Gaussian Processes with Spectral Delta kernel for higher accurate Potential Energy surfaces for large molecules

Rodrigo A. Vargas–Hernández¹,²,³ and Jake R. Gardner³

¹) Chemical Physics Theory Group, Department of Chemistry, University of Toronto,Toronto,Ontario M5S 3H6, Canada.
²) Vector Institute for Artificial Intelligence, Toronto, Canada.
³) Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA.

(Dated: 30 September 2021)

The interpolation of high-dimensional potential energy surfaces (PESs) is commonly done with physically-inspired deep-neural network models. In this work, we illustrate that Gaussian Processes (GPs) are also capable of interpolating high-dimensional complex physical systems. The accuracy of GPs depends on the robustness of the kernel function, and a boost in the accuracy is achieved by linearly combining kernel functions. In this work, we proposed an alternative route by parametrizing the kernel function through Bochners’ theorem. We interpolated the PES of various chemical systems achieving a global accuracy of $< 0.06$ kcal/mol for Benzene, Malonaldehyde, Ethanol, and protonated Imidazole dimer using only 15 000 training points. Additionally, for Aspirin, we achieved a global error of 0.063 kcal/mol with 20 000 points. Given these results, we believe this kernel function is system-agnostic and could allow GPs to tackle a wider variety of high-dimensional physical systems.

I. INTRODUCTION

Interpolation of potential energy surfaces (PESs) is one of the most common applications of supervised machine learning (ML) methods. Commonly, this is done with deep neural networks (DNNs) parametric models, or kernel models, such as Gaussian Processes (GPs) or kernel ridge regression (KRR). While both methodologies have proven to be flexible enough, GPs require less "tuning" compared to NNs, where the search for an optimal architecture could be computationally demanding. One of the advantages of GPs is their ability to quantify the uncertainty in their prediction, which is commonly used in applications where noisy data or the optimization of f black-box functions. The latter application has proven to be a successful tool in physical-chemistry.

GPs and KRR mainly depend on two factors: training data, and the kernel function. The latter quantifies the similarity between a pair of points. If the kernel function can generalize a similarity metric, GPs are efficient regression algorithms that require less training data than NNs, and are also capable of extrapolating functions beyond the training data regimen. To achieve more robust kernel functions, one can simply combine different simple kernel functions. However, to automate the search for an optimal kernel combination, one can use Bayesian model selection to search over possible priors, where a prior is determined by the kernel function form. Given the combinatorial nature of the space of possible kernel functions, sampling this distribution is hard and one must rely on heuristic search approaches. One such approach is the so called Automatic Bayesian Covariance Discovery (ABCD) described in greater detail below. Recently, people have proposed the use of Bayesian optimization for this as well.

Another technique that has proven to construct accurate PES is based on KRR using the atomic gradient information, also known as gradient-domain machine learning (GDML). Here, the kernel complexity is usually not modified. However, a combination of kernels could be easily implemented given the analytic derivatives of them; however, the training procedure is based on all possible combinations of kernels and the selection of them is through a cross-validation scheme. The computational complexity of GDML also scales cubically with the number of training data. In some cases, for large covariance matrices the Nyström approximation is applied.

Although kernel structure search has been shown to accurately interpolate of PESs, they are not often applied to study high-dimensional physical systems due to a lack of accuracy and significant computational cost of $O(N^3)$, where $N$ is the number of training points. Typically, approximations must be used to achieve the required scalability, which can degrade the predictive performance of GP models even further. For systems that require a higher number of training data points, multiple GPs must be trained in order to find the optimal kernel combination.

In this work, we present a different approach to construct kernel functions based on Bochners’ theorem that are both more robust and computationally efficient. In the following sections, we briefly introduce GPs and
review how complex kernels can be constructed by using Bochner’s theorem, a strategy known as spectral kernel learning. We interpolate the PES of four different chemical systems, e.g., protonated Imidazole dimer, Benzene, Malonaldehyde, and Ethanol. We also illustrate that by using GPytorch and KEOPS, modern deep-learning libraries, combined with GPUs we can efficiently train full GPs with 20,000 training points without relying on the Nyström approximation.

II. GAUSSIAN PROCESSES

Gaussian Processes are one of the most used non-parametric probabilistic ML models in physical sciences. A GP is specified by its mean (μ) and covariance (σ) functions, f(x) ~ GP(μ, σ). One of the key components in a GP model is the kernel function, k(x_i, x_j), used to parametrize the covariance matrix. For prediction, GPs have an exact posterior distribution whose mean and standard deviation are,

\[
\mu(x^*) = k(x^*, X)^T [K(X, X) + \sigma_n I]^{-1} y
\]

\[
\sigma(x^*) = k(x^*, x^*) - k(x^*, X) [K(X, X) + \sigma_n I]^{-1} k(x^*, X)
\]

where X, for PESs, contains the geometry configurations and y is the energy for each configuration. K(·, ·) is known as the design or covariance matrix, whose elements are given by the kernel function, K_{ij} = k(x_i, x_j). x^* is the point where prediction is aimed, and k(x^*, X) is the similarity vector between x^* and all the training data.

Given the nature of GPs, it is possible to find the marginal distribution in closed form, p(y|X). It is common to use the logarithm of the marginal likelihood distribution, LMLL, to optimize the free-parameters of the kernel function (θ),

\[
\log p(y|X, \theta) = -\frac{1}{2} y^T [K(X, X) + \sigma_n I]^{-1} y - \frac{1}{2} \log |K(X, X) + \sigma_n I| - \frac{N}{2} \log 2\pi
\]

Some of the most common kernel functions are,

\[
k_{RBF}(x_i, x_j) = \exp \left[ -\frac{1}{2} r^2(x_i, x_j) \right]
\]

\[
k_{RQ}(x_i, x_j) = \left( 1 + \frac{r^2(x_i, x_j)}{2\alpha} \right)^{-\alpha}
\]

\[
k_{MAT_{1/2}}(x_i, x_j) = \left( 1 + \sqrt{3} r(x_i, x_j) \right) \exp \left[ -\sqrt{3} r(x_i, x_j) \right]
\]

\[
k_{MAT_{3/2}}(x_i, x_j) = \left( 1 + \sqrt{5} r(x_i, x_j) + \frac{5}{3} r^2(x_i, x_j) \right) \exp \left[ -\sqrt{5} r(x_i, x_j) \right]
\]

where r^2(x_i, x_j) = (x_i - x_j)^T M(x_i - x_j), where M is a diagonal matrix parametrized with a ℓ_d length scale for each dimension of x. For more details about GPs we refer the reader to Ref. [39].

In previous works, it has been shown the accuracy of GPs increase by simply combining kernels to account for more complicated functions. One of the most common approaches was proposed in Refs. [34, 36, 38] where the combination of kernels was selected by criteria such as the maximum LMLL. To avoid selecting a kernel (M_i) with a large number of parameters, a term that penalizes the number of free parameters (|M_i|) in the kernel was included,

\[
\text{BIC}(M_i) = \log p(y|X, \theta, M_i) + \frac{1}{2} |M_i| \log N.
\]

This methodology is commonly named as the Bayesian information criterion (BIC). These kernel structure discovery methods have been demonstrated to extrapolate physical observables accurately enough to detect phase transitions. Additionally, GPs with complex kernels have shown the possibility to predict accurate energies for PES trained only with low energy points. The computational complexity of Gaussian processes compounded with the fact that kernel search requires training many such models has resulted in an inability to use full available datasets. Also, by considering highly complex kernel combinations, the optimization of the kernel parameters becomes harder, and given the greedy search strategy in practice one could end up with non-optimal kernels. Finally, raising the complexity of the kernel combination in practice hits a plateau in the learning capacity meaning, adding more kernels do not increase the accuracy of the models.

Another possibility to construct kernel functions automatically is through Bochner’s theorem.

\[
k(\tau) = \int_{\mathbb{R}^D} e^{2\pi i s^T \tau} S(s) \, ds,
\]

where \(\tau = x_i - x_j\) and \(S(s)\) is the spectral density of the k(·, ·). S(·) and k(·) are Fourier duals. In order for k(·) to be a valid kernel, S(·) must be integrable. For example, the spectral density of the SE kernel is also a Gaussian function. This theorem ensures that the kernel k(τ) parametrizes a positive-definite covariance matrix. The Matern family of covariance functions can also be derived using the Bochner’s theorem.

In Ref. [43] Lázaro-Gredilla et al. proposed a novel way to construct S(s) by assuming it is proportional to a probability measure, S(s) ∝ p_S(S). By doing so, the integral over the frequency domain can be computed by Monte Carlo,

\[
k(x_i - x_j) = \int_{\mathbb{R}^D} ds \, e^{2\pi i s^T (x_i - x_j)} S(s)
\]

\[
= \sigma_0^2 \int_{\mathbb{R}^D} ds \, e^{2\pi i s^T x_i} e^{-2\pi i s^T x_j} p_S(s)
\]

\[
= \sigma_0^2 \mathbb{E}_{p_S} \left[ e^{2\pi i s^T x_i} e^{-2\pi i s^T x_j} \right],
\]
where $\sigma_0^2$ is a normalization constant, and $E_{ps}$ is the expectation value with respect to $ps$. The samples from $ps$ used to compute $k(\cdot, \cdot)$ are known as spectral points.

It is possible to cancel the imaginary part of $k(\cdot, \cdot)$ in Eq. 10 by sampling a pair of $\{s_r, -s_r\}$. This Monte Carlo procedure is valid given that $S(s)$ is symmetric around zero. By taking this into account, $k(\cdot, \cdot)$ has the following closed form,

$$k(x_i - x_j) \approx \frac{\sigma_0^2}{2|\delta|} \sum_{r=1}^{\delta} \left[ e^{2\pi i s_r^T x_i} e^{-2\pi i s_r^T x_j} + e^{-2\pi i s_r^T x_i} e^{2\pi i s_r^T x_j} \right]$$

$$= \frac{\sigma_0^2}{|\delta|} \sum_{r=1}^{\delta} \cos(2\pi s_r^T (x_i - x_j)),$$

(11)

where $|\delta|$ is the total number of spectral points or samples used to approximate Eq. 10. $s_r$ are the frequencies that will be learned by maximizing Eq. 3. Because this kernel function corresponds to an explicit finite basis expansion, inference can be done in $O(1)$ time and space. As it is stated in Ref. 48, this approximation is similar to a set of Dirac deltas with amplitude $\sigma_0^2$ which are distributed accordingly to $p(s)$. For this work, we denote this kernel as the spectral delta kernel, $k_{SD}(\cdot, \cdot)$. Figure 1 depicts a simple example on how to approximate $k_{RBF}$ with 100 spectral points. It should be noted that $k_{SD}(\cdot, \cdot)$ is not the only possible kernel that can be derived from the Bornhuet’s theorem. In Ref. 19 it was shown that by assuming $S(s)$ as a linear combination of two Gaussians centered at $s$ and $-s$, the integral in the Fourier space has a close form where $k(\tau) = \sum q \omega_q e^{-2\pi \tau^T M_q} \cos(2\pi \tau^T \mu_q)$, $\omega_q$, $M_q$ and $\mu_q$ are the free parameters of this kernel.

III. RESULTS AND DISCUSSION

Here, we first compare the accuracy of the PES for the protonated imidazole dimer,15 interpolated with GPs with various kernel functions, i.e., Matern (Eq. 7), Spectral Delta (Eq. 11), and kernel combination through the BIC method. For the latter, we used the kernel that was optimized through the BIC method, with 5 000 points, in Ref. 15

$$k_H(x_i, x_j) = a_0 k_{MAT}^{s/2} + a_1 k_{MAT}^{3/2} + a_2 k_{RBF} + a_3 k_{RBF},$$

(12)

where all $a_i$s and the internal parameters of all the kernels were optimized by maximizing Eq. 3. The PES for the protonated imidazole dimer is a 51D surface with a range of energy points spanning from $[0, 35 000]$ cm$^{-1}$. For the $k_{SD}$, we considered a various number of spectral points, $|\delta|$, to study the impact on the accuracy of the model’s prediction. We quantified the accuracy of each model by computing the root-mean-square error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i} (y_i - \hat{y}_i)^2},$$

(13)

where $y_i$s are the predicted values with each GP, and $\hat{y}_i$ are the exact values computed at the MP2/6-31++G(d,p) level of theory.25 The test data set consists of 10 000 points spread throughout the same energy-range of the training points. All kernel parameters for each different GP were optimized by maximizing LMLL in the GPytorch suite using the Adam optimizer. To speed up the computation of the GPs with the $k_{MAT}$ and $k_{RBF}$ kernels we used the KEOPS library. All calculations were carried in a single GPU, V100 SXM2 32GB. For a GP with $k_H$, 18 000 training points was the maxi-
mmum number of points that we could used before running out of memory. The RMSEs for all different models, as a function of training points, are depicted in Fig. 2.

We found that in the low limit of training data, a GP with a simple kernel can predict a more accurate PES than a GP with the $k_{SD}$ kernel regardless of the number of spectral points. However, as the number of training points increases so does the accuracy of a GP with $k_{SD}$. This correlates with the idea that more points contain more information therefore more robust kernels can be designed. There is a significant difference between a $k_{SD}$ with 500 and 1 000 spectral points, where the RMSE is almost half of the most simple $k_{SD}$. The most accurate PES was achieved with a GP trained with 20 000 points using the $k_{SD}$ kernel with 5 000 spectral points, RMSE = 0.176 kcal/mol; twice more accurate than a GP with the $k_H$ kernel and 18 000 points (RMSE = 0.547 kcal/mol).

For fragments 1 and 2, individual molecule of Imidazol, the optimal kernel found with the BIC method for 5 000 energy points was,

$$k_{12}(\mathbf{x}_i, \mathbf{x}_j) = k_M(\mathbf{x}_i, \mathbf{x}_j) = a_0 k_{MAT}^{5/2} + a_1 k_{RQ} + a_2 k_{RBF}.$$  

\hspace{1cm} (14)

The third fragment describes the bridge between fragments 1 and 2. The optimal kernel optimized by the BIC method is,

$$k_{12}(\mathbf{x}_i, \mathbf{x}_j) = \left(a_0 k_{MAT}^{5/2} \times k_{RQ} + a_1 k_{RBF}\right) \times k_{MAT}^{15/8}.$$  

\hspace{1cm} (15)

We found that, for $|\delta| = 15 000$ and $N = 15 000$ we achieved a total RMSE of 0.056 kcal/mol, where for fragments 1 and 2 the average error is 0.032 kcal/mol and an error of 0.086 kcal/mol for the fragment that describe the interaction between both Imidazole molecules. For 5 000 training points, the $k_{SD}$ kernel can interpolate a more accurate PES than with a GP with the Matern kernel, even with 3 times more training points; Table II.

We also consider the interpolation of the Benzene, Malonaldehyde, Ethanol, and Aspirin systems. Our results illustrate that GPs, trained with the spectral density kernel, interpolate with high accuracy the PESs for high-dimensional molecular systems. For all four systems, we computed the mean absolute error (MAE) (Eq. 16) for the entire data set, including the training set.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$  

\hspace{1cm} (16)

All calculations for GPs were carried in a single GPU, Tesla T4 16 GB. In Table II we report the number of training and spectral points that lead to the most accurate GP. For all four systems, except for Malonaldehyde, we found that GPs with $k_{sd}$ interpolate more accurately that state-of-the-art deep-learning models, e.g., deep-tensor NN (DTNN) and Cormorant, and KRR methods combined with gradients, e.g., GDI,[12] and sGDI.[11] Figure 3 displays the value of the optimized $k_{SD}$ for each pair of geometries in the data sets for all four molecules.

| $|\delta|$ | $N$ | $\xi_1$ | $\xi_2$ | $\xi_{12}$ | Total |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 10 000 | 5000 | 0.074 | 0.064 | 0.205 | 0.131 |
| 10000 | 0.048 | 0.041 | 0.117 | 0.077 |
| 15000 | 0.041 | 0.038 | 0.087 | 0.059 |
| 20000 | 0.062 | 0.039 | 0.073 | 0.060 |
| 15 000 | 5000 | 0.073 | 0.064 | 0.206 | 0.131 |
| 10000 | 0.044 | 0.038 | 0.113 | 0.073 |
| 15000 | 0.034 | 0.030 | 0.086 | 0.056 |
| $k_{MAT}^{5/2}$ | 15 000 | 0.227 | 0.162 | 0.102 | 0.171 |
| $k_{1,2}, k_{12}$ | 5000 | 0.124 | 0.099 | 0.124 | 0.116 |

$^a$ From Ref. [13], the RMSE is 0.1815 kcal/mol.

FIG. 2. RMSE of the protonated Imidazole dimer interpolated with a GP with different kernels. We considered the $k_{MAT}^{5/2}$, $k_{RQ}$, $k_{1,2}$, and $k_{RBF}$ with a different number of $|\delta|$. The RMSE for each model was computed with 10 000 energy points that were not including in the training set.

TABLE I. The RMSE of each fragment and total PES with 10,000 test energy points spanned in the same energy range from the training data, as a function of the spectral, $|\delta|$, and training points $N$. 

$^b$ Eqs. (14, 15).
TABLE II. The lowest error for four different molecular systems computed with a GP using the $k_{SD}$ kernel. The total error was computed for the entire data set, including the training points.

|                 | $k_{SD}$ | MAE [meV] | MAE [kcal/mol] |
|-----------------|----------|-----------|----------------|
|                 | $N$      | $|\delta|$ | $N$            | $|\delta|$ |
| Benzene         | 10 000   | 0.31      | 0.41           | 0.0071    |
| Malonaldehyde   | 20 000   | 2.91      | 4.44           | 0.067     |
| Ethanol         | 20 000   | 1.87      | 4.16           | 0.043     |
| Aspirin         | 20 000   | 2.73      | 3.65           | 0.063     |

TABLE III. MAE, reported in kcal/mol, for different ML models. The total error was computed for the entire data set, including the training points.

|                 | $N$ [10$^3$] | ML model | MAE [kcal/mol] |
|-----------------|--------------|----------|----------------|
| Benzene         | 50           | SchNet$^a$ | 0.070           |
|                 | 50           | Cormorant$^b$ | 0.020           |
|                 | 10           | GP $k_{SD}(|\delta|=5K)$ | 0.007           |
| Malonaldehyde   | 15           | GDML$^c$  | 0.08            |
|                 | 15           | sGDML$^d$ | 0.074           |
|                 | 15           | PhysNet$^e$ | 0.072           |
|                 | 15           | GP $k_{SD}(|\delta|=6K)$ | 0.079           |
| Ethanol         | 15           | GDML$^c$  | 0.058           |
|                 | 15           | sGDML$^d$ | 0.051           |
|                 | 15           | PhysNet$^e$ | 0.050           |
|                 | 15           | GP $k_{SD}(|\delta|=6K)$ | 0.049           |
| Aspirin         | 15           | GDML$^c$  | 0.151           |
|                 | 50           |           | 0.130           |
|                 | 15           | sGDML$^d$ | 0.131           |
|                 | 15           | PhysNet$^e$ | 0.124           |
|                 | 50           |           | 0.121           |
|                 | 50           | SchNet$^a$ | 0.12            |
|                 | 50           | Cormorant$^b$ | 0.098           |
|                 | 20           | GP $k_{SD}(|\delta|=5K)$ | 0.063           |

$^a$ Ref.[9], $^b$ Ref.[11], $^c$ Ref.[41], $^d$ Ref.[51], $^e$ Ref.[12].

Aspirin is the largest system considered here, a 57D PES described with 210 features. We found that with 15 000 training points and 2 000 spectral points, the MAE of this GP is 0.127 kcal/mol. By increasing both, $N$ and $|\delta|$ we managed to reduce the error to 0.063 kcal/mol, Table II.

For Benzene, we achieved a RMSE = 0.31 meV with a GP trained with only 10 000 points and 5 000 spectral points. However, a GP with only $N = 5$ 000 and $|\delta| = 200$ is capable of predicting a more accurate PES than GDML and DTNN. All results are displayed in Fig. 4.

FIG. 3. For each molecule, we computed the value of spectral density covariance function, $k_{SD}$, for each pair of geometries $(x_i, x_j)$ in the data set. $k_{SD}$ was optimized by maximizing the LMLL, Eq. 3, and the number of deltas for each system is reported in Table II. The displayed values were normalized for convenience.

The results for Malonaldehyde are presented in Fig. 5. For a GP trained with 15 000 points and 5 000 spectral points, the PES’s accuracy is almost the same as the one predicted with GDML. However, by increasing the number of training points to 18 000, we achieved a more accurate prediction, MAE = 3.05 kcal/mol more accurate than the one with GDML. Any PES interpolated with a GP trained with $|\delta| \geq 500$ is more accurate than the one with DTNN.

For Ethanol, the predicted surface of a GP with $N = 15$ 000 and 1 000 spectral points is comparable with the one computed with GDML, see Fig. 6. However, a GP with a larger number of $|\delta|$ is capable of interpolating this system more accurately; for a GP with $|\delta| = 6$ 000 and $N = 15$ 000 points, the MAE is still the same as the one
We have presented an accurate GP model capable of interpolating high dimensional PESs, i.e., 51D for protonated imidazole dimer, 61D for Benzene, 9D for both Malonaldehyde and Ethanol (36 features for both), and 57D for Aspirin surfaces (210 features). While combining simple kernels have proven to be a successful route, here we show an alternative path to enhance the accuracy of GPs based on the Bochners’ theorem. This methodology can lead to highly accurate GPs capable of interpolating PESs for a variety of chemical systems. The spectral delta kernel (Eq. 11), derived from the Bochners’ theorem, can also be optimized by maximizing the log-marginal likelihood as is commonly done for vanilla GPs. For the protonated Imidazole dimer, we managed to predict a PES with a RMSE of 0.22 kcal/mol using 15 000 training and 2 000 spectral points. By splitting the total PES into three fragments, we achieved a PES with a total RMSE of 0.06 kcal/mol with $|\delta| = [5000, 10000]$. We found that $k_{SD}$ is only more accurate than the BIC method or simple kernels for systems with more than 5 000 points. Additionally, the optimal value of $|\delta|$ depends on the number of training points too; for example, for Benzene $|\delta| \geq 1$ 000 with $N = 10$ 000 produce surfaces with MAEs lower than 1 meV ($\approx 0.023$ kcal/mol). The largest dimensional system considered here was Aspirin. For this system, we managed to predict a global surface with a MAE lower than 0.07 kcal/mol with 20 000 points, which compared with deep-learning models is more accurate. By eye-inspection, we found that the optimal $|\delta|$ is between $|\delta| = [N/3, N/2]$. A possible future work is to study how well GPs with the $k_{SD}$ kernel can extrapolating quantum observables, e.g., the high energy points for
In physical sciences, the interpolation of high-dimensional landscapes is not commonly done by GPs unless the full covariance matrix is approximated by a low-rank matrix, for example using the Nyström approximation, or using deep-NN which architecture must be optimized for different systems. Here, we illustrate that by combining modern deep-learning libraries such as GPytorch and KEOPS with GPUs, GPs are robust supervised ML algorithms capable of approximating high-dimensional complex functions without approximating the covariance matrix and still being training points efficient. The work presented here makes GPs more suitable for applications.

The work presented here makes GPs more suitable for applications.
31. M. Fritz, M. Fernández-Serra, and J. M. Soler, “Optimization of an exchange-correlation density functional for water,” The Journal of Chemical Physics 144, 224101 (2016).

32. Z. Hou and K. Tsuda, “Bayesian optimization in materials science,” in Machine Learning Meets Quantum Physics, edited by K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller (Springer International Publishing, Cham, 2020) pp. 413–426.

33. R. A. Vargas-Hernández, C. Chuang, and P. Brumer, “Multi-objective optimization for retinal photoisomerization models with respect to experimental observables,” (2021), arXiv:2106.10390 [physics.chem-ph].

34. R. A. Vargas-Hernández, J. Sous, M. Berciu, and R. V. Krems, “Extrapolating quantum observables with machine learning: Inferring multiple phase transitions from properties of a single phase,” Phys. Rev. Lett. 121, 255702 (2018).

35. R. A. Vargas-Hernández and R. V. Krems, “Physical extrapolation of quantum observables by generalization with gaussian processes,” in Machine Learning Meets Quantum Physics, edited by K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller (Springer International Publishing, Cham, 2020) pp. 171–194.

36. D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, and G. Zoubin, “Structure discovery in nonparametric regression through compositional kernel search,” in Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 28, edited by S. Dasgupta and D. McAllester (PMLR, Atlanta, Georgia, USA, 2013) pp. 1166–1174.

37. A. Jasinski, J. Montaner, R. C. Forrey, B. H. Yang, P. C. Stancl, N. Balakrishnan, J. Dai, R. A. Vargas-Hernández, and R. V. Krems, “Machine learning corrected quantum dynamics calculations,” Phys. Rev. Research 2, 032051 (2020).

38. D. K. Duvenaud, H. Nickisch, and C. Rasmussen, “Additive gaussian processes,” in Advances in Neural Information Processing Systems Vol. 24, edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (Curran Associates, Inc., 2011) pp. 226–234.

39. C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning (MIT Press Ltd, United States, 2005).

40. G. Malkomes, C. Schaff, and R. Garnett, “Bayesian optimization for automated model selection,” in Workshop on Automatic Machine Learning (PMLR, 2016) pp. 41–47.

41. S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, “Machine learning of accurate energy-conserving molecular force fields,” Science Advances 3 (2017), 10.1126/sciadv.1603015. https://advances.sciencemag.org/content/3/5/ea603015.full.pdf

42. S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, “Towards exact molecular dynamics simulations with machine-learned force fields,” Nature Communications 9, 3887 (2018).

43. K. Asnaashari and R. V. Krems, “Gradient domain machine learning with complex kernels: improving the accuracy of pes and force fields for large molecules,” (2021), arXiv:2107.04779 [physics.chem-ph].

44. J. Hensman, N. Fusi, and N. D. Lawrence, “Gaussian processes for big data,” in Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI’13 (AUAI Press, Arlington, Virginia, USA, 2013) p. 282–290.

45. H. Liu, Y. S. Ong, X. Shen, and J. Cai, “When gaussian process meets big data: A review of scalable gps,” IEEE Transactions on Neural Networks and Learning Systems 31, 4405–4423 (2020).

46. J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, “Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration,” (2019), arXiv:1809.11165 [cs.LG].

47. B. Charlier, J. Feydy, J. A. Glaunès, F.-D. Collin, and G. Durif, “Kernel operations on the GPU, with autodiff, without memory overflows,” arXiv preprint arXiv:2004.11127 (2020).

48. M. Lázaro-Gredilla, J. Quiñero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, “Sparse spectrum gaussian process regression,” Journal of Machine Learning Research 11, 1865–1881 (2010).

49. A. Wilson and R. Adams, “Gaussian process kernels for pattern discovery and extrapolation,” in Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 28, edited by S. Dasgupta and D. McAllester (PMLR, Atlanta, Georgia, USA, 2013) pp. 1067–1075.

50. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” (2017), arXiv:1412.6980 [cs.LG].

51. S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, “Towards exact molecular dynamics simulations with machine-learned force fields,” Nature Communications 9, 3887 (2018).