions at A and B sites via the ligand $L1$. This observation confirms the reasonability of the relations summarized in the graph between features of the SMMs. It is worth noting that the obtained graph shows a high impact α and d_{AB} in the determination of the exchange coupling J_{AB}/k_B. This result hints us to use α and d_{AB} as intermediate indicators for designing SMMs. However, these structural features are computationally expensive and it is hard to predict accurately the values of α and d_{AB} from the features such as the electron negativities and ionization energies of the constituent ligands in which include no information about the coordinating properties of the ligands with metal ions. Therefore, computationally cheap and ligand coordinating properties inclusive features should be added to improve the representability of the feature set and the predictive power of the regression model.

We design a series of artificial molecules which consist of three $MnCl_2$ groups connected by a ligand L (Fig. 8a).

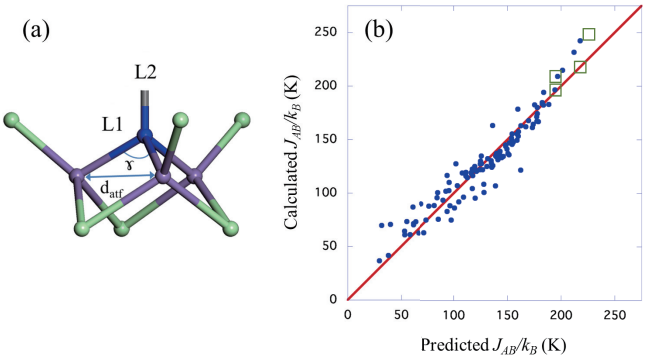

FIG. 8: (a) Schematic geometric structure of the designed artificial molecules with general chemical formula $[(Mn^{2+}Cl_2)_3L1L2]$. Color code: Mn (violet), Mn^{3+} (purple), L1 (blue), Cl (light green). (b) Predicted (by data mining using electronic features and substitutional structural features of ligands) and calculated (by DFT) exchange couplings J_{AB}/k_B for the 114 (blue solid circles) and the newly designed four (open green squares) distorted cubane $Mn^{4+}Mn_3^{3+}$ single molecular magnets. The red line represents the ideal correlation between predicted and calculated results.

The designed artificial molecules have a general chemical formula $[(Mn^{2+}Cl_2)_3L]$ with the same $L(= L1L2)$ as we used for designing the SMMs. The constructed molecular structures were optimized by using the same computational method. We use the distance between Mn ion sites d_{atf} and the angle γ formed between two links between Mn ion sites and $L1$ as two additional features (feature (18), (19)) for describing the coordinating properties of ligand L. Due to the simplicity in the structure of the artificial molecules, these features are computationally much cheaper than the α and d_{AB} of the SMMs.

We then examine whether the additional features can improve the accuracy of the prediction of the exchange coupling J_{AB}/k_B from properties (features (1) - (4), (18), (19)) of the constituent ligands. It is found that the exchange coupling J_{AB}/k_B can be predicted quite well by a linear model using χ_X, χ_{Z1}, χ_{L1}, E_L^{EA}, and d_{atf} as explanatory variables with an average relative error of less than 8% ($R = 0.95$) as shown in figure 8. This result implies that the additional features extracted from the geometrical structure of the designed artificial molecules can be used instead of the computationally expensive geometrical structure features to predict the exchange coupling J_{AB}/k_B of SMMs.

From the obtained linear regression model, we can propose a strategy for selecting ligands among those that preserve the core structure to design the SMMs with high J_{AB}/k_B as follows:

- Ligand at X site with a high electron negativity
- Ligand at Z1 site with a low electron negativity
- Ligand L site with a stable sp^3 electron system and form a short d_{atf} distance

Further, variations of the constituent of the ligand at the Z site may modify slightly the structure of the Mn_4 core. By using this strategy, we designed newly and calculate the J_{AB}/k_B for 4 molecules: $Mn^{4+}Mn_3^{3+}(\mu_3\text{-}(NCH_2\text{-}SiH_3)^{2-})_3(\mu_3\text{-}F^-)(MeC(CH_2\text{-}NOCMe)_3)_3^-(CH(CHO)_2)_3^-$ and $Mn^{4+}Mn_3^{3+}(\mu_3\text{-}L^{2-})_3(\mu_3\text{-}F^-)(N(CH_2\text{-}NOCMe)_3)_3^-(CH(CHO)_2)_3^-$ with L = NCH_2-SiH_3, NCH_2-Si_3H_7, NCH_2-Si_4H_9. The exchange couplingJ_{AB}/k_B of the newly designed molecules can be accurately predicted by the regression model with an average relative error of 6% as shown in figure 8b. The DFT calculation shows that all the four newly designed SMMs are in the group of the SMMs that have the highest values of J_{AB}/k_B. Further, the newly designed molecule $Mn^{4+}Mn_3^{3+}(\mu_3\text{-}(NCH_2\text{-}Si_3H_7)^{2-})_3(\mu_3\text{-}F^-)(N(CH_2\text{-}NOCMe)_3)_3^-(CH(CHO)_2)_3^-$ has a J_{AB}/k_B higher than all the designed SMMs. We also carried out DFT calculations for these new 4 structures within a non-collinear magnetic framework [42] and confirmed the collinearity in their magnetic properties. It is worth to note that the design strategy is derived by mining the data calculated within a collinear magnetic framework and applicable for the purpose of designing SMMs with high J_{AB}/k_B since the SMMs with higher J_{AB}/k_B are expected to have higher collinearity in magnetic properties. For a materials system in which the non-collinear magnetic interactions are dominant, a data representation method that include much of information for estimating the spin-orbit coupling effect is required. Further development of the data representation method and applications of the designing method to materials systems with non-collinear magnetic interactions are promising.

V. CONCLUSION

A combination of data mining and first principles calculation is used to study the structural properties and magnetic properties of 114 distorted cubane $Mn^{4+}Mn_3^{3+}$ single molecule magnets. We demonstrate that the exchange couplings between Mn^{4+} ion and Mn^{3+} ions of all the SMMs can be predicted with a median relative error of 5%, just by using a simple form of sparse regression with their electronic features of constituent ligands and structural features. By using a learning method that consists of several sparse regression processes, all the relations between the structural features and the magnetic properties of the SMMs are quantitatively and consistently summarized in a visual presentation. An effective approach using calculated results for structural properties of simpler artificial molecules instead of computationally expensive properties is proposed to improve the capability of the method. Inferences on the properties of the materials and the suggestion for materials design are discussed based on the obtained graph. A trial of designing new SMMs was made to assess the capability of the method. The accquired results indicate that a first principle calculation-based data mining approach can be

applied to accelerate the understanding and designing of materials.

Acknowledgments

We are thankful for several valuable discussions with K. Q. Than. HC. Dam and TB. Ho thank the support in aid commissioned by the MEXT, JAPAN (No.24700145 and 23300105). The computations presented in this study were performed at the Center for Information Science of the Japan Advanced Institute of Science and Technology.

[1] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, Multiscale Model. Simul. **7**, 842 (2008).
[2] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nature Materials **5**, 641 (2006).
[3] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorg. Chem. **50**, 656 (2011).
[4] A. P. Bartoók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[5] C. M. Handley and P. L. A. Popelier, J. Chem. Theory Comput. **5**, 1474 (2009).
[6] M. Rupp, A. Tkatchenko, K. Muller, and O. A. Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2011).
[7] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Lilienfeld, A. Tkatchenko, and K. Muller, J. Chem. Theo. Comput. **9**, 3404 (2013).
[8] R. Tibshirani, J. R. Statist. Soc. B **58**, 267 (1996).
[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Annals of Statistics **32**, 409 (2004).
[10] R. Kohavi, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence **2**, 11371143 (1995).
[11] R. Sessoli, H.-L. Tsai, A. R. Schake, S. Wang, J. B. Vincent, K. Folting, D. Gatteschi, G. Christou, and D. N. Hendrickson, J. Am. Chem. Soc. **115**, 1804 (1993).
[12] L. Thomas, F. Lionti, R. Ballou, D. Gatteschi, R. Sessoli, and B. Barbara, Nature **383**, 145 (1996).
[13] J. R. Friedman, M. P. Sarachik, J. Tejada, and R. Ziolo, Phys. Rev. Lett. **76**, 3830 (1996).
[14] M. N. Leuenberger and D. Loss, Nature **410**, 789 (2001).
[15] M. Murugesu, M. Habrych, W. Wernsdorfer, K. A. Abboud, and G. Christou, J. Am. Chem. Soc. **126**, 4766 (2004).
[16] C. J. Milios, A. Vinslava, W. Wernsdorfer, S. Moggach, S. Parsons, S. P. Perlepes, G. Christou, and E. K. Brechin, J. Am. Chem. Soc. **129**, 2754 (2007).
[17] R. Clérac, H. Miyasaka, M. Yamashita, and C. Coulon, J. Am. Chem. Soc. **124**, 12837 (2002).
[18] D. Gatteschi and R. Sessoli, Angew. Chem., Int. Ed. **42**, 268 (2003).
[19] R. J. Glauber, J. Math. Phys. **4**, 294 (1963).
[20] J. S. Bashkin, H. Chang, W. E. Streib, J. C. Huffman, D. N. Hendricson, and G. Christou, J. Am. Chem. Soc. **109**, 6502 (1987).
[21] S. Wang, K. Filting, W. E. Streib, E. A. Schmitt, J. K. McCusker, D. N. Hendrickson, and G. Christou, Angew. Chem., Int. Ed. **30**, 305 (1991).
[22] S. Wang, H. Tsai, E. Libby, K. Folting, W. E. Streib, D. N. Hendrickson, and G. Christou, Inorg. Chem. **35**, 7578 (1996).
[23] H. Andres, R. Basler, H. Gudel, G. Aromí, G. Christou, H. Buttner, and B. Rufflé, J. Am. Chem. Soc. **122**, 12469 (2000).
[24] W. Wernsdorfer, N. Aliaga-Alcalde, D. N. Hendrickson, and G. Christou, Nature **416**, 406 (2002).
[25] N. A. Tuan, N. H. Sinh, and D. H. Chi, J. App. Phys. **109**, 07B105 (2011).
[26] N. A. Tuan, N. T. Tam, N. H. Sinh, and D. H. Chi, IEEE Trans. Mag. **47**, 2429 (2011).
[27] P. Hohenberg and W. Kohn, Phys. Rev. **136** (1964).
[28] B. Delley, Chem. Phys. **92**, 508 (1990).
[29] I. J. Bradley Efron, Trevor Hastie and R. Tibshirani, Ann. Statist. **32**, 407 (2004).
[30] B. Hammer, L. Hansen, and J. Nrskov, Phys. Rev. B **59** (1999).
[31] B. Delley, Int. J. Quantum Chem. **69**, 423 (1998).
[32] D. N. Hendrickson, G. Christou, E. A. Schmitt, E. Libby, J. S. Bashkin, S. Wang, H. Tsai, J. B. Vincent, P. D. W. Boyd, J. C. Huffman, et al., J. Am. Chem. Soc. **114**, 2455 (1992).
[33] M. W. Wemple, H. Tsai, K. Folting, D. N. Hendrickson, and G. Christou, Inorg. Chem. **32**, 2025 (1993).
[34] R. S. Mulliken, J. Chem. Phys. **23**, 1833 (1955).
[35] R. S. Mulliken, J. Chem. Phys. **3**, 573 (1935).
[36] A. James and M. Lord, *Macmillan's Chemical and Physical Data* (Macmillan, London, UK, 1992).
[37] The electron affinity of a ligand is a measure of the tendency of that ligand to attract electrons [35] which calculated by using the same DFT method [28, 29].
[38] N. A. Tuan, S. Katayama, and D. H. Chi, Phys. Chem. Chem. Phys **11**, 717 (2009).
[39] The R^2 factor of the predition.
[40] N. Meinshausen and P. Buhlmann, Ann. Statist. **34**, 1436 (2006).
[41] H. L. Schlafer and G. Gliemann, *Basic Principles of Ligand Field Theory* (Wiley Interscience, New York, USA, 1969).
[42] Non-collinear DFT calculations were carried out by using OpenMX code [43] with localized pseudo-atomic orbitals basis set and Ceperley-Alder exchange-correlation functional [44] parameterized by Perdew and Zunger [45]. J-dependent pseudo potentials with full relativistic effect and spin-orbit coupling [46] were used for all calculations.
[43] Http://www.openmx-square.org/.
[44] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).
[45] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[46] A. H. MacDonald and S. H. Vosko, J. Phys. C **12**, 2977 (1979).