PAPER • **OPEN ACCESS**

Optimal design of experiments in the context of machine-learning inter-atomic potentials: improving the efficiency and transferability of kernel based methods

To cite this article: Bartosz Barzdajn and Christopher P Race 2025 *Modelling Simul. Mater. Sci. Eng.* **33** 025011

OPEN ACCESS
IOP Publishing
Modelling and Simulation in Materials Science and Engineering
Modelling Simul. Mater. Sci. Eng. **33** (2025) 025011 (20pp) https://doi.org/10.1088/1361-651X/ada050

Optimal design of experiments in the context of machine-learning inter-atomic potentials: improving the efficiency and transferability of kernel based methods

Bartosz Barzdajn[1,*] **and Christopher P Race**[2]

[1] The University of Manchester, Oxford Rd, M139PL Manchester, United Kingdom
[2] The University of Sheffield, Western Bank, Sheffield S102TN, United Kingdom

E-mail: b.barzdajn@outlook.com

Received 16 May 2024; revised 2 December 2024
Accepted for publication 17 December 2024
Published 27 January 2025

Abstract

Data-driven machine learning (ML) models of atomistic interactions are often based on flexible and non-physical functions that can relate nuanced aspects of atomic arrangements to predictions of energies and forces. As a result, these potentials are only as good as the training data (usually the results of so-called *ab initio* simulations), and we need to ensure that we have enough information to make a model sufficiently accurate, reliable and transferable. The main challenge stems from the fact that descriptors of chemical environments are often sparse, high-dimensional objects without a well-defined continuous metric. Therefore, it is rather unlikely that any ad hoc method for selecting training examples will be indiscriminate, and it is easy to fall into the trap of confirmation bias, where the same narrow and biased sampling is used to generate training and test sets. We will show that an approach derived from classical concepts of statistical planning of experiments and optimal design can help to mitigate such problems at a relatively low computational cost. The key feature of the method we will investigate is that it allows us to assess the quality of the data without obtaining reference energies and forces—a so-called offline approach. In other words, we are focusing on an approach that is

* Author to whom any correspondence should be addressed.

easy to implement and does not require sophisticated frameworks that involve automated access to high performance computing.

Keywords: interatomic potentials, machine learning, optimal desing, material science, GAP

1. Introduction

Inter-atomic potentials are surrogate models replacing complex quantum-mechanical (QM) calculations with fast-to-evaluate functions, directly or indirectly dependent on the positions of atoms, returning forces and energies. With such potentials, we can simulate millions of atoms at timescales of nanoseconds or microseconds; something beyond the reach of QM models.

When regarded as a regression problem, the development of these potentials is challenging. Firstly, the problem cannot be easily formulated using a space with a comprehensive metric (e.g. $\mathbb{R}^{3\times N}$, N being a number of atoms). A useful potential needs to be applicable to a variety of configurations requiring different numbers of atoms in a computational cell. Hence, the map, linking to energies and forces, will be defined (implicitly) on a collection of positions of various sizes. For this reason alone, formulation of potentials will be facilitated by descriptors of chemical environments—functions that represent a collection of atoms as a vector, tensor or a set (with the simplest form being a list of pairwise distances). As a result, we have two, rather than one, consecutive and non-trivial relationships. The first maps the positions of atoms to their abstract representatives and is defined by our decision to use a particular descriptor. The second is defined by a function that accepts a descriptor as an input and outputs energies and/or forces. This has been illustrated on the example of the embedded-atom model (EAM) (see e.g. A.F. Voter in [1, 2]) in figure 1.

For example, when we wish to train an inter-atomic potential, we need to choose a suitable descriptor and determine the right model complexity. This way, we can maximise the precision of the fit without introducing a bias. However, it is easy to make a mistake by selecting a simple descriptor that cannot differentiate between many distinct chemical environments. At the same time, we can define a very flexible regression model that can accommodate all the differences in the training data. As a result, we will encounter the problem of over-fitting despite defining a model with the right overall capacity. This is not an unusual problem in development of EAM potentials.

In response to such challenges, researchers are turning to machine-learning (ML) methods, such as, after [3], neural network potentials [4], kernel based methods like the Gaussian approximation potential (GAP) – the framework for 'atomistic' Gaussian process (GP) regression [5] and Gradient-domain machine learning (ML) models [6], moment tensor potentials (MTPs) [7], methods with a physically inspired basis like atomic cluster expansion (ACE) [8], and many others [9, 10].

In this paper we focus on methods that are coupled with extensive, or rather 'expressive', descriptors, such as the smooth overlap of atomic positions (SOAP) ([11]), designed to be applicable to many materials, distinguish between alloying elements and reflect nuanced many-body interactions. In other words, models flexible enough to be as good as the data provided. The price we will have to pay for high levels of accuracy is susceptibility to insufficient information. Once we decide to use descriptors and regression models that are sufficiently complex to be error-free and universal, we need to make sure that the training data is sufficient as well. This way interpolations and extrapolations will not fail as soon as we try to make predictions on examples moderately distinct from the training set.

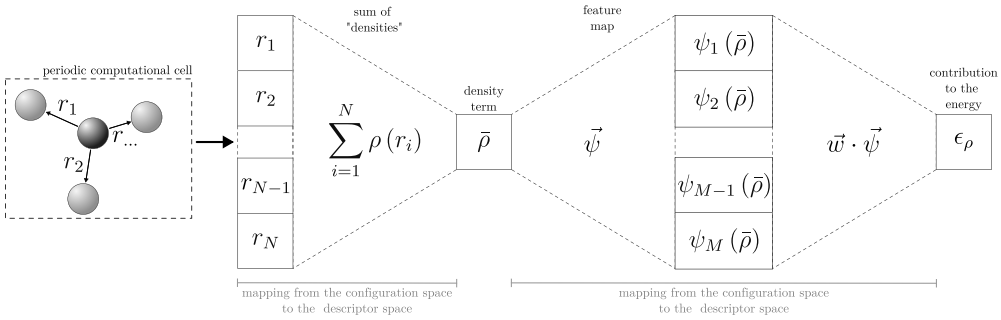

Figure 1. Illustration of the problem of double mapping using the example of the EAM model. Here we focus only on the contribution of a specific atom to the total energy and consider only the embedding representing the many-body interactions. The density $\bar{\rho}$ in principle refers to the local electronic density and consists of contributions $\rho(r_i)$ from neighbouring atoms, where ρ is a non-linear function that can also depend on adjustable parameters, while r_i represents the distance to a specific neighbour. We also assume that considering the first N neighbours provides almost complete information. The quantity $\bar{\rho}$ can be considered as a one-dimensional descriptor of the local environment. The contribution to the energy will be a non-linear function of this quantity. However, in this example we assume that it can be expressed in a linear basis to illustrate the model complexity represented by the number of features M. An example of a feature can be $\bar{\rho}$ raised to the kth power in the polynomial representation of the mapping.

It is deceptively tempting to manage this problem using brute force and integrate into the training set as many examples as possible. However, it can be inefficient, unreliable, and even with access to high-performance computing (HPC) facilities prohibitively expensive.

The choice of training examples are often motivated by physics and intuition. But what is a distinct training example from the perspective of a researcher, might not be distinguishable by the model or descriptor. What is informative may not be representative. For instance, a training set might be improved significantly by the inclusion of unrealistic configurations that provide better coverage of possible descriptor values. Likewise, polluting a data set with contradicting (e.g. when descriptor cannot distinguish between physically different examples) or irrelevant information might result in a loss of accuracy or performance respectively.

We need a systematic approach, as we have to navigate an extensive set of possible atomic configurations, through at least two maps (configuration - descriptor and descriptor - predictions of energies, etc), and most likely without a metric. It is not a surprise that the development of a potential is such an involving and risky task.

In practice, researchers often resort to using some form of active learning, like the classical Csányi *et al* learn-on-the-fly approach [12], where model, or models, trained on partial training sets are used to query if new candidates are a good addition to the current set. As reported by Jinnouchi *et al* in their overview, these strategies can result in a significant reduction in the time required for training [13]. Albeit, they still require access to vast computational resources and complex software infrastructure. Given that the training and assessment will likely involve shared HPC resources, with strict management policies, and infrastructure that might be difficult to implement. Furthermore, initial queries are made using models that can be significantly flawed unless the original database was already extensive at the beginning. Which brings us right back to the initial point.

Candidate searches can be more efficient if they can be done without labelling, i.e. obtaining values for associated quantities of interest, which usually means estimating energies and forces

with *ab initio* QM calculations. With kernel methods, for example, we can use the pivoted low-rank approximations to mark a subset of candidates for labelling. After all, the idea behind this approximation is to select a subset of data that is most representative of the whole, i.e. giving the closest approximation to the full kernel matrix. This feature of labelling is even included in the original implementation of the GAP [5]. Likewise, as shown by Podryabinkin and Shapeev in their active learning scheme [14], or more recently by Lysogorskiy *et al* [15], we can apply in the context of training of potentials criteria developed for statistical design of experiments, i.e. using the rigorous language of statistics to decide in advance what information to add or start with.

We argue that classical concepts of statistical planning, such as optimal design of experiments [16], many of which were developed when even supercomputers could not match the speed of today's smartphones, can indeed provide a basis for efficient solutions to the problems outlined above. Optimality in this context means that for a given number of training examples (budget), we can choose those that minimise the uncertainty associated with the model parameters. Optimal designs are unique to linear basis function models and thus to kernel methods such as GAP [16]. Associated methods allow us to manage model uncertainty and verify the feasibility of learning before even the first calculation is made. They allow rapid exploration and sifting through a large number of training examples. At the same time, they provide access to well-established measures of quality that are independent of labels/outputs. These measures are one of the main advantages of statistical planning and offer a definite improvement over 'vanilla' active learning.

In our work, we will present how to optimise data for on kernel-based methods rather than 'weight-space' models like ACE or MTP, as the associated optimality criteria and algorithms are less known and more difficult to implement. We also aim for solutions that provide optimal, or more realistically optimised, sets without the need for retraining and estimation of empirical errors. In this respect, our work is related to that of Karabin and Perez [17]. However, we focus on methods that aim to find optimal solutions by directly considering optimality criteria rooted in statistical theory.

2. Methodology

To emphasise the key characteristic of designs we are focusing on, we refer to them as *a priori* designs. As mentioned before, an optimal training set can be constructed before any *ab initio* calculation or experiment is made. The key is to choose an appropriate measure. For example, in linear regression with constant normally distributed 'noise', the variance-covariance matrix of regression coefficients is $\Sigma = \sigma^2 \left(X^\top X\right)^{-1}$, where σ^2 is the variance of the 'noise', and X is the design matrix—our usual assembly of inputs associated with features. It immediately transpires that by selecting appropriate inputs to X, we can minimise Σ and we do not need any knowledge about the σ^2 or outputs (the y's in our data set). More details can be found in appendix A.

There are many well-established criteria for optimality of designs that are more suitable than the direct optimisation of Σ. The most common is the D—optimality that seeks to maximise the determinant of the observed Fisher information matrix, which is given by $\sigma^{-2}X^\top X$ for linear models with the constant 'noise' (notice the similarity with the formulation of Σ). This particular criterion was used in earlier work by Podryabinkin and Shapeev to mark candidates for evaluation/labelling [14]. However, we have to keep in mind, that each family of models, be it ordinary least-squares or kernelised ridge regression (KRR, [18]), will require different

strategies and criteria of optimality. Moreover, non-linear models will have only local optimal designs for a specific range of parameters [16].

In our work, we focus on the GAP model framework of GP regression (GPR). This framework can be categorised as state-of-the-art atomistic modelling that can give predictions almost as good as the training data.

To reiterate argument from the introduction, high-quality data, usually in the form of atomic configurations and associated energies and forces, are expensive to generate, even with modern hardware and efficient implementations of the density functional theory (DFT) [19]. Furthermore, we want to use the model to make predictions on large computational cells that are beyond the reach of the DFT method, rather than performing a significant number of less demanding calculations. Hence, we cannot test the model by comparing it to a reference, nor can we rely on the prediction variance of the model, as it is unreliable in highlighting errors due to poor definition of the descriptor. All this renders any form of active learning a less appealing choice.

To address design criteria for GPR, and as such GAP, we need to refer to the fundamentals. A GP is a collection of random variables with joint Gaussian distribution

$$f(\vec{x}) \sim \mathcal{GP}\left(m(\vec{x}), k(\vec{x})\right),$$

specified by mean function m and covariance function k. The model, specified by rules to evaluate m and k, is defined by the posterior distribution conditioned on the data ([20], chapter 2). In the GPR framework, each element of the training set and each extrapolation point corresponds to a degree of freedom of a multivariate Gaussian distribution. The expectation and variance of this distribution are defined by a kernel matrix—a matrix of inner products between all data points (training and predictions). Formulation of GPR and the framework in the context of atomistic modelling can be found in [21] and [22].

The GPR can be regarded as a linear method with respect to weights (model parameters) if we choose to formulate it in the so-called weight-space view. In the function-space view, it is strongly related to KRR[3]. Both methods share the same estimator of expectation if the penalty corresponds to the prior variance (e.g. data 'noise').

Relationships between these methods suggest that optimal designs are to some extent transferable. However, we need to be careful. In our experiments, when we tried to minimise the maximum prediction variance of the KRR model in an active learning framework, we created a dataset that was performing worse, when used to train a GAP model, than the source-sampling. Here, the source-sampling is a method that is applied to generate a large number of candidates, that are later reduced by the algorithm to create an optimised training set.

The most straightforward design for GPR is the maximum entropy (MaxEnt) sampling introduced by Shewry and Wynn in [23]. We refer here to the entropy defined as the expectation of information content, also known as the Shannon information. In principle, it is a different quantity the entropy defined in physics. The optimality criterion authors propose (section 4) is

$$\max \log\det \mathbf{K}, \tag{1}$$

where $\mathbf{K}$ is the covariance matrix of the multivariate Gaussian. This is the part of the covariance matrix associated with the training data in the GPR framework.

[3] Kernel methods replace the explicit evaluation of features with their dot products, which we can calculate using fast-to-evaluate formulas – kernel functions. This way, we can implicitly work with a large, or infinite, number of features. However, full matrices used in regression will be as large as the dataset.

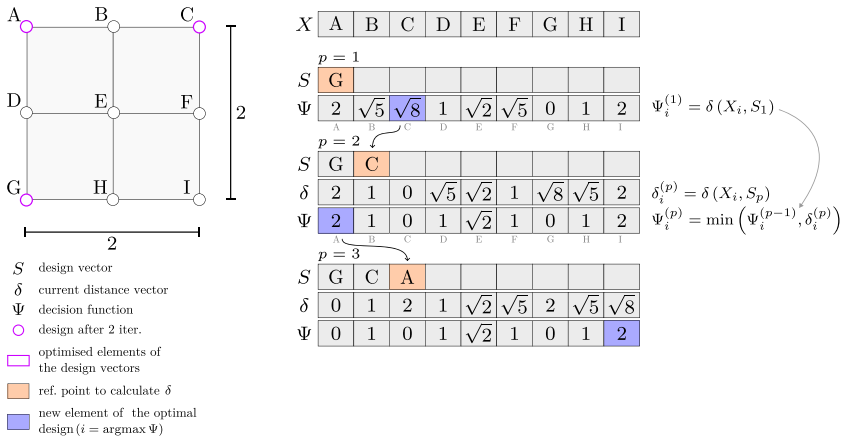

Figure 2. Illustration of the Golchi and Loeppky algorithm [25] for max-min designs, which maximises the minimum distance, using the example of nine points on a square grid. The state in each iteration is defined by the design vector S and vector Ψ. The vector Ψ can be considered as a 'decision' function while X represents the pool of candidates. Initially, Ψ consists of distances between the first element and the remaining elements. In the following iterations, Ψ is updated element-wise according to $\Psi_i^{(p)} = \min\left(\Psi_i^{(p-1)}, \delta_i^{(p)}\right)$, where p is the iteration index, $\delta_i^{(p)}$ is the distance between pth and ith candidate and $\delta(\cdot)$ is the distance function. In each iteration we find the maximum value of update Ψ. Position of this element indicates the new optimal design point.

As discussed by the authors, this criterion attempts to maximise the variability in the training set. In other words, their distinctiveness and the coverage of the domain. Initially, we used this criterion in combination with a modified exchange algorithm [24], i.e. an algorithm that swaps candidates in the training set with the pool of potential replacements until the set becomes an optimal representation of the domain. However, we were concerned with its low efficiency of exploration as the algorithm required, to be computationally efficient, generating a covariance matrix for all the data, limiting the number of candidates we can consider at the same time.

For this reason, we decided to apply a conditional max–min design with an algorithm presented in [25] that performed well in comparison with our previous solutions. This design aims to maximise the minimum distance between samples. The algorithm is illustrated in figure 2. A simple and efficient implementation of this algorithm can be found in the appendix B, as well as in [26] along with its applications in other contexts.

It proceeds in a greedy manner, i.e. candidates are ordered in terms of importance, starting with the most informative examples, in other words the most distanced. As the distance, we choose the squared kernel distance ([27])

$$D_K^2 = k(p,p) + k(q,q) - 2k(p,q), \tag{2}$$

where K is the covariance function between configurations p and q. We chose a measure based on the similarity kernel in order to create a design that is directly related to the regression method. Here, we used a complete covariance, as described in [21], consisting of radial-basis functions for pair-potentials and a polynomial kernel with the SOAP descriptor for many-body interactions. Defining the distance in such a way also is consistent with optimality criterion 1 (see figure 3).

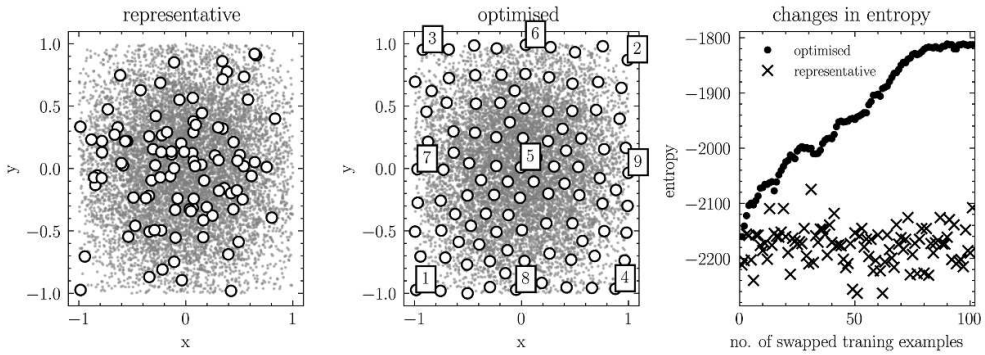

Figure 3. Application of the Golchi and Loeppky algorithm [25] for max-min designs. In this example, we are optimising the training set for Gaussian process regression. The distance measure follows the definition 2 and is based on a Gaussian kernel with unit scale parameter. The training points are selected from a pool of candidates generated using strongly biased sampling. This pool consist of 10^4 samples from the normal distribution $\mathcal{N}(0, 0.5)$. On the plots they are represented by small grey points with opacity. Such an example simulates conditions under which we have to select atomistic configuration for training of ML potentials. The first plot on the left illustrates a representative design, which is a random sub-sample of 100 candidates. This plot also illustrates how biased the underlying sampling is. Next, in the middle, we show the optimised training set, also consisting of 100 examples. The index indicates the order of addition of the first nine optimal points, revealing the algorithm's strategy of filling empty spaces after enclosing a domain. Finally, the plot on the right shows how the entropy changes when we replace the random/representative examples with the optimised ones. Entropy for representative sets corresponds to different realisations of this design. Here, the entropy of the resulting kernel matrix K is proportional to $\log(\det K)$ [23]. Note that the aim is not to provide a uniform importance sampling, but optimal training points for a given regression method. In this case, however, these objectives coincide so that we can visually assess the quality of the solution. Other models may have a different solution. See the example (a) in figure 7. Finally, this results can be easily recreated using the algorithm 2 from appendix B.

According to our experiments with the Euclidean metric and highly biased sampling, the algorithm generates space-filling designs that embrace the whole domain. For linear least-squares regression, such designs also tend to minimise maximum prediction variance, and as such, they are consistent with the aims of G–optimality [28, 29]. Figure 3 demonstrates the test of the algorithm on a simple example of two-dimensional Cartesian space.

The distance is the same as in 2. However, in this example we selected the Gaussian kernel. As such, the solution is an optimised set with respect to the performance of the GPR regression, rather than optimal filling of the space. Although, for a Gaussian kernel with a large scale parameter these goals will coincide.

3. Generation of optimal training sets and evaluation methodology

Zirconium is one of the metals of particular interest to our research group, and we could greatly benefit from a robust and well-designed database of DFT calculations that we could use to train atomistic ML models. Naturally, our focus here will be on the GAP method.

It is reasonable to begin with a collection of examples that will inform the model about elastic deformations (see e.g. [30]). This first step is an excellent case study that is both realistic and simple. In other words, we will build an optimised database that we can use to reliably and efficiently predict the elastic energy density of a pure Zr lattice. The main objective is to assess whether the methodology, i.e. this particular form of optimal design, is advantageous compared to more direct and common approaches.

The basic premise is to develop a representative sampling, equivalent to a random uniform sampling in Euclidean space, and use it to generate optimal and representative training sets (the procedure is described in the following subsections). These sets will be used to train ML models with a consistent method, that allows for comparison, which then will be evaluated on all available data. To generate reference values (labels/outcomes), we will use the DFT implemented in VASP, with optimised basis and standard convergence criteria (more details will be provided later). Rather than relying only on abstract measures, as we did in the first examples presented in figure 3, we decided that we can make a much more convincing case by demonstrating that optimal designs lead to better models with the same training budget.

In this section we will discuss how optimal sets can be created, starting with a description of the overall framework, the design of the sample generation, the selection of optimal candidates, and the strategy for evaluation.

It is important to emphasize here that while we will quantitatively evaluate the performance of the models, at this stage we are not trying to create the best possible model for a given set, nor are we trying to show which optimality criteria have the greatest utility. Our aim is to demonstrate convincingly that the introduction of optimal design is beneficial in the context of development of ML potentials.

Returning to the main subject, this case study provides an excellent opportunity, as it is conceivable to achieve an adequate sampling for elastic deformations in a straightforward manner. This allows an easy comparison between a more 'traditional', albeit not naive, approach to generation of training sets and methods rooted in statistical planning and optimal design.

We also allowed ourselves to simplify the framework for fitting ML models. For example, rather than investing in prior evaluation of the uncertainty associated with DFT calculations, which is by no means a trivial task, we decided that the variance of the training data may be treated as any other hyperparameter and included in their optimisation[4]. However, we acknowledge that incorporating such prior information into the model would lead to more reliable estimators and could help to avoid issues such as overfitting.

On the other hand, our models are representatives/proxies of the data, and we prefer to avoid situations in which the fitting procedure compensates for deficiencies of the data with prior information, e.g. by helping to 'filter-out' the 'noise' and equalise highly perturbed data with superior, less 'noisy' counterparts. In other words, we have emphasised the simplicity and, above all, comparability. In this context we prefer a systematic and consistent approach rather than a more precise although biased (by prior knowledge) solutions. Finally, we also believe that a more forgiving set is better overall, although we are aware that any simplifications introduced here will open up our result to interpretation.

[4] Representing the uncertainty of the training data as a parameter or set of parameters is one of the most compelling features of GAP and GPR.

3.1. Generating optimal training sets

In order to create optimised training sets that allow us to train more robust models, we first generate as large a set of candidates as possible. The core idea behind optimisation is to sift through these candidates and select those that will optimise the performance measure for a given budget. As described in the previous sections, we will use a max–min design, which maximises the minimum distance defined by carefully selected descriptors and kernels.

It is important to remember that an empirical model is only as good as the data it interpolates (if we exclude all prior information). Therefore, we need to ensure that the space is adequately covered and that we have a sufficiently good source sampling—one that can draw from each possibility, here defined by the structure and strain state, with a reasonable, but not necessarily optimal, probability.

Each time, the transformation matrix, used to deform a randomly picked structure, is generated from a random uniform vector of eigenvalues (we have essentially reversed the eigenvalue decomposition) and rotated using Euler angles chosen indiscriminately from the group of 3D rotations. This method was carefully chosen to avoid the generation of highly biased sets and overlapping candidates.

Furthermore, without correct constraints, successful optimisation is not guaranteed. When using high-dimensional descriptors and kernel-based distances, the algorithm may favour solutions where all examples are orthogonal to each other. This would mean that the training set is useless for models that are supposed to be interpolators and require at least some correlations to be non-zero. We can fall into this trap by allowing the selection of extremely deformed examples. Therefore, it is necessary restrict the search space accordingly. In our case we did this by limiting the maximum and minimum eigenvalues of the deformation matrix.

Initially, we generated the deformation using uniform sampling of the elements of the strain tensor. However, in our experiments, this sampling was severely underperforming. The optimised sets were always significantly better and it was difficult to judge the results against the 'random'/representative sampling. As a result, we could see the benefits of statistical planning right from the start, as it forced us to enhance the framework. As we observed a significant improvements in models trained on representative sets, we asserted that this method provided a suitable benchmark for evaluating the performance of optimal designs.

In the case study, we used a sufficiently large candidate set size of 100 000 (necessary size was determined in convergence studies), from which we selected 500 examples using the framework discussed earlier, i.e. the conditional max-min design with the greedy algorithm [25]. Note that with the improved sampling, models based on optimal and representative data would essentially converge for training sets larger than 1000–2000 elements, at least in terms of energy predictions. However, these estimates are very specific and depend on the species and phases we wish to include.

At this point we would like to deviate from the current analysis and broaden the scope of the discussion for a moment. For sets consisting of configurations with more than one unique descriptor, such as those containing thermalised lattices or defects, it would be very difficult, if not nearly impossible, to design an appropriate source sampling based only on measures of deviations from equilibrium atomic positions. The most important consideration here is that descriptors representing realistic configurations, whether we use SOAP or otherwise, often by design occupy only a fragment of a much 'larger' high-dimensional space (the requirement of universality). Therefore, it is recommended that when designing source-sampling, we limit ourselves to examples that are within the potential application domain. Otherwise, it may be impossible to fill the relevant subspace and find a good compromise between the ability to

discriminate between 'similar' configurations and the ability to appropriately propagate correlations. It is a dilemma similar to that faced by an artist when choosing a brush size to match the scale of the canvas and at the same time provide a sufficient level of detail.

In our experience, a good solution is to use very low-quality *ab initio* molecular dynamics (MD) at temperatures sufficient to explore all the distances between pairs of atoms that we expect to encounter. The high precision of the forces is unimportant here, since the only reason to perform MD is to obtain realistic constraints on the atomic configurations. The advantage of this approach is that at temperatures above the melting point and sufficiently large computational cells, we can in principle explore all possible realistic descriptors of local chemical environments. The disadvantage is that it lacks any quality of uniformity, which means that it may only be efficient in conjunction with discriminative statistical designs such as the one discussed in this paper. The irony here is that in order to train reliable ML models—which inherently have no pre-existing physical relationships, waiting for data to imprint them—sometimes we must first impose physically-based constraints. This process essentially injects prior physical knowledge that these models, by design, do not initially reflect. With respect to distance/dissimilarity measures, we have to consider that these configurations would consist of multiple unique descriptors. In such a case, we can either use a single local chemical environment to represent the configuration, e.g. one that is most different from the current selection, or take advantage of kernels that relate entire configurations [31].

We can now return to the main discussion. As planned for the final atomistic model, the kernel matrix, and thus the measure of distance, for the data set optimisation consisted of two components. We used the Gaussian kernel to estimate pair-wise interactions, with a scaling factor of 0.8 and a length scale of 0.3, while for many body interactions we selected the SOAP descriptor and 4th-order polynomial kernel with a 0.3 scaling factor. The SOAP descriptor was initiated with the following string defining the parameters: `soap cutoff = 6.8 l_max = 10 n_max = 10 normalize = T atom_sigma = 0.15 n_Z = 1 Z = {40}`. The cut-off of the pair-wise distance descriptor was set to 7.0. Parameters were selected by attempting to minimise the number of overlapping (kernel distance close to zero) and orthogonal (normalised kernel distance close to 1) examples in the trial candidate set. In other words, we selected a descriptor that can distinguish between the most similar chemical environments and, at the same time, that will not be too sensitive to changes in atomic positions and lose the ability to quantify similarity of most distinct configurations (clipping effect).

Reference data (labels for training sets) were evaluated using the DFT method implemented in VASP [32–35]. We choose the APW (augmented plane-wave) basis with an energy cut-off of 600 eV, automatic determination of number of k-points and Methfessel–Paxton smearing method [36]. We relied on default settings (VASP 5) with respect to other parameters.

3.2. ML models representing different designs

As mentioned earlier, ML (GAP) models fitted to different training sets will be used as proxies to empirically assess the quality of the designs, rather than as the best possible models for a given training set. In this context, we should be aware that with kernel-based methods such as GAP and GPR, it is very easy to achieve very high precision in the representation of the training data. The off-sample performance, however, is strongly influenced by the choice of hyperparameters. In order to achieve consistency in terms of the quality of predictions, while keeping the procedure relatively simple, instead of relying on sophisticated cross-validation methods,

we decided to take advantage of the GPR formulation and select the kernel hyperparameters θ using the maximum likelihood estimation (MLE), which aims to optimise the log-likelihood:

$$\mathcal{L}(\theta) = -\frac{1}{2}\left(y^{\top}Q^{-1}y + \ln\det(Q) + n\ln(2\pi)\right),$$

where $Q = K + \sigma I$, Q is the kernel matrix of the additive model (pair-potential and the SOAP kernel) and σ is the prior variance of the data. The key information is that the latter is also treated as a hyperparameter and subject to optimisation. Therefore, as mentioned before, we simplify the procedure by treating the uncertainty of data as something to be determined. The MLE estimates are known to suffer from slight overfitting. However, they provide good computational efficiency while resulting in well regularised models with respect to the data. As a result, we believe that we do not need to be concerned about 'runaway' interpolation errors, as is often the case with higher order polynomials.

Finally, we will only use energies in the optimisation, as we are considering perfect lattices and we essentially have a single unique descriptor per energy label.

For the sake of computational efficiency, the optimisation procedure used our own simple Python implementation of the GAP, based on the QUIP library ([5]) for evaluation of SOAP vectors, and the covariance matrix adaptation evolution strategy ([37]) with a population of 64 vectors and three independent restarts.

Using the optimised hyperparameters, we trained a GAP model using the `gap_fit` program from the QUIP library. In contrast to the optimisation of the hyperparameters, we used both forces and virials at this stage, although the forces acting on the atoms were always zero due to the crystal symmetry. While the methodology focuses on energies—the optimisation relies only on the direct representation of descriptors (directly related to energy estimates), if we can still observe advantages of the design with additional information, we can make a stronger case. The main reason for this choice was to associate the data with more realistic models.

Finally, we have not used any sparsification methods. That is, all data points are used to define estimates. The reason for this is that sparsification will have overlapping utility in many respects. For example, using the k-means clustering method implemented in the referenced library to replace similar training examples with a single representative is in itself a form of design. However, it cannot be considered optimal in a strict sense.

4. Numerical experiments, conclusions and future work

In this section, we quantitatively evaluate the methodology used to construct optimal training sets and assess the implementation of the algorithm. The analysis begins by examining the impact of the algorithm on the similarity between training examples. In particular, we compare the randomly selected ('representative') deformed structures with those from the optimised set. The results are shown in figure 4.

It is immediately apparent that, also in this case, the algorithm works by removing overlapping examples and emphasising diversity, as indicated by the shift of the entire distribution, including the minimum, towards larger distances. The opposite of overlap is when the examples are orthogonal and the dissimilarity, i.e. the kernel distance, is 2 in our case, as we have omitted the normalisation in equation (2).

While this demonstrates that the methodology improves the quality of a training sets with respect to abstract measures, we aim to show that this change is significant in practical scenarios. Following the procedure described in the previous section, we generated two optimised sets, OPT1 and OPT2, consisting of 500 examples selected from two independent pools of

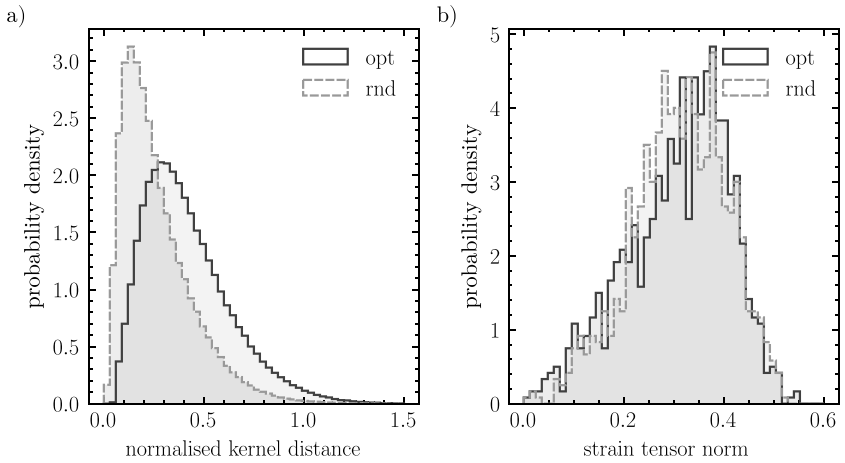

Figure 4. Comparison of pairwise distances within a random and an optimised training set. Figure (a) shows normalised probability histograms of the kernel distance. Figure (b) shows the distribution within a training set of the l_2-norm of the deformation tensor, which we use to quantify its magnitude. The lack of a strong difference between the locations of these distributions shows that the reduction in overlap has not been achieved by simply pushing the 'magnitude' of deformations.

candidates – 100 000 per pool. Additionally, from each pool we selected randomly and indiscriminately a subset of 500 candidates, creating two representative training sets: RND1 and RND2. Here, OPT stands for 'optimised' and RND for 'random', i.e. representative of the source-sampling. As discussed previously, the aim is to demonstrate the benefits of statistical planning and optimal design by constructing a simple but convincing test, rather than to develop a method with the greatest accuracy or utility. However, we believe that a broader comparison, using improved reference data, incorporating different optimality criteria and optimisation algorithms, should be a focus of future research.

The idea is to quantify the performance in a way that is closely related to cross-validation. In the assessment we used all DFT energies and GAP predictions made on all configurations. As we have four training sets and therefore four models, we can make a total of 16 comparisons. For example, we take the GAP model trained on OPT1, evaluate energies on configurations from OPT1, OPT2, RND1, RND2 and compare GAP predictions with DFT references. When we test the model predictions on its own training set, we are essentially evaluating the goodness-of-fit. Otherwise, we inquire about the off-sample performance. Results are presented in figures 5 and 6.

It immediately transpires that the optimised training sets deliver a more consistent and overall better performance than their representative (random) counterparts. Here, we measure the performance as the worst off-sample performance given by the highest, among all test sets, 0.99 quantile of the absolute error.

An interesting observation is that optimal sets tend to perform better when evaluated against other optimal sets. Likewise, representative sets perform better when tested within the same class. Since optimal sets are systematically generated and demonstrate a more consistent performance, the slight advantage of RND sets over OPT sets when tested against RND sets suggests that we may have improved transferability at the expense of some precision in more specific contexts. In other words, optimised sets are likely to provide better coverage of the

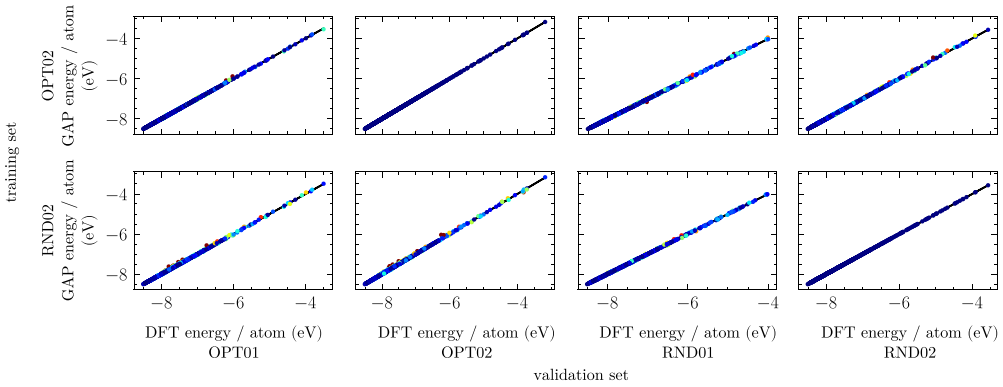

Figure 5. Examples of energy per atom prediction performance on training and validation sets. The rainbow colour map shows the absolute value of the prediction error from figure 6 (the blue colour indicates near-zero prediction error, while the red colour indicates near-maximum prediction error). Due to the nature of the GAP and GPR methods and the choice of how we treat data uncertainty, only the off-sample results show meaningful errors. The main purpose of these plots is to illustrate that in all cases, with a few exceptions, we have achieved good results, indicating that there are no significant problems with the fitting procedure. Here we also provide a sense of scale for the values presented in figure 6.

entire domain, but with fewer samples concentrated around the mode of the source distribution. This result is exactly what we would expect from the optimal design.

When interpreting the results, there are two main considerations. First, the underlying generation of training examples has already been improved as a result of the initial analysis in the context of statistical planning. While this may sound repetitive, it is not always easy or possible to improve the source sampling, as in the case of elastic deformations of perfect lattices. Therefore, we can expect that in many scenarios the result of the optimisation will be more substantial. Second, we defined the distance measure using a kernel matrix associated only with energies. However, the final training sets also consisted of forces and virials as part of the associated labels/outputs (we were concerned that otherwise fitting scenarios would be considered unrealistic). Therefore, we optimised only a part of the information inflow. In conclusion, we believe that we can make a stronger case than one based solely on a straightforward analysis of the data.

We now discuss potential limitations. First, the uncertainty in the reference data was treated as a hyperparameter and optimised within the framework of MLE. In other words, it was incorporated as a standard component of the ML model. This could lead to overfitting. However, despite the potential drawbacks of this approach, discussed earlier in section 3, relatively high differences in energy between examples (from -9 eV to -3 eV) suggest that the discrepancies in the DFT predictions are likely to be smaller than the observed improvements. Note that in the case of GPR and GAP, when the parameters are adequately optimised, even with near-perfect interpolation of the training data, we are unlikely to encounter the ‘runaway’ interpolation errors commonly associated with fitting of high-order polynomials. Furthermore, the performance differences across all training set categories were consistent across 12 comparisons (excluding the goodness-of-fit).

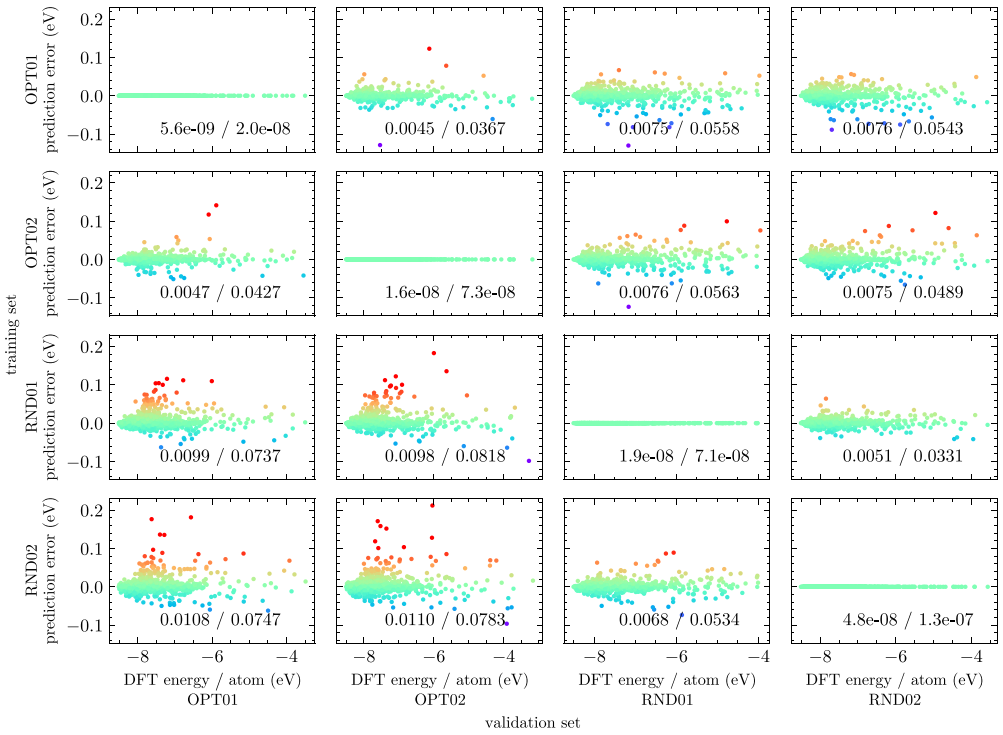

Figure 6. Results of cross-validation. Here (GAP) models are defined by their corresponding training sets and organised in rows. Each column corresponds to the validation set on which the model was tested. The diagonal plots illustrate the performance of the fit. The colour indicates the position on the *y*-axis. While this may be considered redundant, it helps to highlight differences in performance. Summary statistics are: standard deviation (first number in the label) of the error (DFT results vs. GAP prediction) and 0.99 quantile of the absolute error (second number in the label).

Secondly, although not detailed for the sake of brevity, the stresses required to evaluate the elastic constants for the hcp, bcc and fcc structures were also monitored during development. It is important to note that the associated deformation patterns were not included in the training data. We also tracked extrapolations to a dimer. In the best cases, the stresses from the model based on the optimised set essentially overlapped with the DFT results, and although some inconsistencies were observed in other cases, they remained within reasonable limits.

While we recognise that our approach is open to interpretation, we are confident that the results are sufficiently robust to support our hypothesis. Indeed, the optimal design has the potential to significantly improve established methodologies, in particular with regards to the reliability and generalisability of data-driven ML models.

An additional, but quite important, result of this work is that we have shown that empirical methods of performance evaluation, such as the well-known test-train-split framework, should be applied with caution. It is relatively easy to overestimate the predictive power of a model and fall into the trap of confirmation bias. This problem is clearly illustrated by the drop in the off-sample performance of representative (random) models as we move to evaluation versus optimised test sets.

For example, if the method used to generate training examples probes only a narrow part of a domain, we obtain 'dense' sampling, which can give us deceptively good interpolations in that region. While this is not a problem in Euclidean spaces, where it is easy to define indiscriminate sampling, with high-dimensional descriptors it can be extremely difficult, or even impossible, to define a metric that allows us to sample the domain 'uniformly'. In other words, we must always keep in mind that we run the risk of replicating the biases introduced by the source sampling. The statistical planning of experiments may help us to overcome these challenges. However, only when certain conditions are satisfied.

First and foremost, we need an efficient method to adequately sample the application domain. In this paper we decided to use what we like to call '*a priori*' design, i.e. optimal sets were generated without any input of reference values (DFT energies). The difficulty here was the massive RAM requirements resulting from the goal of max-min design, which required all descriptors to be generated beforehand. For other models and/or optimality criteria, the challenges may be different. For example, solutions based on active learning and posterior variance evaluation do not have such requirements. Although they may not be as reliable and require extremely expensive evaluation of energies, forces and virials. For so-called parametric models, we could instead use the classical D-optimality criterion (section A), with well-established procedures that can be adopted to use on-the-fly candidate generation in *a priori* design [38]. However, the computational efficiency of these algorithms may then become a constraint.

The application of optimal design in the context of ML interatomic potentials is still relatively unique. There are therefore many areas for potential development. One is the need for robust ways to define the underlying sampling, especially for many body representations. We suggest that developments around defining similarity measures and even metrics [31, 39, 40] in particular are aligned with this goal and can be used to ensure efficient coverage of the descriptor space. Another important part of the methodology is an efficient optimisation algorithm. Here we would recommend focusing on lazy algorithms that allow continuous exploration of the configuration spaces. However, these algorithms must include appropriate constraints, which may be difficult to determine without input from physical models. As a final suggestion, we encourage readers to explore example implementations of max-min and D-optimal designs in Python, applied to much simpler scenarios [26]. We believe that these examples may help to facilitate the adoption of these methodologies in one's own framework.

Data availability statement

The data can be easily recreated using the methods presented in the paper. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We would like to kindly acknowledge The Engineering and Physical Sciences Research Council (EPSRC) for funding the MIDAS project (Mechanistic understanding of Irradiation Damage in fuel Assemblies – reference EP/S01702X/1). C P Race was funded by a University Research Fellowship of the Royal Society. Calculations were performed on a computational cluster, maintained by the Computational Shared Facility, The University of Manchester.

5. Author contributions

Bartosz Barzdajn: Writing—Original draft, Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization. **Christopher Race**: Writing - Review & Editing, Conceptualization, Methodology, Validation, Formal analysis, Supervision, Project administration, Funding acquisition.

Appendix A. Introducing the concept of the optimal design

We will illustrate the concept of optimal design using the simplest regression model

$$y = Xw + e,$$

where y is a vector of observed values (labels), X is the design matrix with ith row representing observation x_i and defined by a vector-valued feature map $\phi(x_i)$ (e.g. in polynomial regression $\phi_k(x_i) = x_i^k$ with k ranging from 0 to the order of a polynomial), w is a vector of model parameters and e is a disturbance term (assume symmetric uni-modal distribution of e and $\mathrm{cov}(e) = \sigma^2 I$). Obviously, we want to find the model parameters with a minimum number of training examples and a minimum uncertainty.

Consider the example of D-optimality which aims to maximise the determinant of the Fisher information matrix (FIM). For least-square estimators FIM is simply given as $\mathscr{I} = \sigma^{-2} X^\top X$. It is related to the covariance of w which is given by $\Sigma = \sigma^2 \left(X^\top X\right)^{-1}$. Hence, by maximising the information $\mathscr{I}$ we minimise the uncertainty Σ. We see that the objective can be achieved without referring to y's at all and that we can quantify the effect of selecting different training examples without σ, i.e. the noise/uncertainty associated with the data. Hence, the optimal design will be a function solely of X [5]. An example of applying an optimal design to a simple polynomial model can be found in figure 7.

In this example, we intentionally used random uniform sampling rather than the equal division of the domain. While the latter seems like a natural choice, when the feature space is high-dimensional, some forms of random sampling are often considered to be adequate. Therefore, the results can be viewed as a representative illustration of the benefits that optimal design can provide. Furthermore, by selecting low-order polynomials, we can show that optimal designs can differ from the usual intuition. While after some careful consideration we could conclude that the best design for linear models is to concentrate our whole budget on the edges of the domain, the 4×2 (four points sampled twice) design from the second example might be more difficult to foresee without the formal analysis. Particularly, it can be difficult in high-dimensional and non-Euclidean spaces.

Note that there are other popular criteria such as the A-optimality – minimise $\mathrm{tr}\Sigma$, E-optimality—maximise minimal eigenvalue of $\mathscr{I}$ or G-optimality which seeks to minimise the maximum prediction variance [16]. These are just few examples and an appropriate selection will depend on the family of models and expectations with respect to the performance of estimators.

[5] Additionally, if the problem is ill-conditioned, the matrix $X^\top X$ will have zero determinant and will not be invertible.

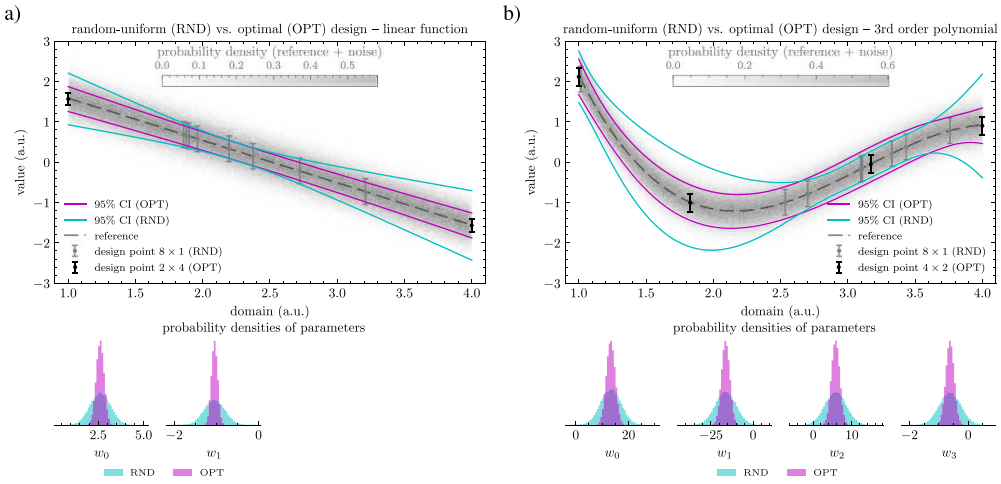

Figure 7. Comparison of optimal design (D-optimality) for polynomial regression with randomly selected point from a domain. Note that all points were selected before making estimates. Confidence interval (CI) and density of parameters are estimated using Monte Carlo (MC) method with 5×10^4 samples. Error bars represent design points, i.e. points where 'measurements' were made in each MC iteration. Functions and their parameters are defined as follows: $f(x) = w_0x^0 + w_1x^1 + \ldots w_Nx^N$. Note that the space-filling design will yield similar results to the optimal design. Significant difference in the performance is emphasised by the unfavourable sample.

Algorithm 1. Generation of the uniform grid of points.

```
x1, x2 = np.mgrid[−1:1.05:.05, −1:1.05:0.05]
x1, x2 = x1.flatten(), x2.flatten()
X = np.vstack((x1, x2)).T
rng = np.random.default_rng()
X = rng.permutation(X, axis = 0)
```

Appendix B. Example implementation in Python

In this section we present a simple implementation of the conditional max-min design (figure 2) using the example of two-dimensional Euclidean space. In this case, it will result in a space-filing design, which is an overall well performing design and is close to optimal for high degree polynomials or kernel based methods with a Gaussian kernel.

We start by generating the pool of candidates. It will be a dense, uniform grid of points covering the entire domain $[-1, 1] \times [-1, 1]$. The Python code using the NumPy library, imported under the 'np' label, can be found in the listing 1.

In the context of ML potentials, in realistic scenarios we do not have well-defined metric, and even random uniform sampling is impossible or impractical, let alone the importance sampling. However, such an idealised test case allows for an easier assessment of the algorithm and its implementation.

The main optimisation loop, presented in listing 2, has a very simple implementation that takes advantage of vectorised element-wise operations.

Algorithm 2. Selection of optimal training points.

```
def delta(x, y, p = 2):
    return np.linalg.norm(x - y, axis = 1, ord = p)

i = np.argmax(np.linalg.norm(X, axis = 1))
X[[0, i]] = X[[i, 0]]
S = [0] Psi = delta(X, X[S[0]])
for k in range (40):
    Psi = np.minimum(Psi, delta(X, X[S[-1]]))
    S.append(np.argmax(Psi))
```

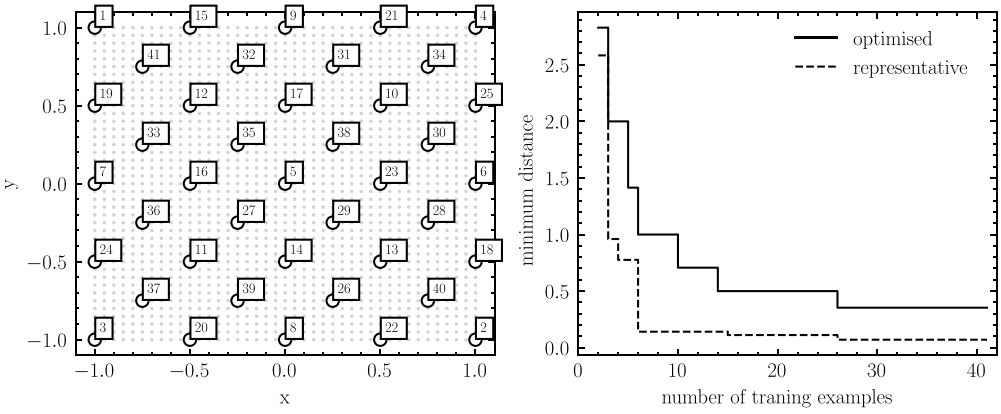

Figure 8. Illustration of the max-min design. On the left, optimal points selected form a candidates distributed on a regular grid. Ordering is 'random'. Labels indicate in which order points were added to the design. On the right, minimum distance of the max-min design compared to random, indiscriminate selection of candidates.

The distance between elements is defined by the Euclidean distance (l_2-norm). The results of the optimisation are presented in Figure 8.

Here, we will omit implementation of the code used to generate figures and calculate the minimum distance.

As in the example from figure 3, given the budget, the algorithm tries to cover the whole space evenly and embrace the whole domain. This is an important feature, as it allow us to adjust the budget simply by selecting first n candidates.

ORCID iDs

Bartosz Barzdajn https://orcid.org/0000-0002-3081-4131
Christopher P Race https://orcid.org/0000-0002-9775-687X

References

[1] J H Westbrook and R L Fleischer 1994 *Intermetallic Compounds* (Wiley)
[2] Voter A F The Embedded-Atom Method (available at: https://public.lanl.gov/afv/VoterEAM chapter.pdf)
[3] Kocer E, Ko T W and Behler J 2022 Neural network potentials: a concise overview of methods *Annu. Rev. Phys. Chem.* **73** 163–86
[4] Lorenz S, Groß A and Scheffler M 2004 Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks *Chem. Phys. Lett.* **395** 210–5
[5] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104** 136403
[6] Chmiela S, Sauceda H E, Müller K R and Tkatchenko A 2018 Towards exact molecular dynamics simulations with machine-learned force fields *Nat. Commun.* **9** 3887
[7] Alexander V S 2016 Moment tensor potentials: a class of systematically improvable interatomic potentials *Multiscale Model. Simul.* **14** 1153–73
[8] Drautz R 2019 Atomic cluster expansion for accurate and transferable interatomic potentials *Phys. Rev.* B **99** 014104
[9] Mishin Y 2021 Machine-learning interatomic potentials for materials science *Acta Mater.* **214** 116980
[10] Friederich P, Häse F, Proppe J and Aspuru-Guzik A 2021 Machine-learned potentials for next-generation matter simulations *Nat. Mater.* **20** 750–61
[11] Albert P B, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115
[12] Gabor Csányi T A, Payne M C and De Vita A 2004 Learn on the Fly': a hybrid classical and quantum-mechanical molecular dynamics simulation *Phys. Rev. Lett.* **93** 175503
[13] Jinnouchi R, Miwa K, Karsai F, Kresse G and Asahi R 2020 On-the-fly active learning of interatomic potentials for large-scale atomistic simulations *J. Phys. Chem. Lett.* **11** 6946–55
[14] Podryabinkin E V and Shapeev A V 2017 Active learning of linearly parametrized interatomic potentials *Comput. Mater. Sci.* **140** 171–80
[15] Lysogorskiy Y, Bochkarev A, Mrovec M and Drautz R 2023 Active learning strategies for atomic cluster expansion models *Phys. Rev. Mater.* **7** 043801
[16] Rasch D and Herrendörfer G Statystyczne Planowanie DośWiadczeń, Wydawnictwo Naukowe, PWN Sp. z o.o
[17] Karabin M and Perez D 2020 An entropy-maximization approach to automated training set generation for interatomic potentials *J. Chem. Phys.* **153** 094110
[18] Kevin P M 2012 *Machine Learning: A Probabilistic Perspective Adaptive Computation and Machine Learning Series* (MIT Press)
[19] Kohn W and Sham L J 1965 Self-consistent equations including exchange and correlation effects 140 A1133–8
[20] Edward Rasmussen C and Williams C K I 2006 *Gaussian Processes for Machine Learning Adaptive Computation and Machine Learning* (MIT Press)
[21] Albert P B and Csányi G 2015 Gaussian approximation potentials: a brief tutorial introduction *Int. J. Quantum Chem.* **115** 1051–7
[22] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 Gaussian process regression for materials and molecules *Chem. Rev.* **121** 10073–141
[23] Shewry M C and Wynn H P 1987 Maximum entropy sampling *J. Appl. Stat.* **14** 165–70
[24] Miller A J and Nguyen N-K 1994 Algorithm AS 295: a fedorov exchange algorithm for D-Optimal Design *Appl. Stat.* **43** 669
[25] Golchi S and Loeppky J L Monte Carlo based designs for constrained domains (arXiv:1512.07328)
[26] Barzdajn B 2024 Implementing conditional max-min designs in python (Zenodo) (available at: https://zenodo.org/records/11191241)
[27] Phillips J M and Venkatasubramanian S A gentle introduction to the kernel distance (arXiv:1103. 1625)
[28] Johnson M E, Moore L M and Ylvisaker D 1990 Minimax and maximin distance designs *J. Stat. Plan. Inference* **26** 131–48
[29] Werner G M, Pronzato L and Waldl H 2011 Beyond space-filling: an illustrative case *Proc. Environ. Sci.* **7** 14–19

[30] Szlachta W J, Bartók A P and Csányi G 2014 Accuracy and transferability of gaussian approximation potential models for tungsten *Phys. Rev.* B **90** 104108
[31] Sandip D, Bartók A P, Csányi G and Ceriotti M 2016 Comparing molecules and solids across structural and alchemical space *Phys. Chem. Chem. Phys.* **18** 13754–69
[32] Kresse G and Furthmüller J 1996 Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set *Comput. Mater. Sci.* **6** 15–50
[33] Kresse G and Furthmüller J 1996 Efficient iterative schemes for ph *ab initio* total-energy calculations using a plane-wave basis set *Phys. Rev.* B **54** 11169–86
[34] Kresse G and Hafner J 1993 *Ab Initio* molecular dynamics for liquid metals *Phys. Rev.* B **47** 558–61
[35] Kresse G and Joubert D 1999 From ultrasoft pseudopotentials to the projector augmented-wave method *Phys. Rev.* B **59** 1758–75
[36] Methfessel M and Paxton A T 1989 High-precision sampling for Brillouin-zone integration in metals *Phys. Rev.* B **40** 3616–21
[37] Hansen N *et al* 2024 CMA-ES/pycma: r4.0.0
[38] Meyer R K and Nachtsheim C J 1995 The coordinate-exchange algorithm for constructing exact optimal experimental designs *Technometrics* **37** 60
[39] Widdowson D and Kurlin V 2021 Pointwise distance distributions of periodic point sets (arXiv:2108.04798)
[40] Widdowson D, Mosca M M, Pulido A, Cooper A I and Kurlin V 2021 Average minimum distances of periodic point sets - foundational invariants for mapping periodic crystals *MATCH Commun. Math. Comput. Chem.* **87** 529–59