Learning formation energy of inorganic compounds using matrix variate deep Gaussian process

Saket Mishra

6Sense Insights Inc,
Bangalore, INDIA

Piyush Tagade

Next Gen Research,
Samsung Advanced Institute of Science and Technology,
Samsung R& D Institute-Bangalore, INDIA

Abstract

Future advancement of engineering applications is dependent on design of novel materials with desired properties. Enormous size of known chemical space necessitates use of automated high throughput screening to search the desired material. The high throughput screening uses quantum chemistry calculations to predict material properties, however, computational complexity of these calculations often imposes prohibitively high cost on the search for desired material. This critical bottleneck is resolved by using deep machine learning to emulate the quantum computations. However, the deep learning algorithms require a large training dataset to ensure an acceptable generalization, which is often unavailable a-priory. In this paper, we propose a deep Gaussian process based approach to develop an emulator for quantum calculations. We further propose a novel molecular descriptor that enables implementation of the proposed approach. As demonstrated in this paper, the proposed approach can be implemented using a small dataset. We demonstrate efficacy of our approach for prediction of formation energy of

inorganic molecules.

Keywords: Quantum chemistry, High throughput screening, Deep learning, Variational inference, Deep Gaussian process

---

1. Introduction

Identification and design of novel materials is critical for advancement of many engineering applications. Example of such applications include lithium ion batteries [1], photovoltaic cells [2], energy [3], storage materials [4], scintillator materials [5], etc. A typical materials discovery process relies heavily on trial and error, wherein, candidate materials are selected based on prior expertise and intuition, and subsequent experimental investigation is carried out to determine the material properties. However, cost and time-frame involved in this process significantly slows down the materials discovery process.

Automated high throughput screening (HTS) has a potential to significantly accelerate the materials discovery process [6, 7]. HTS uses quantum chemistry computations [8], that numerically solves the steady state Schrodingers equation to determine the material properties. Subsequently, experimental investigations are carried out for few potential candidates. Although efficient compared to the experimental investigations, computational cost of quantum chemistry computations limits this approach to screening few thousand candidate materials. Recent advancements in the machine learning methodologies can be exploited to further accelerate the materials discovery process.

A typical machine learning algorithm uses a large training dataset consisting of possibly multidimensional input and output datapoints to learn an optimal functional form for the input to output mapping [9, 10, 11]. Such an approach can be

used to establish a correlation between mathematical representation of molecular structures (known as fingerprints) and the corresponding molecular properties. Accuracies comparable to the quantum chemistry calculations can be achieved by using more advanced deep learning algorithms [12]. However, accuracies of these algorithms is dependent on two critical requirements:

1. choice of appropriate molecular fingerprint
2. availability of large training dataset

In this paper, we make two contributions. First, we propose a new molecular fingerprinting methodology. The proposed fingerprinting uses atomic position of to obtain radial and angular distribution functions. Subsequently, we project these distribution functions on a set of orthogonal Gegenbauer polynomials, the resulting coefficients are used as molecular fingerprints. Second, we develop a matrix variate deep Gaussian process approach to learn the molecular structure-properties correlation. The proposed approach can be implemented with a small dataset compared to the state of the art deep learning approaches. We demonstrate effectiveness of our approach for prediction of formation energy of crystal structures, which is an important indicator of molecular stability [13].

2. Proposed Methodology

2.1. Matrix variate deep Gaussian process

For a given dataset of input datapoint $\boldsymbol{x}$ and corresponding outputs $\boldsymbol{y}$, state of the art machine learning algorithms like deep neural networks assumes a parametric functional form for the input-output mapping, and subsequently uses an optimization algorithm for parameter estimation [10]. On the contrary, Gaussian

process regression (GPR) approach treats the input-output mapping as a random function with Gaussian process prior [14, 15, 16], and uses an optimization algorithm to estimate parameters of this Gaussian process prior. For an input-output mapping

$$\boldsymbol{y} = f(\boldsymbol{x}), \tag{1}$$

the Gaussian process prior is defined as

$$p(f(\boldsymbol{x}); \boldsymbol{\theta}) = \mathcal{GP}(\mu, \Sigma), \tag{2}$$

where the unknown parameters $\boldsymbol{\theta}$ are estimated by maximizing the log-likelihood of the dataset, given by

$$\mathcal{L} = \log\left[p(\boldsymbol{y}|\boldsymbol{f})\, p(f; \boldsymbol{\theta})\right]. \tag{3}$$

Present state of the art uses a variant of the GPR, known as a sparse Gaussian process (SGP) [17]. The SGP formulation uses pseudo-points, a set of assumed input datapoints $\boldsymbol{Z}$, such that

$$\boldsymbol{u} = f(\boldsymbol{Z}). \tag{4}$$

Subsequently using the marginalization, the log-likelihood is given by

$$\log\left[p(\boldsymbol{y})\right] = \log \int \left[p(\boldsymbol{y}|f)\, p(f, \boldsymbol{u})\right] df d\boldsymbol{u}. \tag{5}$$

However, evaluation of the log-likelihood is analytically intractable, thus, a variational inference [18] is used for maximization of the log-likelihood. For a given variational distribution $q(f, \boldsymbol{u})$, Eq. 5 is given by

$$\log\left[p(\boldsymbol{y})\right] = \log \int \left[\frac{p(\boldsymbol{y}|f)\, p(f, \boldsymbol{u})\, q(f, \boldsymbol{u})}{q(f, \boldsymbol{u})}\right] df d\boldsymbol{u}. \tag{6}$$

Using Jensen's inequality, the lower bound on the log-likelihood is obtained as

$$\mathcal{L} = \int \log \left[\frac{p\left(\mathbf{y}|f\right) p\left(f|\boldsymbol{u}\right) p\left(\boldsymbol{u}\right)}{q\left(f,\boldsymbol{u}\right)} \right] q\left(f,\boldsymbol{u}\right) df d\boldsymbol{u}. \tag{7}$$

On using $q\left(f,\boldsymbol{u}\right) = p\left(f|\boldsymbol{u}\right) q\left(\boldsymbol{u}\right)$, Eq. 7 simplifies to

$$\mathcal{L} = \int \log \left[p\left(\mathbf{y}|f\right) q\left(f\right) \right] df - \mathcal{KL}\left[q\left(\boldsymbol{u}\right) \| p\left(\boldsymbol{u}\right) \right] \tag{8}$$

where

$$q\left(f\right) = \int p\left(f|\boldsymbol{u}\right) q\left(\boldsymbol{u}\right) du, \tag{9}$$

and $\mathcal{KL}\left[q\left(\boldsymbol{u}\right) \| p\left(\boldsymbol{u}\right)\right]$ is a Kulback-Liebler divergence between $q\left(\boldsymbol{u}\right)$ and $p\left(\boldsymbol{u}\right)$. The lower bound $\mathcal{L}$ is evaluated by using samples from $q\left(f\right)$.

Similar to a deep neural network, a deep variant of the GPR formulation can be obtained by using a set of functions $f_1, f_2, ..., f_n$, such that the output of f_{i-1} is provided as an input to f_i. Similar to the SGP formulation, variational inference is used to obtain the lower bound as

$$\mathcal{L} = \int \log \left[p\left(\mathbf{y}|f_n\right) q\left(f_n\right) \right] df - \sum_{i=1}^{n} \mathcal{KL}\left[q\left(\boldsymbol{u}_i\right) \| p\left(\boldsymbol{u}_i\right) \right]. \tag{10}$$

In Young et al. [19], the DGP formulation is extended by using a matrix variate Gaussian process [20, 21, 15] prior for f_i and $\boldsymbol{u}_i$. We use the resultant Matrix-variate deep Gaussian process (MVDGP) to obtain structure-property correlation for inorganic molecules.

2.2. Molecular fingerprinting

Molecular fingerprinting procedure proposed in this paper is shown in Figure 1. Fingerprinting is used to mathematically represent a molecular structure. In this paper, we utilize radial and angular distribution functions for molecular

fingerprinting. We first obtain atomic position in a unit cell, as shown in Figure 1 (a), and subsequently use lattice vector to create a supercell with periodic boundary conditions (see Fig. 1 (b)). The atomic positions in the Cartesian coordinates are then converted to the spherical coordinate system (Fig. 1 (c)), and finally, radial and angular distribution functions are obtained using Gaussian kernel density estimation [22], as shown in Fig. 1 (d).

In the final step, we project the distribution functions on a suitably selected Gegenbauer polynomials [23, 24] to obtain the molecular fingerprints. Gegenbauer polynomials are a set of orthogonal polynomials on an interval $[-1, 1]$. The orthogonality is defined by

$$\int_{-1}^{1} C_n(r)C_m(r)(1-r^2)^{\alpha-\frac{1}{2}}dr = 0. \tag{11}$$

For a radial (or angular) distribution function $g(r)$, appropriately scaled in the interval $[-1, 1]$, orthogonal projection of $g(r)$ on Gegenbauer polynomials is given by

$$g_n = \int_{-1}^{1} g(r)C_n(r)(1-r^2)^{\alpha-\frac{1}{2}}dr = 0. \tag{12}$$

The coefficients g_n are used as molecular descriptors. A typical set of the coefficients is shown in Figure 1 (e).

3. Results

We demonstrate the effectiveness of our proposed approach for prediction of formation energy of inorganic compounds. For demonstration, we consider a dataset of 28000 inorganic molecules from the open quantum materials database [25]. Figure 2 investigates the distribution of the database using a violin plot. The violin

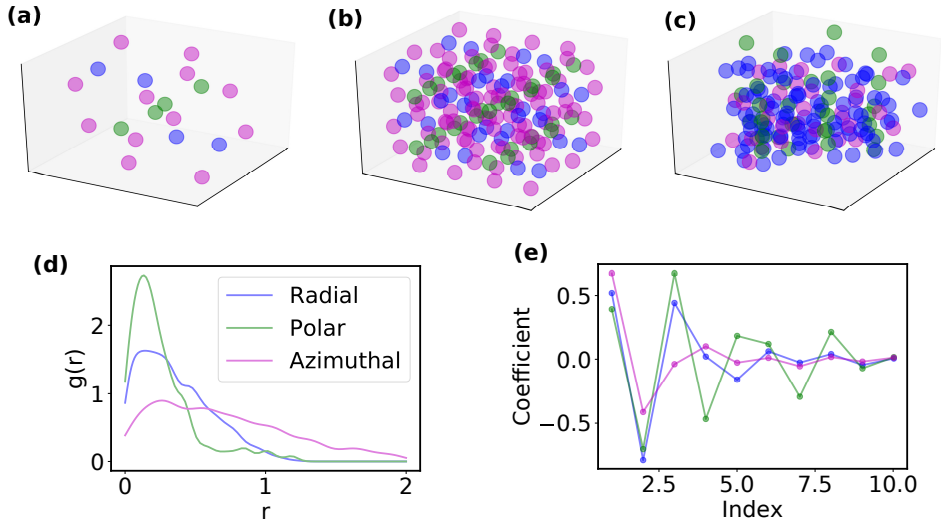

Figure 1: Molecular fingerprinting. Figure (a) a unit cell. Figure (b) shows a super cell obtained from a unit cell. Figure (c) shows atomic positions of the super cell in spherical co-ordinates. Figure (d) shows radial and angular distribution functions and figure (e) shows Gegenbauer coefficients of the distribution functions.

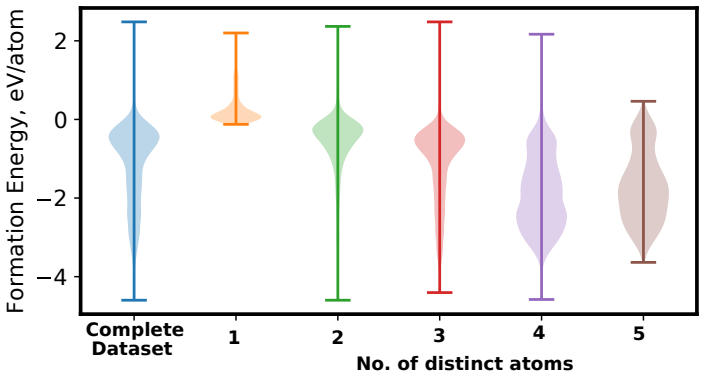

Figure 2: This figure investigates distribution of formation energy using violin plot. Leftmost plot shows distribution of formation energy of the complete dataset, while, the remaining plots show formation energy distribution as a function of number of distinct atoms in the inorganic molecule.

plot shows the PDF of the database obtained using a kernel density estimate. For better visualization, the distribution is shown symmetrically.

Leftmost plot shows distribution of the formation energy of the complete database. Subsequent five plots shows the distribution of the formation energy for molecules with number of distinct atoms. As can be observed from the figure, the distribution of formation energy depends on the number of different type of atoms present in the crystal structure. Inorganic molecules are expected to have negative formation energy, with majority of the molecules have formation energy greater than $-1.0eV/atom$. However, there is a noticeable tail to the distribution with formation energy $<= -1.0eV/atom$ (see leftmost violin plot of Fig. 2). For crystal structures consisting of single type of atoms, formation energy is primarily positive. The violin plot for crystal structures formed using two atoms closely resembles the violin plot for the complete dataset, however, heavy tail is not present in the violin plot. Although the violin plots of crystal structures formed using two and three atoms are similar, the distribution of formation energy for three atom structures has a noticeably heavier tail compared to the two atom structures. As can be observed from the Fig. 2, heavy tail of the distribution of formation energy observed for the complete database primarily consist of crystal structures formed using more than two atoms.

We use the OQMD database to train the MVDGP model. The MVDGP is implemented in TensorFlow [26] with the Adam optimizer [27] for parameter estimation. We first randomly split the OQMD database into 80% training and 20% test data. We randomly select 1000 molecules from the testing dataset. This dataset of 1000 molecules is used for validation of all the models presented in this paper. We randomly select 100 to 4000 molecules from the training dataset

to train MVDGP. For all the MVDGP models presented in this paper, we use five nodes in the hidden layer, while, we use 200 pseudo-points for input layer and 50 pseudo-points for the hidden layers. Figure 3 shows the mean absolute error (MAE) in prediction for validation dataset as a function of training dataset size and number of hidden layers in the MVDGP. The MAE decreases with increase in dataset size and number of hidden layers, with the MAE less than $0.3eV/atom$ is obtained for MVDGP with two hidden layers and 4000 training data.

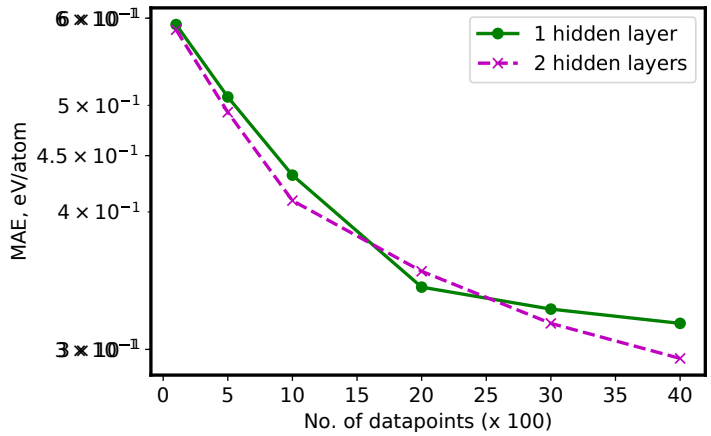

Figure 3: Figure shows mean absolute error in prediction for test data as a function of number of hidden layers and dataset size.

Figure 4 shows a Q-Q plot for training (subplot (a)) and testing (subplot (b)) dataset. Prediction for both the training and testing dataset falls near the 45^o line, showing high prediction accuracy of the MVDGP. Coefficient of determination $R^2 = 0.69$ is obtained for the training dataset and $R^2 = 0.72$ is obtained for the validation dataset.

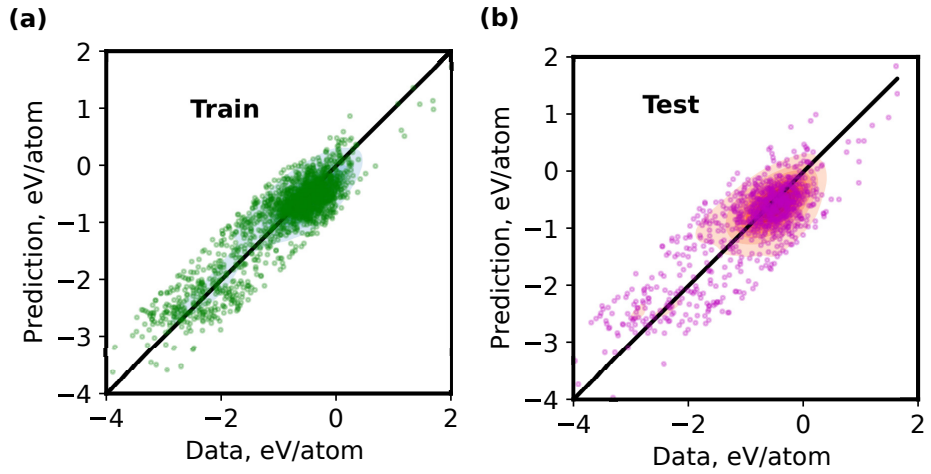

Figure 4: Quantile-quantile plot for MVDGP prediction. Figure (a) shows the Q-Q plot for the training dataset and figure (b) shows the Q-Q plot for the testing dataset.

4. Concluding Remarks

In this paper, we have proposed a matrix-variate deep Gaussian process model for prediction of the formation energy of inorganic molecules. We have also proposed a new fingerprinting approach for the inorganic molecules. Effectiveness of the proposed approach is demonstrated for OQMD database. Using MVDGP, MAE less than 0.3 and $R^2 = 0.72$ is obtained for the validation dataset. Accuracy of a machine learning approach critically depends on a training dataset. However due to a heavy tailed distribution of formation energy of crystal structures of OQMD dataset, randomly selected training dataset fail to adequately represent the complete chemical space of the crystal structures. In the future, we will extend the work presented in this paper to develop an algorithm for concurrent design of experiments and formation energy prediction of the crystal structures.

References

[1] M. S. Park, Y.-S. Kang, D. Im, S.-G. Doo, and H. Chang, "Design of novel additives and nonaqueous solvents for lithium-ion batteries through screening of cyclic organic molecules: an ab initio study of redox potentials," *Physical Chemistry Chemical Physics*, vol. 16, no. 40, pp. 22391–22398, 2014.

[2] C. Wadia, A. P. Alivisatos, and D. M. Kammen, "Materials availability expands the opportunity for large-scale photovoltaics deployment," *Environmental science & technology*, vol. 43, no. 6, pp. 2072–2077, 2009.

[3] E. O. Pyzer-Knapp, K. Li, and A. Aspuru-Guzik, "Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery," *Advanced Functional Materials*, vol. 25, no. 41, pp. 6495–6502, 2015.

[4] L.-C. Lin, A. H. Berger, R. L. Martin, J. Kim, J. A. Swisher, K. Jariwala, C. H. Rycroft, A. S. Bhown, M. W. Deem, M. Haranczyk, *et al.*, "In silico screening of carbon-capture materials," *Nature materials*, vol. 11, no. 7, p. 633, 2012.

[5] C. Ortiz, O. Eriksson, and M. Klintenberg, "Data mining and accelerated electronic structure theory as a tool in the search for new functional materials," *Computational Materials Science*, vol. 44, no. 4, pp. 1042–1049, 2009.

[6] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nature materials*, vol. 12, no. 3, p. 191, 2013.

[7] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," *npj Computational Materials*, vol. 3, no. 1, p. 54, 2017.

[8] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, "A high-throughput infrastructure for density functional theory calculations," *Computational Materials Science*, vol. 50, no. 8, pp. 2295–2310, 2011.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[10] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[11] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[12] G. B. Goh, N. O. Hodas, and A. Vishnu, "Deep learning for computational chemistry," *Journal of computational chemistry*, vol. 38, no. 16, pp. 1291–1307, 2017.

[13] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical review letters*, vol. 108, no. 5, p. 058301, 2012.

[14] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced lectures on machine learning*, pp. 63–71, Springer, 2004.

[15] P. M. Tagade, B.-M. Jeong, and H.-L. Choi, "A gaussian process emulator approach for rapid contaminant characterization with an integrated multizone-cfd model," *Building and Environment*, vol. 70, pp. 232–244, 2013.

[16] P. M. Tagade and K. Sudhakar, "Inferencing component maps of gas turbine engine using bayesian framework," *Journal of Propulsion and Power*, vol. 27, no. 1, pp. 94–104, 2011.

[17] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate gaussian process regression," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 1939–1959, 2005.

[18] Y. Gal, "Uncertainty in deep learning," *University of Cambridge*, 2016.

[19] Y.-J. Park, P. M. Tagade, and H.-L. Choi, "Deep gaussian process-based bayesian inference for contaminant source localization," *arXiv preprint arXiv:1806.08069*, 2018.

[20] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*. Chapman and Hall/CRC, 2018.

[21] P. Tagade, K. S. Hariharan, S. Basu, M. K. S. Verma, S. M. Kolake, T. Song, D. Oh, T. Yeo, and S. Doo, "Bayesian calibration for electrochemical thermal model of lithium-ion cells," *Journal of Power Sources*, vol. 320, pp. 296–309, 2016.

[22] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[23] D. Gottlieb and C.-W. Shu, "On the gibbs phenomenon and its resolution," *SIAM review*, vol. 39, no. 4, pp. 644–668, 1997.

[24] P. M. Tagade and H.-L. Choi, "Mitigating gibbs phenomena in uncertainty quantification with a stochastic spectral method," *Journal of Verification, Validation and Uncertainty Quantification*, vol. 2, no. 1, p. 011003, 2017.

[25] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The open quantum materials database (oqmd): assessing the accuracy of dft formation energies," *npj Computational Materials*, vol. 1, p. 15010, 2015.

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[27] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.