



Predicting structure-dependent Hubbard U parameters via machine learning

Guanghui Cai^{1,2,7} , Zhendong Cao^{1,2,7}, Fankai Xie^{1,2}, Huaxian Jia⁵, Wei Liu⁵, Yaxian Wang^{1,2}, Feng Liu⁶, Xinguo Ren^{1,3,*}, Sheng Meng^{1,2,3,*} and Miao Liu^{1,3,4,*} 

¹ Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

² School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100190, People's Republic of China

³ Songshan Lake Materials Laboratory, Dongguan, Guangdong 523808, People's Republic of China

⁴ Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

⁵ Tencent AI Lab, Tencent, Shenzhen 518057, People's Republic of China

⁶ Department of Materials Science and Engineering, University of Utah, Salt Lake City, UT 84112, United States of America

E-mail: renxg@iphy.ac.cn, smeng@iphy.ac.cn and mliu@iphy.ac.cn

Received 2 November 2023, revised 21 December 2023

Accepted for publication 27 December 2023

Published 24 January 2024



Abstract

DFT + U is a widely used treatment in the density functional theory (DFT) to deal with correlated materials that contain open-shell elements, whereby the quantitative and sometimes even qualitative failures of local and semi-local approximations can be corrected without much computational overhead. However, finding appropriate U parameters for a given system and structure is non-trivial and computationally intensive, because the U value has generally a strong chemical and structural dependence. In this work, we address this issue by building a machine learning (ML) model that enables the prediction of material- and structure-specific U values at nearly no computational cost. Using Mn–O system as an example, the ML model is trained by calibrating DFT + U electronic structures with the hybrid functional results of more than 3000 structures. The model allows us to determine an accurate U value (MAE = 0.128 eV, R^2 = 0.97) for any given Mn–O structure. Further analysis reveals that M–O bond lengths are key local structural properties in determining the U value. This approach of the ML U model is universally applicable, to significantly expand and solidify the use of the DFT + U method.

Supplementary material for this article is available [online](#)

Keywords: DFT + U , machine learning, Bayesian optimization

⁷ G Cai and Z Cao contributed equally to this work.

* Authors to whom any correspondence should be addressed.



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

1. Introduction

The complex nature of many-body interactions makes it a long-standing challenge to further improve the exchange–correlation (XC) energy functional for density functional theory (DFT). The semi-local XC functionals, such as the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof (PBE) [1] and many others [2, 3], have well-known self-interaction issues, failing to describe the energy bands of many compounds correctly, especially those ionic compounds with open-shell elements. One viable solution is to add intra-atomic interactions between electrons to mitigate the self-interaction error intrinsic to the local or semi-local XC functionals, namely the Hubbard U correction. The DFT + U method, which was first proposed by Anisimov *et al* [4] and further developed by Dudarev *et al* [5], introduces an on-site Coulomb interaction term to penalize partial occupation of the localized orbitals, which can correctly predict behaviors of strongly correlated systems, e.g. Mott insulators [6]. However, finding appropriate U values for a given material system is generally a challenging task. Previously, Wang *et al* fitted one U value for a given open-shell transition-metal (TM) element based on experimental chemical reaction enthalpies of TM oxides [7], but the one- U -value predicted reaction enthalpies are less satisfactory [8]. On the other hand, first-principles approaches, such as the linear response method [9] and constrained random phase approximation (cRPA) [10–12], were developed to self-consistently determine the U value for a given system, but the computational cost is increased by at least 10 times. Recently, the machine learning (ML) method was employed, specifically the Bayesian optimization (BO), to extract the U value according to higher-level *ab initio* results [13]. Such a method has been also successfully applied to interface [14, 15] and superlattice [16], with extra computational overhead.

One major challenge to find an appropriate U lies in its strong dependence on local stoichiometry and structure, since by its very nature, the Hubbard U correction represents a short-range electronic interaction which is implicitly associated with local charge density and orbital symmetry. Therefore, the U value has poor transferability. It depends strongly on stoichiometry or d -valence, as reflected, for example, by a value of 6.7 eV in Ce_2O_3 but 5.13 eV in CeO_2 [17]. It also has a strong structural dependence, as demonstrated by its variation in pressure-induced phase transitions of non-magnetic [18] and magnetic structures [19]. This in turn calls for deriving an accurate U value in order to obtain a more reliable potential energy surface [20]. Especially, in general, one has to derive a specific U value, one at a time, for a given system of the specific chemical and structural environment, which is not only a redundant process but also computationally costly and time-consuming. Therefore, it is highly desirable and useful to establish an efficient approach, once for all, which allows one to predict an accurate U value for any given system of any chemical and structural environment on the fly.

In this work, we tackle this problem by employing our in-house high-throughput BO-based workflow, and using the

Mn–O compounds as our model system. The model is primarily designed to predict the U value for GGA-PBE when utilizing VASP code, aiming to yield electronic structures closely aligned with the HSE06 level. We employ more than 3000 Mn–O configurations whose band gaps and energy bands are fitted to the high-level hybrid functional result (Heyd–Scuseria–Ernzerhof functional [21, 22], HSE) as for the abundant structure availability of the Mn–O system. This step essentially follows what has been done in [13]. In a second, and more important step, we carry on to employ a supervised random forest ML algorithm [23, 24] to train a predictive Hubbard U model, which then predicts the U value with sufficient accuracy and efficiency for any given Mn–O structure, even those not included in the training dataset. The obtained ML model shows a remarkable accuracy and reaches the coefficients of determination $R^2 = 0.97$ and mean absolute error (MAE) of 0.128 eV for U values. More importantly, it is unraveled by the regression that the U value is primarily associated with the bond length, which is consistent with the cRPA theory.

2. Methods

2.1. Vienna Ab initio Simulation Package (VASP)

The first-principles electronic structure calculations are done using VASP codes [25] based on DFT with the PBE XC functional [1] and the projector-augmented-wave approach [26, 27]. The energy cutoff of the plane-wave basis in the calculation is set to be 520 eV, which suffices for accurately describing the energetics discussed in this work. The Γ -centered K -points grid density of 125 k -points/ \AA^{-3} is adopted for all the calculations, and materials were modeled as ferromagnetic ordering. All the DFT + U results are performed by VASP using the rotationally invariant DFT + U approach introduced by Dudarev *et al* [5]. We note that different DFT + U implementations may yield different results, furthermore, as shown in [28], PBE has the closest result between Dudarev approximation and Lichtenstein form [29] than local density approximation (LDA) and PBE revised for solids (PBEsol). We stress that the Hubbard U values are not transferable among different DFT implementations.

2.2. FHI-aims

The hybrid DFT calculation is performed by the full-potential, all-electron, numeric atomic orbital-based FHI-aims code [30–32]. The real-space band structures for the HSE [21, 22] reference have been performed using a standard *tier 1* basis set and applying *intermediate* integration grids. The K -points grid density is set to 5 in units of \AA^{-1} for obtaining reasonable results. The high-symmetry points are generated by the HighSymmKpath module in the pymatgen library [33] for all the band structure calculations, including VASP and FHI-aims.

2.3. BO

The BO is carried out by the Bayesian Optimization library [34]. The upper confidence bound (UCB) acquisition function is used to predict the value that would be generated by the evaluation of the objective function at a new point and decide what value of \vec{U} to sample in the n th iteration:

$$\vec{U}_n = \arg \max_{\vec{U}} \mu(\vec{U}) + \kappa \sigma(\vec{U}). \quad (1)$$

The hyperparameter κ controls the trade-off between exploration and exploitation. Here, we set $\kappa = 5$. Based on the HSE06 band structure, we include the top 6 valence bands and the bottom 4 conduction bands in the optimization (the top 4 valence bands and the bottom 4 conduction bands are considered for the calculations when PBE results do not have enough valence bands). More details of the BO method can be seen in [13]. The UCB acquisition function is selected as it converges quicker than the probability of improvement (PI) and the expected Improvement (EI) acquisition functions [35].

2.4. ML model

We employ the Random Forest Regression (RFR) implemented in the scikit-learn library [36] to extract the structure-Hubbard U relationship and rank the relative importance of descriptors. Nearly 600 descriptors are generated for each structure following [37], and the model training process sorts out the best ten descriptors for the final model construction. More details about the descriptors are provided in Supplementary. The model is primarily designed to predict the U value for GGA-PBE when utilizing VASP code, aiming to yield electronic structures closely aligned with the HSE06 level.

3. Results

3.1. Workflow

The process including data generation, structural distortion, BO, and ML is schematically illustrated in figure 1. The thermodynamic stability of the compounds is evaluated by the physical quantity of energy above hull (E_{hull}), which is the reaction enthalpy required to decompose a material to other stable compounds [38, 39]. We start from selecting the thermodynamically stable ($E_{\text{hull}} < 200 \text{ meV/atom}$) and Mn–O chemical systems with reasonable size ($N_{\text{atom}} < 20$) from 312 possible structures generated from the Atomly [40] materials database and end up with 67 individual compounds. To enlarge the size of our structural space, we apply uniaxial, biaxial, and triaxial strain of $-2\% < \varepsilon < 2\%$ to the selected 67 compounds and obtain 3724 Mn–O structures. The strain deformation step is employed to modify the screened local Coulomb interaction U [41, 42] and ensures that our model surveys among a large enough structural space to yield a good model extrapolation. The BO method, which is developed by Yu *et al.*, is employed to determine the Hubbard U parameter within the PBE + U

method by fitting to the HSE06 band gap and band structures for all the structures. In this way, a U value for Mn 3d state can be fitted for each structure. We note that materials were modeled as ferromagnetic ordering. In principle, higher-level methods, such as the GW [43] or coupled cluster singles and doubles [44]-level of the calculation, can be used too if one has enough computational resources, but we use the HSE06 as the ‘ground truth’ for its viable efficiency. Finally, a ML model is constructed by harnessing the RFR to directly predict the optimal Hubbard U parameters for any Mn–O structures. The computational details can be found in supplementary.

3.2. Model validation

Random forests are a combination of decision trees that individually make predictions on each input and the overall prediction is determined by a majority voting process. We evaluate the prediction ability of our RFR model by plotting the out-of-bag error, which can be analogous to the conventional cross-validation error but provides a global error estimated for all data points, shown in figure 2(a) along with the detailed data distribution. Our model can predict the Hubbard U values fairly accurately with $\text{MAE} = 0.128 \text{ eV}$ and $R^2 = 0.97$, meaning that the predicted U value falls into a small error range of $\pm 0.128 \text{ eV}$ statistically. The distribution of Hubbard U values shows two peaks due to the uneven distribution of Mn–O bond length which will be discussed later in the paper. It is noticeable that for some structures, their electronic structures are insensitive to the change of Hubbard U parameters, and hence the heavily-structure-dependent model does not apply so well to these compounds, causing the deficiency of the ML model to some extent, but the model overall has good accuracy. To gain a better insight into the physical connection between the Hubbard U parameters and the materials’ properties, we sort out the 10 most important descriptors for the U value prediction, as shown in figure 2(b). There are CB , $\Delta\chi$, P , T , and CN (see supplementary, table S1 for their definitions). Further, these descriptors can be divided into four categories: chemical bond (CB), electronegativity difference ($\Delta\chi$), line surface angle (P , T), and coordinate number (CN). It is obvious that, other than the electronegativity, nearly all the decisive descriptors, the CB , line surface angle, and coordinate number, are primarily a function of the atomistic structure of a compound, indicating that the U is primarily a structure-dependent parameter. The CB length is the most important factor for the predictions (50.7%), which is computed as the (normalized) total reduction of the criterion brought by that feature.

Upon categorizing these 10 factors into four aspects, these parameters have a close connection with each other. Therefore, we look into the Pearson correlation matrix (figure 2(c)) of the 10 descriptors. It can be seen that the choice of the descriptors is fairly orthogonal as those descriptors are weakly coupled with each other, except for the CB descriptor (CB_u^v) and the electronegativity descriptor ($\text{AR } \Delta\chi_u^k$, AR represents Allred–Rockow electronegativity). This is expected considering that the electronegativity descriptors themselves are essentially derived from the crystal structure and the atom species.

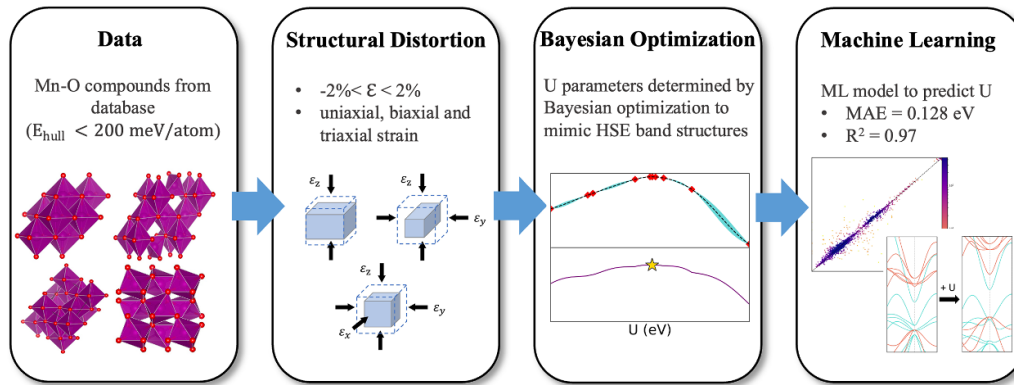


Figure 1. Computational process. The flow chart of the machine learning process to create a structure-dependent Hubbard U model.

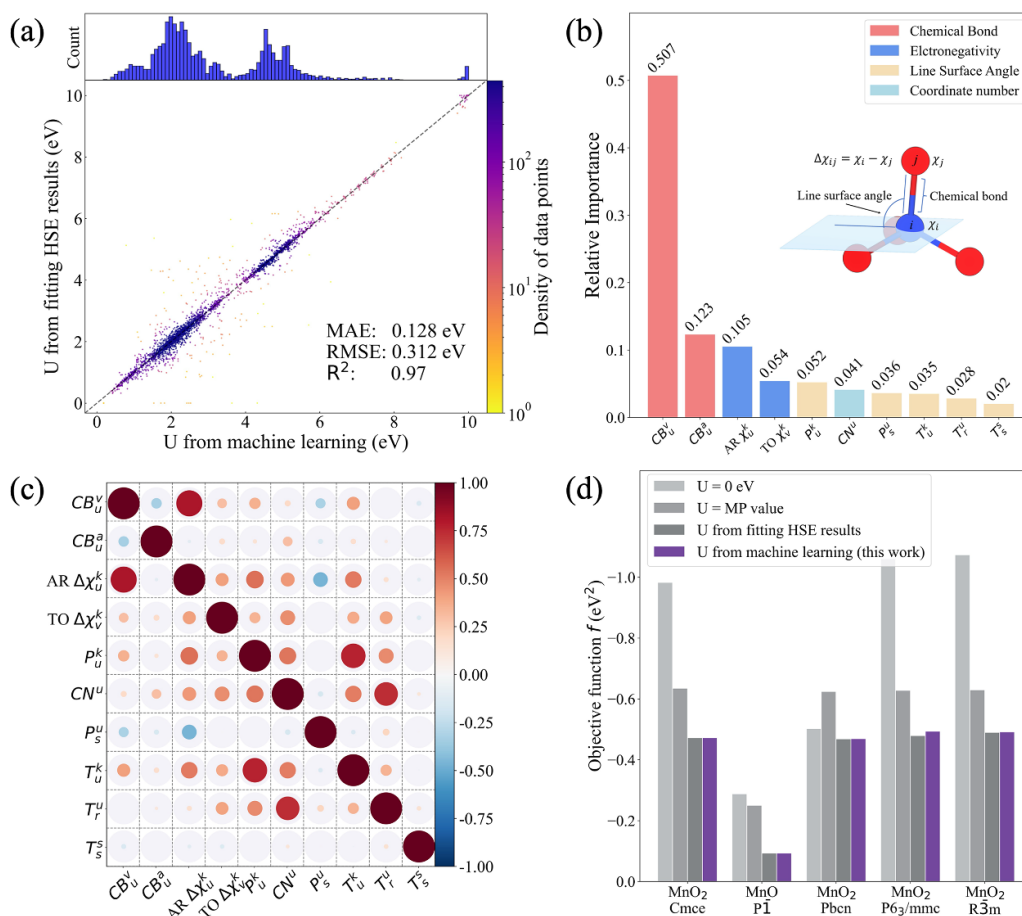


Figure 2. Machine learning model and its applications. (a) The comparison of Hubbard U values from fitting the HSE06 results (Bayesian optimization) and machine learning. (b) The importance ranking of all descriptors in prediction, highlighting the most significant factor from the bond length. Different kinds of descriptors are colored in different colors, as depicted in the legend. The superscript and subscript of the descriptor denote the inter-site and intra-site mathematical operation, respectively (see supplementary). (c) Pearson correlation coefficient matrix of all descriptors. The radii of circles represent the absolute magnitude of coefficients. (d) The performance of different methods to determine the Hubbard U . The objective function is employed as the evaluation of the performance of different methods. Here five structures are chosen from the Atomly database with their chemical formulas and space groups listed by the horizontal axis.

On the other hand, the strength of a CB should depend on the electronegativity difference ($\Delta\chi$) between the two bonding atoms. The $\Delta\chi$ between atoms bonded together will greatly affect the charge density distribution of the local structure, thereby influencing the local electronic screening. Therefore,

the correlation between electronegativity and the CB length further suggests it is viable to assign a U value to a compound based on its structure.

Figure 2(d) presents the performance of the ML methods in comparison with the direct energy calculation with and

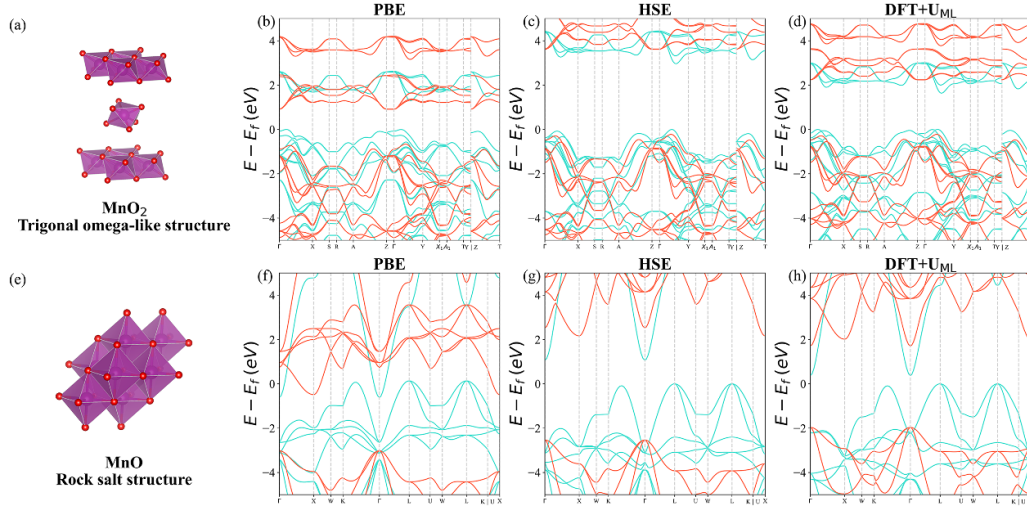


Figure 3. Crystal structures and band structures of MnO and MnO₂ obtained using different methods. (a) The crystal structure of MnO₂. Band structures of MnO₂ obtained in different methods: (b) PBE; (c) HSE06; (d) PBE with ML predicted U . (e) The crystal structure of MnO. Band structures of MnO were obtained in different methods: (f) PBE; (g) HSE06; (h) PBE with ML predicted U . The red line and the cyan line represent the spin-up and spin-down bandstructure, respectively and materials were modeled as ferromagnetic ordering.

without corrections. The objective function to evaluate the performance of a given U value is defined as [13]:

$$f(\vec{U}) = -\alpha_1 (E_g^{\text{HSE06}} - E_g^{\text{PBE}+U})^2 - \alpha_2 (\Delta\text{Band})^2. \quad (2)$$

Here, $\vec{U} = [U^1, U^2, \dots, U^m]$ is the vector of U values applied to different atomic species. E_g^{HSE06} and $E_g^{\text{PBE}+U}$ represent the band gaps calculated by the HSE06 and PBE + U functionals. ΔBand is defined as the mean squared deviation of the PBE + U band structures with respect to their HSE06 counterparts, similar to [45]:

$$\Delta\text{Band} = \sqrt{\frac{1}{N_E} \sum_{i=1}^{N_k} \sum_{j=1}^{N_b} (\epsilon_{\text{HSE06}}^j[k_i] - \epsilon_{\text{PBE}+U}^j[k_i])^2}, \quad (3)$$

where N_E represents the total number of eigenvalues ϵ . The summation goes through N_k K -points and N_b selected bands, and obviously $N_E = N_k * N_b$. The coefficients α_1 and α_2 are the control parameters that assign different weights to the band gaps and band structures. We set $\alpha_1 = 0.25$ and $\alpha_2 = 0.75$ as default in agreement with [13]. The closer the objective function is to 0, the better the PBE + U calculations reproduce the HSE06 results. It can be observed that our model outperforms the traditional treatment which used a fixed U value for a given compound and reaches the same accuracy as that of the BO method developed by Yu *et al* [13]. For all data points, the MAE difference of the objective function between ours and that of Yu *et al* is about 0.01 eV² (supplementary, figure S1). Our method can predict the U value and reproduces the band gap and band structures obtained from HSE06 without running the U parameter calculations, thus making it easier and more efficient for performing DFT + U calculations for strongly correlated systems.

Figure 3 demonstrates the performance of PBE with predicted U for MnO₂ and MnO compounds. It can be found

that the PBE functional is unable to fully capture the electronic structure of the Mn–O compounds (and in fact also other transition metal oxides) due to the incomplete self-interaction error cancellation [46–49], and yields an overestimation of Coulomb repulsion. For MnO₂ (figure 3(a)), the PBE band gap of 0.91 eV (figure 3(b)) is considerably underestimated compared to the HSE06 result of 2.96 eV (figure 3(c)). Moreover, the locations of the conduction band minimum (CBM) and valence band maximum (VBM), as well as the spin-up and spin-down channels, from PBE are drastically different from those from HSE06. Using the Hubbard U predicted in this work, we obtain the spin-polarized band structure that matches the HSE06 result well (figure 3(d)). For MnO, the usage of our predicted Hubbard U successfully corrects the band position closer to the HSE06 values, and more importantly results in a band gap opening for this compound. This material is a conductor according to PBE (figure 3(f)) and a semiconductor with a band gap of 0.39 eV in PBE + U (figure 3(g)), amending the PBE result significantly. We note that adding a proper U correction still underestimates the band gap to some extent, e.g. in MnO₂, the ML U correction increases the gap to 1.63 eV, whereas the HSE06 band gap is 2.96 eV; in MnO, the ML U correction increases the gap to 0.39 eV, whereas the HSE06 band gap is 1.09 eV (figure 3(h)). It reflects that the Hubbard U correction cannot completely capture the features of exchange interactions in HSE06. Overall, our ML model predicts reliable Hubbard U values that apply well to the PBE + U calculations and reproduces the qualitative features of HSE06 band structures.

3.3. Structural dependence

As discussed in the previous session, our ML model indicates that the Hubbard U parameter is greatly structure-dependent. In order to investigate the correlation between the Hubbard U parameter and Mn–O bond length, their distributions among

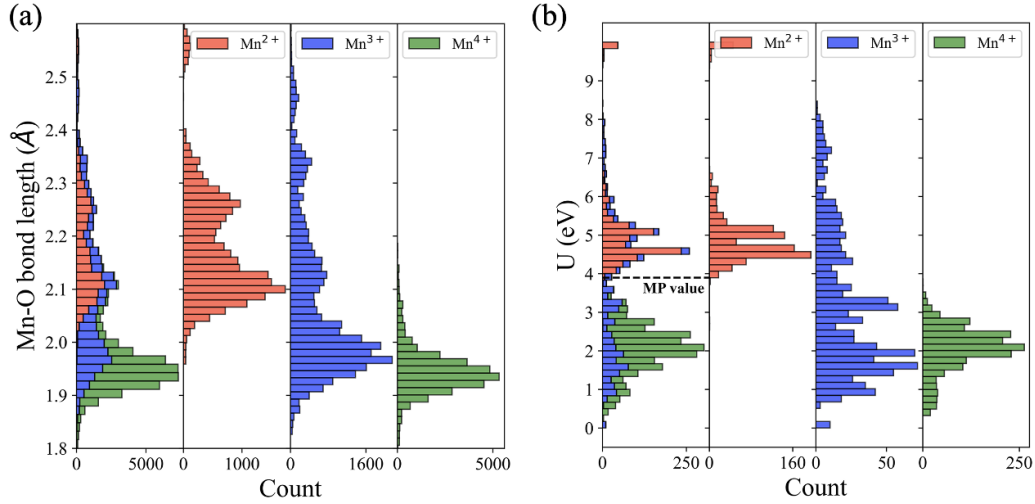


Figure 4. Distribution of bond length and Hubbard U . The (a) bond length and (b) Hubbard U distributions with the contribution of different valence states (Mn^{2+} , Mn^{3+} , and Mn^{4+}) are colored in red, blue, and green, respectively. The dashed line denotes the U value used in the Materials Project.

different valence states are plotted in figure 4. It can be seen that the distribution of Mn–O band length is shape-wise similar to that of the U parameters. For example, Mn^{3+} exhibits a wide distribution of bond lengths, while its Hubbard U spreads over a wider range (from 0 to 10 eV) compared to the Mn^{2+} and Mn^{4+} cases. Furthermore, the overall larger Mn–O bond lengths in Mn^{2+} compounds correspond to their overall larger U values, whereas the smaller Mn–O bond lengths in Mn^{4+} compounds lead to smaller U values. The dashed line in figure 4(b) represents the single U value used in the Materials Project (MP) [50] as obtained by fitting experimental data, showing an apparent discrepancy with the distribution of Hubbard U values from this work. Given that, it implies that the Mn–O bond length is a sensible parameter to describe the variation in U , suggesting the necessity of further investigating the relationship between the Mn–O bond length and the Hubbard U .

The Hubbard U parameter can be calculated by the cRPA method as follows:

$$U_{lm,lm'} = \int d\mathbf{r} d\mathbf{r}' |\phi_{l,m}^{3d}(\mathbf{r})|^2 \tilde{W}(\mathbf{r}, \mathbf{r}') |\phi_{l,m'}^{3d}(\mathbf{r}')|^2, \quad (4)$$

where $\phi_{l,m}^{3d}(\mathbf{r})$ is the wavefunction of a Mn $3d$ orbital at the site l with magnetic angular momentum m . $\tilde{W} = \epsilon^{-1}v = (1 - v\tilde{X}_0)^{-1}v$ denotes the effective screened Coulomb interaction with v denoting the bare Coulomb interaction. \tilde{X}_0 is the polarization function, and its magnitude depends on the electronic response property of the system, which is in turn governed by the atomic structure, in particular the bond lengths between neighboring atoms. Within the cRPA scheme, one needs to compute the microscopic polarization function and screened Coulomb interaction for a given atomic structure, which is *de facto* much more costly than the DFT + U calculation itself. Here, our model directly delivers U values from the atomic structure, circumventing the cumbersome step of calculating the microscopic polarization function, yet capturing

the same essential physics, namely, the U value is ultimately determined by the local chemical environment.

4. Discussion

This work showcases a ML model for predicting Hubbard U to skip the expensive first-principles U value calculation process without sacrificing accuracy. Although the Mn–O system is selected as the model system, the out-of-box models can be created for the community for all the open-shell elements as the U value is essentially local-structure dependent, which is consistent with the findings reported in previous work [17, 20]. This method has the advantage that the model allows one to assign an appropriate Hubbard U parameter to a system prior to the DFT calculation and yields improved results that are close to the higher-level methods such as HSE06 or GW. The GW-level of accuracy is also tangible once such a GW dataset is available. When applying the DFT + U method to structural relaxations or molecular dynamics simulations, it would be ideal to adjust the U parameter to appropriate values on the fly as the structure evolves [18, 51, 52]. However, this will become prohibitively expensive if the U value is determined using the conventional first-principles approaches, such as the linear response or the cRPA schemes. The pre-trained ML model, as demonstrated in the present work, will make all this readily happen.

Another advantage is that this approach can be extended to several other properties of systems other than the energy band difference. For example, the model can also calibrate the adhesive energy of the system by including the energies in the objective function. Also, the intersite interaction parameter, V [53], can be also incorporated into the model to further improve its predictive power, which we hope to spark a future investigation.

Moreover, the accuracy and robustness of our ML model can be further enhanced with the reinforced dataset. The

purpose of this paper is to showcase this approach, while we are aware that with the hybrid-functional-level treatment adopted, only a small dataset is produced (3724 data points) due to computational cost. If the size of the dataset is presumably enlarged by one or two orders of magnitude, the ML model could evolve into a deep neural network, meanwhile, the model accuracy, extrapolation, and generalization can be greatly enhanced.

Finally, our work demonstrates that the Hubbard U parameter is local-structure dependent to some extent. However, the U value we used in this work is a kind of global measure of the electronic screening effect, which may be not sensitive to the change in local structure. One solution can be to assign the U values for every inequivalent site, which requires a tremendous amount of calculation resources and a fairly large dataset for model training. To this end, we hope efforts can be made by the entire community to collaboratively carry forward this method to generally reliable and efficient models to predict Hubbard U values for all open shell elements.

In summary, we developed a data-driven method for predicting the value of Hubbard U for DFT calculations. Specifically, a ML model is constructed to predict the Hubbard U for Mn–O systems, which can accurately assess the U value of a system without running costly first-principles calculations. It is also demonstrated that the predicted Hubbard U can reproduce hybrid functional-level band gap and band structures without actual hybrid functional-level runs. In addition, our ML model reveals the bond length which shares similar distribution with the Hubbard U is the most decisive factor in determining the U value, which can be justified by cRPA theory. Developing a ML model that can accurately yield appropriate U values for a given structure, without actually running expensive and sophisticated electronic-structure calculations, is a long-sought goal. We demonstrate in this work that this is indeed possible, at least in a given type of system. More work is needed to extend the present model from Mn–O systems to general atomic species and structures, but we do not expect essential difficulties that prevent us from eventually achieving this goal. Our ML model not only opens up a new avenue to calculate Hubbard U values for all open-shell elements, but also provides insights into the physical correlation between the U parameters and local structure in condensed matter, which is relevant to many other important physical questions, such as metal–insulator transitions, superconductivity, magnetic phase transition, etc.

As mentioned by the editor, we noticed that a recent paper [54] introduced a methodology for fitting the $U + V$ parameters within a specific system as an improvement to Yu *et al* [13]. We would point out that this paper is distinctively different from these two papers [13, 54] as they developed methods to fit U or $U + V$ parameters for the studying system, but our method does not need the extra fitting once the ML model is trained for Mn–O as demonstrated. Thus the two papers [13, 54] can be employed as a tool to generate data for our ML model.

4.1. Future perspectives

Accurately predicting Hubbard U parameters has been a long-standing pursuit due to its relevance to a wide range of physical inquiries concerning the fractional, magnetic, lattice, and charge excitations fundamental to quantum materials and devices. Conventional methods for fitting the Hubbard U parameters, such as the linear response method and cRPA method, requires circumventing the cumbersome steps of first-principles calculations. This work has provided significant physical insights, revealing the close association between the U value and the local environment of the open-shell cations, including factors such as bond lengths, coordination numbers, and more. Consequently, it is proposed that a ML model can be developed to reliably predict structure-specific U values for any given structure, if one have enough data to establish the relationship between U and local structure. The study, then, has successfully demonstrated the development of an ML model for the Mn–O system, capable of predicting Hubbard U parameters without relying on expensive linear response method or cRPA method based on a structure-specific U dataset with 3000 datapoints. Importantly, this work suggests the potential to develop a universal U -predicting model directly from the atomistic structures of compounds, which could significantly reshape the ways of DFT + U calculations.

Data availability statement

Data will be available upon request. The code for BO and ML in this work can be found at: <https://github.com/zdcao121/ml4dftu>

Acknowledgments

We would acknowledge the financial support from the Chinese Academy of Sciences (Grant Nos. XDB33020000, CAS-WX2023SF-0101, ZDBS-LY-SLH007, and YSBR047), National Key R&D Program of China (2021YFA1400200, and 2021YFA0718700), and National Natural Science Foundation of China (Grand Nos. 12025407, 12134012 and 12188101). The computational resource is provided by the Platform for Data-Driven Computational Materials Discovery of the Songshan Lake materials laboratory.

Author contributions

M L proposed and led this project. Z C and G C wrote the code. Z C performed the calculations and analyzed the results. X R and S M copiloted the project with important intellectual contributions. F X, H J, and W L provided assistance with the ML algorithm. Z C and M L wrote the manuscript. Y W, F L, X R, and S M reviewed and revised the manuscript. G C and Z C contributed equally to this work.

Conflict of interest

The authors declare no competing interests.

ORCID iDs

Guanghui Cai  <https://orcid.org/0009-0006-7978-2050>

Miao Liu  <https://orcid.org/0000-0002-1843-9519>

References

- [1] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8
- [2] Becke A D 1988 Density-functional exchange-energy approximation with correct asymptotic behavior *Phys. Rev. A* **38** 3098–100
- [3] Lee C, Yang W and Parr R G 1988 Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density *Phys. Rev. B* **37** 785–9
- [4] Anisimov V I, Zaanen J and Andersen O K 1991 Band theory and Mott insulators: Hubbard U instead of Stoner I *Phys. Rev. B* **44** 943–54
- [5] Dudarev S L, Botton G A, Savrasov S Y, Humphreys C J and Sutton A P 1998 Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study *Phys. Rev. B* **57** 1505–9
- [6] Han M J, Ozaki T and Yu J O 2006 (N) LDA + U electronic structure calculation method based on the nonorthogonal pseudoatomic orbital basis *Phys. Rev. B* **73** 045110
- [7] Wang L, Maxisch T and Ceder G 2006 Oxidation energies of transition metal oxides within the GGA + U framework *Phys. Rev. B* **73** 195107
- [8] Zhou F, Cococcioni M, Marianetti C A, Morgan D and Ceder G 2004 First-principles prediction of redox potentials in transition-metal compounds with LDA + U *Phys. Rev. B* **70** 235121
- [9] Cococcioni M and de Gironcoli S 2005 Linear response approach to the calculation of the effective interaction parameters in the LDA + U method *Phys. Rev. B* **71** 035105
- [10] Aryasetiawan F, Karlsson K, Jepsen O and Schönberger U 2006 Calculations of Hubbard U from first-principles *Phys. Rev. B* **74** 125106
- [11] Miyake T and Aryasetiawan F 2008 Screened Coulomb interaction in the maximally localized Wannier basis *Phys. Rev. B* **77** 085122
- [12] Şaşıoğlu E, Friedrich C and Blügel S 2011 Effective Coulomb interaction in transition metals from constrained random-phase approximation *Phys. Rev. B* **83** 121101
- [13] Yu M, Yang S, Wu C and Marom N 2020 Machine learning the Hubbard U parameter in DFT+U using Bayesian optimization *npj Comput. Mater.* **6** 180
- [14] Yu M, Moayedpour S, Yang S, Dardzinski D, Wu C, Pribyl V S and Marom N 2021 Dependence of the electronic structure of the EuS/InAs interface on the bonding configuration *Phys. Rev. Mater.* **5** 064606
- [15] Yang S, Dardzinski D, Hwang A, Pikulin D I, Winkler G W and Marom N 2021 First-principles feasibility assessment of a topological insulator at the InAs/GaSb interface *Phys. Rev. Mater.* **5** 084204
- [16] Popov M N, Spitaler J, Rومانer L, Bedoya-Martínez N and Hammer R 2021 Bayesian optimization of Hubbard U's for investigating InGaN superlattices *Electron. Mater.* **2** 370–81
- [17] Lu D and Liu P 2014 Rationalization of the Hubbard U parameter in CeO_x from first principles: unveiling the role of local structure in screening *J. Chem. Phys.* **140** 084101
- [18] Hsu H, Umemoto K, Cococcioni M and Wentzcovitch R 2009 First-principles study for low-spin LaCoO₃ with a structurally consistent Hubbard U *Phys. Rev. B* **79** 125124
- [19] Tsuchiya T, Wentzcovitch R M, da Silva C R S and de Gironcoli S 2006 Spin transition in magnesiowüstite in Earth's lower mantle *Phys. Rev. Lett.* **96** 198501
- [20] Kulik H J and Marzari N 2011 Accurate potential energy surfaces with a DFT+U(R) approach *J. Chem. Phys.* **135** 194105
- [21] Heyd J, Scuseria G E and Ernzerhof M 2003 Hybrid functionals based on a screened Coulomb potential *J. Chem. Phys.* **118** 8207–15
- [22] Heyd J, Scuseria G E and Ernzerhof M 2006 Erratum: 'Hybrid functionals based on a screened Coulomb potential' [J. Chem. Phys. 118, 8207 (2003)] *J. Chem. Phys.* **124** 219906
- [23] Amit Y and Geman D 1997 Shape quantization and recognition with randomized trees *Neural Comput.* **9** 1545–88
- [24] Ho T K 1998 The random subspace method for constructing decision forests *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 832–44
- [25] Kresse G and Furthmüller J 1996 Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set *Phys. Rev. B* **54** 11169–86
- [26] Kresse G and Joubert D 1999 From ultrasoft pseudopotentials to the projector augmented-wave method *Phys. Rev. B* **59** 1758–75
- [27] Blöchl P E 1994 Projector augmented-wave method *Phys. Rev. B* **50** 17953–79
- [28] Tavadze P, Boucher R, Avendaño-Franco G, Kocan K X, Singh S, Dovale-Farelo V, Ibarra-Hernández W, Johnson M B, Mebane D S and Romero A H 2021 Exploring DFT+U parameter space with a Bayesian calibration assisted by Markov chain Monte Carlo sampling *npj Comput. Mater.* **7** 182
- [29] Liechtenstein A I, Anisimov V I and Zaanen J 1995 Density-functional theory and strong interactions: orbital ordering in Mott-Hubbard insulators *Phys. Rev. B* **52** R5467–70
- [30] Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X, Reuter K and Scheffler M 2009 Ab initio molecular simulations with numeric atom-centered orbitals *Comput. Phys. Commun.* **180** 2175–96
- [31] Ren X, Rinke P, Blum V, Wieferink J, Tkatchenko A, Sanfilippo A, Reuter K and Scheffler M 2012 Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions *New J. Phys.* **14** 053020
- [32] Levchenko S V, Ren X, Wieferink J, Johanni R, Rinke P, Blum V and Scheffler M 2015 Hybrid functionals for large periodic systems in an all-electron, numeric atom-centered basis framework *Comput. Phys. Commun.* **192** 60–69
- [33] Ong S P, Richards W D, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier V L, Persson K A and Ceder G 2013 Python materials genomics (pymatgen): a robust, open-source python library for materials analysis *Comput. Mater. Sci.* **68** 314–9
- [34] Fernando 2022 Bayesian optimization
- [35] Okazawa K, Tsuji Y, Kurino K, Yoshida M, Amamoto Y and Yoshizawa K 2022 Exploring the optimal alloy for nitrogen activation by combining Bayesian optimization with density functional theory calculations *ACS Omega* **7** 45403–8
- [36] scikit-learn/scikit-learn Scikit-learn: machine learning in Python (available at: <https://github.com/scikit-learn/scikit-learn>)
- [37] Liang Y et al 2022 A universal model for accurately predicting the formation energy of inorganic compounds *Sci. China Mater.* **66** 343–51

- [38] Ong S P, Wang L, Kang B and Ceder G 2008 Li–Fe–P–O₂ phase diagram from first principles calculations *Chem. Mater.* **20** 1798–807
- [39] Jain A, Hautier G, Ong S P, Moore C J, Fischer C C, Persson K A and Ceder G 2011 Formation enthalpies by mixing GGA and GGA + U calculations *Phys. Rev. B* **84** 045115
- [40] Liu M and Meng S 2022 Atomly.net materials database and its application in inorganic chemistry *Sci. Sin.-Chim.* **53** 19–25
- [41] Tomczak J M, Miyake T and Aryasetiawan F 2010 Realistic many-body models for manganese monoxide under pressure *Phys. Rev. B* **81** 115116
- [42] Ivashko O et al 2019 Strain-engineering Mott-insulating La₂CuO₄ *Nat. Commun.* **10** 786
- [43] Hedin L 1965 New method for calculating the one-particle Green's function with application to the electron-gas problem *Phys. Rev.* **139** A796–823
- [44] Purvis G D and Bartlett R J 1982 A full coupled-cluster singles and doubles model: the inclusion of disconnected triples *J. Chem. Phys.* **76** 1910–8
- [45] Huhn W P and Blum V 2017 One-hundred-three compound band-structure benchmark of post-self-consistent spin-orbit coupling treatments in density functional theory *Phys. Rev. Mater.* **1** 033803
- [46] Ye L-H, Luo N, Peng L-M, Weinert M and Freeman A J 2013 Dielectric constant of NiO and LDA + U *Phys. Rev. B* **87** 075115
- [47] Pask J E, Singh D J, Mazin I I, Hellberg C S and Kortus J 2001 Structural, electronic, and magnetic properties of MnO *Phys. Rev. B* **64** 024403
- [48] Deng H-X, Li J, Li S-S, Xia J-B, Walsh A and Wei S-H 2010 Origin of antiferromagnetism in CoO: a density functional theory study *Appl. Phys. Lett.* **96** 162508
- [49] Dufek P, Blaha P, Sliwko V and Schwarz K 1994 Generalized-gradient-approximation description of band splittings in transition-metal oxides and fluorides *Phys. Rev. B* **49** 10170–5
- [50] Jain A et al 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002
- [51] Sit P H-L, Cococcioni M and Marzari N 2006 Realistic quantitative descriptions of electron transfer reactions: diabatic free-energy surfaces from first-principles molecular dynamics *Phys. Rev. Lett.* **97** 028303
- [52] Sit P H-L, Cococcioni M and Marzari N 2007 Car–Parrinello molecular dynamics in the DFT+U formalism: structure and energetics of solvated ferrous and ferric ions *J. Electroanal. Chem.* **607** 107–12
- [53] Leiria Campo V Jr and Cococcioni M 2010 Extended DFT + U + V method with on-site and inter-site electronic interactions *J. Phys.: Condens. Matter* **22** 055602
- [54] Yu W et al 2023 Active learning the high-dimensional transferable Hubbard U and V parameters in the DFT + U + V scheme *J. Chem. Theory. Comput.* **19** 6425–33