Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Fast Bayesian Optimization of Needle-in-a-Haystack Problems using Zooming Memory-Based Initialization

Alexander Siemenn (✉ asiemenn@mit.edu)
Massachusetts Institute of Technology https://orcid.org/0000-0001-8841-7887
Zekun Ren
Singapore-MIT Alliance for Research and Technology https://orcid.org/0000-0002-6987-1715
Qianxiao Li
National University of Singapore (NUS)
Tonio Buonassisi
MIT

Article

Keywords: rare materials discovery, efficient algorithms, adaptive acquisition functions, trust regions, active learning, extremely imbalance data, auxetic materials, thermoelectric materials

Posted Date: September 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2003069/v1

License: 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Fast Bayesian Optimization of Needle-in-a-Haystack Problems using Zooming Memory-Based Initialization

Alexander E. Siemenn[*]
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
asiemenn@mit.edu

Zekun Ren
Department of Electrical and Computer Engineering
Singapore-MIT Alliance for Research and Technology
Singapore 138602, Singapore
Xinterra
Singapore 139949, Singapore

Qianxiao Li
Department of Mathematics
National University of Singapore
Singapore 138602, Singapore

Tonio Buonassisi
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

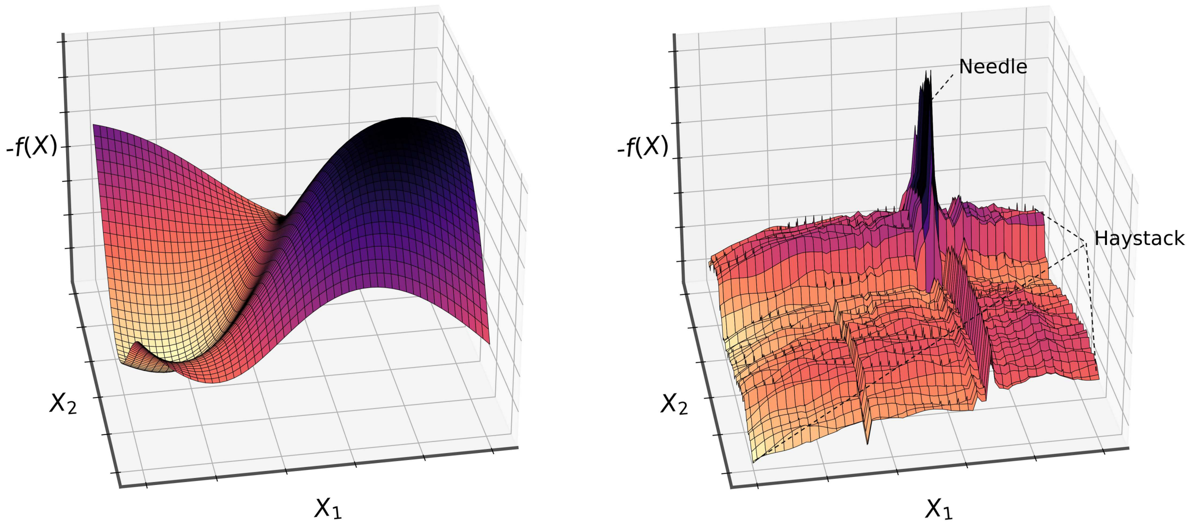
Abstract

Needle-in-a-Haystack problems exist across a wide range of applications including rare disease prediction, ecological resource management, fraud detection, and material property optimization. A Needle-in-a-Haystack problem arises when there is an extreme imbalance of optimum conditions relative to the size of the dataset. For example, only 0.82% out of 146k total materials in the open-access Materials Project database have a negative Poisson's ratio. However, current state-of-the-art optimization algorithms are not designed with the capabilities to find solutions to these challenging multidimensional Needle-in-a-Haystack problems, resulting in slow convergence to a global optimum or pigeonholing into a local minimum. In this paper, we present a Zooming Memory-Based Initialization algorithm, entitled ZoMBI, that builds on conventional Bayesian optimization principles to quickly and efficiently optimize Needle-in-a-Haystack problems in both less time and fewer experiments by addressing the common convergence and pigeonholing issues. ZoMBI actively extracts knowledge from the previously best performing evaluated experiments to iteratively zoom in the sampling search bounds towards the global optimum "needle" and then prunes the memory of low-performing historical experiments to accelerate compute times by reducing the algorithm time complexity from $O(n^3)$ to $O(1)$, as the number of experiments sampled increases. Additionally, ZoMBI implements two custom acquisition functions that use active learning to further guide the sampling of new experiments towards the global optimum. We validate the algorithm's performance on two real-world 5-dimensional Needle-in-a-Haystack material property optimization datasets: discovery of auxetic Poisson's ratio materials and discovery of high thermoelectric figure of merit materials. The ZoMBI algorithm demonstrates compute time speed-ups of 400x compared to traditional Bayesian optimization as well as efficiently discovering materials in under 100 experiments that are up to 3x more highly optimized than those discovered by current state-of-the-art algorithms.

***K*eywords** rare materials discovery · efficient algorithms · adaptive acquisition functions · trust regions · active learning · extremely imbalance data · auxetic materials · thermoelectric materials

1 Introduction

Current optimization algorithms achieve good results on low-dimensional problems that are smooth and have wide basins of attraction. Examples of smooth manifolds with wide basins of attraction within material science include process- and recipe-optimization problems such as tuning perovskite manufacturing variables to achieve higher efficiency [1], optimizing microfluidics flow parameters to achieve ideal droplet formation [2], optimizing silver nanoparticle recipes for optical properties [3], and tuning perovskite compositions with physics-based constraints to maximize stability [4]. Optimization techniques like Bayesian optimization (BO) are well-suited to model these simple manifolds using a Gaussian Process (GP) surrogate [5, 6, 7, 8, 9]. However, the performance of this BO with a GP breaks down as the manifold complexity increases. Material property optimization problems that have high technological significance, such as discovering materials with rare properties or materials with a specific combination of properties, have search space manifolds that more closely resemble a *Needle-in-a-Haystack* [10], shown in Figure 1(b), rather than a smooth or convex space. This Needle-in-a-Haystack (NiaH) problem arises when only few optimum conditions exist within the entire dataset, resulting in an extreme imbalance. Interpolating the parameter space of an imbalanced dataset with an estimation function, such as a GP, results in smoothing over the optimum or over-predicting the properties of the materials found near the optimum [11, 12, 13]. Examples of NiaH materials optimization problems include discovering auxetic materials (*i.e.*, materials that have a highly negative Poisson's ratio, ν) for energy absorptive medical devices or protective armor [14, 15, 16] and discovering materials that have a combination of high electrical conductivity and low thermal conductivity (*i.e.*, a highly positive thermoelectric figure of merit, ZT) used from improving sensor technology to enable ubiquitous solid-state cooling [17, 18, 19]. Both of these rare material optimization problems are examples where an extreme data balance exists in the dataset because only a fraction of the total number of materials exhibit these rare properties [14, 20, 21, 22, 23]. This NiaH optimization challenge of extremely imbalanced datasets is largely applicable to many fields, not just materials science, including the fields of ecological resource management [24], fraud detection [25, 26], and rare diseases [27, 26].



(a) Process Optimization Manifold

(b) Materials Optimization Manifold

Figure 1: **Archetype Manifolds in Materials Science Optimization.** (a) Smooth and wide basin of attraction landscape that is common to process optimization problems. This 2D projected manifold is adapted from the 6D perovskite process optimization problem by Liu *et al.* [1]. (b) Rough and narrow basin of attraction landscape that is typical of material property optimization problems. This 2D projected manifold is obtained from the 5D negative Poisson's ratio optimization problem presented in this paper [20, 21].

Several challenges exist for the current landscape of computational tools that inhibit effective optimization of these complex NiaH problems. Firstly, the "needle" makes up only a small percentage of the total manifold search space, resulting in a weak correlation between the measured input parameters and the target property of interest, inhibiting discovery of the region containing the needle [28, 29, 11]. This challenge requires the development of an algorithm that can more quickly determine the plausible region of the manifold where the needle exists. The second challenge for algorithms, such as BO, to optimize NiaH manifolds is in the nature of the acquisition function to pigeonhole sampling into local minima because of the narrowness of the needle's basin of attraction [30, 31]. Standard BO acquisition functions, including expected improvement (EI) [32] and lower confidence bound (LCB) [7, 12], are static sampling techniques that only adjust sampling based on the output of the surrogate model, which enacts smoothing of the needle [11, 5, 6]. To overcome this challenge, active learning-based tuning of the acquisition function hyperparameters can be implemented to improve the sampling quality and avoid pigeonholing. Lastly, there exists a computing challenge for

NiaH problems where, typically, several thousands of samples must be observed to find an optimum when using an algorithm that is poorly-suited to tackle NiaH manifolds [10]. The compute time of BO using a GP surrogate scales with the complexity $O(n^3)$, where n is the number of experiments sampled, hence, the compute time of traditional BO blows up as more data is required to find the optimum [33, 34, 35, 5, 6, 36, 37]. To solve this computing challenge, an algorithm must be designed that both efficiently optimizes the space in as few experiments as possible and reduces the effect of compounding compute times over the length of the optimization procedure.

In recent literature, algorithms have been developed to address some of these challenges individually, but not all of them together. The first class of solutions bound the search space using a trust region approach to sample regions with higher probability of containing the optimum. Uber AI develop `TuRBO` [38] that compiles a set of independent model runs, using separate GP surrogate models to compute a new, smaller search region, narrowed in on the target optimum. Regis develops `TRIKE` [39] that utilizes maximization of the EI acquisition function to bound a trust region containing the global optimum. Diouane *et al.* develop `TREGO` [40], which interleaves sampling between global and local search regions, where the local search regions are defined by the single best historical experiment sampled. Although these methods offer solutions to one of the three challenges presented, each method has its downfalls when optimizing NiaH problems. For example, `TuRBO` requires the computation of several GP model runs, which increases compute time and also does not guarantee that the needle will be resolved due to interpolation effects; `TRIKE` is inflexible to the use of other acquisition functions as it locks the user in to only using EI, which may pigeonhole into local minima; `TREGO` uses only the best sampled experiment to define its search regions, which will yield inconsistent or sub-optimal results when the needle consists of a fractional region of the manifold and single point is unlikely to land in its basin of attraction. The second class of solutions to the challenges presented in this paper are designed to decrease the computing time required to run an optimization procedure. A common method for reducing the compute time of BO with a GP surrogate is to introduce a sparse GP [5, 41, 36]. A sparse GP uses a small subset of pseudo data, often denoted as m, to reduce the GP time complexity from $O(n^3)$ to $O(nm^2)$ [42]. However, the process of selecting a useful subset requires minimizing the Kullback-Leibler divergence between the sparse GP and true posterior GP, which is often a computationally intensive procedure of using variational inference [43]. In addition to sparse GPs, new algorithms have been developed in literature to improve the compute time of optimization in various ways. Van Stein *et al.* develop `MiP-EGO` [44], which parallelizes the function evaluations of efficient global optimization (EGO) to discover optima faster and in fewer experiments using derivative-free computation [45]. Joy *et al.* [46] use directional derivatives to accelerate hyperparameter tuning by 100x and achieve higher accuracy than the `FABOLAS` baseline by Klein *et al.* [47]. Zhang *et al.* develop `FLASH` [48] to achieve optimization speed-ups of 50% by using a linear parametric model to guide algorithm search within high-dimensional spaces. Snoek *et al.* [13] design a neural network-based parametric model that reduces the overall time complexity of BO to $O(n)$ compared to the complexity of $O(n^3)$ of standard BO with a GP surrogate model. These existing methods from literature within the class of solutions for accelerating compute time are generally introducing external models necessary to perform optimization, such as neural networks, variational inference, or parameteric models. While these external models do speed-up compute time, they often lack the predictive capabilities to capture the weak correlation between measured input parameters and the target property of interest in NiaH problems. We illustrate this mechanic later in the paper when comparing the optimization results on two materials science NiaH problems of a fast algorithm `MiP-EGO` with that of `TuRBO`, an algorithm better suited for discovering optima within narrow basins of attraction.

Although these methods from existing literature address some of the challenges in optimizing NiaH problems, none of them have been designed specifically to quickly and efficiently discover a needle-like optimum within a haystack of sub-optimal points, resulting in all of them falling short of full solution. Therefore, in this paper, we design an algorithm that addresses all three of the challenges faced when optimizing NiaH problems by (1) zooming in the manifold search bounds iteratively and independently for each dimension based on m number of best memory points to quickly converge to the plausible region containing the global optimum needle, (2) anti-pigeonholing into local minima by using actively learned acquisition function hyperparameters to tune the exploitation-to-exploration ratio, (3) relieving compute utilization by pruning the low-performing memory points not being used to zoom in the search bounds. The proposed algorithm, entitled [Zo]oming [M]emory-[B]ased [I]nitialization (`ZoMBI`), combines these three contributions into a method that efficiently optimizes NiaH problems quickly. In essence, this process of scanning broadly and then focusing in on points of interest based on memory was inspired by the way we humans solve similar problems, but stands in contrast to the way standard BO methods with static acquisition functions solve problems. We demonstrate the performance of this algorithm on two NiaH materials science datasets: (1) discovery of materials with negative Poisson's ratio and (2) discovery of materials with both high electrical conductivity and low thermal conductivity. The performance of the proposed `ZoMBI` algorithm is compared against standard BO with static acquisition functions and two state-of-the-art (SoTA) algorithms, one from each of the two classes of partial NiaH solutions: (1) `TuRBO` (bounded search space) and (2) `MiP-EGO` (faster compute).

2 Methodology

In this paper, we develop two major contributions: (1) the ZoMBI algorithm and (2) active learning acquisition functions. Through the combination of these two contributions, the optimum region of a NiaH manifold can be quickly discovered in fewer experiments without pigeonholing into local minima. Thus, the three challenges of optimizing NiaH problems are addressed: (1) the challenge of finding a hypervolume within the manifold that contains the needle-like optimum [28, 29, 11], (2) the challenge of avoiding pigeonholing into local minima [9, 1, 30, 31], (3) the challenge of the polynomially increasing compute times of BO using a GP surrogate [34, 35, 5, 6, 36, 37]. We demonstrate the implementation of ZoMBI on a 5D analytical Ackley function, a 5D dataset of materials with Poisson's ratios, and a 5D dataset of thermoelectric materials, all of which exhibit a NiaH problem. The experiment data points are denoted as X and for the materials datasets, each X-dimension corresponds to a continuous-valued material property, such as density or formation energy, indicated in Table 1. For each of the three problems, the objective is to find the target value, y, with either the lowest or highest value depending on if the problem is minimization or maximization. This optimum y-value resembles a needle for each problem because it located within a narrow and steep basin of attraction. Precisely, the needle optimum for each problem has a value of $y = 0$ for the Ackley function (minimization), $y = -1.7$ for Poisson's ratio dataset (minimization), and $y = 1.9$ for the thermoelectric merit dataset (maximization).

2.1 Zooming Memory-Based Initialization (ZoMBI) Algorithm

The ZoMBI algorithm has two key features: (1) iterative inward bounding of proceeding search spaces using the m number of best-performing memory points within the prior search space and (2) iterative pruning of low-performing historical search space memory. The newly computed search space bounds are unique for each dimension, such that optimum basin of attraction of complex, non-convex NiaH manifolds can be discovered. This algorithm leverages these two key features to guide the acquisition of new data towards more optimal regions while only fitting the surrogate within the suggested optimum region to resolve more detail of the space of interest, as shown in Figure 2. This process subsequently reduces the compute time significantly compared to the compute of a GP in a standard BO procedure, as shown in Figure 4.

We define m as the number of retained memory points during an activation of ZoMBI. The m memory points are saved to memory while all other data are erased from memory. These are the historical data points that achieve the m lowest (for minimization) target values, y, and they are used to zoom in the search bounds. Using these memory points, the multi-dimensional upper and lower bounds of the zoomed search space are computed for each dimension, d. Let $\mathbf{X} := \{X_1, X_2, \ldots, X_n\}$ be a set of data points, where $X_j \in \mathbb{R}^d$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the objective function. We first assume that the points in $\mathbf{X}$ are in general position so that $f(\mathbf{X})$ contains unique elements. Then, for each $m \leq n$ define $\mathbf{X}^{(m)} = \{X_{\pi(1)}, \ldots, X_{\pi(m)}\}$ where π is a permutation on $\{1, \ldots, n\}$ so that $\{f(X_{\pi(j)})\}$ is in ascending order. If $f(\mathbf{X})$ contains repeated elements, we may first remove the points with repeated f values and apply the definition above. Then, for each d, the bounds are defined as:

$$\begin{aligned} \mathcal{B}_d^l &= \min_{X \in \mathbf{X}^{(m)}} \{[X]_d\} \\ \mathcal{B}_d^u &= \max_{X \in \mathbf{X}^{(m)}} \{[X]_d\}, \end{aligned} \tag{1}$$

where $\mathcal{B}_d^l$ and $\mathcal{B}_d^u$ computed lower and lower bounds for each dimension, d, respectively. The bounds $[\mathcal{B}_d^l, \mathcal{B}_d^u]$ constrain the proceeding acquisition of new data as well as the computation of a GP, such that sampling cannot occur outsides of the bounded region. This constraining process operates independently for each dimension, such that each dimension has a unique lower and upper bound. To initialize the algorithm with data from the constrained space, i data points are sampled from the bounded region using Latin Hypercube Sampling (LHS). LHS splits a d-dimensional space into $i * d$ equally spaced strata, where i is the number of points to sample uniformly over d dimensions with low variability, unlike random sampling that has high sampling variability [49]. A GP surrogate model is retrained on these i LHS points sampled from the constrained space and then for every proceeding experiment sampled from the space, denoted as a forward experiment, the surrogate model is retrained. Thus, the GP is only being trained on information within the constrained region and as the constrained region iteratively zooms inward and decreases in hypervolume, so does the region computed by the GP. This process allows for more information to be resolve within regions plausibly containing the global optimum basin of attraction. Up to ϕ forward experiments are sampled in serial, where $\{X_i\} \cup \{X_\phi\} \subseteq \{X_n\}$. These forward experiments are sampled by maximizing an acquisition value, $a \in [0, 1]$, computed by a user-selected acquisition function from one of the four functions EI, EI Abrupt, LCB, and LCB Adaptive, introduced in Section 2.2. Once $i + \phi$ number of experiments are sampled, the bounds are re-constrained using the m best performing experiments, i new experiments are sampled from the zoomed-in space using LHS, and then the memory is pruned. The process of collecting ϕ forward experiments is repeated. A complete constraining-resetting iteration is denoted as an activation, α. This iterative zooming and pruning process over several α significantly speeds up compute time, discussed further in Section 3.3. Implementation of `ZoMBI` is shown in Algorithm 1.

Algorithm 1: Zooming Memory-Based Initialization (`ZoMBI`)

```
Input  :  $\mathbf{X}$: Set of data points $\{X_1, X_2, \dots, X_n\}$, where $X_j \in \mathbb{R}^d$,
          $\mathbf{y}$: Set of target values $\{y_1, y_2, \dots, y_n\}$, where $y_j \in \mathbb{R}$,
          $\alpha$: Number of ZoMBI activations,
          $\phi$: Number of forward experiments per activation,
          $\boldsymbol{\gamma}$: Set of acquisition function hyperparameters $\{\beta, \xi, \epsilon, \eta\}$,
          $AF$: An acquisition function selected by the user
Output :  The next experimental condition $X_{n+1} \in \mathbb{R}^d$ and measured target value $y_{n+1} \in \mathbb{R}$
 1 for $\alpha$ activations do
 2     Compute bounds $\{\mathcal{B}_d^l, \mathcal{B}_d^u\} \leftarrow \{\min, \max\}_{X \in \mathbf{X}^{(m)}}\{[X]_d\}$
 3     Initialize with $i$ LHS data points $\{X_i\} := \{X_1, X_2, \dots, X_i\}$, where $X_j \in \mathbb{R}^d$, $[\mathcal{B}_d^l, \mathcal{B}_d^u]$
           and target values $\{y_i\} := \{y_1, y_2, \dots, y_i\}$, where $y_j \in \mathbb{R}$
 4     Overwrite memory $\mathbf{X} \leftarrow \{X_i\}$ and $\mathbf{y} \leftarrow \{y_i\}$
 5     for $\phi$ forward experiments do
 6         Let $n = i + \phi$
 7         Retrain surrogate model $f(\mathbf{X})$ using target values $\mathbf{y}$
 8         Extract set of surrogate means $\{\mu\}$ and variances $\{\sigma\}$
 9         Compute set of acquisition values $\{a\} \leftarrow AF(\{\mu\}, \{\sigma\}; \boldsymbol{\gamma})$
10         Find the best new experimental condition $X_{n+1} \leftarrow \arg\max\big[\{a\}\big]$
11         Measure target value of new experimental condition $y_{n+1} \leftarrow f(X_{n+1})$
12         Append outputs to sets $\mathbf{X}$.append$(X_{n+1})$ and $\mathbf{y}$.append$(y_{n+1})$
13     end
14 end
```

2.2 Active Learning Acquisition Functions

Traditional BO acquisition functions, such as EI and LCB, use the computed means and variances from a surrogate model to compute an acquisition value; maximizing this acquisition value guides sampling of the manifold [7, 32, 12]. However, these traditional acquisition functions are static, such that they do not actively use any information about the performance of previously sampled experiments to guide sampling. Hence, we implement an active learning approach into the acquisition functions to develop two novel functions, EI Abrupt and LCB Adaptive, that dynamically adapt their sampling based on the quantity and quality of previously sampled experiments. In contrast to a static acquisition function, these dynamic acquisition functions are initialized with an initial set of hyperparameter values to guide their search but then tune these values as sampling progresses. The developed EI Abrupt and LCB Adaptive functions are used within the `ZoMBI` framework to further accelerate optimization and avoid pigeonholing, see line 9 of Algorithm 1.

EI Abrupt uses actively learned information about the quality of previously measured experiment target values, y, to change sampling policies. For example, if the value of y plateaus for three or more experiments in a row, EI Abrupt will abruptly switch from a greedy sampling policy to a more explorative sampling policy. Specifically, this information feedback received by the function determines if the current round of sampling should exploit the surrogate mean values, $\mu(X)$, or explore the surrogate variances, $\sigma(X)$. EI Abrupt computes an acquisition value, $a \in [0, 1]$, for a given X, wherein the X with the highest a is selected by the acquisition function as the next suggested experiment to measure. EI Abrupt is implement for a minimization problems as:

$$
\begin{aligned}
a_{\text{EI Abrupt}}(X, y; \beta, \xi, \eta) &= \begin{cases} \big(\mu(X) - y^* - \xi\big)\,\Phi(Z) + \sigma(X)\psi(Z), & \text{if } |\nabla\{y_{n-3\dots n}\}| \leq \eta \\ \mu(X) - \beta\sigma(X), & \text{otherwise} \end{cases} \\
Z &= \frac{\mu(X) - y^* - \xi}{\sigma(X)},
\end{aligned}
\tag{2}
$$

where y^* is the lowest measured target value thus far (*i.e.*, the running minimum), $\Phi(\cdot)$ is the cumulative density function of the normal distribution, $\psi(\cdot)$ is the probability density function of the normal distribution, and $|\nabla\{y_{n-3\dots n}\}|$

is the absolute value of the gradient of the set of target values of the last three sampled experiments; a gradient of 0 indicates a plateau. Moreover, $\beta = 0.1$, $\xi = 0.1$, and $\eta = 0$ are hand-tuned initialization hyperparameters used for the rest of the paper for EI Abrupt. These hyperparameters were selected based on *a priori* domain knowledge of EI Abrupt performance on a variety of different problems. The most important hyperparameter for efficient sampling is β, whose ideal value is non-obvious, but it is found that $\beta = 0.1$ allows EI Abrupt to switch into an explorative sampling policy while still having a strong weight on the surrogate means, implying that exploration does not veer far.

LCB Adaptive uses actively learned information about the quantity of previously sampled experiments, n, to tune its hyperparameter. For example, as the n increases, LCB Adaptive decays its β hyperparameter value to become less explorative and more exploitative. Specifically, this information feedback received by the function determines the contribution of both $\mu(X)$ and $\sigma(X)$ to the acquisition value, a. Similar to EI Abrupt, LCB Adaptive computes an acquisition value, $a \in [0, 1]$, for a given X, wherein the X with the highest a is selected by the acquisition function as the next suggested experiment to measure. LCB Adaptive is implemented for a minimization problem as:

$$a_{\text{LCB Adaptive}}(X, n; \beta, \epsilon) = \mu(X) - \epsilon^n \beta \sigma(X), \tag{3}$$

where n is the number of experiments sampled, and $\beta = 3$ and $\epsilon = 0.9$ are hand-tuned initialization hyperparameters selected based on *a priori* domain knowledge of the function's performance on a variety of different problems. Having a large β and an ϵ close to 1 supports a gradual decay from very explorative to very exploitative, rather than a rapid decay. In the following section (Section 3.2), the dynamic EI Abrupt and LCB Adaptive are shown to both discover optima in fewer experiments and avoid pigeonholing into local minima better than their static counterparts by actively balancing the ratio of exploitation to exploration using learned information about the quality and quantity of previously sampled experiments.

3 Demonstration of ZoMBI Mechanics using an Ackley Function

3.1 Zooming Bounds

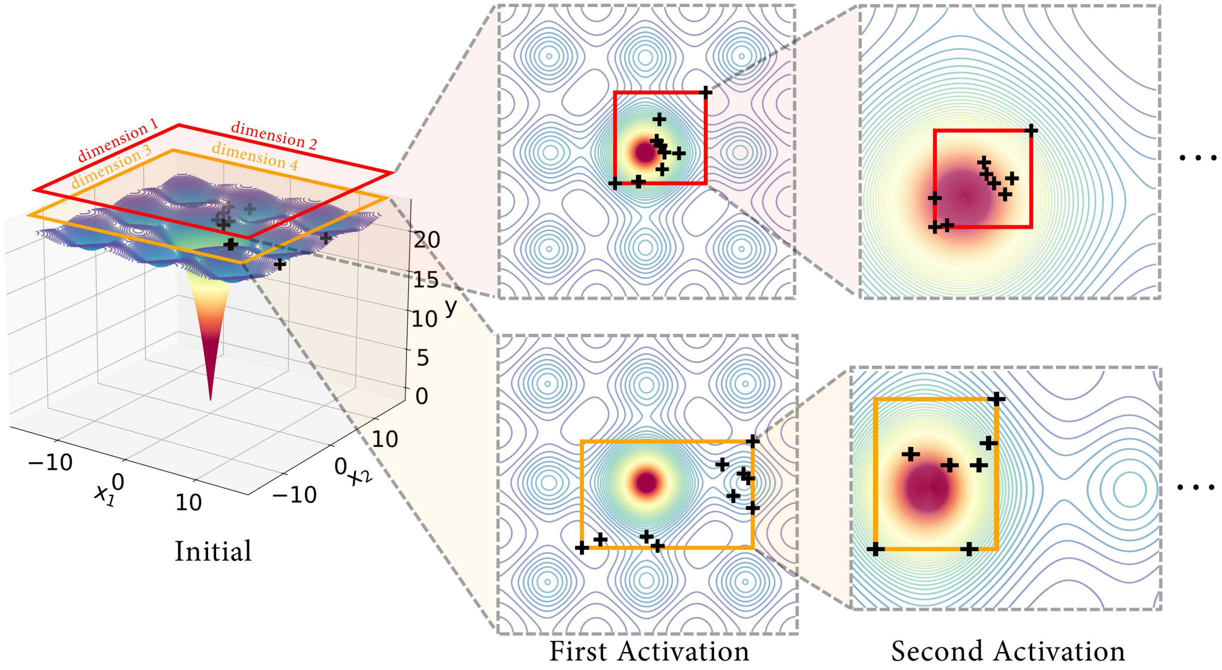
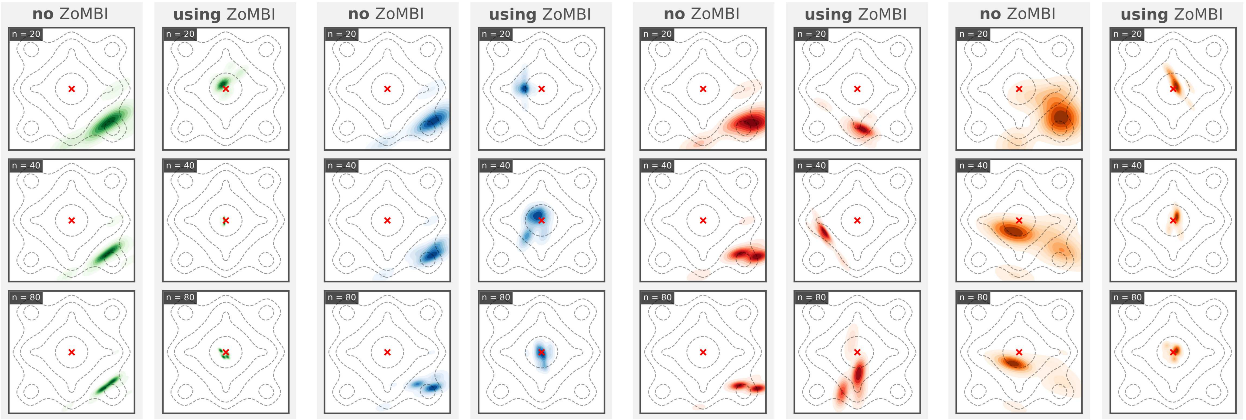


Figure 2: **Zooming Search Bounds.** For every activation of ZoMBI, the search bounds are zoomed inward based on the prior best-performing memory points. A 4D Ackley function manifold is projected in 2D. The bounding regions of each 2D slice are illustrate by the red and orange boxes. The ϕ number forward experiments sampled are illustrated as black markers. The global optimum is indicated by the red region of the heatmap.

Zooming in the search bounds on the manifold addresses challenge number one of optimizing NiaH problems, which is the challenge of finding the general hypervolume region that contains the needle-like optimum. Figure 2 illustrates how the ZoMBI algorithm iteratively zooms in the search bounds based on the number of activations, α. An Ackley function is used as a simulated example due to its non-convexity and needle-like global optimum [50, 51]. For each activation, m prior points that achieved the lowest target values, y, are retained in memory and used to zoom the search bounds in. This zooming occurs independently across each dimension and is based on the minimum and maximum values of the m memory points along each dimension, as shown in Equation 1. The red and orange rectangles illustrate the evolution of the bounds over space and time. Initially, sampling occurs across the entire manifold for ϕ forward experiments per activation, shown by the black markers. However, by using the best performing memory points to zoom in the search bounds, pigeonholing into local minima can also be avoided as the search bounds are pulled away from these trap minima and move closer towards the global minimum basin of attraction. The iterative zooming of ZoMBI does not guarantee convergence on the global optimum, but if a sufficient initialization set is obtained, convergence often gets close to the global optimum as shown across several examples in Figures 3, 7, and9.

3.2 Anti-pigeonholing



(a) LCB (b) LCB Adaptive (c) EI (d) EI Abrupt

Figure 3: **Acquisition Function Sampling Density.** The colored heatmaps indicate the regions of a 2D slice from a 5D Ackley function where sampling density is high for each respective acquisition function: (a) LCB, (b) LCB Adaptive, (c) EI, and (d) EI Abrupt. The contour lines indicate the manifold topology with local minima as the circular and pointed regions of the contours. The red "x" indicates the global minimum. For each acquisition function, the left panel shows the sampling density after $n = \{20, 40, 80\}$ evaluated experiments without the use of ZoMBI while the right panel shows the sampling density after $n = \{20, 40, 80\}$ evaluated experiments with the use of ZoMBI.

Pigeonholing into the local minima of a function occurs when an optimization algorithm has insufficient learned knowledge of the manifold topology to continue exploring potentially profitable regions or when the algorithm's hyperparameters are improperly tuned, leading to overly exploitative tendencies [1, 9]. The ZoMBI algorithm's anti-pigeonholing capabilities are two-fold: (1) the zooming search bounds help the acquisition function to quickly stop sampling local minima once a better performing data point is found and (2) actively learned acquisition function hyperparameters use knowledge about the domain to help exit a local minimum. Figure 3 demonstrates the anti-pigeonholing capabilities of ZoMBI on optimizing a 5D Ackly function with both static and dynamic acquisition functions, compared to that of traditional BO. The needle-like global minimum is indicated by the red "x" and the local minima are indicated by the circular and pointed regions of the contour lines. The sampling density of each acquisition function is illustrated by the heatmap, where the darker colors indicate higher sampling density regions. The goal is to get high sampling density near the red "x". It is shown that without ZoMBI being activated, the LCB, LCB Adaptive, and EI acquisition functions all end up pigeonholing into local minima. However, EI Abrupt initially pigeonholes into a local minima but then switches from an exploitative to an explorative mode to jump out of the local minimum and converge closer to the global. Conversely, when running the optimization procedure with ZoMBI active, all of the acquisition functions except the most exploitative, EI, converge onto the global minimum fast. LCB Adaptive and EI are shown to initially start sampling towards a local minima, but as ZoMBI is iteratively activated, the search bounds zoom in closer to the global minimum. Thus, with the combination of active learning dynamic acquisition functions and zooming search bounds, pigeonholing into sub-optimal local minima can be more readily avoided while optimizing NiaH problems, although avoidance is not guaranteed, as shown by the sampling density of EI.

3.3 Memory Pruning

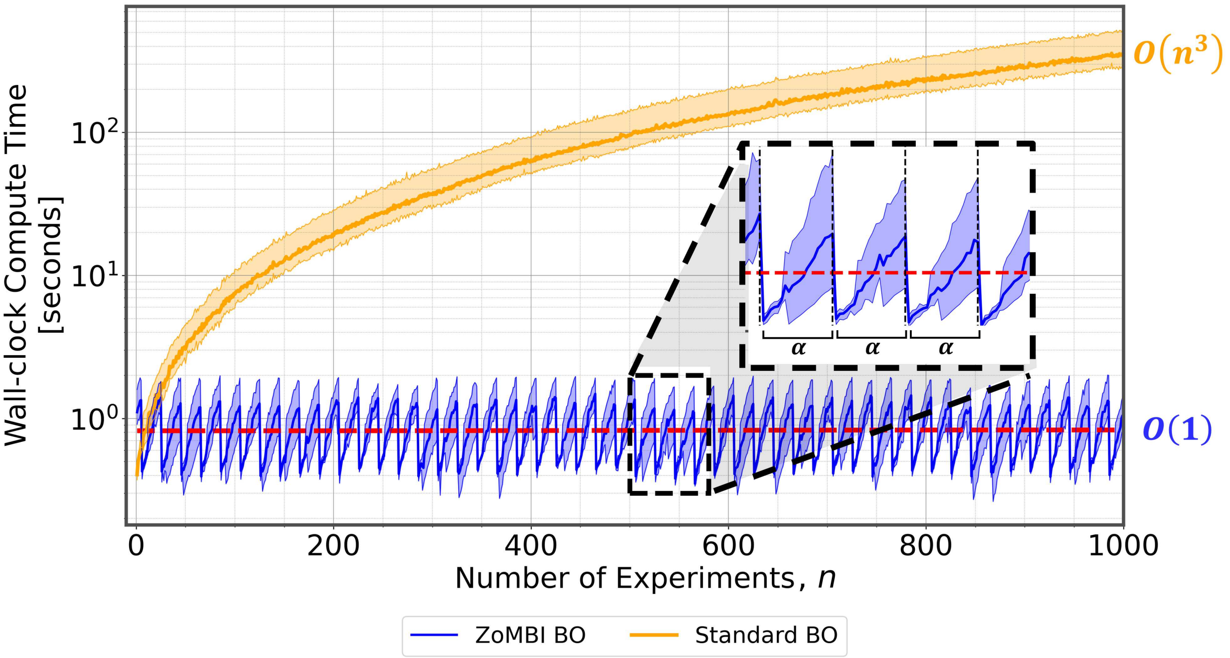


Figure 4: **Wall-clock Compute Time.** The compute time per experiment is illustrate for traditional BO with a GP surrogate (orange) and for `ZoMBI` with a GP surrogate (blue) with the y-axis in log-scale. Five independent trials of each method were run to optimize a 5D Ackley function with a narrow basin of attraction using an NVIDIA Tesla Volta V100 GPU [52]. The averages of the trials are shown as solid orange and blue lines while the shaded regions indicate the maximum and minimum compute times bounds. The red dashed line indicates the trend of the `ZoMBI` compute times. The measured compute time includes the time to compute the GP surrogate model and the time to acquire an experiment from the surrogate.

As more experiments are amassed and committed to memory to run traditional BO by computing a GP and an acquisition value, the compute time increases polynomially, following the $O(n^3)$ time complexity of GP matrix multiplication [33, 5, 6, 53, 36, 37]. This complexity is unfavorable as it leads to compounding compute times as more experiments are run. Therefore, we implement a memory pruning feature into the `ZoMBI` algorithm that iteratively selects which prior data points to keep and which ones to prune from the memory during each activation, α. Via memory pruning, the number of experiments used to train the GP surrogate varies between $[i, i + \phi]$ for every α, rather than being proportional to n. This is computationally favorable because $\{X_i\} \cup \{X_\phi\} \subseteq \{X_n\}$. Thus, for a single α, the time complexity is $O((i + \phi)^3)$. However, since ϕ resets back to zero after each α, a non-increasing sawtooth pattern in compute time is exhibited, hence, as $\alpha, n \to \infty$, the complexity approaches $O(1)$. Figure 4 illustrates that the sawtooth compute time pattern maps to the resetting interval of ϕ, which trends towards a constant, non-increasing value over many α and n. After collecting 1000 experiments, the compute time of traditional BO trend towards > 400 seconds, whereas after 1000 experiments, the compute time `ZoMBI` trends towards a constant 1 second. Therefore, the memory pruning feature of `ZoMBI` accelerates the optimization compute time by over 400x at $n = 1000$ and achieves further relative acceleration as n increases. The combination of the three foundational features of `ZoMBI`, (1) zooming bounds, (2) anti-pigeonholing, and (3) memory pruning, drives fast optimization of NiaH problems without sacrificing the ability to converge on the global optimum.

4 Experiments

4.1 Varying Basin of Attraction Width

The `ZoMBI` algorithm is designed specifically to tackle NiaH problems where the basin of attraction containing a global minimum is narrow [30, 31]. In our first experiment we explore the question, *how does basin of attraction width affect*

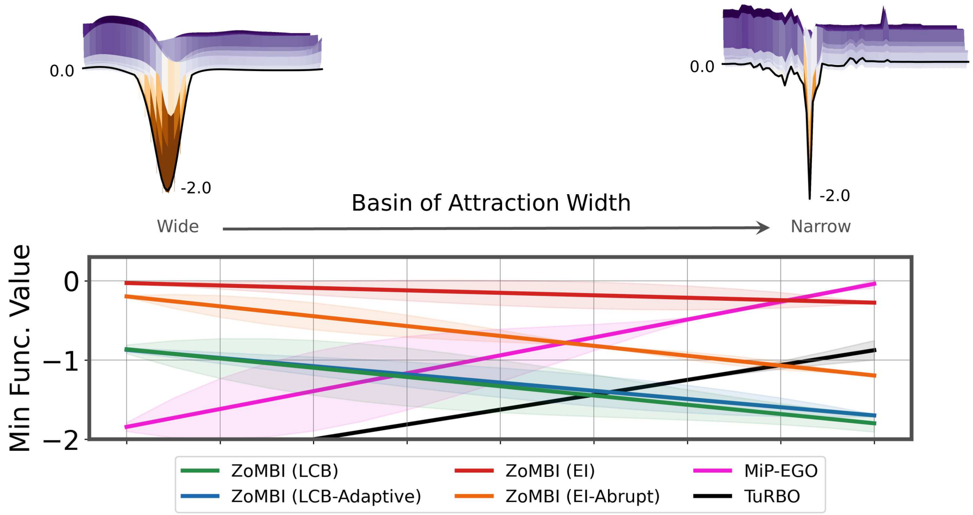


Figure 5: **Synthetically Varying Basin of Attraction Width.** The colored lines indicate the median trend lines of 12 independent runs for the final experimental evaluation of each algorithm after 100 sampled experiments for a given basin width. The spread of the data from the trend line is indicated by the shaded regions. The datasets with different width basins are obtained by synthetically convolving the original 5D Poisson's ratio dataset topology with Gaussian noise of increasing magnitudes. The objective is to find the minimum function value, $f_{\min} = -2$. In wide basin datasets, conventional SoTA BO methods, TuRBO and MiP-EGO, perform the best. However, for NiaH problems with narrow basins, the ZoMBI implementations perform better.

the ability of ZoMBI to find the global minimum? We present a synthetic example of a 5D noisy and non-convex dataset with a global minimum of $f_{\min} = -2$ whereby applying a Gaussian smoothing function to the dataset, the basin of attraction width is synthetically widened. The original, unmodified 5D manifold is based on the Poisson's ratio materials optimization dataset explored later in this paper [20, 21]. Figure 5 illustrates the trend lines of optimization performance of ZoMBI with each of its possible acquisition functions, relative to two SoTA methods from literature, TuRBO [38] and MiP-EGO [44]. The trend lines show that for wider basin of attraction problems, TuRBO and MiP-EGO significantly outperform all implementations of ZoMBI with different acquisition functions. However, as the basin of attraction narrows, the problem becomes a NiaH. The trend in performance of ZoMBI improves as the basin width decreases, the algorithm becomes more capable of finding the global minimum. Conversely, TuRBO and MiP-EGO become less capable of discovering the global minimum as the manifold transitions into a NiaH problem. As the Gaussian smoothing is lessened, all global and local minima basins narrow, resulting in less of the manifold space containing optimal regions, which accelerates the inward zooming of the search bounds using ZoMBI. Additionally, the more explorative acquisition functions are shown to perform better than the exploitative functions, such as EI, because pigeonholing is more readily avoided, as shown previously in Figure 3. This synthetic example demonstrates that the ZoMBI method has potential to perform better than SoTA to optimize NiaH problems. In the following sections, two real-world materials science NiaH datasets are optimized using ZoMBI and the performance is compared to SoTA.

4.2 Material Property Optimization: Negative Poisson's Ratio

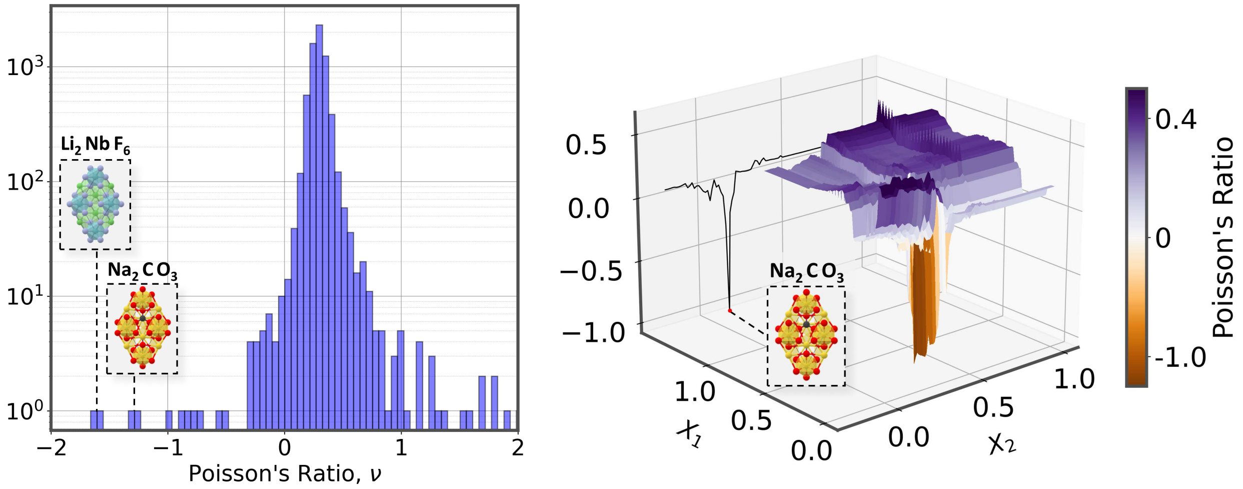
We demonstrate the ability of the ZoMBI algorithm to optimize Needle-in-a-Haystack problems on two real-world datasets. The first dataset consists of 146k materials and the objective is to find the material with the minimum negative Poisson's ratio, ν. The second dataset consists of 1k materials and the objective is to find the material with the maximum thermoelectric merit, ZT, *i.e.*, a material with high electrical conductivity and low thermal conductivity. Both of these datasets are 5-dimensional and are obtained from the open-access Materials Project database [20]. Table 1 describes the five continuous training variables and the target variable to be optimized. These ν- and ZT-datasets provide examples of the ZoMBI algorithm's ability to optimize relevant real-world NiaH problems.

The ν dataset exhibits a Needle-in-a-Haystack problem due to very few materials having negative ν values [14, 20, 21, 15]. A positive $\nu > 0$, describes a material that expands when a compressive load is applied to the orthogonal direction [54, 55]. Conversely, a negative $\nu < 0$ describes a material that contracts rather than expands when compressed in the orthogonal direction, denoted as an auxetic material [14, 23] – a rare phenomenon that occurs in only 0.82% of materials within the Materials Project database [20, 21]. Auxetic materials with highly negative Poisson's ratios have energy

Table 1: Description of variables from the two real-world Needle-in-a-Haystack materials science datasets [20].

Training Variable	**Units**	**Description**
Density	g/cm^3	Density of the entire molecule.
Formation Energy	eV/atom	Normalized change of energy to form target phase.
Energy Above Hull	eV/atom	Normalized energy to decompose into stable phase.
Fermi Energy	eV	Highest energy level at absolute zero.
Band Gap	eV	Valence to conduction band electron excitation energy

Target Variable	**Units**	**Description**
Poisson's Ratio, ν	Unitless	Mechanical deformation perpendicular to the loading direction.
Thermoelectric Merit, ZT	Unitless	Electrical and thermal potential of a material to produce current.



(a) Raw Dataset Histogram

(b) Computed Random Forest Manifold

Figure 6: **Histogram and Random Forest of Poisson's Ratio Dataset.** (a) The Poisson's ratio histogram of all 146k materials in the Materials Project Dataset [20, 21] with y-axis in log-scale. The two needles are called out, indicating the locations of minimum Poisson's ratio values $\nu_{\text{min}} = \{-1.2, -1.7\}$. (b) The noisy, non-convex manifold topology generated by an RF regression of 500 trees on the Poisson's ratio dataset. A projected 2D slice of the 5D space is illustrate with z-axis and colorbar indicating the Poisson's ratio. The slice of space shown indicates the narrow basin of attraction region containing the $\nu_{\text{min}} = -1.2$ needle.

absorptive properties, which are ideal materials for wearable medical devices and protective armor that must absorb the energy of large impacts to keep bones from shifting or to inhibit the penetration of the protective layer [15, 16]. Thus, for this NiaH problem, the objective is to discover the material with the lowest ν value. Figure 6 illustrates the spread of ν values within the raw dataset as a histrogram as well as a manifold generated by a Random Forest (RF) regression on the raw dataset using 500 trees. The search space generated by the RF is noisy and non-convex with narrow basins of attraction containing each optimum, resulting in a challenging NiaH optimization problem. The ground truth "needle" materials with the lowest ν values are Li_2NbF_6 with $\nu \approx -1.7$ and Na_2CO_3 with $\nu \approx -1.2$.

Figure 7 illustrates the performance of `ZoMBI` in discovering the lowest ν-value material, compared to the SoTA `TuRBO` and `MiP-EGO` algorithms. The `ZoMBI` algorithm is run with each of the four acquisition functions: LCB, LCB Adaptive, EI, and EI Abrupt. In under 100 evaluated experiments, LCB and LCB Adaptive discover one of the needles within the dataset (Li_2NbF_6) and, similarly, EI Abrupt discovers the other needle (Na_2CO_3). The distribution of ν values for the final experiment across all ensemble runs is illustrated for each method to highlight the sampling density and general rate of success. LCB Adaptive and EI Abrupt are the first two implementations of `ZoMBI` to discover a $\nu < 0$ material because of their ability to actively tune their sampling hyperparameters. After 30 experiments, the `ZoMBI` search bounds have zoomed inward enough for the explorative LCB acquisition function to discover a region of the manifold containing highly negative ν material, eventually leading to the global minimum needle. These three implementations of `ZoMBI`: LCB, LCB Adaptive, and EI Abrupt, have a steep drop in the discovered ν value, allowing these methods to discover an optimum fast, in fewer experiments than both SoTA methods. Overall, LCB and LCB Adaptive implementations of `ZoMBI` discover the most optimum minimum $\nu \approx -1.7$, while the SoTA algorithms `TuRBO` and `MiP-EGO` only discover

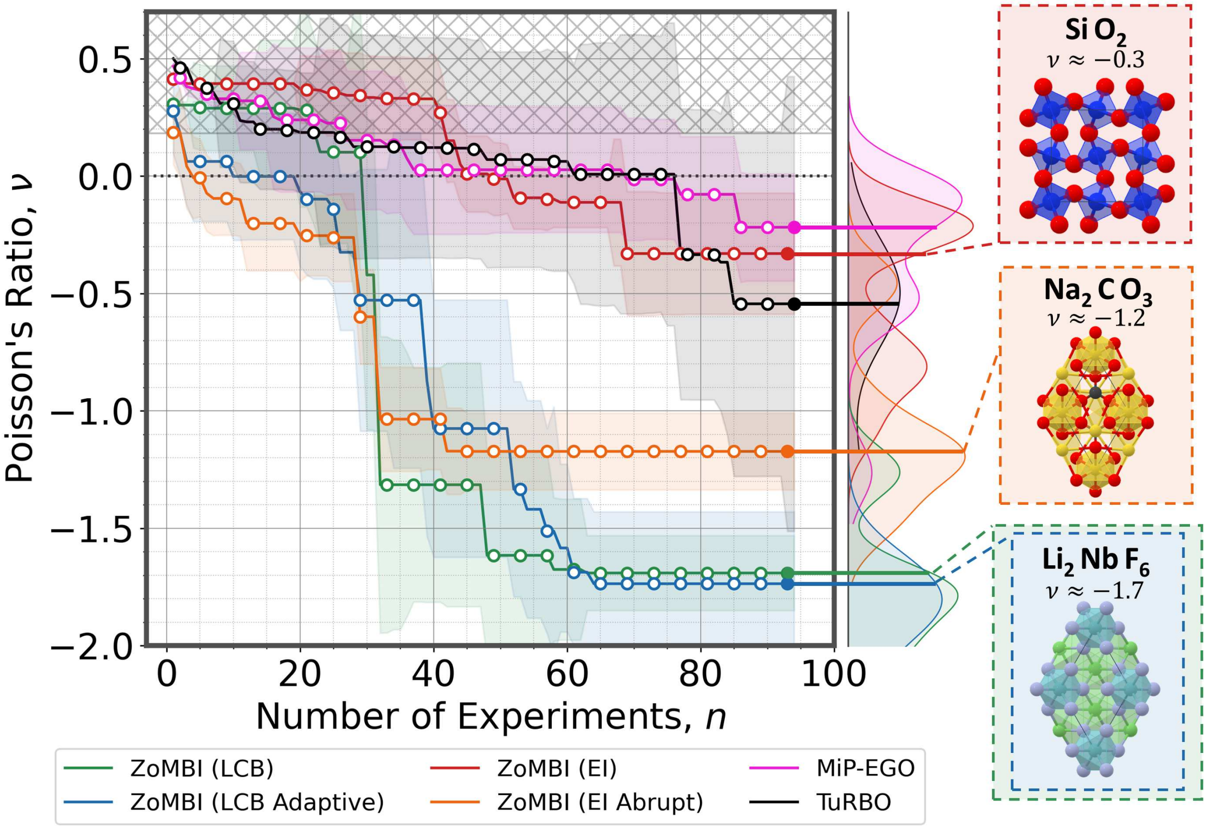


Figure 7: **Discovery of Rare Negative Poisson's Ratio Materials.** The optimization objective is to find the material with the minimum Poisson's ratio, $\nu_{\min}$, in 100 experiments from the dataset presented in Figure 6. The green, blue, red, and orange lines indicate the median best running evaluated sample of ZoMBI using the LCB, LCB Adaptive, EI, and EI Abrupt acquisition functions, respectively. The pink and black lines indicate the median best running evaluated sample of the SoTA methods, MiP-EGO and TuRBO, respectively. The median for each method is taken over the best 12 independent model runs. The shaded regions indicate the variance between model runs. The crosshatched region indicates the space discovered by standard BO methods, without the use of ZoMBI. The dashed black line indicates the $\nu = 0$ inflection point. The distribution of the final sampled ν value for each method at the 100th experiment is shown as a kernel density estimation with a 0.5 smoothing factor. The materials formulae and unit cells that have the closest evaluated ν value discovered by each ZoMBI method at the end of the 100 experiments are illustrated.

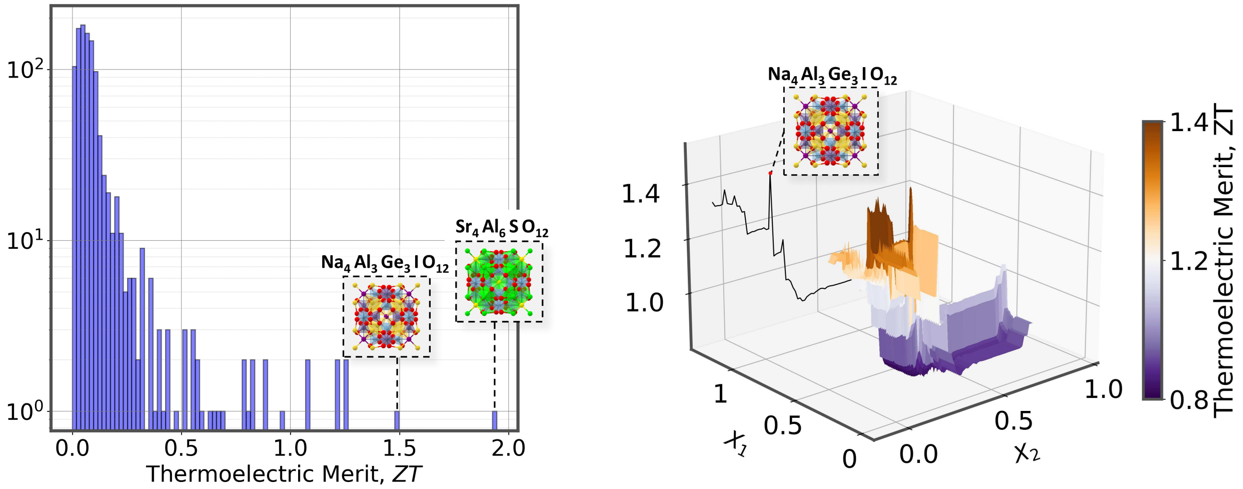
$\nu \approx -0.55$ and $\nu \approx -0.20$, respectively. These results demonstrate that with proper selection an acquisition function, ZoMBI achieves better performance and a higher success rate than SoTA on optimizing this real-world materials science NiaH problem.

4.3 Material Property Optimization: Thermoelectric Merit

The ZT dataset exhibits a Needle-in-a-Haystack problem, similar to the ν dataset because very few materials have high ZT values [20, 10]. However, rather than ZT being a directly measurable mechanical material property like Poisson's ratio, ZT must be computed using a combination of several thermal and electrical material properties [57]:

$$ZT = \frac{S^2\sigma}{\kappa}T, \tag{4}$$

where S is the Seebeck coefficient, σ is electrical conductivity, T is the average temperature, and κ is thermal conductivity. The ZT is computed for each material in the Materials Project database using BoltzTraP [56]. Of the initial 146k materials, 1k of them have the required thermal and electrical properties to compute a ZT value. ZT is a common figure of merit used to describe the thermal-to-electrical or electrical-to-thermal conversion efficiency of



(a) Raw Dataset Histogram

(b) Computed Random Forest Manifold

Figure 8: **Histogram and Random Forest of Thermoelectric Merit Dataset.** (a) The Thermoelectric Merit, ZT, histogram of all 1k materials in the dataset computed by BoltzTraP [56, 20] with y-axis in log-scale. The two needles are called out, indicating the locations of maximum ZT values $ZT_{\text{max}} = \{1.4, 1.9\}$. (b) The noisy, non-convex manifold topology generated by an RF regression of 500 trees on the ZT dataset. A projected 2D slice of the 5D space is illustrate with z-axis and colorbar indicating the ZT value. The slice of space shown indicates the narrow basin of attraction region containing the $ZT_{\text{max}} = 1.4$ needle.

thermoelectric materials [58, 59, 60, 61]. A higher ZT indicates that the material is better able to convert a thermal gradient into an electrical current [57]. Materials with large ZT values have a range of applications from usage as solid-state cooling devices to being used as sensors that when heated up, will produce an electrical signal [17, 18, 19]. For this NiaH problem, the objective is to discover the material with the highest ZT value. Figure 8 illustrates the spread of ZT values within the raw dataset as a hisitogram, as well as a manifold generated by an RF regression on the raw dataset using 500 trees. Similar to the ν manifold, the ZT manifold is noisy and non-convex with narrow basins of attraction [20, 56]. The ground truth "needle" materials with the highest ZT values are $Na_4Al_3Ge_3IO_{12}$ with $ZT \approx 1.4$ and $Sr_4Al_6SO_{12}$ with $ZT \approx 1.9$.

Figure 9 illustrates the performance of `ZoMBI` in discovering the highest ZT-value material, compared to the SoTA `TuRBO` and `MiP-EGO` algorithms. Initially, we see `TuRBO` outperform all other algorithms, but then it is unable to accelerate its sampling towards the needle basins of attraction. Similarly, `MiP-EGO` gets trapped in a local minimum and is unable to escape. Conversely, after 50 evaluated experiments, `ZoMBI` LCB Adaptive and EI Abrupt supersede `TuRBO` and quickly discover high ZT materials, illustrating the advantage of active learning acquisition functions. Although the active learning acquisition functions prove to be more successful than the SoTA algorithms, none of the tested algorithms are able to discover the maximum global needle, $Sr_4Al_6SO_{12}$, only the second best needle, $Na_4Al_3Ge_3IO_{12}$. This result is likely due to the data imbalance being too extreme that far out on the tail of the ZT dataset, in turn, generating an RF manifold complexity too high, even for `ZoMBI`. Hence, indicating that there are limitations in the manifold complexity that `ZoMBI` can optimize, and further illustrating that convergence on the global optimum needle is not guaranteed using this method. However, for the ZT dataset, the LCB Adaptive implementation of `ZoMBI` discovers the second best needle, $Na_4Al_3Ge_3IO_{12}$ with $ZT \approx 1.4$, while the SoTA algorithms `TuRBO` and `MiP-EGO` only discover $ZT \approx 0.65$ and $ZT \approx 0.45$, respectively. Thus, LCB Adaptive demonstrates the highest performing optimization results across both of the real-world NiaH datasets, discovering the most optimal materials the fastest for both the ν and ZT datasets.

5 Summary & Conclusions

In this paper, we proposed the [Zo]oming [M]emory-[B]ased [I]nitialization (`ZoMBI`) algorithm that builds on the principles of Bayesian optimization to accelerate the optimization of Needle-in-a-Haystack problems by two-fold, firstly by requiring fewer experiments to achieve a better optimum than state-of-the-art, and secondly by pruning the memory of low-performing historical experiments to speed-up compute time. The `ZoMBI` algorithm exceeds state-of-the-art performance on optimizing Needle-in-a-Haystack datasets by (1) using the values of the m best performing previously

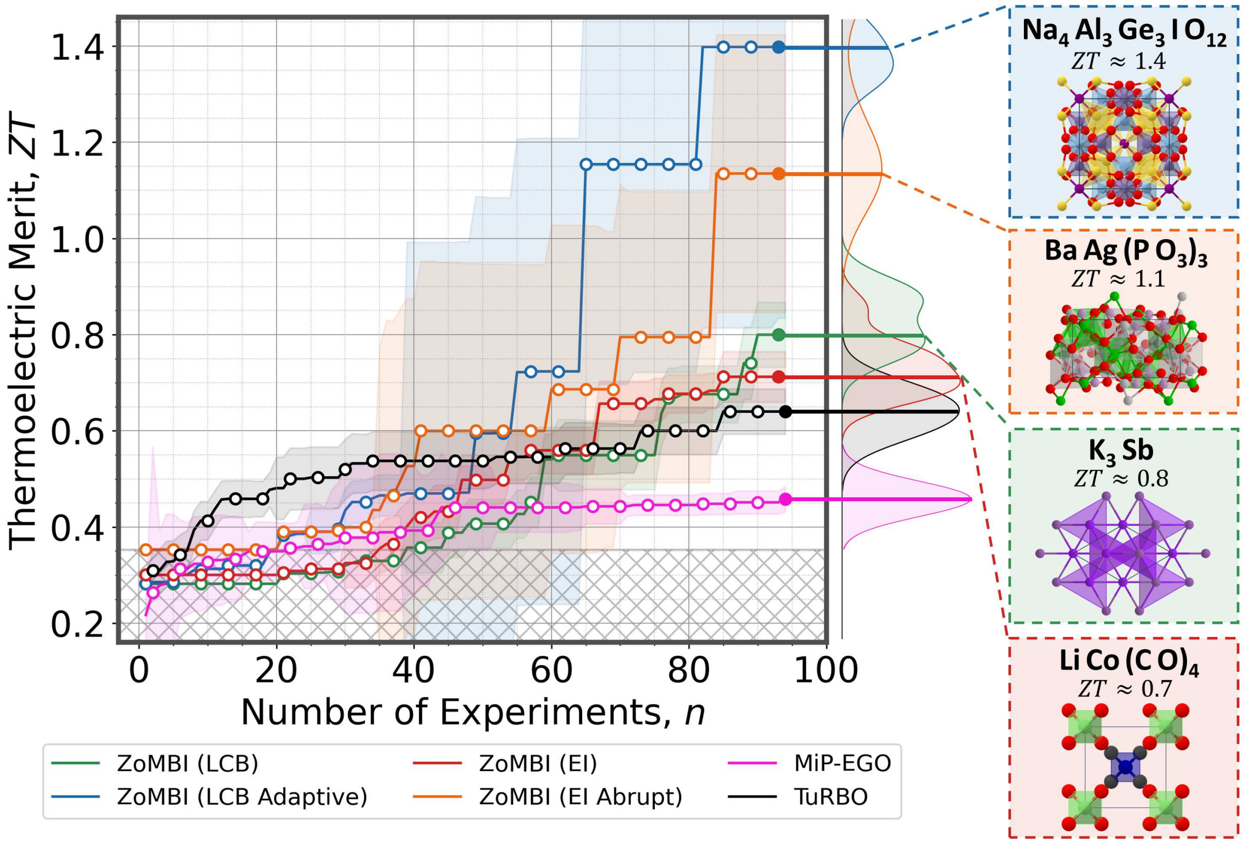


Figure 9: **Discovery of Rare Positive Thermoelectric Merit Materials.** The optimization objective is to find the material with the maximum thermoelectric merit, ZT_{max}, in 100 experiments from the dataset presented in Figure 8. The green, blue, red, and orange lines indicate the median best running evaluated sample of ZoMBI using the LCB, LCB Adaptive, EI, and EI Abrupt acquisition functions, respectively. The pink and black lines indicate the median best running evaluated sample of the SoTA methods, MiP-EGO and TuRBO, respectively. The median for each method is taken over the best 12 independent model runs. The shaded regions indicate the variance between model runs. The crosshatched region indicates the space discovered by standard BO methods, without the use of ZoMBI. The distribution of the final sampled ZT value for each method at the 100th experiment is shown as a kernel density estimation with a 0.5 smoothing factor. The materials formulae and unit cells that have the closest evaluated ZT value discovered by each ZoMBI method at the end of the 100 experiments are illustrated.

sampled memory points to iteratively zoom in the search bounds of the manifold uniquely on each dimension and (2) implementing two custom acquisition functions, LCB Adaptive and EI Abrupt, that actively learn information about the manifold during optimization to tune the sampling of new experimental conditions from a surrogate. The main contributions of this algorithm solve three fundamental challenges of optimizing non-convex Needle-in-a-Haystack problems: (1) the challenge of locating the hypervolume region of the manifold containing the narrow global optimum basin of attraction [28, 29, 11] is alleviated by introducing iterative search bounds based on learned knowledge of the manifold; (2) unwanted pigeonholing into local minima [30, 31, 5, 6] is avoided by both the zooming mechanics of ZoMBI as well as using the two acquisition functions developed in his paper, LCB Adaptive and EI Abrupt, that tune their hyperparameters through active learning; (3) the challenge of polynomially increasing compute times of BO using a GP surrogate [33, 34, 35, 5, 6, 36, 37] is addressed by actively pruning the retained memory of the algorithm after each activation, α, in turn, reducing the time complexity from $O(n^3)$ to $O(1)$ as $\alpha, n \to \infty$. By developing the ZoMBI algorithm to solve these challenges, it becomes possible to quickly and efficiently find optimal solutions to complex Needle-in-a-Haystack problems in fewer experiments.

Solving a Needle-in-a-Haystack problem that arises from extremely imbalanced data is a significant challenge that has important implications in science and engineering, especially within the field of materials science [10, 28]. In this paper, we use ZoMBI to discover the optimum materials in two real-world materials science Needle-in-a-Haystack datasets where only a small fraction of the entire search space consists of the target optimum conditions. In the first dataset, we discover a material with a highly negative Poisson's ratio, ν, [20, 21] and in the second dataset, we discover a material with a highly positive thermoelectric figure of merit, ZT [56, 20], both rare material properties. For the first dataset, the ZoMBI algorithm with the LCB and LCB Adaptive custom acquisition function both discover the material with the minimum $\nu \approx -1.7$, while the state-of-the-art algorithms TuRBO [38] and MiP-EGO [44] only discover $\nu \approx -0.55$ and $\nu \approx -0.20$, respectively. For the second dataset, the ZoMBI algorithm with the LCB Adaptive custom acquisition function discovers the material with the maximum $ZT \approx 1.4$, while the state-of-the-art algorithms TuRBO [38] and MiP-EGO [44] only discover $ZT \approx 0.65$ and $ZT \approx 0.45$, respectively. These results demonstrate that the ZoMBI algorithm is more well-suited to tackle Needle-in-a-Haystack optimization problems than current state-of-the-art methods, however, the selection of an acquisition function for ZoMBI is important because the EI acquisition function often performed worse than state-of-the-art. Thus, this manual selection of an appropriate acquisition function is a current limitation of the algorithm and demonstrates that convergence to a global optimum is not always guaranteed while using ZoMBI, but given the correct set of conditions, ZoMBI exceeds state-of-the-art. Moreover, we demonstrated that ZoMBI often does not outperform state-of-the-art for non-Needle-in-a-Haystack problems, such as those with wide basins of attraction. Hence, the use of the ZoMBI algorithm and selection of an appropriate acquisition function should be considered based on the nature of the optimization problem as it is not always guaranteed to find a more optimal solution than other methods.

Alas, the significance of developing the ZoMBI algorithm is to quickly and efficiently tackle difficult Needle-in-a-Haystack optimization problems in extremely imbalance datasets. In this paper, we showcased the ability of the developed algorithm to discover rare materials with highly-optimized properties in a short period of time and in few number of experiments. Discovering rare materials quickly and efficiently enables the widespread access to a new range of materials applications from engineering high-performance medical devices to ubiquitous solid-state cooling systems [15, 16, 17, 18, 10, 19]. However, the application space for ZoMBI to accelerate the efficient discovery of highly-optimized solutions extends past materials science and is generally applicable for many Needle-in-a-Haystack problems, including those found in ecological resource management [24], fraud detection [25, 26], and rare disease prediction [27, 26]. Ultimately, we aim for this contribution to support the elimination of the time and resource barriers previously inhibiting the throughput of optimizing complex and challenging Needle-in-a-Haystack problems across a broad range of application spaces.

Acknowledgements

Basita Das is thanked for help in naming the algorithm. Xiaonan Wang is thanked for initial discussions for this study. John Dagdelen, Hongbin Zhang, and Shyam Dwaraknath are thanked for discussion of and reference to different Needle-in-a-Haystack problems within materials science. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper.

Data Availability

Implementation of the ZoMBI algorithm, the experimental dataset analyzed during the current study, the simulated data and labeled data supporting the findings of this study, and the data comprising the figures in this paper are all available in the following GitHub repository: https://github.com/PV-Lab/ZoMBI.

Author Contributions

A.E.S., Z.R., and T.B. conceived of and designed the study. Q.L. and T.B. provided guidance on machine learning methods, benchmark functions, and datasets. A.E.S. and Z.R. wrote the code. A.E.S. performed the machine learning modeling and analysis. A.E.S. wrote the paper, while all co-authors reviewed the manuscript.

Conflicts of Interest

Although our laboratory has IP filed covering photovoltaic technologies and materials informatics broadly, we do not envision a direct COI with this study, the content of which is open sourced. Two of the authors (Z.R. and T.B.) own equity in Xinterra Pte Ltd, which applies machine learning to accelerate novel materials development.

References

[1] Zhe Liu, Nicholas Rolston, Austin C. Flick, Thomas W. Colburn, Zekun Ren, Reinhold H. Dauskardt, and Tonio Buonassisi. Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing. *Joule*, 6(4):834–849, 2022.

[2] Alexander E. Siemenn, Evyatar Shaulsky, Matthew Beveridge, Tonio Buonassisi, Sara M. Hashmi, and Iddo Drori. A Machine Learning and Computer Vision Approach to Rapidly Optimize Multiscale Droplet Generation. *ACS Applied Materials & Interfaces*, 14(3):4668–4679, 2022.

[3] Flore Mekki-Berrada, Zekun Ren, Tan Huang, Wai Kuan Wong, Fang Zheng, Jiaxun Xie, Isaac Parker Siyu Tian, Senthilnath Jayavelu, Zackaria Mahfoud, Daniil Bash, Kedar Hippalgaonkar, Saif Khan, Tonio Buonassisi, Qianxiao Li, and Xiaonan Wang. Two-step machine learning enables optimized nanoparticle synthesis. *npj Computational Materials 2021 7:1*, 7(1):1–10, 2021.

[4] Shijing Sun, Armi Tiihonen, Felipe Oviedo, Zhe Liu, Janak Thapa, Yicheng Zhao, Noor Titan P. Hartono, Anuj Goyal, Thomas Heumueller, Clio Batali, Alex Encinas, Jason J. Yoo, Ruipeng Li, Zekun Ren, I. Marius Peters, Christoph J. Brabec, Moungi G. Bawendi, Vladan Stevanovic, John Fisher, and Tonio Buonassisi. A data fusion approach to optimize compositional stability of halide perovskites. *Matter*, 4(4):1305–1322, 2021.

[5] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian Processes using Pseudo-inputs. In Y Weiss, B Schölkopf, and J Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

[6] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[7] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. 2010.

[8] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian Optimization Of Machine Learning Algorithms. pages 1–12, 2001.

[9] Qiaohao Liang, Aldair E. Gongora, Zekun Ren, Armi Tiihonen, Zhe Liu, Shijing Sun, James R. Deneault, Daniil Bash, Flore Mekki-Berrada, Saif A. Khan, Kedar Hippalgaonkar, Benji Maruyama, Keith A. Brown, John Fisher, and Tonio Buonassisi. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *npj Computational Materials 2021 7:1*, 7(1):1–10, 2021.

[10] Yoolhee Kim, Edward Kim, Erin Antono, Bryce Meredig, and Julia Ling. Machine-learned metrics for predicting the likelihood of success in materials discovery. *npj Computational Materials*, 6(131), 2020.

[11] Ioan Andricioaei and John E Straub. Finding the needle in the haystack: Algorithms for conformational optimization. *Computers in Physics*, 10:449, 1996.

[12] Matthias Seeger. Gaussian processes for machine learning. *International journal of neural systems*, 14(2):69–106, 2004.

[13] Jasper Snoek, Oren Ripped, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Md Mostofa Ali Patwary, Prabhat, and Ryan P. Adams. Scalable Bayesian optimization using deep neural networks. *32nd International Conference on Machine Learning, ICML 2015*, 3:2161–2170, 2015.

[14] John Dagdelen, Joseph Montoya, Maarten De Jong, and Kristin Persson. Computational prediction of new auxetic materials. *Nature Communications*, 8(1):1–8, 2017.

[15] Krishna Kumar Saxena, Raj Das, and Emilio P. Calius. Three Decades of Auxetics ResearchMaterials with Negative Poisson's Ratio: A Review. *Advanced Engineering Materials*, 18(11):1847–1870, 2016.

[16] Q Liu. Literature Review: Materials with Negative Poisson's Ratios and Potential Applications to Aerospace and Defense. Technical report, Australian Government Department of Defense, 2006.

[17] Wael A Salah and Mai Abuhelwa. Review of Thermoelectric Cooling Devices Recent Applications. *Journal of Engineering Science and Technology*, 15(1):455–476, 2020.

[18] Ran He, Gabi Schierning, and Kornelius Nielsch. Thermoelectric Devices: A Review of Devices, Architectures, and Contact Optimization. *Advanced Materials Technologies*, 3(4):1700256, 2018.

[19] Jun Mao, Gang Chen, and Zhifeng Ren. Thermoelectric cooling materials. *Nature Materials*, 20(4):454–461, 2020.

[20] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[21] Maarten De Jong, Wei Chen, Thomas Angsten, Anubhav Jain, Randy Notestine, Anthony Gamst, Marcel Sluiter, Chaitanya Krishna Ande, Sybrand Van Der Zwaag, Jose J. Plata, Cormac Toher, Stefano Curtarolo, Gerbrand Ceder, Kristin A. Persson, and Mark Asta. Charting the complete elastic properties of inorganic crystalline compounds. *Scientific Data*, 2(1):1–13, 2015.

[22] Amir Yeganeh-Haeri, Donald J. Weidner, and John B. Parise. Elasticity of α-Cristobalite: A Silicon Dioxide with a Negative Poisson's Ratio. *Science*, 257(5070):650–652, 1992.

[23] Rod Lakes and K. W. Wojciechowski. Negative compressibility, negative Poisson's ratio, and stability. *Physica Status Solidi (B) Basic Research*, 245(3):545–551, 2008.

[24] Lisa J Rew, Bruce D Maxwell, Frank L Dougher, and Richard Aspinall. Searching for a needle in a haystack: evaluating survey methods for non-indigenous plant species. *National Park Biological Invasions*, 8:523–539, 2006.

[25] Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, Jiahang Chen, W Wei, J Li, L Cao, Y Ou, and J Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4):449–475, 2012.

[26] Neil G Marchant and Benjamin I P Rubinstein. Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance. *KDD '21, August 14–18, 2021, Virtual Event, Singapore*, page 11, 2021.

[27] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11(1):1–13, 2011.

[28] Koby Crammer and Gal Chechik. A Needle in a Haystack: Local One-Class Optimization. *Proceedings of the 21st International Conference on Machine Learning, Banff, Canada*, 2004.

[29] Haixiang Liu, Yuanming Hu, Bo Zhu, Wojciech Matusik, and Eftychios Sifakis. Narrow-Band Topology Optimization on a Sparsely Populated Grid. *ACM Transactions on Graphics*, 37(6):1–14, 2018.

[30] Helena E. Nusse and James A. Yorke. Basins of Attraction. *Science*, 271(5254):1376–1380, 1996.

[31] George Datseris and Alexandre Wagemakers. Effortless estimation of basins of attraction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(2):023104, 2022.

[32] Philipp Hennig and Christian J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13:1809–1837, 2012.

[33] Belyaev Mikhail, Burnaev Evgeny, and Kapushev Yermek. Exact Inference for Gaussian Process Regression in case of Big Data with the Cartesian Product Structure. 2014.

[34] Cheng Li, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh, and Alistair Shilton. High Dimensional Bayesian Optimization Using Dropout. *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, 2017.

[35] Zi Wang, Chengtao Li, Stefanie Jegelka, and Pushmeet Kohli. Batched High-dimensional Bayesian Optimization via Structural Kernel Learning. *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR*, 70, 2017.

[36] Thang D Bui, Josiah Yan, and Richard E Turner. A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation. *Journal of Machine Learning Research*, 18:1–72, 2017.

[37] Gongjin Lan, Jakub M Tomczak, Diederik M Roijers, and A E Eiben. Time Efficiency in Optimization with a Bayesian-Evolutionary Algorithm. 2020.

[38] David Eriksson, Michael Pearce, Jacob R Gardner, Ryan Turner, and Matthias Poloczek. Scalable Global Optimization via Local Bayesian Optimization. 2020.

[39] Rommel G. Regis. Trust regions in Kriging-based optimization with expected improvement. *Engineering Optimization*, 48(6):1037–1059, 2015.

[40] Y Diouane, V Picheny, R Le Riche, A Scotto, and Di Perrotolo. TREGO: a Trust-Region Framework for Efficient Global Optimization. 2021.

[41] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[42] Felix Leibfried, Vincent Dutordoir, S T John, and Nicolas Durrande. A Tutorial on Sparse Gaussian Processes and Variational Inference. 2021.

[43] R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In D. Barber, T. Cemgil, and S. Chiappa, editors, *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press, 2011.

[44] Bas van Stein, Hao Wang, and Thomas Back. Automatic configuration of deep neural networks with parallel efficient global optimization. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019.

[45] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13:455–492, 1998.

[46] Tinu Theckel Joy, Santu Rana, Sunil Gupta, and Svetha Venkatesh. Fast hyperparameter tuning using Bayesian optimization with directional derivatives. *Knowledge-Based Systems*, 205:106247, 2020.

[47] Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets. 2017.

[48] Yuyu Zhang, Mohammad Taha Bahadori, Hang Su, and Jimeng Sun. FLASH: Fast Bayesian Optimization for Data Analytic Pipelines. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[49] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

[50] D H Ackley. A connectionist machine for genetic hillclimbing. 1987.

[51] E P Adorio. MVF - Multivariate Test Functions Library in C for Unconstrained Global Optimization , 2005.

[52] Albert Reuther, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Vijay Gadepally, Michael Houle, Matthew Hubbell, Michael Jones, Anna Klein, Lauren Milechin, Julia Mullen, Andrew Prout, Antonio Rosa, Charles Yee, and Peter Michaleas. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2018.

[53] Elon S Correa and Jonathan L Shapiro. Model Complexity vs. Performance in the Bayesian Optimization Algorithm. In T P Runarsson, H-G Beyer, E Burke, J J Merelo-Guervos, L D Whitley, and X Yao, editors, *Parallel Problem Solving from Nature*, pages 998–1007. Springer, 2006.

[54] Hoss Belyadi, Ebrahim Fathi, and Fatemeh Belyadi. Rock mechanical properties and in situ stresses. *Hydraulic Fracturing in Unconventional Reservoirs*, pages 215–231, 2019.

[55] Yuriy M. Poplavko. Mechanical properties of solids. *Electronic Materials*, pages 71–93, 2019.

[56] Georg K.H. Madsen and David J. Singh. BoltzTraP. A code for calculating band-structure dependent quantities. *Computer Physics Communications*, 175(1):67–71, 2006.

[57] B. Hinterleitner, I. Knapp, M. Poneder, Yongpeng Shi, H. Müller, G. Eguchi, C. Eisenmenger-Sittner, M. Stöger-Pollach, Y. Kakefuda, N. Kawamoto, Q. Guo, T. Baba, T. Mori, Sami Ullah, Xing Qiu Chen, and E. Bauer. Thermoelectric performance of a metastable thin-film Heusler alloy. *Nature*, 576(7785):85–90, 2019.

[58] Hee Seok Kim, Weishu Liu, Gang Chen, Ching Wu Chu, and Zhifeng Ren. Relationship between thermoelectric figure of merit and energy conversion efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, 112(27):8205–8210, 2015.

[59] Wei Hsin Chen, Po Hua Wu, Xiao Dong Wang, and Yu Li Lin. Power output and efficiency of a thermoelectric generator under temperature control. *Energy Conversion and Management*, 127:404–415, 2016.

[60] H. Julian Goldsmid. Bismuth telluride and its alloys as materials for thermoelectric generation. *Materials*, 7(4):2577–2592, 2014.

[61] Pedro M. Rodrigo, Alvaro Valera, Eduardo F. Fernandez, and Florencia M. Almonacid. Annual Energy Harvesting of Passively Cooled Hybrid Thermoelectric Generator-Concentrator Photovoltaic Modules. *IEEE Journal of Photovoltaics*, 9(6):1652–1660, 2019.