# Utilizing Machine Learning for Efficient Parameterization of Coarse Grained Molecular Force Fields

James L. McDonagh,[†] Ardita Shkurti,[‡] David J. Bray,[*,‡] Richard L. Anderson,[‡] and Edward O. Pyzer-Knapp[*,†]

†*IBM Research UK, Hartree Centre, SciTech Daresbury, Warrington, Cheshire WA4 4AD, U.K.*

‡*STFC, Hartree Centre, SciTech Daresbury, Warrington, Cheshire WA4 4AD, U.K.*

E-mail: david.bray@stfc.ac.uk; epyzerk3@uk.ibm.com

**Abstract**

We present a machine learning approach to automated force field development in Dissipative Particle Dynamics (DPD). The approach employs Bayesian optimization to parameterize a DPD force field against experimentally determined partition coefficients. The optimization process covers a discrete space of over $40,000,000$ points, where each point represents the set of potentials that jointly form a force field. We find that Bayesian optimization is capable of reaching a force field of comparable performance to the the current state-of-the-art within 40 iterations. The best iteration during the optimization achieves an $R^2$ of 0.78 and an RMSE of 0.63 log units on the training set of data, these metrics are maintained when a validation set is included, giving $R^2$ of 0.8 and an RMSE of 0.65 log units. This work hence provides a proof-of-concept, expounding the utility of coupling automated and efficient global optimization with

a top down data driven approach to force field parameterization. Compared to commonly employed alternative methods, Bayesian optimization offers global parameter searching and a low time to solution.

# Introduction

Molecular simulation has become a valuable technique for gaining insight into the behaviour of chemical and materials systems. Many different techniques exist ranging from approximate and simulation coupled quantum mechanics,[1,2] classical atomistic approaches such as Molecular Dynamics (MD)[3,4] and mesoscopic methods such as Coarse Grained Molecular Dynamics (CGMD)[5,6] and Dissipative Particle Dynamics (DPD).[7–9]

For simulations to provide useful insights, the molecular interactions need to be suitably modeled. This is achieved through inter-particle potentials, which are idealized mathematical descriptions of the interactions between particles. Typically, these potentials operate between particular atoms or groups and represent a particular interaction (e.g. bond stretching, bond bending and twisting, Van der Waals interaction, electrostatic attraction/repulsion etc), such that the collection of all interactions present in a simulation is governed by a collection of potentials. The collection of such potentials is referred to as a force field.

Force fields require extensive parameterization to ensure an optimal description of the molecular interactions. This is generally achieved by testing how accurately a set of trial force field parameters reproduces relevant physical observations, often from experiment or higher level theory, such as quantum chemical data. This can require a large number of trials, while one navigates to an appropriate set of parameters. This searching of the parameter space is a laborious and typically expensive task. Many optimization attempts employ manual fitting procedures,[10,11] which are extremely expensive in terms of research time, as well as requiring considerable existing insight on the part of the researcher. Others employ local optimization techniques, which are typically restricted to search only in the vicinity of the starting position.[12] Less commonly used are global optimization strategies, which are the

focus of this work. These methods enable searching of comparatively vast parameter spaces, which are not bound to a local minima. However, global optimization procedures are often too computationally expensive to utilize in force field development.

The use of the parameterization methods introduced above can be seen in the development of a plethora of force fields. MD is one of the clearest examples of the complexity and difficulty involved in accurate force field generation. Many of the leading MD force fields, such as CHARMM,[3] AMBER,[4] OPLS,[13,14] are the result of years of incremental improvements and extensions through a variety of fitting methods. In fact these force field names more correctly refer to a set of force fields with specific parameterizations for particular applications. Within the MD community, there is an increasing interest in developing automated methodologies to speed up the process of parameterization.[15–20] For example, new parameters in CHARMM can be developed using semi-automated tools.[15,16] In addition to automation, a number of *smarter* methods are being incorporated into model parameterization to help accelerate the process and to find better solutions. Examples of these methods include the application of Gaussian processes to learn or parameterize inter-atomic potentials and investigate their generalizability,[21–24] along with a variety of other machine learning approaches such as neural networks,[24,25,25–29] which have shown promising results in recent years.

Within the MD community accurate parameters and reliable models have been generated for a variety of chemical classes; such as small drug molecules, proteins and hydrocarbons. However, there are many industrially relevant fields not well served by the current MD force fields. The reasons for this are two fold:

- the chemical constituents of products often lie outside the well parameterized chemical classes;
- the scale (time and number of particles) for industrially relevant simulations often exceeds what is currently tractable in MD on commonly employed computer clusters.

This can result in a significant barrier to wide-scale adoption of modeling and simulation within some sectors of the chemical industry.

In this work, we introduce the application of an efficient global optimization methodology, Bayesian optimization, to automated top down parameterization of DPD force fields. This strategy aids in alleviating the barriers detailed previously, as the force field can be optimized automatically, on relevant data, in a reasonable time period. Additionally, the DPD method, allows one to access time and length scales inaccessible via MD whilst maintaining coarse molecular features.

Several groups have developed methods based around machine learning and Bayesian models for bottom up fitting of DPD force fields utilizing the outputs of MD simulations.[30–33] Liu *et al.*[32] provided some of the first applications in this area using force matching between MD simulation and coarse graining with a Bayesian inference refinement. More recent work by Dequidt and Solano Canchaya[30] also utilizes a force matching methodology, where by an atomistic trajectory is sampled and coarse grained coordinates determined in relation to the all atom models. Bayesian modeling was then applied to locate the most likely DPD force field parameters to reproduce the sampled coarse grained coordinates. More common approaches to CGMD and DPD force field parameterization are methods such as, Iterative Boltzmann Inversion (IBI) and stochastic parametric optimization.[34] IBI is typically used to determine parameters, which reproduce a reference Radial Distribution Function (RDF). The method involves the calculation of initial CGMD parameters approximated as the potential of mean force between a pair of coarse grained particles at a given distance. This initial estimate is iteratively refined by a correction factor. The method can suffer from practical limitations such as the selection of an interaction cut-off distance.[34–37] The stochastic parametric approach involves the selection of an empirical function with a number of free parameters which can be fit to reproduce target properties.[34,38]

Whilst these bottom up fitting methods for DPD show significant promise, the accuracy of the underlying atomistic force field determines the accuracy of the DPD model. Here there

is a danger that the atomistic force field is not well suited to model the behavior or chemicals of interest. In this article, we tackle this issue by performing a top-down parameterization of a DPD force field directly to experimental data, hence one can select the data best suited to their problem. As an example, here we are using data on partition coefficients ($\log P$). In this work, $\log P$ has been adopted because of the abundance of curated data available in the literature, the wide community interest in this property[39–41] and it has been used previously in work by Anderson *et al.*[11,42] to manually fit DPD force fields.

The proposed Bayesian optimization method enables an efficient automated search, which learns a probabilistic approximation of the parameter landscape via a machine learning technique known as a Gaussian process. The search is carried out over a range of DPD parameters, which form a high dimensional grid. These ranges are defined prior to the optimization process. Such a prior definition allows for expert knowledge to be encoded into the optimization process. To the authors knowledge this is a novel approach.

We focus on a modest set of molecules made up of alkanes and primary alcohols. Our aim in this work is to provide a proof-of-concept, hence we focus on well understood chemistry which has open experimental literature data available for validation. Experimental $\log P$ data, obtained from the literature, is used as a reference in this work.

We calculate a predicted $\log P$ from DPD simulations, using the protocol of Anderson *et al.*.[11] Whilst DPD is not the most efficient method for calculating $\log P$, we have found that $\log P$ is a good parameter for top down fitting of DPD parameters. Therefore we apply existing and documented methods for calculating $\log P$ using DPD in this work.[11] The Bayesian optimization process then efficiently guides the search to regions of parameter space which minimizes the error between the experimental values and the simulation values. Noting that the optimization in the present case is over a discrete high dimensional grid the hypothesis here is that the Bayesian optimization will locate an optimal region of parameter space very efficiently. It may be possible to further refine this result considering a continuous parameter space.

The remainder of the article is arranged as follows: an outline of the general methods adopted in order to develop an automated parameterization process; the presentation of a proof-of-concept test case using $\log P$ as the target of the optimization process; a discussion and evaluation of the results; finally, we summarize the main findings of the paper and discuss potential extensions of the proposed method.

## Methods

The automated Bayesian parameterization procedure described in this work requires several elements to be connected in a workflow. The workflow consists of simulations, analysis and optimization. We have developed a workflow engine named CAROL,[43] which is used to orchestrate the simulation and analysis process. In this section we describe the automated procedures and inputs including: the Bayesian optimization scheme, the DPD simulation details and literature data.

## Bayesian Optimization Scheme

Bayesian optimization is machine learning approach, capable of efficiently balancing the exploration exploitation dilemma that is commonly encountered in optimization problems.[44–46] The optimization of DPD force field parameters is achieved through a python based Bayesian optimization library. The Bayesian optimization library constructs a probabilistic model of the objective function utilizing scikit-learn's[47] Gaussian process regression with a square-exponential kernel.

$$k(x, x') = \sigma^2 exp\left(\frac{(x - x')^2}{2l^2}\right) \tag{1}$$

Equation 1 defines the squared exponential kernel where $x$ and $x'$ denote a pair of inputs $l$ is the length scale parameter and $\sigma$ is the variance of function values from the mean. The hyper-parameters are optimized by maximizing the log-marginal-likelihood. The Bayesian

optimization library employs an expected improvement (EI) acquisition function to select the next point to sample. EI is here used with an epsilon parameter, that assists in balancing the exploration and exploitation of the parameter space by defining a minimum magnitude of improvement, which is expected of a new sample point. The improvement is calculated from the utility function ($\gamma(x)$), commonly referenced as the improvement function:[48,49]

$$\gamma(x) = \frac{\mu(x) - f^* - \epsilon}{\sigma^2(x)}, \tag{2}$$

In the improvement function, defined in equation 2, $\mu(x)$ is the mean, $\sigma^2(x)$ is the variance $\epsilon$ defines the minimum improvement margin and $f^*$ is the currently best located point assigned as the lowest loss. The EI acquisition function can then be defined in terms of $\gamma(x)$ thus:

$$EI(x) = \mu(x) - f^*\Phi(\gamma) + \sigma(x)\phi(\gamma) \tag{3}$$

Equation 3 defines the expected improvement acquisition function in terms of gamma. $\Phi(\gamma)$ is the cumulative distribution function and $\phi(\gamma)$ is the probability density function.[48] This is a commonly applied embodiment of Bayesian optimization that has been deployed in other areas.[50,51]

In this study, we optimize a DPD force field by minimizing the error in calculated $\log P$ data compared to experimental $\log P$ data for nine solute molecules over three different solvent pairs defining $\log P$. The combination of the nine solutes and three different solvent pairs results in 15 $\log P$ data points for training.

The standard DPD force field is purely repulsive and soft-core in nature. The primary parameter in these DPD force fields, which governs the chemical interactions, is the so called conservative repulsion parameter $A_{ij}$. This parameter governs the strength of the repulsion between beads of type $i$ and $j$. The unique pairwise interactions and the ranges of their conservative potential repulsion parameters are shown in Table 1.

Table 1: The pairwise bead-bead interactions optimized in this work: each interaction pair is bounded within the intervals defined by $A_{ij} \in [A_{ij}^{Lower}, A_{ij}^{Upper}]$. We note here that $2H_2O - 2H_2O$ interaction acts as reference and remains fixed at 25.0.

| Interaction pairing | $A_{ij}^{Lower}$ | $A_{ij}^{Upper}$ |
|---|---|---|
| $2H_2O - 2H_2O$ | 25.0 | 25.0 |
| $2H_2O - CH_3$ | 35.0 | 47.0 |
| $2H_2O - CH_2$ | 39.0 | 51.0 |
| $2H_2O - CH_2OH$ | 14.0 | 26.0 |
| $CH_3 - CH_3$ | 24.0 | 36.0 |
| $CH_3 - CH_2$ | 19.0 | 31.0 |
| $CH_3 - CH_2OH$ | 41.0 | 53.0 |
| $CH_2 - CH_2$ | 9.0 | 21.0 |
| $CH_2 - CH_2OH$ | 22.0 | 34.0 |
| $CH_2OH - CH_2OH$ | 25.0 | 37.0 |

Seven potential values in the ranges specified in Table 1 were selected with a spacing of 2 DPD units, inclusive of the lower and upper bounds. The sampling grid which specifies the parameters is generated as a Cartesian product over the sets of interaction values resulting in a sampling grid of $40,353,607$ ($7^9$) potential unique combinations of conservative repulsion parameters.

Our workflow engine, CAROL, manages the execution of the DL MESO DPD simulation engine and the UMMAP analysis package, which calculates the $\log P$ from the simulation data. The Bayesian optimization library and CAROL communicate passing predicted $\log P$, experimental data and trail force fields where required. Within the optimizer the calculated $\log P$ data is used to perform an evaluation of force field performance, before selecting the next point in parameter space for sampling. The optimizer finally updates the simulation inputs to reflect its new choice of trial force field parameters. This process is carried out iteratively for a fixed number of iterations.

At the end of each iteration, the accuracy of the parameters with respect to $\log P$ is determined using the L2-norm, defined as in equation 4. The L2-norm serves as a loss function within the model optimization process.

$$L2 - norm = \sqrt{\sum_{n=1}^{N} (|x_n^{exp} - x_n^{model}|^2)} \tag{4}$$

where $N$ is the number of samples, $x_n^{exp}$ is the real observable for the $n^{\text{th}}$ experiment and $x_n^{model}$ is the corresponding model observable for the $n^{\text{th}}$ experiment run with the current parameter set. In our implementation the observable $x$ will be $\log P$. The loss function is calculated from the 15 chemical systems that we simulate in the training set and defines a single metric for the utility of the parameters for a single iteration. With this loss function a value of 0 would represent a complete fit.

For the optimization process to begin, the Gaussian process model, which acts as a probabilistic surrogate for the real objective function, is initialized with known data from previous calculations. This is achieved by selecting at random 10 force fields and running simulations to obtain the loss (as defined in equation 4), which are then provided as training data to the Gaussian process. Once complete the real optimization process begins and runs in a sequential manner, assessing the suitability of the force field parameters one set at a time for a total of 30 iterations. This number of iteration was selected as it represents approximately one week of compute time for the current proof-of-concept case. The choice of number of iterations is somewhat arbitrary and could be increased or decreased as one desires dependent on resources and the accuracy required of the force field.

For each iteration of the workflow, simulations and analysis tools were run over 15 nodes, each with 16 cores. The Bayesian optimization library on the other hand is computationally inexpensive and is executed on a single core.

Appropriate error handling and monitoring of the simulation and analysis stages are performed to ensure suitable feedback can be provided at each iteration. This prevents the automation process faulting and failing to complete. Three types of major failures are detected and handled within this workflow: simulations may not run or fail to complete; the force field parameters are bad such that the simulated system has only a single liquid phase

rather than two (a requirement for measuring $\log P$[11]); or, the model observable lies outside the measurable range (i.e. there is poor sampling). The workflow handles failed simulations by assigning an extreme value to the observable, i.e. $\log P$ values of $\pm 10.0$. This is also the *modus operandi* where the force field fails such as when the simulated water and organic solvent partition has collapsed or significantly mixed. In this event there is no clear way to define the regions to sample and hence calculate the concentration of solute within a region.

## DPD model definition

In our approach the DPD beads represent molecular fragments comprising 1–2 'heavy atoms' (specifically C and O in this work), with the exception of water ($H_2O$) which is treated super-molecularly. This provides the potential for a wide variety of both aqueous and non-aqueous systems to be modeled by combining these fragments in different configurations. Crucially this approach also leads to an extensible parameter set, since the molecular palette is easily enlarged to include other chemical groups.

In our model we adopt four different bead types representing water, two representing alkanes and one bead for alcohol functionality (Table 2). To establish the basis for the coarse grained (CG) scheme, we first follow Groot and Rabone in defining a water *mapping number*, in our case $N_m = 2$ so that each water bead ($2H_2O$) corresponds, on average, to two water molecules.[52] Following well established protocols we also assert that the density of water in our model (in reduced DPD units) corresponds to $\rho r_c^3 = 3$, where $r_c$ is the cutoff distance for an interaction.[7] We can then use the mapping number tautology $\rho N_m v_m \equiv 1$, where $v_m \approx 30\,\text{Å}^3$ is the molecular volume of liquid water, to determine that $r_c \approx 5.65\,\text{Å}$. This underpins the conversion of all lengths and molecular densities in the model.

Alkane molecules are constructed from connected (bonded) beads comprising (i) $CH_2$ groups of atoms and (ii) $CH_3$, a terminal methyl group. Similarly alcohol molecules are constructed by bonding together alkane beads and a specific bead containing an alcohol functionality, *e. g.* comprised of the $CH_2OH$ group of atoms. Atom to beaded structures for

the molecules explored in this work are given in detail in Table 2.

Table 2: Coarse grained (CG) representations of molecules considered in the present work. The CG bead content is denoted by the contents of square brackets.

| molecule | SMILES code | $n$ beads | CG bead mapping |
|---|---|---|---|
| n-hexane | cccccc | 6 | $[CH_3][CH_2]_4$ $[CH_3]$ |
| n-heptane | ccccccc | 7 | $[CH_3][CH_2]_5$ $[CH_3]$ |
| methanol | co | 1 | $[CH_2OH]$ |
| ethanol | cco | 2 | $[CH_3][CH_2OH]$ |
| 1-propanol | ccco | 3 | $[CH_3][CH_2]$ $[CH_2OH]$ |
| 1-butanol | cccco | 4 | $[CH_3][CH_2]_2[CH_2OH]$ |
| 1-pentanol | cccco | 5 | $[CH_3][CH_2]_3[CH_2OH]$ |
| 1-hexanol | ccccco | 6 | $[CH_3][CH_2]_4[CH_2OH]$ |
| 1-heptanol | cccccco | 7 | $[CH_3][CH_2]_5[CH_2OH]$ |
| 1-octanol | ccccccco | 8 | $[CH_3][CH_2]_6[CH_2OH]$ |
| 1-nonanol | cccccccco | 9 | $[CH_3][CH_2]_7[CH_2OH]$ |
| butan-1,4-diol | occcco | 4 | $[CH_2OH][CH_2]_2[CH_2OH]$ |

Having decided on the CG level, the next part of the model definition is to specify the bonded interactions between beads: once set these interactions will not be optimized by the parameterization procedure. Here we take an approach motivated by previous work and our own experience. A simple harmonic potential $\phi_b = \frac{1}{2}k_b(r_{\alpha\beta} - r_0)^2$ was chosen to represent bonds between connected DPD beads, where $r_{\alpha\beta}$ is the distance between bonded beads $\alpha$ and $\beta$. The nominal bond length, $r_0$, is set as 0.3 for bonds $[CH_2] - [CH_{2/3}]$ and 0.35 for bonds $[CH_2OH] - [CH_{2/3}]$. This minimizes the number of parameters to be fitted in this initial exploration of applying Bayesian optimization in this manner. A single bond constant $k_b = 150$ was adopted throughout (in units of $k_BT$). Note that, contrary to the usual practice in MD, we (and others) do not exclude the 1-2 and 1-3 non-bonded interaction between two bonded DPD beads.

In our model we explicitly introduce an element of rigidity by including a harmonic angular potential between conjoining pairs of bonds. This has been demonstrated to be essential for the correctness of molecular models at the level of coarse graining used here.[53,54]

We here adopt the same three-body angular potential used by Smit and collaborators,[53,55] *viz.* $\phi_a = \frac{1}{2}k_a(\theta_{\alpha\beta\gamma} - \theta_0)^2$ where $\theta_{\alpha\beta\gamma}$ is the angle between the bonds $\vec{\alpha\beta}$ and $\vec{\beta\gamma}$ of bonded bead triplet $\alpha, \beta$ and $\gamma$. We use $\theta_0$ of 105° for angles $[CH_{2/3}] - [CH_2] - [CH_{2/3}]$, 125° for angles $[CH_2OH] - [CH_2] - [CH_{2/3}]$ and $k_a = 5$ (in units of $k_BT$) for all angles.

For the non-bonded interactions between beads $i$ and $j$ we take the standard DPD pairwise soft repulsion, $\phi = \frac{1}{2}A_{ij}(1 - r_{i,j}/R_{ij})^2$ for $r \leq R_{ij}$ and $\phi = 0$ for $r_{i,j} > R_{ij}$, where $A_{i,j}$ is the interaction strength, $R_{i,j}$ the interaction cut-off distance and $r_{i,j}$ the distance between centers of bead $i$ and $j$.[7] We set the self-interaction of water beads to be equal to $A_{H_20,H_20} = 25.0$. The optimization of all other $A_{ij}$ ($i \neq j$ and $i = j$) is the central problem addressed in the present work.

## Literature Data

In this work we have collected experimental data from the literature for three definitions of $\log P$: Octanol/Water ($\log P_{(Oct/H_2O)}$); Hexane/Water ($\log P_{(Hex/H_2O)}$) and Heptane/Water ($\log P_{(Hep/H_2O)}$). Previous attempts towards parameterizing DPD force fields to $\log P$ data focused exclusively on $\log P_{(Oct/H_2O)}$.[11] From compilations of experimental data on $\log P_{(Oct/H_2O)}$ measurements it is clear that in some cases the quality of the data is quite variable. For simple alkanes and alcohols (simple alcohol meaning here an alkane chain with a single OH substitution), the variance between experimental procedures can be as large as 0.6 log units, considering direct measurement methods only.[56] If indirect measurement methods are included, the variance in reported $\log P$ values for these molecules can be as large as 0.7 log units (see the range in reported $\log P$ values for 1-butanol and methanol in Sangster 1997[57]). Typical estimates of experimental errors on a single determination for $\log P$ range ≈0.05 - 0.25 log units.[56] Similar experimental errors are reported for direct methods measuring $\log P_{(Hex/H_2O)}$.[58,59] The difficulties in assessing the accuracy of experimental data for force field parameterization has been expounded for other properties of industrial interest.[60] This places bounds on the accuracy our force field can realistically achieve. Given that the target

data used here comes from a variety of experimental sources we would consider a good result to have errors in the region of $\approx \pm 0.7$ log units. This level of error is consistent with other current state-of-the-art $\log P$ prediction methods, which generally come from Quantitative Structure Activity Relationships (QSAR).[61,62]

Table 3: $\log P$ data from the literature curated for use in this work with the source reference given by each value.

| Solute | $\log P_{(\mathrm{Oct/H_2O})}$ | $\log P_{(\mathrm{Hex/H_2O})}$ | $\log P_{(\mathrm{Hep/H_2O})}$ |
|---|---|---|---|
| 1-propanol | 0.25[57] | -1.48[63] | -1.52[64] |
| 1-butanol | 0.84[57] | -0.78[65] | -0.70[64] |
| 1-pentanol | 1.51[57] | -0.40[63] | -0.40[64] |
| 1-hexanol | N/A | 0.45[65] | 0.45[64] |
| 1-heptanol | N/A | 1.21[63] | 0.98[64] |
| 1-octanol | N/A | N/A | 1.62[64] |
| 1,4-butanediol | -0.80[11] | N/A | N/A |

## Bayesian Optimization Workflow

Figure 1 visually depicts the connections between the various workflow elements adopted in this work. The simulations are all run with the DL MESO DPD[66] simulation engine and the analysis, which predicts the $\log P$ from the simulations, is provided by the UMMAP program.[67] The sections in orange outline the areas where the workflow engine CAROL[43] orchestrates the simulations and analysis. The sections in blue outline the areas controlled by the Bayesian optimization library, which evaluates the relative success of a set of force field parameters, selects the next force field to trial and produces updated input files for the simulations.

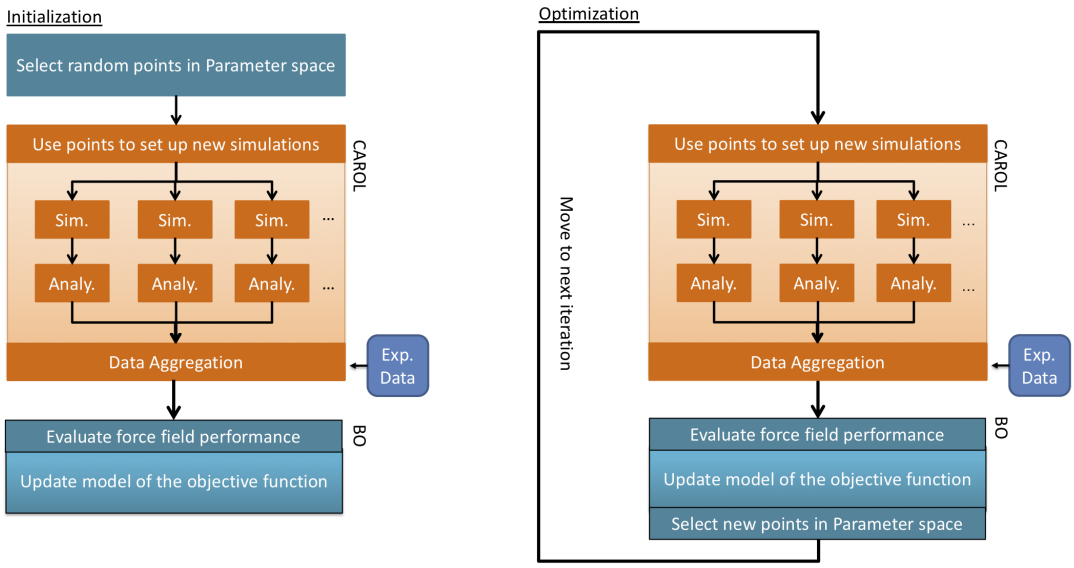


Figure 1: Flow charts representing stages that compose CAROL and the Bayesian optimization library. Arrows in the flowchart show the data flow direction. Left, the initialization process provides initial training data sampled randomly without replacement. Right, the optimization process, once the optimizer is trained, it selects a new sampling point based upon its current knowledge of the parameter space.

On the left hand side of Figure 1, the initialization process is expounded, which provides initial training data for the optimizer upon which to initialize its Gaussian process. In this phase, force field parameters are selected at random from a sampling grid constructed by the optimizer based upon user input. Once complete, the optimizer has some information on the parameter space it is operating within and thus can begin to guide the parameter optimization process.

The optimization process is detailed on the right hand side of Figure 1. In this phase the optimizer calculates a loss at the end of each iteration which is added to the Gaussian process training data. The Gaussian process is retrained and the acquisition function recalculated using the retrained Gaussian process. The maximum value of the acquisition function is found and the force field corresponding to the maxima in the acquisition function is then trialed.

# Results and discussion

We present here the results of the automated DPD force field parameterization using Bayesian optimization. We begin by assessing the performance and overall utility of the optimization method using summary statistics considering all training set systems. We follow this with a microscopic investigation, exploring the choices of parameters from a physical perspective. Finally, we validate the force field using examples from outside the training set.

## Bayesian Optimization Vs Random Sampling

Bayesian optimization searches the parameter space in an intelligent manner balancing the need for exploration in a global search with local exploitation. As an initial benchmark we compare our results against several instantiations of a random search. This tests that intelligent navigation provides a benefit above simple random trail and error. The parameterization engine was initialized with 10 random samples across the grid. The engine was then run for 30 optimization iterations. This process was independently repeated three times using different initial training sets. The purpose of such repetition, was to test the methods performance with variations in the starting data, given a reasonable limit in the number of optimization iterations which can be performed in a timely manner (30 optimization iterations in the current work represents approximately one week over 15 compute nodes).

Bayesian optimization does not operate in the same manner as a conventional gradient based optimization routine. Instead, it operates probabilistically balancing exploration and exploitation, meaning that sequential steps will not necessarily show improvement. As a result, it is much clearer to track the so called regret, which follows the best currently located parameters and associated loss. Figure 2 displays the mean regret trajectory over the three independent Bayesian optimization runs and three independent random sampling runs. The solid lines represent the lowest mean loss encountered at that iteration over the three independent runs of the Bayesian optimization and random sampling. The shaded

regions are the boot strapped (random sampling with replacement) 95% confidence intervals calculated at each point over the three independent runs of the Bayesian optimization and random sampling.

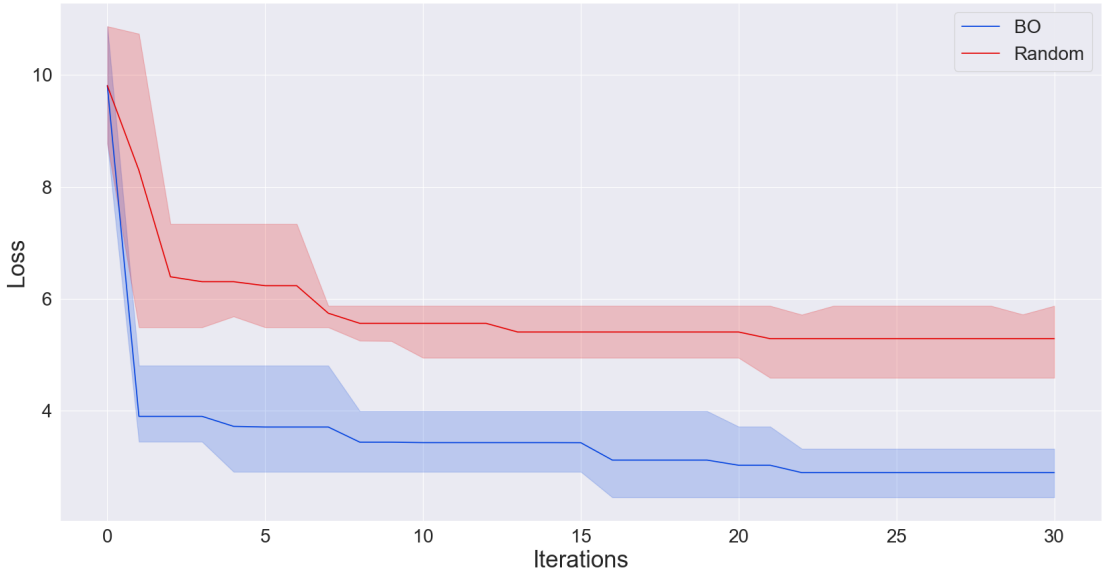


Figure 2: Comparison of optimized and randomly sampled runs. The boot strapped 95% confidence bounds are shown by the shaded areas and the mean regret for each iteration over three independent repetitions of random sampling and Bayesian optimization are displayed as the solid lines. The zeroth iteration is set to the mean of the Bayesian optimization training data for both the random and Bayesian optimization runs.

From Figure 2, the trend is clear, Bayesian optimization consistently achieves a lower loss value (lower regret) faster than random sampling. This is an important first hurdle, displaying the need for intelligent guidance in order to minimize the time to solution. This test also suggests that the parameter space is sufficiently complicated that given the same number of chances, one is unlikely to stumble upon the best solution.

The best force fields found in the independent runs themselves achieved the summary statistics described in Table 4 and 5. The summary statistics clearly demonstrate the superiority of the Bayesian optimization runs, in which the RMSE from the best Bayesian optimization force fields is substantially lower than the best random sampling force fields.

Additionally, the $R^2$ metric reveals better correlations between predicted and experimental data for the Bayesian optimization runs compared to the random sampling runs. These trends remain when the independent runs are considered in isolation rather than on average. The data for each independent Bayesian optimization run is given in Table 5. Data on all runs, including the independent random sampling runs is given in the SI.

Table 4: Summary statistics for the best trialled force field from each random or optimization run over 30 iterations.

| Run name | $R^2$ | RMSE |
|---|---|---|
| Best from Random sample | 0.63 | 1.2 |
| Best from Bayesian optimization | 0.78 | 0.63 |

## LogP Predictions

In this section, we investigate the accuracy of the log $P$ predictions and the optimal parameters are explored for physical significance. Considering firstly all three Bayesian optimization runs, one can see that the best force fields over the three runs appear to sit in two different regions of parameter space. The data is shown in Table 5.

Table 5: The best force fields found in each of the independent Bayesian optimization runs and the best force field found from random sampling. In all cases the dissipative parameter $\gamma_{ij} = 4.5$ for all interactions. The dissipative cutoff for all simulations is set to 1.15. All simulations were run in the NPT ensemble at a DPD temperature of 1 and pressure of 23.7, which correspond to $25^o$ and 1 atmosphere. Bond lengths, $R_{ij}$, are applied to all simulations.

| Interactions | $R_{ij}$ | $A_{ij}$ BO run 1 | $A_{ij}$ BO run 2 | $A_{ij}$ BO run 3 | Best Random $A_{ij}$ |
|---|---|---|---|---|---|
| $CH_2 - CH_2$ | 0.9250 | 21 | 21 | 9 | 15 |
| $CH_2 - CH_2OH$ | 0.9370 | 22 | 22 | 22 | 26 |
| $CH_3 - CH_2$ | 0.9410 | 19 | 19 | 19 | 25 |
| $CH_3 - CH_3$ | 0.9570 | 36 | 36 | 32 | 26 |
| $CH_3 - CH_2OH$ | 0.9525 | 49 | 53 | 49 | 43 |
| $CH_2OH - CH_2OH$ | 0.9800 | 25 | 31 | 25 | 27 |
| $2H_2O - CH_2$ | 0.9625 | 39 | 39 | 39 | 39 |
| $2H_2O - CH_3$ | 0.9785 | 35 | 35 | 35 | 37 |
| $2H_2O - CH_2OH$ | 0.9900 | 14 | 14 | 14 | 16 |
| $2H_2O - 2H_2O$ | 1.0000 | 25 | 25 | 25 | 25 |
| Loss | | 3.3 | 2.45 | 2.91 | 4.59 |
| $R^2$ | | 0.72 | 0.78 | 0.77 | 0.63 |
| RMSE | | 0.86 | 0.63 | 0.75 | 1.19 |

Looking at the results in Table 5 Bayesian optimization runs 1 and 2 seem to occupy a similar region of parameter space with generally similar parameters except $CH_3 - CH_2OH$ and $CH_2OH - CH_2OH$, which differ by 4 and 6 DPD units respectively. However, Bayesian optimization run 3 displays a slightly different force field suggesting potentially two related regions of the parameter space which hold suitable force fields.

Across the three force fields, six interactions are the same; $CH_2 - CH_2OH$, $CH_3 - CH_2$, $2H_2O - CH_2$, $2H_2O - CH_3$, $2H_2O - CH_2OH$ and the reference interaction $2H_2O - 2H_2O$. Interestingly, this includes all four of the water interactions. This may be rationalized as the solvents must remain relatively immiscible in order to provide an interface in the simulation cell between the solvents, which is a requirement for a $\log P$ calculation to be carried out. Therefore, interactions which maintain relatively immiscible solvents could be considered a prerequisite for successful calculations. From the remaining four interactions, which are all between the organic molecule bead types, the variation in performance of these force fields

is derived.

In comparison to similar force fields, generated previously by some of the authors (RLA and DB), published in Anderson *et al*,[11] the present force field offers greater flexibility in terms of molecule definition. This is due to $CH_2$ beads being defined as oppose to coarser $CH_2 - CH_2$ beads, meaning that odd carbon chain lengths can be constructed. The present force field and that of Anderson *et al* 2017, offer a comparable level of accuracy on the training data in terms of $\log P$ predictions. However, this level accuracy is maintained over the validation set in the present case, where as the previous attempt suffered a substantial reduction in accuracy over the validation set. In addition to these scientific points, RLA estimates the parameterization published in Anderson *et al* 2017 took approximately 16 weeks. The present parameterization effort took approximately 1.5 weeks from initialization to termination, hence the present method offers a substantial improvement in time to solution.

Considering the results from Table 4, it is clear that Bayesian optimization run 2 provided the best force field parameters considering the summary statistics. The optimal force field in this run generated DPD $\log P$ predictions across the set of 15 systems with an RMSE of 0.63 log units and an $R^2$ 0.78 on the training set. Comparing this against the best force field located by random sampling, which achieved RMSE 1.19 and $R^2$ 0.63, it is clear that Bayesian optimization has arrived at superior force fields. The results are presented visually in Figure 3.

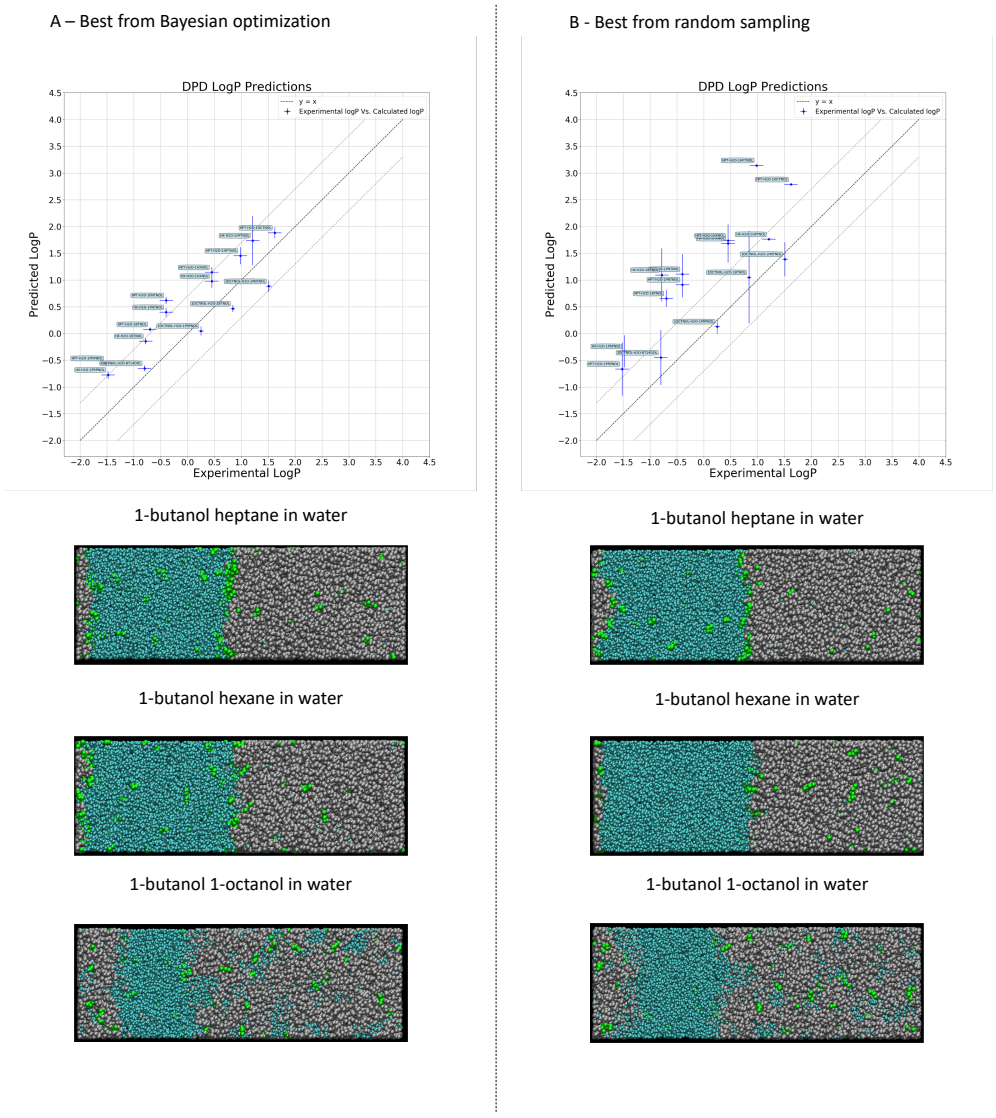


Figure 3: Cyan beads are water, grey beads are the organic solvent and green beads are the solute. Column A, left, represents the $\log P$ predictions over all 15 solutions for the best force field encountered during Bayesian optimization . Column B, right, represents the $\log P$ predictions over all 15 solutions for the best force field encountered during random sampling. A representative example of the simulations employing both force fields is given for the solute 1-Butanol in each of the solvent combinations. Y error bars represent the standard deviation.

Considering Figure 3, it is clear that the best force field located by random sampling is biased towards more positive $\log P$s. In the system snapshots, one can visually verify that fewer solute molecules are present in the aqueous phase, which would lead to a more

positive $\log P$. The grey dashed lines represent $\pm 0.7$ log units, the maximum spread found in the experimental data for some solute molecules. In Figure 3, it is also notable that most predicted points lie within this $\pm 0.7$ log unit error margin for the best force field found using Bayesian optimization , and achieve a smaller standard deviation around this mean value ($y$ error bars). However, applying the best force field from random sampling, most points lie outside of the $\pm 0.7$ log unit error margin and the standard deviation in some cases cannot be calculated due to samples in which no solute molecules are found in one of the solvents. Both force fields however, manage to maintain a solvent boundary between the water and organic solvent, which is something that not all force fields that were trialed achieved.

In the simulations using the best force field, we can see that families of solute molecules in the different solution mixtures follow the expected trends as the carbon chain lengths increase. There is also an interesting observation that the $\log P_{(\mathrm{Oct/H_2O})}$ seems to follow a different trend to the other definitions of $\log P$. This is demonstrated by the separation of the $\log P_{(\mathrm{Oct/H_2O})}$ systems lying clearly on a different trend, with a shallower gradient, to the $\log P_{(\mathrm{Hep/H_2O})}$ and $\log P_{(\mathrm{Hex/H_2O})}$ systems. This is perhaps due to the slight miscibility of water in octanol which is also present in the simulations.

Currently our loss function weights all points equally. This is chosen to insist that the force field is generally applicable across this family of systems. However, this does mean that a failed molecule can dominate the loss function, leading to a suggestion that the force field is poor, which whilst true when considered over all systems, is not necessarily the case for all other systems in the data set. Therefore, naturally the question of whether the accuracy in a defined, potentially small set of molecules, is more important than general applicability could be posed. In this work, there is an additional consideration; if we can automatically generate these force fields in a small enough amount of time does general applicability matter? We could simply generate bespoke force fields optimized for accuracy over generality. We believe this is a topic for further work and consideration with a caveat that, even though the current work shows a substantive reduction in the time to generate a force field compared to some

of the authors previous work, another considerable reduction in the time to solution would be needed to make such a suggestion practical.

## Validation

In this section we validate the force field parameters. Figure 4 shows the results from a validation set of four similar molecules with experimental log $P$'s. One can see clearly that the validation set lies on the same trend lines as the training data as shown in Figure 4. Having added the validation data to the training data the summary statistics remain similar $R^2$ 0.8 and RMSE 0.65, compared to the training set $R^2$ 0.78 and RMSE 0.63. In comparison to some of the authors previous work this model maintains its accuracy from training into the validation data. Previous work showed a notable increase in RMSE when moving from training to validation set.[11]

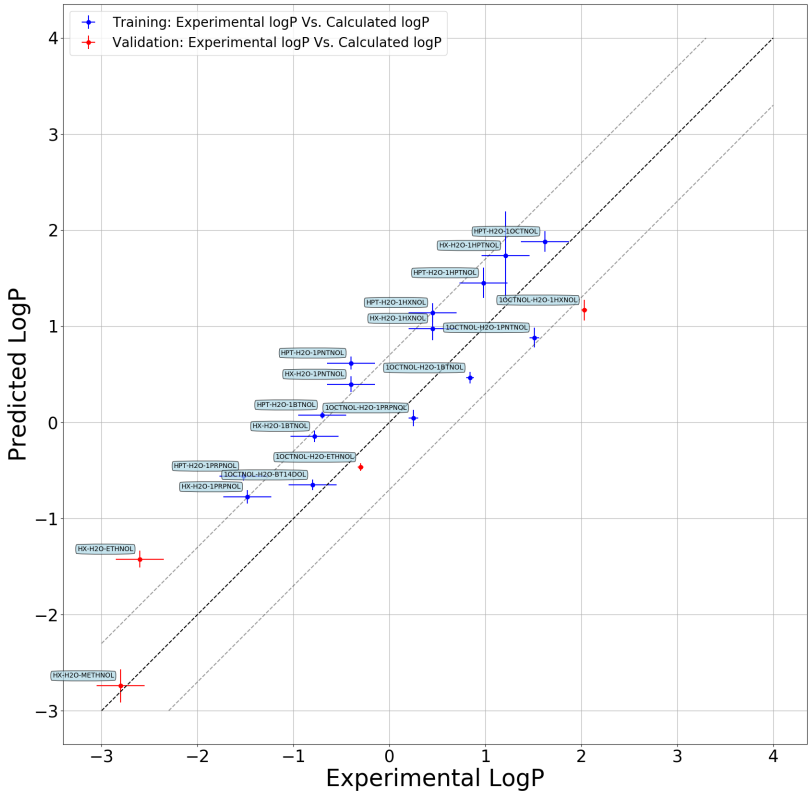


Figure 4: log $P$ validation test. Blue data were included in the training set discussed above. Red data were used as validation set.

## Future work

We believe there are areas where further improvements can be made, which will require subsequent thorough investigations. Of particular note, investigations detailing the effects of different kernel choices for the Gaussian process and acquisition function for Bayesian optimization need to be investigated. Additionally, the effects of varying the loss function should be considered together with the questions of transferability and multi-objective optimization. There are also further questions about how best to access the potential parameters, which rest between the grid points we have defined and how to automatically determine the optimal coarseness of the model. However, we do believe this novel method holds much promise,

owing to its efficiency, relative ease of automation and low compute overheads.

# Conclusions

In conclusion, we show that Bayesian optimization can be effectively applied to optimize a molecular force field, which is treated as a black box function. Having provided 10 randomly selected initial data points Bayesian optimization finds a force field with an RMSE of 0.63 log units and an $R^2$ of 0.78 for the training set in 30 optimization iterations. This level of accuracy is in line with current state of the art cheminformatics models. This force field has been found in approximately 1.5 weeks, with no human intervention once the process had begun, demonstrating Bayesian optimization's ability to automatically discover good regions of parameter space for molecular interaction potentials. This is in comparison to the sixteen week time frame that resulted in the development of parameters in Anderson *et al.*2017[11]

Over several separate instantiations of Bayesian optimization and random sampling, we show that Bayesian optimization locates a superior force field faster. We also note that the observable parameter, in this case $\log P$, is much more poorly predicted by even the best of the randomly sampled force fields. The best force fields from all of the independent Bayesian optimization instantiations provide reasonable predictions with RMSE's and $R^2$ values in line with state-of-the-art models.

Considering the current state-of-the-art force fields available for DPD, and the methods employed to generate these force fields, we believe our approach is arguably one of the most efficient taking approximately 1.5 weeks, start to finish, to generate a good force field over a modest data set with relatively modest compute resources (15 compute nodes). Many state-of-the-art DPD force fields have been generated by hand requiring months, possibly years, of a researcher's time. In this work, we also see that the force field which is found has similar predictive accuracy for $\log P$ compared to other DPD force fields. Additionally, the force field is shown to be stable in its predictive accuracy over a small validation set to

a greater extent than in other state-of-the-art DPD force fields.[11]

## Acknowledgement

The authors thank Michael Johnston, Bill Swope and Kirk Jordan for many stimulating discussions that helped shape the work contained in this article. This work was supported by the STFC Hartree Centre *Innovation: Return on Research programme*, funded by the UK Department for Business, Energy & Industrial Strategy.

## References

(1) Car, R.; Parrinello, M. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

(2) Tuckerman, M. E. Ab initio molecular dynamics: basic concepts, current trends and novel applications. *Journal of Physics: Condensed Matter* **2002**, *14*, R1297.

(3) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. a.; Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* **1983**, *4*, 187–217.

(4) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham III, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995**, *91*, 1–41.

(5) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry B* **2007**, *111*, 7812–7824.

(6) Rudd, R. E.; Broughton, J. Q. Coarse-grained molecular dynamics and the atomic limit of finite elements. *Phys. Rev. B* **1998**, *58*, R5893–R5896.

(7) Groot, P. B., Robert D.Warren Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *The Journal of Chemical Physics* **1997**, *107*, 4423–4435.

(8) Espanol, P.; Warren, P. Statistical Mechanics of Dissipative Particle Dynamics. *Europhys. Lett.* **1995**, *30*, 191.

(9) Español, P.; Warren, P. B. Perspective: Dissipative particle dynamics. *J. Chem. Phys.* **2017**, *146*, 150901.

(10) Foloppe, N.; MacKerell Jr, A. D. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of computational chemistry* **2000**, *21*, 86–104.

(11) Anderson, R. L.; Bray, D. J.; Ferrante, A. S.; Noro, M. G.; Stott, I. P.; Warren, P. B. Dissipative particle dynamics: Systematic parametrization using water-octanol partition coefficients. *The Journal of chemical physics* **2017**, *147*, 094503.

(12) Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S. Building a more predictive protein force field: a systematic and reproducible route to AMBER-FB15. *The Journal of Physical Chemistry B* **2017**, *121*, 4023–4039.

(13) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225–11236.

(14) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic

peptides and crambin. *Journal of the American Chemical Society* **1988**, *110*, 1657–1666.

(15) Vanommeslaeghe, K.; MacKerell Jr, A. D. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *Journal of chemical information and modeling* **2012**, *52*, 3144–3154.

(16) Vanommeslaeghe, K.; Raman, E. P.; MacKerell Jr, A. D. Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *Journal of chemical information and modeling* **2012**, *52*, 3155–3168.

(17) Popelier, P. L. Molecular simulation by knowledgeable quantum atoms. *Physica Scripta* **2016**, *91*, 033007.

(18) Huang, L.; Roux, B. Automated force field parameterization for nonpolarizable and polarizable atomic models based on ab initio target data. *Journal of chemical theory and computation* **2013**, *9*, 3543–3556.

(19) Wu, J. C.; Chattree, G.; Ren, P. Automation of AMOEBA polarizable force field parameterization for small molecules. *Theoretical chemistry accounts* **2012**, *131*, 1138.

(20) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *The Journal of Physical Chemistry Letters* **2014**, *5*, 1885–1891.

(21) Zielinski, F.; Maxwell, P. I.; Fletcher, T. L.; Davie, S. J.; Di Pasquale, N.; Cardamone, S.; Mills, M. J.; Popelier, P. L. Geometry Optimization with Machine Trained Topological Atoms. *Scientific reports* **2017**, *7*, 12817.

(22) McDonagh, J. L.; Silva, A. F.; Vincent, M. A.; Popelier, P. L. Machine Learning of Dynamic Electron Correlation Energies from Topological Atoms. *Journal of chemical theory and computation* **2017**, *14*, 216–224.

(23) Silva, A. F.; Vincent, M. A.; McDonagh, J. L.; Popelier, P. L. The transferability of topologically partitioned Electron correlation energies in water clusters. *ChemPhysChem* **2017**, *18*, 3360–3368.

(24) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters* **2010**, *104*, 136403.

(25) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *The Journal of chemical physics* **2016**, *145*, 170901.

(26) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the $\Delta$-machine learning approach. *Journal of chemical theory and computation* **2015**, *11*, 2087–2096.

(27) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* **2012**, *108*, 058301.

(28) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science* **2017**, *8*, 3192–3203.

(29) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters* **2015**, *6*, 2326–2331.

(30) Dequidt, A.; Solano Canchaya, J. G. Bayesian parametrization of coarse-grain dissipative dynamics models. *The Journal of chemical physics* **2015**, *143*, 084122.

(31) Solano Canchaya, J. G.; Dequidt, A.; Goujon, F.; Malfreyt, P. Development of DPD coarse-grained models: From bulk to interfacial properties. *The Journal of chemical physics* **2016**, *145*, 054107.

(32) Liu, P.; Shi, Q.; Daumé III, H.; Voth, G. A. A bayesian statistics approach to multiscale coarse graining. *The Journal of chemical physics* **2008**, *129*, 12B605.

(33) Ruff, K. M.; Harmon, T. S.; Pappu, R. V. CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. *The Journal of chemical physics* **2015**, *143*, 12B607_1.

(34) Li, Z.; Bian, X.; Yang, X.; Karniadakis, G. E. A comparative study of coarse-graining methods for polymeric fluids: Mori-Zwanzig vs. iterative Boltzmann inversion vs. stochastic parametric optimization. *The Journal of chemical physics* **2016**, *145*, 044102.

(35) Tschöp, W.; Kremer, K.; Batoulis, J.; Bürger, T.; Hahn, O. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polymerica* **1998**, *49*, 61–74.

(36) Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4*, 225–248.

(37) Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *Journal of computational chemistry* **2003**, *24*, 1624–1636.

(38) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry B* **2005**, *109*, 2469–2473.

(39) Schnackenberg, L. K.; Beger, R. D. Whole-molecule calculation of log P based on molar volume, hydrogen bonds, and simulated 13C NMR spectra. *Journal of chemical information and modeling* **2005**, *45*, 360–365.

(40) McDonagh, J.; van Mourik, T.; Mitchell, J. B. Predicting melting points of organic molecules: applications to aqueous solubility prediction using the general solubility equation. *Molecular informatics* **2015**, *34*, 715–724.

(41) Lyubartsev, A. P.; Jacobsson, S. P.; Sundholm, G.; Laaksonen, A. Solubility of Organic Compounds in Water/Octanol Systems. A Expanded Ensemble Molecular Dynamics Simulation Study of log P Parameters. *The Journal of Physical Chemistry B* **2001**, *105*, 7775–7782.

(42) Anderson, R. L.; Bray, D. J.; Del Regno, A.; Seaton, M. A.; Ferrante, A. S.; Warren, P. B. Micelle Formation in Alkyl Sulfate Surfactants Using Dissipative Particle Dynamics. *Journal of Chemical Theory and Computation* **2018**, *14*, 2633–2643.

(43) Shkurti, A. *https://www.scd.stfc.ac.uk/Pages/carol.aspx* **2018**,

(44) Chen, J.; Xin, B.; Peng, Z.; Dou, L.; Zhang, J. Optimal contraction theorem for exploration–exploitation tradeoff in search and optimization. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **2009**, *39*, 680–691.

(45) Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **2002**, *3*, 397–422.

(46) Tan, K. C.; Chiam, S. C.; Mamun, A.; Goh, C. K. Balancing exploration and exploitation with adaptive variation for evolutionary multi-objective optimization. *European Journal of Operational Research* **2009**, *197*, 701–713.

(47) Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(48) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* **2016**, *104*, 148–175.

(49) Brochu, E.; Cora, V. M.; De Freitas, N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599* **2010**,

(50) Jasrasaria, D.; Pyzer-Knapp, E. O. Dynamic Control of Explore/Exploit Trade-Off In Bayesian Optimization. *arXiv preprint arXiv:1807.01279* **2018**,

(51) Groves, M.; Pyzer-Knapp, E. O. Efficient and Scalable Batch Bayesian Optimization Using K-Means. *arXiv preprint arXiv:1806.01159* **2018**,

(52) Groot, R.; Rabone, K. Mesoscopic Simulation of Cell Membrane Damage, Morphology Change and Rupture by Nonionic Surfactants. *Biophysical Journal* **2001**, *81*, 725–736.

(53) Venturoli, M.; Smit, B. Simulating the self-assembly of model membranes. *PhysChemComm* **1999**, *2*, 45–49.

(54) Nagarajan, R. Molecular Packing Parameter and Surfactant Self-Assembly: The Neglected Role of the Surfactant Tail. *Langmuir* **2002**, *18*, 31–38.

(55) Venturoli, M.; Sperotto, M. M.; Kranenburg, M.; Smit, B. Mesoscopic models of biological membranes. *Physics Reports* **2006**, *437*, 1 54.

(56) Sangster, J. Octanol-Water Partition Coefficients of Simple Organic Compounds. *Journal of Physical and Chemical Reference Data* **1989**, *18*, 1111–1229.

(57) Sangster, J. *Octanol-water partition coefficients: fundamentals and physical chemistry*; John Wiley & Sons, 1997.

(58) Bergström, C. A.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharmaceutical research* **2002**, *19*, 182–188.

(59) Schulte, J.; Dürr, J.; Ritter, S.; Hauthal, W.; Quitzsch, K.; Maurer, G. Partition coefficients for environmentally important, multifunctional organic compounds in hexane+ water. *Journal of Chemical & Engineering Data* **1998**, *43*, 69–73.

(60) Swope, W. C.; Johnston, M. A.; Duff, A. I.; McDonagh, J. L.; Anderson, R. L.; Alva, G.; Tek, A. T.; Maschino, A. P. The Challenge to Reconcile Experimental Micellar Properties of the CnEm Nonionic Surfactant Family. *The Journal of Physical Chemistry B* **2019**,

(61) Zang, Q.; Mansouri, K.; Williams, A. J.; Judson, R. S.; Allen, D. G.; Casey, W. M.; Kleinstreuer, N. C. In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *Journal of chemical information and modeling* **2017**, *57*, 36–49.

(62) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *Journal of cheminformatics* **2018**, *10*, 10.

(63) Ruelle, P. The n-octanol and n-hexane/water partition coefficient of environmentally relevant chemicals predicted from the mobile order and disorder (MOD) thermodynamics. *Chemosphere* **2000**, *40*, 457–512.

(64) Tayar, N. E.; Tsai, R.-S.; Testa, B.; Carrupt, P.-A.; Hansch, C.; Leo, A. Percutaneous penetration of drugs: A quantitative structure-permeability relationship study. *Journal of pharmaceutical sciences* **1991**, *80*, 744–749.

(65) Garrido, N. M.; Economou, I. G.; Queimada, A. J.; Jorge, M.; Macedo, E. A. Prediction of the n-hexane/water and 1-octanol/water partition coefficients for environmentally relevant compounds using molecular simulation. *AIChE Journal* **2012**, *58*, 1929–1938.

(66) Seaton, M.; Smith, W. DL MESO USER MANUAL. 2016.

(67) Bray, D. J. *https://www.scd.stfc.ac.uk/Pages/UMMAP.aspx* **2017**,