Data integration for accelerated materials design via preference learning

Xiaolin Sun\textsuperscript{1}, Zhufeng Hou\textsuperscript{1,3}, Masato Sumita\textsuperscript{4,5}, Shinsuke Ishihara\textsuperscript{4}, Ryo Tamura\textsuperscript{1,2,4,5} and Koji Tsuda\textsuperscript{1,2,5,6}

\textsuperscript{1} Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa, Chiba 277-8561, Japan
\textsuperscript{2} Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan
\textsuperscript{3} State Key Laboratory of Structural Chemistry, Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences (CAS), Fuzhou, 350002 Fujian, People’s Republic of China
\textsuperscript{4} International Center for Materials Nanoarchitectonics (WPI-MANA), National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan
\textsuperscript{5} Center for Advanced Intelligence Project, RIKEN, 1-4-1 Nihombashi, Chu-o-ku, Tokyo 103-0027, Japan
\textsuperscript{6} Author to whom any correspondence should be addressed. E-mail: tsuda@k.u-tokyo.ac.jp

Keywords: preference learning, Gaussian process, data integration, surrogate model

Abstract

Machine learning applications in materials science are often hampered by shortage of experimental data. Integration with external datasets from past experiments is a viable way to solve the problem. But complex calibration is often necessary to use the data obtained under different conditions. In this paper, we present a novel calibration-free strategy to enhance the performance of Bayesian optimization with preference learning. The entire learning process is solely based on pairwise comparison of quantities (i.e., higher or lower) in the same dataset, and experimental design can be done without comparing quantities in different datasets. We demonstrate that Bayesian optimization is significantly enhanced via data integration for organic molecules and inorganic solid-state materials. Our method increases the chance that public datasets are reused and may encourage data sharing in various fields of physics.

1. Introduction

A substantial amount of materials data are accumulated in public databases \cite{1,2,3}, and machine-learning-based design of materials is increasingly common in recent years \cite{4,5,6}. The problem of materials design is mathematically formulated as a black-box optimization problem, where a large number of candidates are available and the goal is to find the candidate with best target property via a minimum number of observations. In Bayesian optimization \cite{7}, one of the most prominent methods of black-box optimization, a next candidate to observe is chosen using a Bayesian surrogate model trained with observed candidates. Gaussian process \cite{8} is one of most frequently used surrogate models that provide prediction together with uncertainty quantification. Besides materials discovery, Bayesian optimization has been applied to many physics applications including autonomous x-ray scattering experiments \cite{9}, inverse scattering \cite{10}, crystal structure prediction \cite{11}, design of organic synthesis experiments \cite{12}. Bayesian optimization is employed as a main machine-learning tool in ProjectAda, a Canadian lab-automation project (http://www.projectada.ca/).

Despite progress in materials informatics, machine learning often yields poor results due to shortage of experimental data \cite{13}. The problem may be solved by augmenting the current dataset with an external dataset in public databases or private repositories. However, data about the property of interest may not be available. Even if such a dataset is found, the observations are often done in different conditions or
Figure 1. Data integration with related data. Bayesian optimization is employed to find the materials of longest absorption wavelength from candidate materials A–F, where a machine learning model (ML) recommends a next material for experiments (EXP) repeatedly. In an external dataset, the HOMO–LUMO gap is available for several materials. Among them, B, C and E are included in the candidate materials, i.e., overlap. Our method aims to use the external dataset to accelerate the search of the best material in terms of the wavelength.

Figure 2. Data integration with preference learning. Two datasets are available but the measured values are not directly comparable. In our method, each dataset is separately translated to preference relations. A Gaussian process model is trained from all the relations. The trained model yields probability distributions of latent values at all points in the descriptor space.
no data about the property of interest in most cases. Multi-fidelity learning [13] can combine two datasets about the same property with different magnitude of error, but it is not applicable to incompatible properties.

In this paper, we propose a new calibration-free strategy of data integration without comparing the quantities in different datasets. Figure 2 illustrates our basic idea. First of all, each dataset is described as a set of pairwise relationships. Pairwise comparison is done for every pair of target values and the outcome is summarized as a set of ‘larger-than’ relationships. Then, a Bayesian surrogate model is learned from the two sets of pairwise relationships only. As a result, the learned model has a value range completely different from the original datasets, but it can still be used to select candidates with Bayesian optimization. One can use any preference learning method, but we employed Gaussian process-based method by Chu and Ghahramani [15] in this paper.

In benchmarking our method, we consider two materials search problems. First, we search for organic molecules with longer absorption wavelength [4]. Bayesian optimization is applied to 40 candidate compounds whose absorption wavelength is computed via TD-DFT. Using an external data about the HOMO–LUMO gap, we performed several computational experiments and found significant search acceleration. Second, an oxide with the largest bandgap is sought from 194 candidates [3]. Similar compounds whose absorption wavelength is computed via TD-DFT. Using an external data about the HOMO–LUMO gap, we performed several computational experiments and found significant search acceleration.

2. Methods

A set of candidate materials is represented as \( \{z_i\}_{i=1,...,N} \), where \( z_i \in \mathbb{R}^d \) is a vector of descriptors. The corresponding values of target property are represented as \( \{y_i\}_{i=1,...,N} \). They are initially unknown and revealed by observation. Let us assume that \( k \) observations are already made \( Z = \{(z_i, y_i)\}_{i=1,...,k} \) and we would like to choose a next candidate. In addition, we have an external dataset \( Z' = \{(z'_i, y'_i)\}_{i=1,...,k'} \) at hand. Before merging the datasets, each one is converted to preferences. If \( y_i > y_j \), we denote \( z_i > z_j \), i.e., \( z_i \) is preferred over \( z_j \), for \( 1 \leq i < j \leq k \). Similarly, if \( y'_i > y'_j \), we denote \( z'_i > z'_j \) for \( 1 \leq i < j \leq k' \). After comparing all pairs, \( Z \) and \( Z' \) are converted to preference sets of size \( k(k-1)/2 \) and \( k'(k'-1)/2 \), respectively. A Gaussian process is trained from the merged preference set, and subsequently used to rank the remaining candidates for next observation. Note that no comparison is made across the two datasets.

2.1. Gaussian process preference learning

In this section, we briefly review the preference learning method by Chu and Ghahramani [15]. For notational simplicity, all descriptor vectors in \( Z \cup Z' \) are redefined as \( X = \{x_i\}_{i=1,...,n} \). Let \( D \) denote the merged preference set,

\[
D = \{v_i \succ u_i\}_{i=1,...,m},
\]

where \( v_i, u_i \) are taken from \( X \). After learning from \( D \), the Gaussian process will be able to assign a latent value \( f(x) \) to any vector \( x \in \mathbb{R}^d \). In addition, the variance of a latent value can be inferred. Bayesian optimization will be performed based on these latent values.

The prior probability of \( f(x_i) \) is defined as

\[
P(f) = \frac{1}{(2\pi)^{d/2} |\Sigma|^2} \exp\left(-\frac{1}{2} f^T \Sigma^{-1} f\right),
\]

where \( f = [f(x_1), f(x_2), \ldots, f(x_n)]^T \), and \( \Sigma \) is the covariance matrix defined by a radial basis function kernel [8]. Using Gaussian noise variables \( \delta \sim \mathcal{N}(0, \sigma^2) \), the probability of preference \( v_k \succ u_k \) is described as

\[
P \left( v_k \succ u_k \mid f(v_k) > f(u_k) + \delta_v \right) = \int \int P \left( v_k \succ u_k \mid f(v_k) + \delta_v > f(v_k) + \delta_u \right) \mathcal{N}(\delta_v; 0, 1) \mathcal{N}(\delta_u; 0, 1) d\delta_v d\delta_u.
\]

The probability of data generation is then defined as

\[
P(D|f) = \prod_{k=1}^{m} P \left( v_k \succ u_k \mid f(v_k) > f(u_k) \right).
\]
By using Bayes’ theorem, we can arrive at the posterior probability,

\[ P(f|D) = \frac{P(f) P(D|f)}{P(D)} = \frac{P(f)}{P(D)} \prod_{k=1}^{m} P(\psi_k > u_k|f(v_k), f(u_k)). \]

The maximum \textit{a posteriori} estimate (MAP) of the latent values is defined as \( f^\text{MAP} = \arg \max_P P(f|D) \).

Taking the logarithm of the posterior probability, the solution is obtained by minimizing

\[ S(f) = -\sum_{k=1}^{m} \ln \psi(\psi_k) + \frac{1}{2} f^T \Sigma^{-1} f \quad \text{where} \quad \psi_k = \frac{f(v_k) - f(u_k)}{\sqrt{2}\sigma} \quad \text{and} \quad \Psi(s) = \int_{-\infty}^{s} N(\gamma; 0, 1) d\gamma. \]

To make a prediction at a new sample point \( x^* \), we infer the probability distribution of its latent value as

\[ P(f^*|D) = \int P(f^*|f)P(f|D)df \sim N(f^*; K^T \Sigma^{-1} f^\text{MAP}, K^{**} - K^T(\Sigma + \Lambda^{-1}_{\text{MAP}})^{-1}K^*), \]

where \( K^* = [K(x^*, x_1), K(x^*, x_2), \ldots, K(x^*, x_m)]^T \), \( K^{**} = K(x^*, x^*) \), and \( \Lambda_{\text{MAP}} \) is the Hessian matrix \( \frac{\partial^2 \psi(f)}{\partial f^2} = \Sigma^{-1} \) at \( f = f^\text{MAP} \). The predicted mean and variance of the latent value at \( x^* \) are \( \mu^* = K^T \Sigma^{-1} f^\text{MAP} \) and \( \sigma^* = K^{**} - K^T(\Sigma + \Lambda^{-1}_{\text{MAP}})^{-1}K^* \), respectively. All hyperparameters are set as instructed in [15].

2.2. Bayesian optimization based on preference learning

In Bayesian optimization, the mean latent value \( \mu^* \) and standard deviation \( \sigma^* \) are computed for all remaining candidates. Let \( \mu_{\text{max}} \) denote the maximum value observed so far. The expected improvement of a candidate \( x^* \) is described as follows.

\[ \text{EI}(x^*) = (\mu_{\text{max}} - \mu^*)\Phi \left( \frac{\mu_{\text{max}} - \mu^*}{\sigma^*} \right) + \sigma^* \varphi \left( \frac{\mu_{\text{max}} - \mu^*}{\sigma^*} \right), \]

where \( \Phi \) and \( \varphi \) represent the cumulative distribution function and the probability density function of standard normal distribution, respectively. The candidate with maximum expected improvement is chosen for next observation.

3. Results

3.1. Absorption wavelength of molecules

In this section, we evaluate how much Bayesian optimization is enhanced by exploiting an external dataset via preference learning. To evaluate the performance fairly, we need to be aware of the effect of overlap (figure 3). The overlap in our context is defined as the set of materials included both in the candidate set and the external data. In figure 1, materials B, C and E correspond to the overlap. If all the candidate materials to be observed in current research are included in external examples, the external dataset would provide plenty of information about experimental design (figure 3, right). With small overlap, it may be difficult to accelerate the search (figure 3, left).

We created our own small database of 94 organic molecules with their HOMO–LUMO gaps and absorption wavelength computed via TD-DFT (supplementary table 1). See [4] for computational details. To test the performance statistically, 50 candidate sets are created, each of which has 40 randomly selected molecules. Bayesian optimization is applied to discover the molecule with longest absorption wavelength.

For each candidate set \( C \), we created five types of external datasets, each consists of 50 molecules. For each candidate set \( C \), we created five types of external datasets, each consists of 50 molecules. For \( q = 0, 25, 50, 75 \) and 100, the \( q\%\)-overlap dataset consists of \( \lfloor qN/100 \rfloor \) molecules in \( C \), \( 50 - \lfloor qN/100 \rfloor \) molecules not in \( C \), and their HOMO–LUMO gaps. Since the HOMO–LUMO gap is inversely correlated to the absorption wavelength, the preferences about the HOMO–LUMO gap are flipped.

To see how the Gaussian process model is enhanced due to an external dataset, we evaluated it with ranking accuracy. First, molecules in \( C \) are divided into 80% training set and 20% test set. A Gaussian process model is trained with preferences derived from the training set and an external dataset. As descriptors, 200 dimensional features were obtained using RDKit Descriptors Calculators [16, 17]. The trained model is used to compute latent values of test examples. For the test set, the difference between two rankings due to experimental wavelengths and latent values are measured with a ranking accuracy measure called Normalized Discounted Cumulative Gain (NDCG) [18]. If rankings are completely identical, NDCG is one. A smaller value of NDCG indicates a larger difference in rankings. Figure 4(a) shows the ranking accuracy without any external dataset (i.e., single dataset) and that with a various type of external dataset.
Figure 3. In materials design with an external dataset, we search the best one from a set of candidates (red), using the information from a set of examples in the external dataset (blue). If these two sets have large overlap (right), we can make most of the external data for accelerating the search, while it would be difficult without no overlap (left).

Figure 4. Results for organic molecules about the integration of absorption wavelength and HOMO–LUMO gap data. (a) Ranking accuracy by Gaussian process with preference learning. (b) Success rate of Bayesian optimization with preference learning against the number of iterations.

The accuracy improved as the degree of overlap is increased and the accuracy is almost perfect for 100% overlap. This result indicates that an external dataset with a related but non-identical property can still improve our preference learning model.

Next, we performed Bayesian optimization. First, two molecules are randomly chosen and the selection with Bayesian optimization is applied from the third molecule. For a degree of overlap, we performed 50 runs of Bayesian optimization, where the initial two molecules and the external dataset was resampled in every run. The success rate at iteration $j$ is defined as the fraction of runs where the best molecule was found within $j$ selections of molecules. Figure 4(b) shows the result without any external dataset (i.e., single dataset) and with a different type of external set. Although the result at 0% overlap did not improve the single-set Bayesian optimization, significant acceleration was observed for all the other cases. This is an encouraging result for data sharing and reuse in materials science, because preference learning was shown to enlarge the scope of reusability. Namely, if a scientist makes a dataset about a property (e.g., HOMO–LUMO gap) publicly available, another scientist may be able to solve a materials discovery problem about a related property (e.g., absorption wavelength).

3.2. Bandgap of inorganic materials

The same series of benchmarking experiments is applied to another subject. The online material database of the National Renewable Energy Laboratory (NREL, https://materials.nrel.gov) provides bandgap calculated by Perdue Burke Ernzerhof (PBE) method of 2142 oxides [13]. Among these oxides, 194 oxides have the bandgap data by many-body GW calculation method [3]. GW calculation predicts band gaps more accurately but is far more computationally expensive than PBE [19, 20]. We define a search problem of finding the oxide with largest bandgap in terms of GW. 132 dimensional descriptors are obtained using ElementProperty featurizer of matminer [21]. The candidate set $C$ is determined as the 194 oxides with GW bandgaps and the total set $A$ corresponds to the 2142 oxides.
Figure 5. Results for oxides. (a) Ranking accuracy by Gaussian process with preference learning. (b) Success rate of Bayesian optimization with preference learning against the number of iterations.

Ranking accuracy and Bayesian optimization performances are shown in figures 5(a) and (b), respectively. With an external dataset without overlap, ranking accuracy was worse than that of single dataset. Nevertheless, Bayesian optimization was accelerated in comparison to the single dataset case. As section 3.1, a larger overlap resulted in higher accuracy and better acceleration.

4. Discussion and conclusion

We reported that preference-learning-based data integration works excellently in two kinds of materials datasets. This result is surprising and encouraging at the same time, because the conversion of numerical data to preferences incurs information loss in trade with calibration-free integration. Our method extends easily to deal with more than three datasets. In current materials science, data sharing is not commonly done due to difficulty of integration. Our method may promote cooperation among researchers to save the cost of expensive and time-consuming experiments.

Machine learning is often criticized as a black-box approach [22], and our preference learning method is not an exception. It achieves better prediction, but may not contribute in understanding of the underlying phenomenon. Recently, interpretable machine learning models (explainable AI) are of broad interest [23, 24]. In future work, it would be fruitful to extend our method by incorporating some of these ideas for better interpretability.

In materials sciences, there is wide-spread misunderstanding that machine learning always require a large amount of data. One favorable aspect of our results is that our method worked in small data scenarios (i.e., less than several hundred data points). When users want to use larger datasets, current implementation of our algorithm may not be very scalable, because the computational complexity is $O(m^3)$ [15] where $m$ is the number of preference relations. Recent developments in Gaussian process and preference learning [25, 26] may be beneficial in improving scalability.

Acknowledgments

This work was supported by ‘Materials Research by Information Integration’ Initiative (MI2I) project. XS would like to gratefully acknowledge the financial support from the China Scholarship Council (CSC No. 201809120018). KT is supported by NEDO P15009, SIP (Technologies for Smart Bio-industry and Agriculture), JST CREST JPMJCR1502, and JST ERATO. RT is supported by JST CREST JPMJCR17J2 and SIP (‘Materials Integration’ for Revolutionary Design System of Structural Materials). The authors thank Diptesh Das, Koki Kitai, Jinzhe Zhang and Yuan Yao for discussions.

ORCID iDs

Zhufeng Hou https://orcid.org/0000-0002-0069-5573
Koji Tsuda https://orcid.org/0000-0002-4288-1606
References

[1] Jain A et al 2013 APL Mater. 1 011002
[2] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 Phys. Rev. Lett. 108 058301
[3] Lany S 2013 Phys. Rev. B 87 085112
[4] Sumita M, Yang X, Ishihara S, Tamura R and Tsuda K 2018 ACS Cent. Sci. 4 1126
[5] Ju S, Shiga T, Feng L, Hou Z, Tsuda K and Shiomi J 2017 Phys. Rev. X 7 021024
[6] Krems R V 2019 Phys. Chem. Chem. Phys. 21 13392
[7] Ueno T, Rhone T D, Hou Z, Mizoguchi T and Tsuda K 2016 Mater. Discov. 4 18
[8] Rasmussen C E and Williams C K I 2005 Gaussian Processes for Machine Learning (Cambridge, MA: MIT Press)
[9] Noack M M, Doerk G S, Li R, Fukuto M and Yager K G 2020 Sci. Rep. 10 1525
[10] Vargas-Hernández R A, Guan Y, Zhang D H and Krems R V 2019 New J. Phys. 21 022001
[11] Tsuchi Sato N, Kino H, Miyake T, Tsuda K and Oguchi T 2018 Phys. Rev. Mater. 2 013803
[12] Häse F, Roch L M, Kreisbeck C and Aspuru-Guzik A 2018 ACS Cent. Sci. 4 1134
[13] Pilania G, Gubernatis J E and Lookman T 2017 Comput. Mater. Sci. 129 156
[14] Nakata M and Shimazaki T 2017 J. Chem. Inf. Model. 57 1300
[15] Chu W and Ghalamani Z 2005 Proc. 22nd Int. Conf. Mach. Learn. - ICML 05 (Bonn: ACM Press) pp 137–44
[16] Weininger D 1988 J. Chem. Inf. Comput. Sci. 28 31
[17] RDKit: Open-Source Cheminformatics Software (www.rdkit.org/)
[18] Wang Y, Wang L, Li Y, He D, Chen W and Liu T-Y 2013 Proc. 26th Annnu. Conf. Learn. Theory COLT 2013 p 6
[19] Karlický F and Otrepka M 2013 J. Chem. Theory Comput. 9 4155
[20] Crowley J M, Tahir-Kheli J and Goddard W A 2016 J. Phys. Chem. Lett. 7 1198
[21] Ward L et al 2018 Comput. Mater. Sci. 152 60
[22] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 Phys. Rev. Lett. 114 105503
[23] Rudin C 2019 Nat. Mach. Intell. 1 206
[24] Adadi A and Berrada M 2018 IEEE Access 6 52138
[25] Raykar V C, Duraiswami R and Krishnapuram B 2007 Artif. Intell. Stat. pp 388–95
[26] Ambikasaran S, Foreman-Mackey D, Greengard L, Hogg D W and O’Neil M 2015 IEEE Trans. Pattern Anal. Mach. Intell. 38 252