Gaussian Process-Based Inference and Minimization of Energy and Forces in Molecular Dynamics

Jad Sbaï, Pedro Sousa, Grant Wilkins
Department of Computer Science
University of Cambridge
Cambridge, UK CB3 0FD
{js2837, pmms2, gfw27}@cam.ac.uk

Abstract

This report explores the application of Gaussian processes (GPs) for modelling atomic interactions and predicting energies and forces within molecular systems. Using molecular dynamics simulations of a water system, different GP regression approaches were tested using descriptors such as atomic positions, distances, and SOAP vectors. The results demonstrate that GPs can effectively interpolate potential energies over simulation trajectories, with the SOAP descriptor outperforming raw distances. However, challenges emerged in extending GP predictions to forces with higher noise levels. The report also examines the potential of Bayesian optimization integrated with the GP energy model to efficiently search for optimal low-energy molecular conformations. While the optimizers showed promise in finding energy minima, limitations existed due to the high-dimensional search space. Overall, the project reveals promising accuracy from GPs for emulating simulation data and complexities in applying them for derivative properties like forces. It highlights active research directions in improving model flexibility, incorporating electronic effects, and combining data-driven GP approaches with physics-based constraints. All code can be found at `https://github.com/JadSbai/ASE-GP`.

1 Introduction

The intersection of machine learning (ML) and physical sciences has emerged as an exciting space for further understanding and modelling complex systems. While powerful, traditional computational methods in physics often grapple with limitations regarding computational efficiency and scalability (1). With its ability to learn complex patterns from data, machine learning offers a promising alternative. Specifically, Gaussian Processes (GP), known for their flexibility and robustness in regression tasks, provide a probabilistic framework ideal for modelling uncertainties inherent in physical systems (2; 3). This project explores the application of GP in understanding energy and force relationships within molecular systems, a cornerstone in computational chemistry and materials science. Further, it aims to employ Bayesian Optimization techniques (BO) for tuning the kernel's hyperparameters and minimizing the energy of the studied molecular structure.

Molecular dynamics simulations, traditionally reliant on classical mechanics and quantum mechanical methods, are computationally expensive, especially for large systems. With their inherent ability to model non-linear relationships and quantify uncertainties, Gaussian Processes offer a promising alternative. By treating the problem as a regression task, GPs allow us to interpolate and predict physical properties like forces and energy based on atomic positions and bond angles. We focus on harnessing GP regression to create models that can learn from simulation-generated data. This data encompasses the nuanced interactions between atoms, considering factors such as interatomic distances, bond angles, and the resulting forces and energy states. Implementing GP in this context is not just a computational exercise but a step towards more efficient predictions in molecular dynamics.

It offers a pathway to reduce the computational overhead of traditional simulation methods while maintaining, or even enhancing, the accuracy of the predictions.

In this project, we demonstrate the application of GPs in modelling atomic interactions within a water (H2O) system. We train separate models to predict potential energy and 3D forces based on relative nuclear positions. This encompasses constructing and optimizing GP kernels sensitive to rotations, distances, and angles between bonds. The resulting fitted GP models can rapidly interpolate energies and forces at a limited computational cost. Success here would validate GPs as an efficient surrogate for traditional simulations in molecular modelling, potentially accelerating property predictions across computational chemistry and materials science. Additionally, we leverage the optimized GP kernels for energy minimisation using BO, an important task for understanding the stability and reactivity of molecular structures.

In the subsequent sections, we delve into the details of GPs, our data generation and processing methodology, and how these models are applied to our molecular system and then reused for energy minimization. In Section 2, we present a background and overview of molecular dynamics and existing numerical solutions. Then, in Section 3, we describe the different approaches we use for predicting force and energy for molecular systems and minimizing the energy of simulated structures. Section 4 shows our collected results and how well our predictions align with our given molecular dynamics system.

2 Background

2.1 Molecular Dynamics Simulations

Molecular dynamics (MD) simulation is a proper computational method in modern chemical physics, materials science, and molecular biology (2; 3; 4). It involves numerically integrating Newton's equations of motion for a system of atoms and molecules to reveal thermodynamic, structural, and transport properties.

While classical MD modelling has a long history dating back to the 1950s (5), continued advances in high-performance computing have enabled MD to scale to larger, more complex condensed phase systems. State-of-the-art special-purpose supercomputers can now simulate upwards of 100 million particles (6). Despite this progress, MD still faces challenges regarding computational scaling that limit accessible time and length scales (7). First-principles-based techniques that explicitly model interatomic interactions through quantum mechanical electronic structure theory formulations exhibit steep scaling. The gold standard methods like density functional tight binding scale in $O(N^2)$ to $O(N^3)$ complexity for N particles (8). More approximate yet widely adopted classical force fields reduce this to $O(N)$ or $O(N \log N)$ (9); however, this comes at the cost of reduced accuracy. While classical MD has been transformative, it faces limitations, particularly in accurately modelling complex chemical reactions and long-timescale phenomena. The approximations in classical force fields can lead to inaccuracies in predicting specific physical properties, especially when exploring uncharted chemical spaces or extreme conditions (6).

In addressing these challenges, *ab initio*, or first-principles, MD (AIMD) simulations have emerged as a powerful approach. Unlike classical MD, AIMD does not rely on predetermined force fields but calculates atomic interactions based on quantum mechanics, typically using methods like density functional theory (DFT) (10). However, the computational cost of AIMD is substantially higher than classical MD due to the complexity of solving electronic structure problems at each timestep. The scaling of computational cost with system size in AIMD is typically between $O(N^3)$ and $O(N^4)$, which limits its applicability to relatively small systems or shorter timescales (1). Data-driven methods like Gaussian process regression promise to deliver predictive accuracy without the exponentially growing expenses of electronic structure computations (2; 3). The approach pursued in this work aims to contribute to progress in this emerging research direction.

2.2 GAP and ML Toward Molecular Dynamics

Machine learning potentials (MLPs) present a promising alternative, leveraging the power of data-driven approaches to capture complex atomic interactions. GAP is a machine learning approach that uses Gaussian processes to interpolate between points in the configuration space of atoms, informed by quantum mechanical calculations. It aims to balance the accuracy of *ab initio* methods and the

efficiency of classical force fields. GAP has been successfully applied in various contexts, from predicting the properties of materials to simulating complex chemical reactions (11; 12).

Gaussian process regression (GPR) stands out for its simplicity and the probabilistic framework it provides, which quantifies prediction uncertainties. It has demonstrated exceptional accuracy in MD simulations, successfully capturing intricate atomic interactions with a precision often exceeding that of empirical force fields. Studies have highlighted the potential of GPR models to revolutionize MD simulations by providing high accuracy without the computational burden of electronic structure calculations (2; 3; 13).

2.3 Bayesian Optimization Toward Molecular Dynamics

Bayesian optimization (BO) is increasingly used in molecular dynamics (MD) simulations and related fields like computational chemistry and materials science for optimizing costly black-box functions (14). Typical applications include tuning force field parameters, designing MD simulation workflows, accelerating structure predictions, and optimizing molecular geometries. In these settings, key objectives like calculating lattice energies for different molecular orbitals have no simple closed-form expression and instead must be numerically evaluated through expensive MD simulations (10). Further, gradients often need to be made available or meaningful. BO provides an efficient methodology for navigating these high-dimensional search spaces with limited function calls.

By constructing a posterior surrogate model to emulate the true objective, BO can guide sampling toward promising candidates to optimize properties globally. Gaussian process regression is commonly used given kernel flexibility (2). Recent works have tuned interatomic potentials with BO to discover novel crystal structure materials (15).

In this work, the applications of BO are twofold. Initially, we use BO to fine-tune the kernel hyperparameters in our GP regression models for force and energy (16). This tuning helps capture atomic interactions inherent in the training data. Secondly, we reuse the optimised kernel to experiment with BO to identify the molecular structure with the lowest total energy. Energy minimization is fundamental in studying molecular systems, as it helps identify the most stable configurations of molecules (17). The lower the energy state of a molecular structure, the more stable it is likely to be. This is particularly important in fields like drug design, materials science (18), and protein structure modelling (19), where identifying stable molecular configurations can lead to breakthroughs in developing new materials or pharmaceuticals. By employing Bayesian Optimization for this task, we can efficiently navigate the complex energy landscape of molecular systems to find these minimum energy states.

As simulations remain the primary workflow in computational chemistry, integrating BO will be increasingly crucial for efficiently directing simulations toward meaningful molecular insights and discoveries.

3 Methods

3.1 Atomic Simulations Environment (ASE)

Our project utilizes Molecular Dynamics (MD) simulations using the Atomic Simulation Environment (ASE) Python package. The key to achieving accurate MD simulation results is carefully setting parameters. We initiate simulations at an initial temperature of 293 Kelvin (20 degrees Celsius) and employ the Effective Medium Theory (EMT) calculator within ASE for computing forces and energies. EMT is chosen for its balance between computational efficiency and accuracy. To emulate a larger environment, we adopt a supercell approach, replicating a pattern of molecules, like H2O, across an area to simulate a more extensive space with a density of 1.0 (water).

Our simulations use the Langevin dynamics algorithm to realistically model particle behaviour amid thermal fluctuations, blending deterministic and random thermal forces for natural temperature variation. We set a low friction factor of 0.0002 to simulate a less viscous environment and apply periodic boundary conditions for a continuous space effect. Our primary focus is on potential energy, which is crucial for understanding molecular behaviour. The raw data comprises 3D atomic positions, but to derive more meaningful insights, we use descriptors like distances between atom

pairs and the SOAP (Smooth Overlap of Atomic Positions) descriptor (20), which provides a detailed representation of each atom's local environment.

Data collection during simulations uses a listener that records the molecular system's state every two femtoseconds (fs), yielding 50 data points over a 100 fs simulation. This comprehensive approach in data handling and simulation settings facilitates a deeper understanding of the molecular dynamics under study.

3.2 Designing Gaussian Processes for Force and Energy

In this project, our goal was not merely to replicate GAP (Gaussian Approximation Potentials), coded in FORTRAN, but to delve deeper into the mechanics of energy and force prediction in Python. While GAP provided a foundation and inspiration, we sought the flexibility to experiment with various aspects of our GP model. This included trying different kernels, data processing pipelines, descriptors, and overall design approaches.

Considerable effort went into processing the data from the simulations. It was retrieved, cleaned, and processed after saving the raw data in a specialized `.xyz` file format. As the subsequent sections show, the formatting varied depending on the chosen descriptors. A significant challenge was maintaining the integrity of the original data structure while considering dimensionality reduction to minimize information loss.

Notably, predicting forces in three dimensions (x, y, z) presented a unique challenge, requiring multi-variate outputs. Our first approach was coregionalization, which involves constructing a covariance matrix for correlations between the different output dimensions. Despite its potential, we faced significant numerical stability issues. The covariance matrix was not strictly positive definite, only semi-positive, mainly due to the low force values, often around $10e - 6$. This issue persisted even after normalization. The semi-positive definiteness of the covariance matrix may be due to the limited variance in these low-force values. This variation made it difficult for coregionalization to effectively model these forces.

3.2.1 Kernel Selection

In our project, kernel selection was tailored to each task (energy or force prediction) and descriptor, with each requiring a unique approach to capture the correlation between features and predictions. For the raw descriptor of 3D positions of 81 atoms, we observed periodic behaviour in output energies, leading us to choose a periodic exponential kernel for its balance in capturing smooth, periodic functions and adjustable variance, along with a White kernel to address noise.

We followed the GAP library's precedent for pairwise distances, opting for a Matern kernel for its control over smoothness. The SOAP descriptors, known for their somewhat linear relationship with potential energy, led us to combine a Dot Product kernel with an RBF kernel, effectively capturing both linear and non-linear dynamics.

Predicting forces in our project required fitting separate Gaussian Processes (GPs) for each force component (x, y, z). Initially, we tried modelling all-atom positions for each component using various kernels, but the GPs struggled to correlate atom positions with total force accurately. To address this, we shifted our focus to a single atom, training a GP on its position and force data over time for better generalization across the system. We chose the Matern 3/2 kernel for its smoothness, which effectively captures the relationship between position and force for an individual atom, simplifying the model and allowing for more precise tuning of kernel parameters.

Predicting forces in our project required fitting separate Gaussian Processes (GPs) for each force component (x, y, z). Initially, we tried modelling all-atom positions for each component using various kernels, but the GPs struggled to correlate atom positions with total force accurately. To address this, we shifted our focus to a single atom, training a GP on its position and force data over time for better generalization across the system. We chose the Matern 3/2 kernel for its smoothness, which effectively captures the relationship between position and force for an individual atom, simplifying the model and allowing for more precise tuning of kernel parameters.

3.3 Energy Minimization

Once the optimal kernel function was found for the energy model, we reused it to define a GP emulator to estimate the energy of various molecular structures, making it a valuable tool for experimentation. BO combined the emulator with the Expected Improvement (EI) acquisition function to search for the molecular structure that yielded the lowest energy. It quantifies the expected increase in the objective function based on the current best observation and the uncertainty in model predictions. This makes it adept at guiding the optimization process towards areas likely to yield improvements while encouraging exploration to avoid local optima. `Emukit`'s optimizers were tested to maximize the acquisition function.

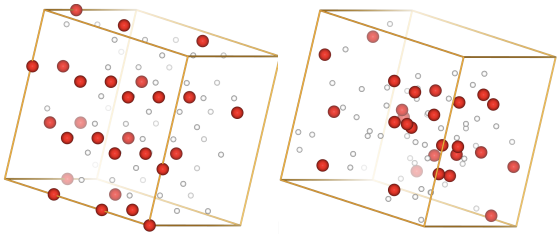
The `GradientAcquisitionOptimizer` utilises a quasi-Newton method, specifically the L-BFGS algorithm (21), known for efficiently handling continuous acquisition functions. It approximates the Hessian matrix, enabling efficient navigation through complex optimization landscapes without requiring the computational cost of the full Hessian. Based on the principle of Variable Neighbourhood Search as described in the context of SMAC (Sequential Model-based Algorithm Configuration) (22), the `LocalSearchAcquisitionOptimizer` begins with multiple local searches from randomly selected points and iteratively evaluates one-exchange neighbourhoods, allowing for a thorough exploration of the parameter space. Finally, the `RandomSearchAcquisitionOptimizer` adopts a more straightforward approach and evaluates the acquisition function at randomly selected points. This valuable method ensures diverse search space sampling, avoiding local optima traps that more deterministic methods might encounter.

3.4 Sensitivity Analysis

This study attempted a detailed sensitivity analysis to assess the influence of each atom's individual coordinates on the predicted energy of molecular structures, hoping to understand our complex simulator's dependencies better. Leveraging `Emukit`'s `ModelFreeMonteCarloSensitivity` and `MonteCarloSensitivity` modules, we conducted Sobol sensitivity analysis (23) by generating random samples from the input space and evaluating them with the surrogate model. The Sobol method then analyses the variance in the model's output to determine the contribution of each input parameter (and combinations thereof) to the total variance. Our experiments yielded main and total effects for each parameter, hoping to elevate our understanding of the individual and interactive impacts of the parameters on the predicted energy.

4 Experiments

The goal of the ASE program and training data generation is to find an optimal energy configuration for a system of atoms. This exploratory process uses an EMT calculator moving from a system like in Figure 1(a) to Figure 1(b). Using this training data, we trained GPs toward energy and force prediction. This section will describe the results to test our GP models and further exploration for BO and sensitivity analysis.



(a) Initial Configuration (b) Optimal Configuration

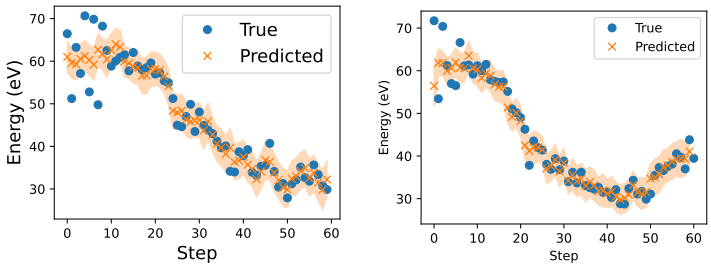
Figure 1: Transformation of H_2O System from ASE + EMT Calculator

4.1 GP Fitting of Energy and Force

The total energy modelling results in Figures 2(b) and 2(a) demonstrate that Gaussian processes can accurately interpolate energies over the simulation trajectory. The energy modelling benchmarks

reveal that Gaussian process (GP) regression is promising for emulating quantum simulation data. Both the SOAP and distance-based descriptors effectively encode molecular structures to enable low-error interpolation. This aligns with literature highlighting that energies are often smooth and easier to model with data-driven approaches (2).

However, upon closer analysis, we observe a moderate accuracy advantage from using the SOAP features over raw distances. SOAP achieves under 13.5eV^2 mean squared error versus 65.1eV^2 for distances. We hypothesize that this is because SOAP encodes the full 3D distribution of neighbouring atoms, making it more sensitive than nuclear coordinates. In contrast, scalar distances discard angular nuances. By providing a better summary statistic of atomic interactions, SOAP offers superior tracking for how a system evolves over simulation time. These initial findings reveal that explicitly encoding richer atomic environments in descriptors can enhance model flexibility. However, distances do still perform reasonably well, explaining the majority of energetic variation. This success highlights GP regression's composition in easily incorporating different input features. Testing alternative representations reveals potentially cheaper workflows for balancing simplicity and description.



(a) SOAP Predictions per Time Step (b) Distances Predictions per Time Step

Figure 2: Predictors for Total Energy of System

Descriptor	Quantity	MSE	RMSE
SOAP / Distances	Energy	13.477 / 65.067	3.671 / 8.066
	Force x	2.35 / 122.759	1.533 / 11.080
	Force y	329.84 / 297.211	18.161 / 17.234
	Force z	1863.80 / 338.096	43.172 / 18.388

Table 1: Error Metrics for GPs for Different Quantities and Predictors

The GP regressor demonstrates effectiveness in emulating quantum simulation energies but shows limitations in force predictions. Our methodology utilises SOAP and nuclear coordinates to create GPs for each force dimension. When training the GP, we focus on force data from a single atom. For distance-based predictions, we validate the model using just one atom. This approach, however, revealed a key limitation: distance predictions across different atoms tended consistently towards zero, indicating very poor performance in generalizing beyond the single atom used for validation. Consequently, the mean squared error (MSE) values presented in 1 must be interpreted with this context in mind—the distances-based MSE reflects predictions for a single atom, while the SOAP-based MSE encompasses predictions across all atoms over multiple timesteps.

Figure 3 shows these challenges more distinctly. While the model captures general local trends, it struggles with higher frequency noise, a common issue since forces are energetic derivatives and are subject to complex physical constraints. In reality, all atoms contribute to the force field, indicating that factors beyond just position data are at play. For this reason, the SOAP descriptor objectively outperforms the nuclear coordinates approach due to its ability to encapsulate neighbourhood density information, offering a more comprehensive view of atomic environments. This feature of SOAP allows for better capturing of both local and global interactions within the molecular system. While nuclear coordinates perform decently well on local, atom-specific trends, SOAP's broader perspective enables the model to generalize effectively across different atomic configurations, as shown in Figure 3(d).

The task of accurately representing molecular force fields, considering their dimensionality and complexity, remains a significant challenge in the field.

4.1.1 Hyperparameter Optimization of Energy & Force

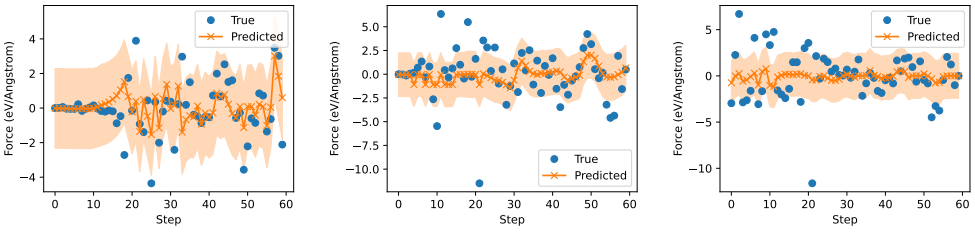
Hyperparameter optimization in the context of GPR helps tailor a model to a specific dataset. In our study, we employed Bayesian optimization to optimize the hyperparameters of the kernel functions for energy and force. Below in Table 2, we outline the different kernels and hyperparameters we optimized for our GPs from different descriptors. Note that we include a WhiteKernel for each of our quantities. We include these values for the sake of reproducibility.

Descriptor	Quantity	Kernel + WhiteKernel(ν)	Optimal Hyperparameters
SOAP	Energy	DotProduct(σ_0) + RBF(l)	$\sigma_0 = 61, l = 0.12, \nu = 0.2$
	Force x	DotProduct(σ_0)	$\sigma_0 = 42, \nu = 0.8$
	Force y	Matern32(σ^2, l)	$\sigma^2 = 12, l = 0.4, \nu = 0.1$
	Force z	Matern32(σ^2, l)	$\sigma^2 = 8, l = 0.6, \nu = 0.1$
Distances/Positions	Energy	Periodic(σ^2)	$\sigma^2 = 9.4, \nu = 0.3$
	Force x	Constant(c) $\times$ RBF(l)	$c = 939, l = 60, \nu = 0.005$
	Force y	Constant(c) $\times$ RBF(l)	$c = 1000, l = 42.9, \nu = 0.4$
	Force z	Constant(c) $\times$ RBF(l)	$c = 967, l = 0.5, \nu = 0.1$

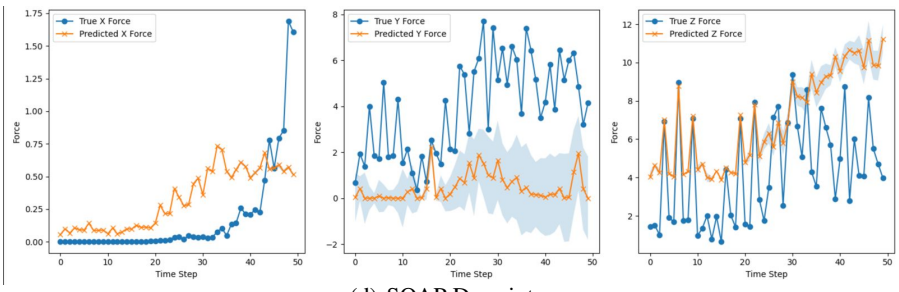
Table 2: Optimal Hyperparameters Used in Evaluation of Metrics

For SOAP energy, the high $\sigma_0 = 61$ in the DotProduct kernel suggests a strong focus on linear inhomogeneities, while the RBF kernel's small $l = 0.12$ indicates a sensitivity to non-linear details. For x force, the $\sigma_0 = 42$ in the DotProduct kernel points to a strong linear dominance. In contrast, the Matern32 kernel for y and z forces, with $\sigma^2 = 12$ and 8, and $l = 0.4$ and 0.6, caters to the different behaviours of these forces—higher variability in y and smoother changes in z. In the Distances/Positions descriptor, the Periodic kernel's $\sigma^2 = 9.4$ for energy emphasizes modelling significant cyclical variations.

As shown in the table, we optimized each spatial dimension (x, y, z) of the force data. This approach is based on the assumption that forces in different dimensions may exhibit distinct characteristics. The objective function we used evaluates the MSE between the predicted and actual values. This dimension-specific approach allows for a more tailored and potentially accurate modelling of the force data in each spatial dimension. However, we note that modelling force fields is still an open problem; therefore, our solution is rudimentary compared to the state-of-the-art (13).



(a) Force x (b) Force y (c) Force z



(d) SOAP Descriptor

Figure 3: GP for Force in Each Direction for Single Atom

4.2 BO for Energy Minimization

The suitability of BO to search for optimal molecular structures was tested using a fresh dataset generated from rerunning the simulation. This allowed us to validate further the previously found optimal kernel's suitability for unseen data. We did not use any descriptors in this BO process since our simulation already involved a high-dimensional input space - with a total of 243 parameters where each sample had 81 atoms, and each atom had 3 spatial coordinates.

We initialised the BO loop by fitting the GP regressor on 5 random points sampled from the simulator dataset and their corresponding energies. Emukit's `LatinDesign` sampler was originally used, but limited computational resources and high dimensionality made it unfeasible. As an alternative, we focused the BO on energy minimization of the fixed dataset of molecular structures generated by the simulator. However, in order to leverage Emukit's `BayesianOptimizationLoop`, we had to define a `ContinuousParameterSpace` based on the maximum and minimum values for each parameter.

In each BO iteration, the next candidate point was proposed by maximizing the acquisition function and then finding its closest point in the fixed dataset to approximate sampling the continuous input space. This was achieved by computing the Euclidean distance (2-norm) as the distance metric between the point returned by the acquisition function and the fixed simulation dataset, ensuring that BO finds a minimum within the dataset. The algorithm continued for 20 iterations, tracking the best-observed energy value, simple regret relative to the true optimal energy, and all observed minimum energies at each step. Figure 4 showcases the resulting energies and regrets for each acquisition optimizer. For benchmarking purposes, we also implemented a manual BO without an optimizer. All loops used the same acquisition function and GP kernel. The different starting energies in the plots reflect the stochastic nature of the starting points' selection process. However, the key to evaluating optimizer performance is to look at the trend of energy reduction over iterations.

The `GradientAcquisitionOptimizer` shows a consistent decrease in energy levels, indicating that it quickly found areas in the parameter space that led to lower energy configurations. It shows a rapid convergence towards the minimum energy, suggesting a strong exploitation capability when the gradient of the acquisition function can be computed. Similarly, `LocalSearchAcquisitionOptimizer` registered a strong decrease for the first six iterations but then plateaued until finally reaching the minimum later in the loop. The flat region is likely related to the over-exploitation of the local neighbourhood around the acquired points. `RandomSearchAcquisitionOptimizer` exhibits a similar trend to `GradientAcquisitionOptimizer`, reaching the minimum in fewer iterations through a steady decay in observed energies. Finally, the manual implementation shows a more erratic energy profile with significant fluctuations, indicating a less systematic search through the parameter space and an emphasis on exploration in the first iterations.

All optimizer-enabled BO loops were able to find the molecular structure that yielded the true lowest energy, 43.76 eV. As expected, their respective regret plots mimic the energy trajectories, eventually equalling zero when the true minimum is found. Unfortunately, the manual implementation converged to 52.65 eV and struggled to find the true minimum, hence not reaching zero regret. This suggests the over-exploitation of a local minimum. The proximity function used to find the nearest point in the dataset to the optimizer's proposed candidate likely affected its energy path, selecting points in areas already explored repeatedly. Such an approach can limit the optimizer's ability to adequately explore the search space, particularly when combined with the random initialization of points. If the initial points are clustered in less optimal regions of the space, the optimizer may spend more time exploiting these areas rather than effectively exploring new regions that could lead to the minimum.

4.3 Sensitivity Analyses of the Energy Model

Once our energy GP regressor was trained and its optimal kernel parameters found, we attempted to perform a detailed sensitivity analysis using the Monte Carlo-based Sobol method with `num_monte_carlo_points` set to one million points. The first parameter showed a main effect of 1.0000088 and a total effect of 1.00026803. In contrast, all the remaining 242 parameters each returned a main effect of -0.00026803 and a total effect of 0.00064406. These results suggest a higher impact on the predicted energy by the x coordinate of the first atom. This could imply that the first parameter captures a critical aspect of the molecular structure that is highly influential in determining its energy. However, in this work, we did not reach the desired level of interpretability for our results. The uniformity in the effects of the other parameters could also hint at redundancy

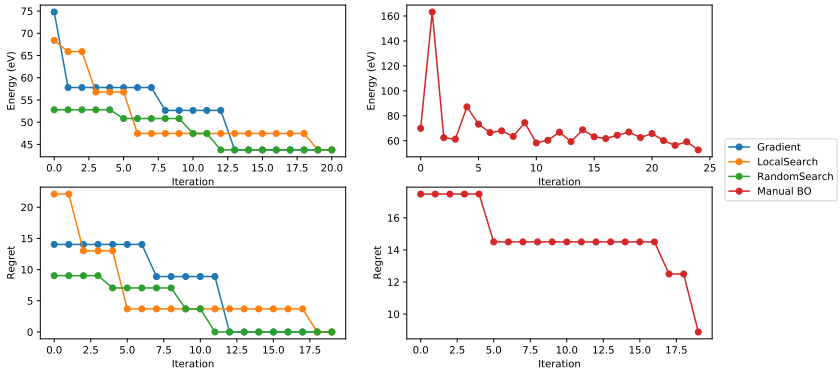


Figure 4: Observed Energy and Regrets for Bayesian Optimization with different optimizers.

or high correlation among these parameters. In high-dimensional spaces, it's common for some parameters to exhibit collinear behaviours, which can dilute the apparent impact of each individual parameter. Further experimentation is required for more insightful conclusions.

5 Discussion and Future Directions

5.1 Limitations

The representations used only leverage nuclear coordinates. This limits the descriptive capacity to model more complex quantum mechanical effects like charge transfer, polarization, and spin dynamics that require incorporating electronic structure. One of the problems with this is the dimensionality required to unpack certain data. For example, bond angles were interesting to us as a way to train a force model. However, the number of bond angles in an N atom system was on the order of N^3, making this intractable for our 81-atom system. Extending to descriptors from higher-fidelity computations or multiple descriptor types could improve our GP but also potentially require a different model architecture.

The training and evaluation of our GP models were constrained to relatively small and simple molecular systems compared to the state-of-the-art (6). This limitation raises concerns about the models' ability to generalize across different types of molecules, particularly those with unique bonding patterns or atomic species not represented in the training set. Furthermore, even though successful for a fixed simulation dataset, BO could be further improved by employing specific techniques to address the high-dimensionality of the parameter space. Specifically, (24) introduced batch BO via structural kernel learning and (25) expanded on learning a lower-dimensional latent space with Variational Autoencoders (26).

Finally, our GP models cannot extrapolate outside of the scope of the simulation without significant uncertainty. This reliance on extrapolation compromises the reliability of predictions in unexplored areas of the chemical space. Combining the GP framework with physics-based priors or constraints can solve this challenge. For example, integrating previously mentioned AIMD knowledge as priors, with the consultation of domain experts, within the GP framework could guide the extrapolation process, imposing physically realistic boundaries on the predictions (10; 27).

5.2 Conclusion

Applying GPs in molecular dynamics simulations shows considerable promise, especially in developing surrogate models to reduce computationally expensive simulations and assist with energy minimization. However, addressing the challenges and limitations above is crucial. Future work and extensions of our project should focus on systematic benchmarking against diverse molecular datasets, employing data augmentation strategies, and exploring hybrid models that combine GPs with physics-based methods. These efforts will be pivotal in fully realizing the potential of GPs in molecular dynamics and assessing their applicability in broader contexts.

References

[1] J. Pan, "Scaling up system size in materials simulation," *Nature Computational Science*, vol. 1, no. 2, pp. 95–95, 2021.

[2] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," *Chemical Reviews*, vol. 121, pp. 10073–10141, 08 2021.

[3] M. Krynski and M. Rossi, "Efficient gaussian process regression for prediction of molecular crystals harmonic free energies," *npj Computational Materials*, vol. 7, no. 1, p. 169, 2021.

[4] M. Ceriotti, "Unsupervised machine learning in atomistic simulations, between predictions and understanding," *The Journal of Chemical Physics*, vol. 150, p. 150901, 04 2019.

[5] B. J. Alder and T. E. Wainwright, "Studies in Molecular Dynamics. I. General Method," *The Journal of Chemical Physics*, vol. 31, pp. 459–466, 08 2004.

[6] D. E. Shaw, P. J. Adams, A. Azaria, and et al., "Anton 3: Twenty microseconds of molecular dynamics simulation before lunch," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, (New York, NY, USA), Association for Computing Machinery, 2021.

[7] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018.

[8] S. Ahnert, G. Csányi, and R. Kondor, "Gaussian processes in molecular dynamics electronic structure." Discussion Group Meeting, July 2005.

[9] J. W. Ponder and D. A. Case, "Force fields for protein simulations.," *Advances in protein chemistry*, vol. 66, pp. 27–85, 2003.

[10] W. Kohn, A. D. Becke, and R. G. Parr, "Density functional theory of electronic structure," *The Journal of Physical Chemistry*, vol. 100, pp. 12974–12980, 01 1996.

[11] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Phys. Rev. Lett.*, vol. 104, p. 136403, Apr 2010.

[12] S. Klawohn, J. P. Darby, J. R. Kermode, G. Csányi, M. A. Caro, and A. P. Bartók, "Gaussian approximation potentials: Theory, software implementation and application examples," *The Journal of Chemical Physics*, vol. 159, p. 174108, 11 2023.

[13] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," *Chemical Reviews*, vol. 121, no. 16, pp. 10142–10186, 2021. PMID: 33705118.

[14] S. Diwale, M. K. Eisner, C. Carpenter, W. Sun, G. C. Rutledge, and R. D. Braatz, "Bayesian optimization for material discovery processes with noise," *Mol. Syst. Des. Eng.*, vol. 7, pp. 622–636, 2022.

[15] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, "Crystal structure prediction accelerated by bayesian optimization," *Physical Review Materials*, vol. 2, no. 1, p. 013803, 2018.

[16] J. Köfinger and G. Hummer, "Empirical optimization of molecular simulation force fields by bayesian inference," *The European Physical Journal B*, vol. 94, no. 12, p. 245, 2021.

[17] K. Roy, S. Kar, and R. N. Das, "Chapter 5 - computational chemistry," in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* (K. Roy, S. Kar, and R. N. Das, eds.), pp. 151–189, Boston: Academic Press, 2015.

[18] R. Catlow, *Energy Minimization Techniques in Materials Modeling*, pp. 547–564. 01 2005.

[19] A. Jabeen, A. Mohamedali, and S. Ranganathan, "Protocol for protein structure modelling," in *Encyclopedia of Bioinformatics and Computational Biology* (S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, eds.), pp. 252–272, Oxford: Academic Press, 2019.

[20] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Physical Review B*, vol. 87, no. 18, p. 184115, 2013.

[21] J. E. Dennis and J. J. Moree, "Quasi-newton methods, motivation and theory," *SIAM Review*, vol. 19, no. 1, pp. 46–89, 1977.

[22] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Learning and Intelligent Optimization* (C. A. C. Coello, ed.), (Berlin, Heidelberg), pp. 507–523, Springer Berlin Heidelberg, 2011.

[23] J. Nossent, P. Elsen, and W. Bauwens, "Sobol' sensitivity analysis of a complex environmental model," *Environmental Modelling & Software*, vol. 26, no. 12, pp. 1515–1525, 2011.

[24] Z. Wang, C. Li, S. Jegelka, and P. Kohli, "Batched high-dimensional bayesian optimization via structural kernel learning," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, p. 3656–3664, JMLR.org, 2017.

[25] A. Grosnit, R. Tutunov, A. M. Maraval, R.-R. Griffiths, A. I. Cowen-Rivers, L. Yang, L. Zhu, W. Lyu, Z. Chen, J. Wang, J. Peters, and H. Bou-Ammar, "High-dimensional bayesian optimisation with variational autoencoders and deep metric learning," 2021.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013.

[27] E. Aprà, E. J. Bylaska, W. A. de Jong, and et al., "Nwchem: Past, present, and future," *The Journal of Chemical Physics*, vol. 152, no. 18, p. 184102, 2020.