TOWARDS ACCURATE MODELING OF DYNAMICS FOR MOLECULAR CRYSTALS BY SCALABLE VARIATIONAL GAUSSIAN PROCESSES

Mikhail Tsitsvero
Institute for Chemical Reaction Design and Discovery (WPI-ICReDD)
Hokkaido University
Sapporo 001-0021, Japan
tsitsvero@icredd.hokudai.ac.jp

Andrey Lyalin
Department of Chemistry, Faculty of Science,
Hokkaido University,
Sapporo 060-0810, Japan
Research Center for Energy and Environmental Materials (GREEN)
National Institute for Materials Science
Namiki 1-1, Tsukuba 305-0044, Japan
lyalin@icredd.hokudai.ac.jp

Mingoo Jin
Institute for Chemical Reaction Design and Discovery (WPI-ICReDD)
Hokkaido University
Sapporo 001-0021, Japan
mingoo@icredd.hokudai.ac.jp

ABSTRACT

During the last decade, machine learning methods enabled the accurate prediction of properties for molecular and solid-state systems at various scales. In particular, Gaussian process (GP) models for modeling potential energy surfaces allowed accurate simulation of the energetics and dynamics of molecules and materials at different scales using moderate amounts of high-cost DFT training data. One of the current limitations hindering the wider applicability of GP models to larger systems is the computational cost associated with the training of complex multi-parametric GP models. In this work, we present a scalable multi-parametric GP model for atomic forces and energies. The presented modeling framework enables training of a wide range of variational GP models, possibly with millions of parameters. We illustrate the proposed methodology by training the collection of the variational GP models for atomic forces for a recently synthesized large molecular crystal system with dynamic gearing parts. We demonstrate that the proposed variational GP model can provide more compact representation of the data with respect to the exact GP models without loss of accuracy.

Keywords Gaussian processes · variational inference · invariant molecular descriptors

1 Introduction

The computational discovery of materials is now a reality due to the availability of high-throughput calculations in conjunction with fast and efficient machine learning techniques. Gaussian process (GP) machine learning (ML) models have been proven to provide excellent fitting quality for the prediction of the local and global, tensorial and scalar properties of materials, along with accurate estimates of uncertainties [2]. However, due to the computational costs

associated with training complex GP models, only simple Gaussian process models are routinely employed in material science. The complexity of the currently used models for predictions of local atomic properties such as forces is limited to low-parametric kernels, which severely limits their representation capacity. Within the field of material science, this means that current models are limited to the accurate treatment of systems with relatively simple dynamics and a limited diversity of atomic environments.

Building on top of our recent works [12, 11], here we present a methodological framework that enables the variational training of scalable multi-parametric Gaussian process models and apply it to fit the atomic forces of molecular crystal. The proposed framework uses the Smooth Overlap of Atomic Positions (SOAP) invariant molecular descriptors [1] as an input set of features and employs the recently proposed predictive log-likelihood approximation to the high-cost marginal log-likelihood objective [5]. Additionally, we propose training a collection of GP models based on symmetries within a molecular crystal, as illustrated in Fig. 1 (a), which allows us to achieve scalability without sacrificing accuracy. We illustrate the variational GP modeling framework by fitting the atomic forces for a recently synthesized molecular crystal system with dynamic gearing parts [6]. We show that variational GP models for atomic forces can provide similar accuracy to the exact GP models, however, using the number of variationally learned high-dimensional inducing points, which is only half of the total number of training data points.

2 Scalable variational Gaussian processes

Gaussian processes (GPs) offer rich non-parametric models for regression and classification tasks [14]. GP is fully specified by a mean function $\mu : \mathbb{R}^d \to \mathbb{R}$ and kernel or covariance function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Choosing a different kernel and mean allows to encode the prior information on the process that generated the data. A typical example of a kernel function that we will use in this work is the scaled Matern kernel with Automatic Relevance Determination (ARD) of features [14]:

$$k_{\text{Matern}}\ (\mathbf{x}_1, \mathbf{x}_2) = \theta_s \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu}d')^\nu K_\nu(\sqrt{2\nu}d') \tag{1}$$

where Γ is a Gamma function, K_ν modified Bessel function of the second kind, $d' = \frac{1}{2}\sum_{i=1}^d \frac{1}{\theta_i^2}(x_i - x_i')^2$ is a weighted distance between points, $\theta_s > 0$ is a scale parameter, $\theta_i > 0$ are the lengthscale parameters, and in this work we use $\nu = 2.5$. This type of kernel automatically learns to assign shorter lengthscales over the irrelevant dimensions in a descriptor space. As an input set of features in this work, we employ the SOAP descriptors and their derivatives with respects to atomic positions[1], which were computed by the LIBRASCAL software library [13].

Typically, all the parameters of GP can be jointly learned with backpropagation procedure by maximizing the marginal likelihood objective

$$p(\mathbf{y} \mid \mathbf{X}) = \int p\left(\mathbf{y} \mid \mathbf{f}, \sigma_{\text{obs}}^2\right) p(\mathbf{f} \mid \mathbf{X})\, \mathrm{d}\mathbf{f}, \tag{2}$$

where $\mathbf{y}$ is the vector with real-valued targets, $\mathbf{f}$ are the latent function values and $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ are the N input locations with $\mathbf{x}_i \in \mathbb{R}^d$. Unfortunately, expression (2) is often too expensive to compute when the number of training samples N is becoming large, since computation of the inverse of the kernel matrix $\mathbf{K}_{NN}^{-1}$ scales as $\mathcal{O}\left(N^3\right)$, where $\mathbf{K}_{NN} = k(\mathbf{X}, \mathbf{X})$. This motivates the design of the approximate methods allowing for scalable GP inference.

Inducing point methods [9], [10], [3] offer a scalable extension for the exact GPs, allowing to train the GP models using a large number of points, when the evaluation of the exact marginal likelihood (2) can be infeasible or when compact representation of the dataset is desirable. In our previous work [11] we compared the impact various training objectives have on the accuracy of uncertainty estimates of potential energies for a synthetic dataset of small molecules. In this work, based on our previous benchmarks for small molecules, we employ the Evidence Lower BOund (ELBO) based predictive log-likelihood approximation to the high-cost marginal log-likelihood objective [5].

The main idea behind the inducing point methods is to replace the large kernel matrix $\mathbf{K}_{NN}$ with its low rank approximation $\mathbf{K}_{MM}$ by introducing the latent variables $\mathbf{u} \in \mathbb{R}^M$ which are evaluated over the set of M inducing points $\mathbf{Z} = \{\mathbf{z}_m\}_{m=1}^M$, with each $\mathbf{z}_m \in \mathbb{R}^d$. Inducing points locations may be set to be learnable variational parameters of the model. In this scenario the GP prior in (2) is augmented by

$$p(\mathbf{f} \mid \mathbf{X}) \to p(\mathbf{f} \mid \mathbf{u}, \mathbf{X}, \mathbf{Z})p(\mathbf{u} \mid \mathbf{Z}). \tag{3}$$

Using the Jensen's inequality we can lower bound the log joint density over the targets and inducing variables

$$\log p(\mathbf{y}, \mathbf{u} \mid \mathbf{X}, \mathbf{Z}) = \log \int p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \mathbf{u})p(\mathbf{u})\, \mathrm{d}\mathbf{f}$$
$$\geq \mathbb{E}_{p(\mathbf{f}|\mathbf{u})}[\log p(\mathbf{y} \mid \mathbf{f}) + \log p(\mathbf{u})]. \tag{4}$$
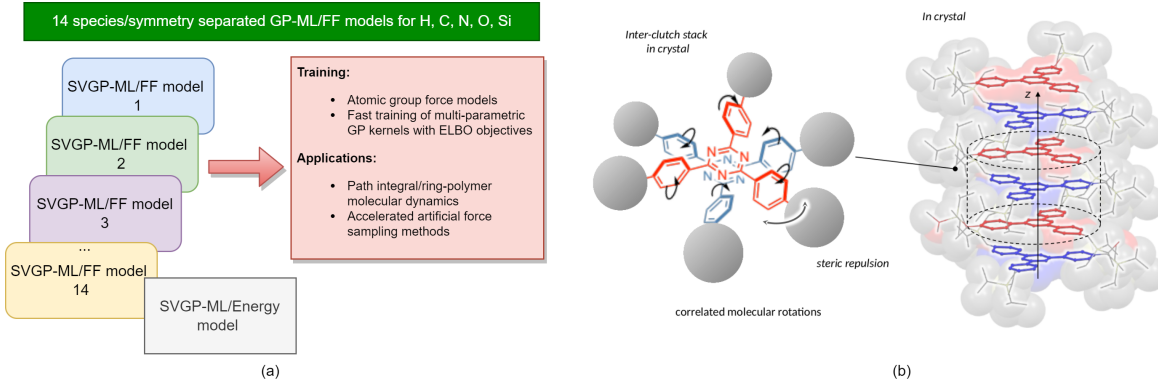
Figure 1: (a) The scalable Gaussian process (GP) framework for training and prediction of the local and global properties: atomic forces and structural energies. Depending on the complexity of local atomic environments, the complexity and representative power of each force or energy, GP model can be selected accordingly. The ensemble approach for training different GP models for different atomic groups enables further scaling to larger molecular systems, such as the molecular crystal system in (a). (b) Molecular crystal system with dynamic gearing parts. Accurate modeling of atomic forces is necessary for simulation of dynamics for such systems and is crucial for understanding the functional property (gearing motions of rings) in the solid-state system.

Finally, by introducing the variational distribution $q(\mathbf{u})$ over the inducing variables, we come up with the following lower bounds for the evidence $p(\mathbf{y})$ [4]

$$
\begin{aligned}
\log p(\mathbf{y}) &\geq \mathbb{E}_{q(\mathbf{u})}[\log p(\mathbf{y}, \mathbf{u} \mid \mathbf{X}, \mathbf{Z})] + H[q(\mathbf{u})] \\
&\geq \mathbb{E}_{q(\mathbf{u})}\left[\mathbb{E}_{p(\mathbf{f}\mid\mathbf{u})}[\log p(\mathbf{y} \mid \mathbf{f})]\right] - \mathrm{KL}[q(\mathbf{u})\|p(\mathbf{u})],
\end{aligned}
\tag{5}
$$

where $H[q(\mathbf{u})]$ is an entropy term.

In the most common scenario, the variational distribution over the inducing variables $\mathbf{u}$ is given by the multivariate normal $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$ parametrized by mean vector $\mathbf{m} \in \mathrm{R}^M$ and covariance matrix $\mathbf{S} \in \mathrm{R}^{M \times M}$. Depending on the desired expressivity of the GP model and complexity of the data, we can choose to learn either the full positive semidefinite symmetric covariance matrix with non-zero non-diagonal elements, use the mean-field approximation by restricting $\mathbf{S}$ to be diagonal, or even collapse the multivariate normal $q(\mathbf{u})$ to a Dirac delta distribution by taking the formal limit $\mathbf{S} \to 0$. In the examples in this work, we use full covariance matrix representation.

The predictive log-likelihood objective, proposed in [5], and which we previously applied to fit energies in a simple molecular dataset [11], aims to improve the uncertainty estimates of the latent function, and is given by

$$
\mathcal{L}_{\mathrm{PLL}} = \sum_{i=1}^{N} \log \mathbb{E}_{q(f_i)}\left[p\left(y_i \mid f_i\right)\right] - \beta\mathrm{KL}[q(\mathbf{u})\|p(\mathbf{u})].
\tag{6}
$$

For the above predictive log-likelihood objective, we can calculate the derivatives efficiently and perform the training in a "black box" manner [8]. In the above approximation of ELBO, the accuracy term, which ensures fitness to the training data, is tightly balanced by the regularization term, which additionally makes the algorithm search for variational distributions with higher entropy, as can be seen from the first bound in (5). It is worth noting that the KL divergence term may also be replaced by the more general Rényi divergence [7], thus allowing for wider classes of variational distributions to be explored.

3   Experiments

In this section, we provide numerical results illustrating the fitting quality of the stochastic variational GP (SVGP) models and compare the variational modeling framework with the exact one. We provide numerical results for fitting the atomic forces of large crystalline system with dynamic gearing parts [6]. Based on the symmetries of local atomic environments within the molecular crystal, we split the 528 atoms of the molecular crystal within a unit cell into 14 groups and trained 14 SVGP and exact GP models as illustrated in Fig. 1. For each SVGP model, we initialized 500 inducing points by randomly sampling the total of 1000 training points for each atomic group. To compare the performance of SVGP models trained with predictive log-likelihood objective (6) and exact GPs trained with objective
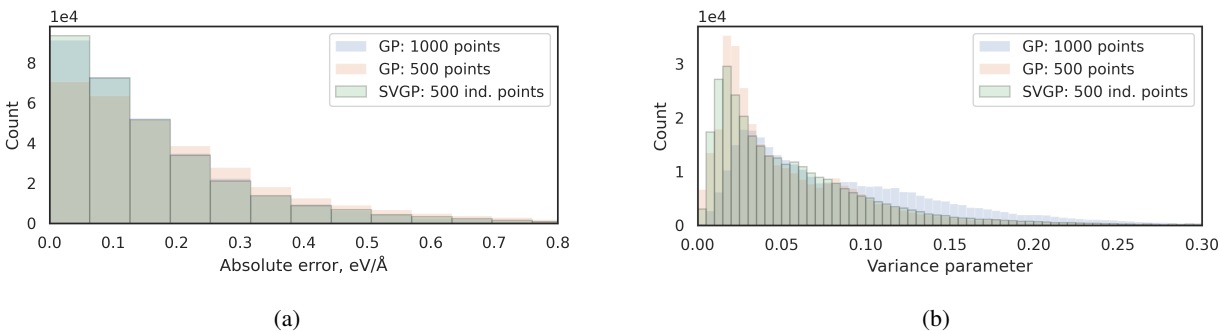
Figure 2: (a) Absolute error for atomic forces on all atoms predicted by the set of SVGP models, each parametrized with 500 inducing points and trained on 1000 points, and for two exact GP models trained with 500 and 1000 points. (b) Distributions of the normalized variance parameter learned by the SVGP and exact GP models.

(2), we additionally trained two sets of exact GP models: with 500 and 1000 force samples. The feature vectors that describe the local chemical environment of each atom and their derivatives with respect to atomic positions were represented by a 1200-dimensional SOAP vectors [1] that we used as input feature vectors.

To train the SVGP and exact GP models, we generated the training dataset by performing the 5000 steps of ab initio molecular dynamics simulation with density functional theory (DFT) with timestep of 1 ps at high temperature $T = 2000$ K. Calculations were carried out using the Perdue-Burke Ernzerhof (PBE) exchange-correlation functional and the projector-augmented wave (PAW) method as implemented in the Vienna ab initio Simulation Package (VASP). A plane wave basis set with an energy cutoff of 500 eV was used. To evaluate the performance of the SVGP and exact GP models, as a test trajectory, we took the trajectory and corresponding atomic forces of correlated barrier crossing by all the gears, see Fig. 1 (b), which we used in the paper [6] to estimate the rotational barriers of gears.

In Fig. 2 we provide numerical results for three modeling settings: (1) collection of 14 SVGP models, each with 500 inducing points and trained on 1000 points; (2) collection of 14 exact GP models trained on 1000 points; (3) collection of 14 exact GP models trained on 500 points. The inducing points of SVGP models were initialized by randomly sampling 500 points out of 1000 training data points. The SVGP and exact GP models were then trained similarly with backpropagation by optimizing the objective functions (2) and (6) respectively. Each model was trained by 1000 iterations with 0.1 learning rate step, followed by tuning 1000 iterations with a reduced learning rate of 0.01.

In Fig. 2 (a) we provide the absolute errors of the predicted atomic forces for three modeling setups. Despite the fact that SVGP for prediction uses 500 variationally learned inducing points, absolute errors for the given test dataset for the SVGP models are on par or even tend to be slightly lower than the absolute errors for the exact GP models. This could be related to the fact that each SVGP model has additional $500 \times 1200 = 600,000$ variationally learned parameters with respect to the exact GP model, allowing to find more optimal representation of the dataset. In Fig. 2 (b), we provide the histogram of the learned GP variance parameter for all atoms within a unit cell. Distribution of the variance parameters differ significantly for the SVGP and exact GP models, with less variability of the variance parameters for SVGP models with respect to the exact GP models trained on 1000 points.

To conclude, in this work we proposed to use a novel variational GP framework for fitting the molecular properties such as atomic forces and energies [11]. We demonstrated the applicability of SVGP method by fitting atomic forces for large molecular crystal system and compared it to the exact GPs. Based on the presented results, in the following, we will apply the presented SVGP modeling framework to simulate the high-cost ring-polymer molecular dynamics with the inherited accuracy of the training DFT data. This will enable simulations of molecular crystals with rotary gears containing up to several thousand atoms within a supercell at timescales up to several nanoseconds with an accuracy of high-cost ab initio calculations.

Acknowledgments

References

[1] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[2] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13):136403, 2010.

[3] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

[4] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR, 2015.

[5] Martin Jankowiak, Geoff Pleiss, and Jacob Gardner. Parametric Gaussian process regressors. In *International Conference on Machine Learning*, pages 4702–4712. PMLR, 2020.

[6] Mingoo Jin, Ryunosuke Kitsu, Natsumi Hammyo, Ayana Sato-Tomita, Motohiro Mizuno, Alexander S. Mikherdov, Mikhail Tsitsvero, Andrey Lyalin, Tetsuya Taketsugu, and Hajime Ito. A steric-repulsion-driven clutch stack of triaryltriazines: Correlated molecular rotations and a thermoresponsive gearshift in the crystalline solid. *Journal of the American Chemical Society*, 145(50):27512–27520, 2023.

[7] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*, 2019.

[8] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[9] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18, 2005.

[10] Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

[11] Mikhail Tsitsvero, Mingoo Jin, and Andrey Lyalin. Learning inducing points and uncertainty on molecular data by scalable variational Gaussian processes. *in review in SIAM/ASA Journal on Uncertainty Quantification, arXiv:2207.07654*, 2024.

[12] Mikhail Tsitsvero, Jenny Pirillo, Yuh Hijikata, and Tamiki Komatsuzaki. NMR spectrum prediction for dynamic molecules by machine learning: A case study of trefoil knot molecule. *The Journal of Chemical Physics*, 158(19), 2023.

[13] Michael J Willatt, Félix Musil, and Michele Ceriotti. Atom-density representations for machine learning. *The Journal of Chemical Physics*, 150(15):154110, 2019.

[14] Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.