Bayesian vs Evolutionary Optimisation in Exploring Pareto Fronts for Materials Discovery

Kai Yuan Andre Low, Eleonore Vissol-Gaudin, *Member IEEE*, Yee-Fun Lim, Kedar Hippalgaonkar

***Abstract*—With advancements in automated experimental setups, material optimisation and discovery can scale to higher throughput with larger evaluation budgets. Two state-of-the-art algorithms with conceptually different multi-objective optimisation strategies (Bayesian and Evolutionary) are compared on synthetic and real-world datasets. Our results show that the Bayesian optimisation strategy, q-Noisy Expected Hypervolume Improvement (qNEHVI) is superior in finding solutions at the Pareto Front rapidly, and when considering hypervolume improvement as a performance indicator. On the other hand, the Evolutionary optimisation strategy, Unified Non-dominated Sorting Genetic Algorithm III (U-NSGA-III), can exploit the Pareto Front and propose a larger pool of optimal solutions, given sufficient evaluation budget, and thus may be a better choice for materials discovery problems where knowing the complete Pareto Front provides greater scientific value to understanding materials space. We discuss the limitations of using hypervolume as a performance indicator for optimisation strategies, alongside hypervolume-based strategies such as qNEHVI, which do not adequately explain the number of solutions at or near the Pareto Front. We also performed a comparison of both optimisation strategies at different batch sizes to consider throughput capabilities.**

***Index Terms*—Bayesian optimisation, constrained multi-objective optimisation, evolutionary algorithm, materials science**

I. Introduction

Materials science as a field is being disrupted with advances in machine learning and automation [1]–[4], where high-throughput experimentation (HTE) capabilities accelerate discovery of materials in more complex search spaces. Users not only save time on experimentation by virtue of automated workflows with faster processing, but also leveraging on equipment with larger batches of experiments to increase throughput and thus minimise experimental time [5], [6]. There have been many successful applications of HTE, particularly in the single objective problem space alongside machine learning-assisted optimisation strategies [7]–[18]. However, many real-world problems are more complex, specifically with multiple conflicting properties to be optimized, for example: strength vs ductility in metal alloys [19], device thickness vs fill factor in photovoltaics [20], or selectivity vs current density in catalysts [21]. In addition, such problems may include constraints that restrict the space of feasible solution. This motivates the need for multi-objective optimisation strategies with constraint handling capabilities to be integrated in HTE setups [22]–[25]. The first step could consist of formulating complex material science problems as constrained multi-objective optimisation problems (CMOPs).

A. Constrained Multi-Objective Optimisation

A CMOP with m objectives and (q+k) constraints, can be defined as:

$$\begin{aligned} \min F(\boldsymbol{x}) &= (f_1(x), \dots, f_m(x))^T \\ st\ g_i(\boldsymbol{x}) &\geq 0, i = 1, \dots, q \\ h_j(\boldsymbol{x}) &= 0, j = 1, \dots, k \\ \boldsymbol{x} &\in R^n \end{aligned} \tag{1}$$

where $F(\mathbf{x})$ defines the multi-dimensional objectives to be optimised, and $g_i(\mathbf{x})$ and $h_j(\mathbf{x})$ define the inequality and equality constraints, respectively. A solution is an *n*-dimensional vector of decision variables, $\boldsymbol{x}$. To determine the objective value of a solution, a Pareto-optimal solution x^1 dominates another solution x^2 if $F(x^1) \leq F(x^2)$ where they are feasible. A total set of all feasible and Pareto-optimal solutions can then be defined as the Pareto Set, or Pareto Front (PF) when mapped onto the objective space. This PF represents all solutions with the optimal trade-off between objectives.

A commonly defined materials discovery problem is usually of combinatorial nature with unexplored regions of objective space, given some mixture of chemicals, precursors, and other process parameters. This problem can be formulated as a CMOP with an unknown PF to be extrapolated to, with minimal

K.H. acknowledges funding from the Accelerated Materials Development for Manufacturing Program at A*STAR via the AME Programmatic Fund by the Agency for Science, Technology and Research under Grant No. A1898b0043. K.H. also acknowledges funding from the NRF Fellowship NRF-NRFF13-2021-0011. *(Corresponding author: K.H).* K.Y.A.L. and K.H. conceived of the research. K.Y.A.L. working with E.V-G. and Y-F.L. developed and tested the algorithms and datasets, with key intellectual contributions from all authors. K.Y.A.L. wrote the manuscript, with input from all co-authors.

K.Y.A.L and E.V-G are with Nanyang Technological University, School of Materials Science and Engineering, Singapore 639798. (emails: kaiyuana001@ntu.edu.sg, eleonore.vg@ntu.edu.sg).

Y-F.L. is with the Institute of Materials Research and Engineering, Singapore 138634. (email: limyf@imre.a-star.edu.sg).

K.H. is with both Nanyang Technological University, School of Materials Science and Engineering, Singapore 639798 and the Institute of Materials Research and Engineering, Singapore 138634. (email: kedar@ntu.edu.sg).

Source code for our work can be found in https://github.com/andrelowky/Constrained-Multi-Objective-Optimisation-for-Materials-Discovery. A supplementary document is also available.

Colour versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

evaluation budget [26]–[29]. This is achieved through selection and evaluation of available solutions $\boldsymbol{x} \in R^n$, where each solution represents the set of experimental input parameters (chemicals, temperature settings etc.) used in the screening. The number of data points is typically low, with most works generally limited to around 10^2-10^3 data points due to practical bottlenecks such as time taken to synthesize and characterise, or simply due to a limited time/cost budget.

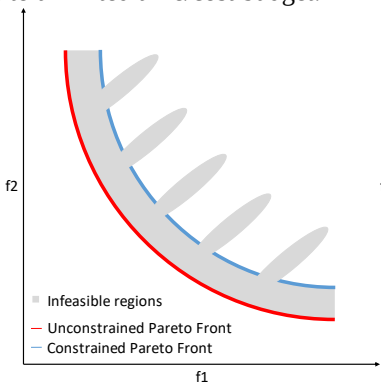

Fig. 1: Illustration of constrained multi-objective (f1 and f2) space for a convex minimization problem in bi-objective space. The addition of infeasible regions in grey shifts the original PF from solid red to blue.

In addition, the PF can be discontinuous with multiple infeasible regions due to underlying property limitations such as phase boundaries/solubility limits, or engineering rules, for example summing mixtures to 100% [30]. Such constraints can also be knowledge-based, where a domain expert with prior knowledge sets them to pre-emptively 'avoid' poor results and converge faster [31]–[33].

CMOPs can be solved in various ways, but recently, two classes of algorithms have shown promises in solving such problems with a high level of success, namely: multi-objective evolutionary algorithms (MOEA) and multi-objective Bayesian optimisation (MOBO).

B. Evolutionary Algorithm

MOEAs [34] work by maintaining and evolving a population of solutions across an optimisation run. For example, Genetic Algorithms (GA) are a specific subset that utilise 'operations' alike biological processes [35]: members of the population are selected to become parents based on a specific selection criterion, and then undergo crossover and mutation to form a children population. Within the field of MOEAs, various constraint handling techniques have been proposed [36]–[38] as well as extensions of MOEAs to many-objective (m>2) problems [39]. MOEAs are well suited to implementations where solutions can be tested in parallel, given their population-based approach, where each generation's population can be treated as a batch. MOEAs have been successfully applied in materials-specific multi-objective problems: experimental data is used to construct a machine learning model which is then treated as a computation optimisation problem to be solved, and the results evaluated physically [40]–[45]. The use of MOEAs relevant to materials science has seen computational and inverse design problems [46]–[52].

C. Bayesian Optimisation

MOBOs leverage on surrogate models to cheaply predict some black-box function, and then utilise an acquisition function to probabilistically compute a predictive function and return the best possible candidate where gain is maximised [53]. The choice of surrogate model can depend on the user, but in recent literature, it has become synonymous with 'kriging' which refers specifically to the use of Gaussian Processes (GP) as the surrogate model, taking advantage of its flexibility and robustness [54]. The extension of MOBO~~s~~ to CMOPs is less mature, with relatively new implementations that cover parallelization, multi-objective and constraints [55]–[59]. On top of these, there are also hybrid variants such as TSEMO [60] or MOEA/D-EGO [61] which integrate the use of MOEAs to improve the prediction quality of the underlying surrogate models. In general, BO as an overarching optimisation strategy has already been established as an attractive strategy for use in both computational design problems [62]–[67], as well as experimentation problems [68]–[74] due to its sample efficient approach.

D. Hypervolume

As previously discussed, the PF defines the set of optimal solutions of a CMOP. For optimisation of CMOPs, hypervolume (HV) is often used as a performance indicator. It defines the Euclidean distance bounded by a point, and the reference point in a single dimension, and a HV in multiple dimensions. It directly shows the quality of the solutions since a solution set with high HV is closer to the true PF and is diverse as it effectively dominates more objective space. An illustration of the HV measure for a multi-objective (two dimensions for illustration) convex minimization problem is presented in Fig. 2, where HV is computed by finding the area of non-dominated solutions, i.e. the solutions closest to PF without any competitor, bounded by a reference point.

Aside from being a performance metric to compare optimisation strategies, HV can also be directly evaluated to guide convergence of various algorithms. Hanaoka et al showed that scalarization-based MOBOs may be best suited for clear exploitation and/or preferential optimisation trajectory of objectives, whereas HV-based MOBOs are better for exploration of the entire search space [75]. Indeed, HV-based approaches empirically show a preference in proposed solutions towards the extrema of a PF [76], [77], and thus can better showcase extrapolation. In contrast, scalarization approaches to reduce multi-objective problems to a single-objective such as hierarchically in Chimera [78] or any user-defined function [79] have limitations: i) it is difficult to determine how to properly scalarize objectives; ii) single objective optimisation methods cannot propose a set of solutions that balance trade-off.

Within the context of multi-objective optimisation and material science implementation, two state-of-the-art algorithms were compared in the present work: q-Noisy Expected Hypervolume Improvement (qNEHVI) [80] and Unified Non-dominated Sorting Genetic Algorithm III (U-

NSGA-III) [81]. They are MOBO and MOEA-based algorithms, respectively. They were chosen based on their reported performance in solving complex CMOPs (with respect to HV score), and the fact that they are capable of highly parallel sampling, making them suitable for integration within an HTE framework.

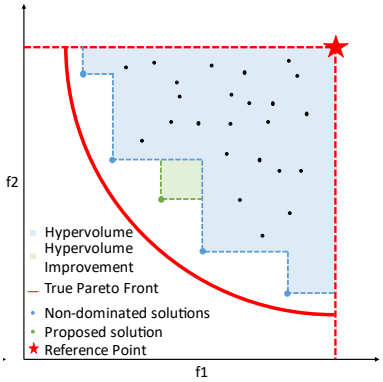

Fig. 2: Illustration of hypervolume for a convex minimization problem in bi-objective space. The red line represents the ground truth PF, while the blue points and region reflect the best-known solutions and their associated hypervolume, respectively. The green point and region are then used to illustrate the contribution of a new evaluated solution. The computation of hypervolume in objective space is performed with respect to a lower bound with a reference point, shown by the red star.

Furthermore, both algorithms are chosen from open-source Python libraries, making them easy to implement and enabling reproducibility of results presented. The main contributions of our work are as follows:

1. We show that qNEHVI is a more sample efficient optimisation strategy which is aligned with the wide adoption of MOBOs for materials experimentation
2. We present 2 alternative means of illustration as metrics to assess the performance of multi-objective optimisation besides HV score
3. We demonstrate the weaknesses of qNEHVI in attaining diverse set of optimal solutions which is an important domain-specific criteria
4. We explore various batch sizes that are relevant to high-throughput experimentation for materials science

We thus provide means for materials scientists to empirically compare and decide upon an appropriate optimisation strategy for experimentation. This paper is organised as follows: the experimental set up, including algorithms and metrics are presented in section II. Benchmark problems and the results obtained are reported and discussed in section III. Finally, section IV concludes this paper and outlines avenues for future work.

II. EXPERIMENTAL SETUP

This section describes the two optimisation strategies, qNEHVI and U-NSGA-III used for this study. We present three metrics of comparison, four synthetic benchmarks and two real-world materials science benchmarks.

A. q-Noisy Expected Hypervolume Improvement

qNEHVI is a HV-based MOBO that utilises expected HV improvement, which was shown to outperform scalarization, entropy-based and even other HV-based approaches like TSEMO. The base acquisition function is defined in [80] as:

$$\alpha_{NEHVI}(\boldsymbol{x}) = \int \alpha_{EHVI}(\boldsymbol{x}|\mathcal{P}_m)\, p(f|\mathcal{D}_m) df \qquad (2)$$

Originally, α_{EHVI} extends the classic Expected Improvement acquisition function [81] to HV as an objective [82]. The integration of $p(f|\mathcal{D}_m)$, which represents the posterior distribution of previously evaluated noisy points, with α_{EHVI} maintains Bayes optimality with noisy observations. Further on, the authors also provided a means of both sequentially and jointly evaluating a batch of points.

Our implementation here relies on the sample code reported in the BoTorch [83] tutorial for constrained multi-objective optimisation, with the significant change being to decrease initial Quasi Monte Carlo (QMC) sampling to 128 (default settings according to API) instead of 512 to improve computational run times. The use of QMC here generates candidates X_{cand} for optimization. We do not foresee significant change in the performance of qNEHVI with this changed hyperparameter.

B. Unified Non-Dominated Sorting Genetic Algorithm III

U-NSGA-III is an improved implementation of many-objective NSGA-III which is better generalisable for single and bi-objective problems. The original NSGA-III [84], [85] relies on reference vectors to maintain diversity but did not include a selection operator to determine fitter and more feasible parents for mating. In comparison, the use of non-dominated ranking for the previous algorithm NSGA-II [86] provides a stronger selection pressure for bi-objective problems. Thus, a new tournament selection operator (Algorithm 1) is introduced that allows U-NSGA-III to take advantage of both reference vector-guided diversity and stronger selection pressure in convergence in single, bi and many objective problems. U-NSGA-III is a suitable MOEA that performs robustly for CMOPs purely without surrogate modelling.

The use of reference points in directing evolution helps to maintain diversity of the entire population, since the reference point is decomposed into multiple reference vectors (depending on number of objectives) in the hyperplane, where points that are closer (smaller Euclidean distance) to a reference vector belong to that niche. In tournament selection, winners are preferred from different niches to preserve diversity. Otherwise, traditional non-dominated sorting is used to determine which parent has a better non-dominated ranking and smaller constraint violation for selection as the winner, p_s, for mating.

We rely upon the implementation found in pymoo [87], setting population size μ, number of children λ and reference points H to be $\mu = \lambda \approx H$, following the original NSGA-III paper [88], [89]. Having $\mu = \lambda$ is analogous to a pure search via U-NSGA-III with no underlying surrogate modelling, since the

total number of proposed candidates is equal to the total sample batch size**.**

Algorithm 1: Pseudo-code for Tournament Selection for U-NSGA-III [90], where p here represents a parent member, and π represents the reference direction associated with that parent

if $\pi(p_1) = \pi(p_2)$ **then**
  if $p_1.rank < p_2.rank$ **then**
    $p_s = p_1$
  else
    if $p_2.rank < p_1.rank$ **then**
      $p_s = p_2$
    else
      if $d_\perp(p_1) < d_\perp(p_2)$
        $p_s = p_1$
      else
        $p_s = p_2$
      end if
    end if
  end if
else
  $p_s = randompick(p_1, p_2)$
end if

C. Metrics

We performed a comparison of qNEHVI and U-NSGA-III on various synthetic and real-world benchmark problems. As a starting point, we took batch size (the number of samples to evaluate per iteration) at 8, following a generally higher range of throughput in materials experimentation. All optimisation runs are initialised with a Sobol sampling of 2*(variables+1), following S. Daulton et al in their implementation of qNEHVI [91].

We compare both approaches based on 3 metrics:

1. Optimisation trajectory – a single optimisation run at high evaluation budget (100 iterations x 8 points per batch) is plotted in objective space to illustrate the trajectory of proposed solutions at each iteration towards the PF.
2. Probability density map – 10 runs at a lower evaluation budget (24 iterations x 8 points per batch) are plotted together with a Gaussian kernel density estimate to illustrate the probability distribution of solutions being proposed in the objective space.
3. Batch sizing – various batch sizes are compared using log HV difference to illustrate their HV improvement.

III. RESULTS AND DISCUSSION

In this section we describe the four synthetic benchmarks and two real-world materials science benchmarks that we apply qNEHVI and U-NSGA-III to, using the three metrics mentioned above to assess their performance. For synthetic benchmarks, we also presented a contour plot to illustrate scaling dimensionality.

A. Synthetic benchmarks

For synthetic benchmarks, we select two-objective scalable problems for comparison as described in Table 1. The ZDT test suite [92] provides a range of PF shapes, while the MW test suite [93] provides constraints and uniquely shaped PFs to challenge the optimisation algorithms. Both test suites rely on a similar construction method for minimization problems: taking a single variable function f_1 against a shape function f_2 as such:

$$\min f_1(x) = x_1 \qquad (3)$$
$$\min f_2(x) = g(x)h(f_1(x), g(x))$$

The single variable function closely resembles certain real-life multi-objective problems where an input is to be minimised against some other objective, for example minimizing process temperature, while achieving a target output [68].

Since the synthetic problems are scalable, we did an additional comparison of both qNEHVI and U-NSGA-III across a range of dimensionality from 2 to 12 to represent possible experimental parameter space in combinatorial screening experiments, as shown in Fig. 3.

U-NSGA-III in Fig. 3 a) and c) shows a more gradual change in colour and did not reach the maximum values for higher dimensions, indicating a slower rate of convergence and poorer HV improvement, respectively, which scale with dimensions. In contrast, results presented in Fig. 3 b) and d) for ZDT1 and ZDT2, respectively, indicate that qNEHVI converges fast at a high HV improvement, as illustrated by the bright yellow coloration which appears early and maintains this up to dim=12 with little loss in initial performance.

qNEHVI, while showing superiority in overall HV score for the ZDT3 and MW7 problem, had a lower rate of convergence and maximum HV improvement as dimensions increase, illustrated in Fig. 3 f) and h) by the colour gradient. Although we note that in other literature, GP models tend to perform poorly at high dimensionalities [94], [95], this was not observed here, to the limit of 12 dimensions. We believe that the underlying stochastic QMC sampling used is what drives the optimisation and hence the performance remains robust.

It should be noted that in Fig. 3 e), U-NSGA-III's HV score on the ZDT3 problem scales inconsistently with dimensionality: dim=5 shows better HV improvement (brighter colour) compared to dim=2 to 4. We attribute this to the disconnected PF being strongly affected by differences in initilisation, where entire regions can be lost as the evolutionary process fails to extrapolate and explore sufficiently. We discuss this further for Fig. 4 and 5 where we visualize the optimization trajectory for both algorithms.

TABLE I
LIST OF SYNTHETIC BENCHMARKS

Name	Definition	PF	Range of x_i	n_obj	n_constr	ref_pt[1]
ZDT1	$f_1(x) = x_1$ $g(x) = 1 + \frac{9}{n-1}\sum_{i=2}^{n} x_i$ $h(f_1, g) = 1 - \sqrt{f_1/g}$	convex	[0, 1]	2	0	[11, 11]
ZDT2	$f_1(x) = x_1$ $g(x) = 1 + \frac{9}{n-1}\sum_{i=2}^{n} x_i$ $h(f_1, g) = 1 - (f_1/g)^2$	concave	[0, 1]	2	0	[11, 11]
ZDT3	$f_1(x) = x_1$ $g(x) = 1 + \frac{9}{n-1}\sum_{i=2}^{n} x_i$ $h(f_1, g) = 1 - \sqrt{f_1/g} - (f_1/g)\sin(10\pi f_1)$	disconnected	[0, 1]	2	0	[11, 11]
MW7	$f_1(x) = g_3 x_1$ $f_2(x) = g_3\sqrt{1 - (f_1/g_3)^2}$ $c_1(x) = (1.2 + 0.4\sin(4l)^{16})^2 - f_1^2 - f_2^2 \geq 0$ $c_2(x) = (1.15 + 0.2\sin(4l)^8)^2 - f_1^2 - f_2^2 \leq 0$ $l = \arctan(f_2/f_1)$	disconnected	[0, 1]	2	2	[1.2, 1.2]

Lastly, we observe in Fig. 3 g) for MW7 that U-NSGA-III performs significantly worst as compared to qNEHVI, regardless of dimensionality. The presence of more complex constraints in the problem means that many solutions are likely to be infeasible and require more iterations to evolve to feasibility according to the evolution mechanism. Infeasible solutions do not contribute to HV improvement at all, and we note that this is one of the limitations of plotting using HV as a metric, where feasibility management is not clearly reflected.

In order to investigate why qNEHVI presented a higher HV improvement for qNEHVI, we then proceed to plot the optimization trajectory to observe solutions in objective space, as shown in Figure 4. We set the number of dimensions to 8. This is representative of a range of experimental parameters that materials scientists would consider practical. We first performed a single optimisation run of 100 iterations x 8 points per batch. The evaluated solutions are plotted onto the objective space and coloured by their respective iteration from dark to bright.

The general observations in Fig. 4 a)-h) comparing qNEHVI to U-NSGA-III are consistent with results previously reported in Fig. 3, specifically in terms of HV scores and convergence rate. In all sub figures, qNEHVI was able to propose solutions at the PF within the first 20 iterations, as shown by the darker colour of points along the red line (true PF). This suggests that it is very sample efficient. However, it was unable to fully exploit the region of objective space close to the PF, and solutions in later iterations are non-optimal. In fact, in Fig. 4 b) and d), ZDT1 and ZDT2 respectively, a large portion of solutions lie along the f1=x1=0 line. This is explained by the choice of reference point, which we explore in more detail in SI 1.

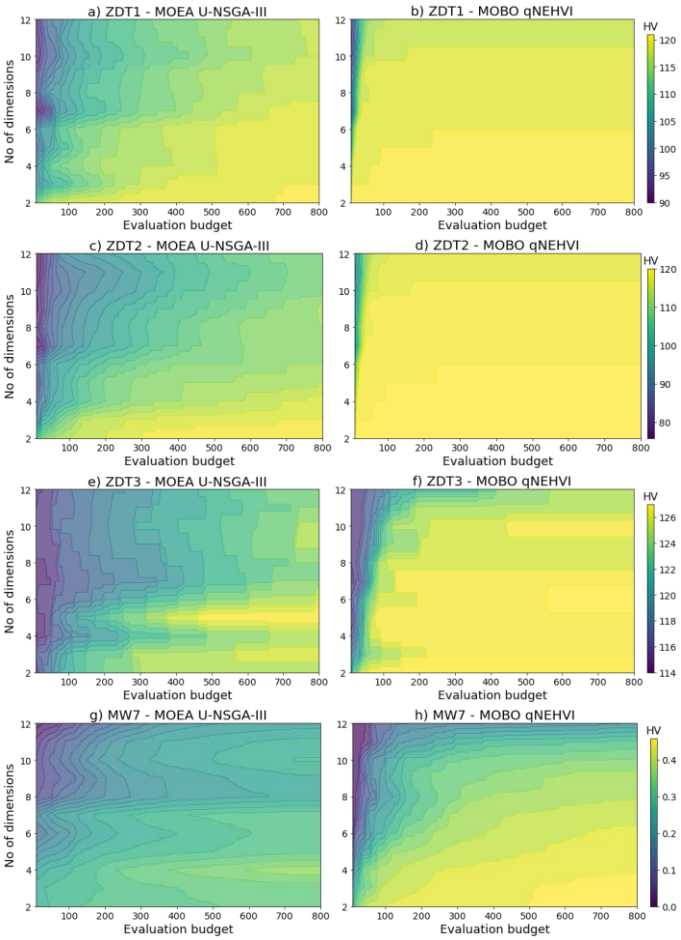

Fig. 3: Contour plots for dimension vs evaluation budget. a-b) ZDT1. c-d) ZDT2. e-f) ZDT3. g-h) MW7. The colour bar illustrates the mean cumulative HV score with respect to cumulative evaluations, over a total evaluation budget of 100 iterations x 8 points per batch. Results are averaged over only 5 runs due to high computational cost of searching over many dimensions. The results here show that qNEHVI is a far superior method when looking at only HV as a performance metric.

[1] Choice of reference point is taken from BoTorch's API.

We hypothesize that qNEHVI is unable to identify multiple bi-objective points along the PF because the underlying GP surrogate model did not accurately model the PF for ZDT1-3. As for MW7, despite the algorithm being able to propose many solutions near the unconstrained PF, it failed to overcome the constraints, as seen by the failure to adjust to the new dotted red line. We observed that qNEHVI's superior HV score (Fig. 3) could be attributed to the stochastic nature of QMC sampling, which is used to provide a pool of candidates for the surrogate model and acquisition function to determine the next 'best' batch of points to evaluate. This hypothesis is supported by results reported in SI 2, where it can be observed that the GP model did not fully learn the objective function.

In contrast, U-NSGA-III, while requiring a significantly larger number of iterations to reach the PF, had a more consistent optimisation trajectory towards the PF, as seen by the gradual colour gradient in Fig. 4 a), c), e), g). This suggests that there are less wasted evaluations for MOEAs, as the latter iterations are targeted towards the PF. However, despite having more solutions near the PF, the HV score is lower for U-NSGA-III than qNEHVI. This is a limitation of using HV as a performance metric: it strictly rewards non-dominated solutions across the entire search space, i.e. a handful of solutions at the PF extrema are preferred, as shown previously in Fig. 3 where U-NSGA-III showed poorer HV improvement compared to qNEHVI for ZDT1, ZDT3 and MW7.

Notably, we observe in Fig. 4 e) and g) that the disconnected PFs for ZDT3 and MW7 can lead to entire regions of objective space being omitted. This is clearly seen in both sub-figures where solutions only have a single trajectory towards the nearest PF region. We previously made the statement, based on results reported in Fig. 3 c) and d), for the same synthetic problems, that the disconnected spaces are strongly influenced by initilisation, where U-NSGA-III's mechanism of tournament selection rewards immediate gain over coverage, i.e. exploitation over exploration. This is both a strength and weakness of U-NSGA-III in comparison to qNEHVI, where the stochastic QMC sampling enables greater exploration of the overall search space, but not the PF.

In addition to observing the optimisation trajectory in the objective space, a way to visualise the efficiency of the algorithms is plotting a probability density map, which is computed by plotting all the sampled points over multiple optimisation runs and computing the probability density function with a Gaussian kernel estimate. This is used to illustrate the likelihood of the same point being sampled in different runs, which is shown with a brighter colour. Similar to Fig. 4, we take a dimensionality of dim=8, but limit the evaluation budget to 24 iterations x 8 points per batch due to the computation cost of multiple runs, as well as that of computing the probability density function.

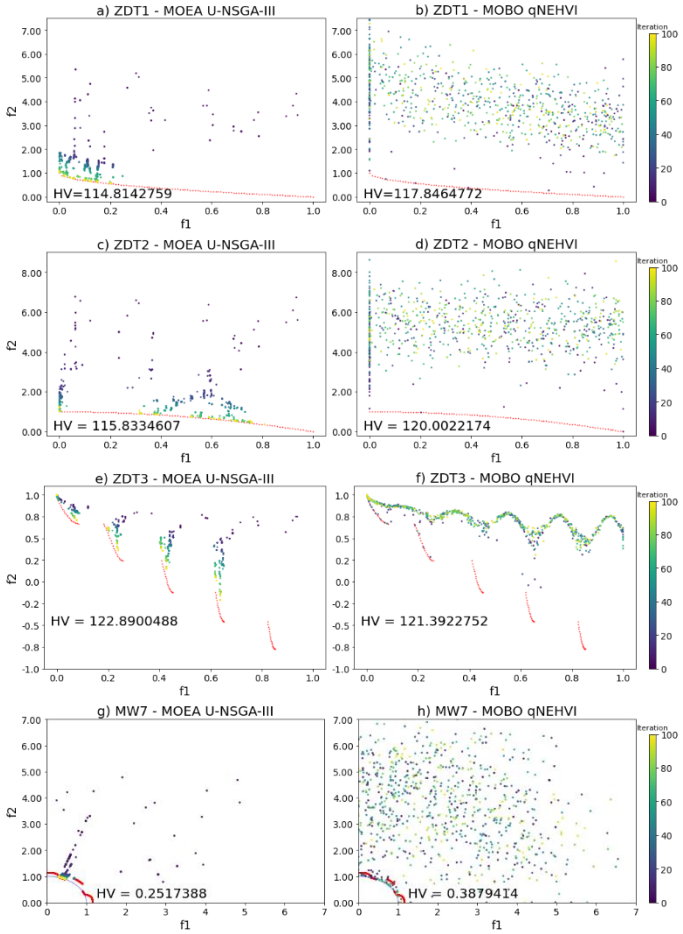

Fig. 4: Optimisation trajectory in objective space for a single optimisation run of 100 iterations x 8 points per batch. a-b) ZDT1. c-d) ZDT2. e-f) ZDT3. g-h) MW7. The red line represents the true PF, while MW7 being a constrained problem has an additional blue line to show the unconstrained PF. The colour of each experiment refers to the number of iterations. All problems clearly show a more gradual evolution of results as the number of iterations progress in U-NSGA-III whereas qNEHVI rapidly approaches PF and then fails to converge further.

Results reported in Fig. 5 further reinforce the observation that qNEHVI produces a large pool of non-optimal solutions for all benchmarks problems, where many points exist away from the PF. Additionally, the darker coloration for qNEHVI in Fig. 5 b), d), f) and h) indicates a much lower probability of occurrence, which reinforces our hypothesis, that HV improvement can be partially attributed to the stochastic nature of QMC sampling. Additionally, Fig. 5 b) and d) for ZDT1 and ZDT2 respectively also show that there were many solutions being proposed at the extrema of f1=x1.

This is the same behaviour as that observed for a single run in Fig. 4 b) and d), and we further elaborate upon it in SI 1. In contrast, the heuristic nature of U-NSGA-III provides more consistency between optimisation runs, which is shown by the brighter regions of points near the PF in Fig. 5a), c), e) and g). indicating a higher probability density. Notably, the bright regions are not spread across objective space evenly. There is a preference for the lower range of f1=x1 since it is easily tunable, i.e. it is simple to derive improvement by simply decreasing x1. This is in line with our previous discussions

based on results reported in Fig. 4, where U-NSGA-III prefers solutions with immediate improvement. Furthermore, we observe that the bright regions are concentrated near the PF, which indicates that U-NSGA-III was able to consistently approach the PF and maintain a larger pool of near-Pareto solutions over the optimisation runs, despite the limited evaluation budget.

In contrast, qNEHVI had relatively few points, although they are lying directly on the PF, which is then shown as a higher mean HV compared to U-NSGA-III. In a real-world context, the larger pool of near-Pareto solutions could have scientific value, especially for users looking to build a materials library and further understand the PF. However, this is not reflected by the HV performance indicator.

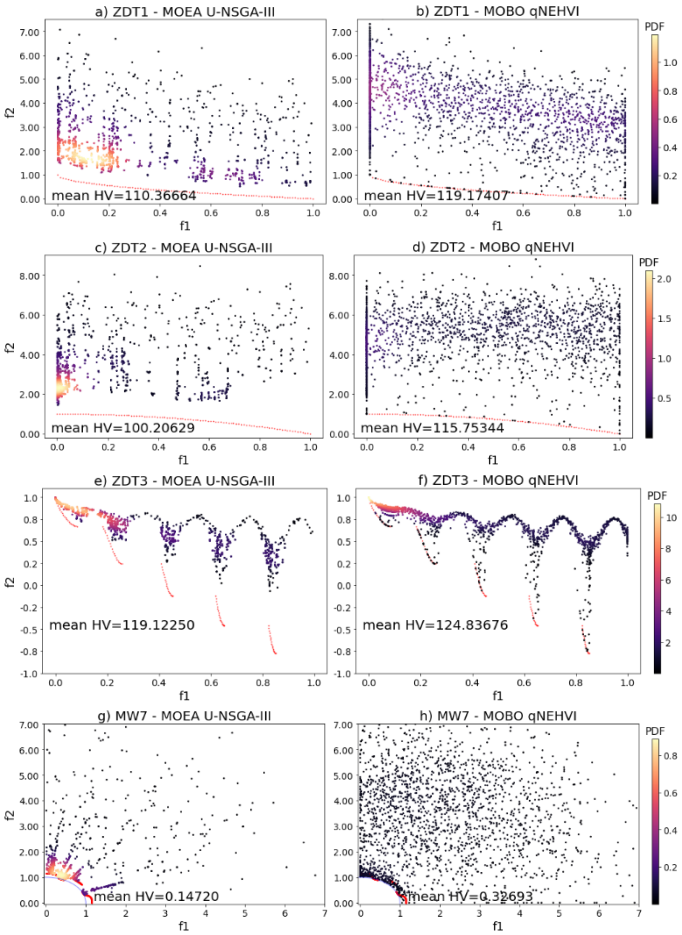

Fig. 5: Probability density maps in objective space for 10 runs of 24 iterations x 8 points per batch. a-b) ZDT1. c-d) ZDT2. e-f) ZDT3. g-h) MW7. The evaluated data points are plotted with a Gaussian kernel density estimate using SciPy to illustrate the distribution of points across objective space. The colour bar represents the numerical value of probability density. Results are averaged over the 10 runs and highlight the lower diversity of points and consistency in optimisation trajectory for qNEHVI compared to U-NSGA-III.

The choice of batch size is another important parameter to consider for materials scientists. It can be tuned when attempting to scale up for HTE. A larger batch size is usually ideal since it provides higher throughput, and thus more time savings since lesser iterations are required. However, batch size affects the performance of optimisation strategies, potentially reducing the number of iterations needed in a run. We thus perform optimisation on the same synthetic problems for different batch sizes, keeping dimensionality at dim=8 and with the same evaluation budget of 192 points and 10 runs as mentioned earlier. We then recorded the hypervolume metric across the run and plotted it as a function of log hypervolume difference, taking log10(HVmax - HVcurrent), taking HVmax from the known PF in pymoo.

The authors of qNEHVI hypothesised that it operates better at small batch sizes by providing a smoother gradient descent in sequential optimisation [80]. Results reported in Fig. 6 a), b) and d) for ZDT1, ZDT2 and MW7, respectively, support this hypothesis, and we clearly observe that the lowest batch size setting of 2, as represented by the pink line, has the best performance overall.

Interestingly, this is also the case for U-NSGA-III where the lowest batch size of 2 tends to give better HV for ZDT1-3 as seen by the blue line. This is also empirically shown in literature where, given a total budget, higher populations may impede convergence as it effectively limits the number of iterations [96]–[98]. It is suggested that the same did not apply for MW7 since the disconnected PF was often not fully explored due to differences in initialisation and how the heuristic search operated, which we discuss previously for Fig. 4 and 5. Instead, a larger batch size i.e. larger population is beneficial in maintaining solutions across disconnected regions of objective space, as seen by the red line in Fig. 6d). We also explain why this did not apply to ZDT3: since the initial sampling was generally able to cover the search space well, there are relatively little ‘lost’ regions as seen from Fig. 4c). Additionally, we provide optimisation trajectory plots for U-NSGA-III at different batch sizes in the SI 3 to illustrate this.

Furthermore, we also observe that qNEHVI has greater variance in log HV difference, compared to U-NSGA-III. This further reinforces our hypothesis that the performance of qNEHVI is in part due to the stochastic QMC sampling, whilst the heuristic nature of U-NSGA-III means that the evolution of solutions is more consistent.

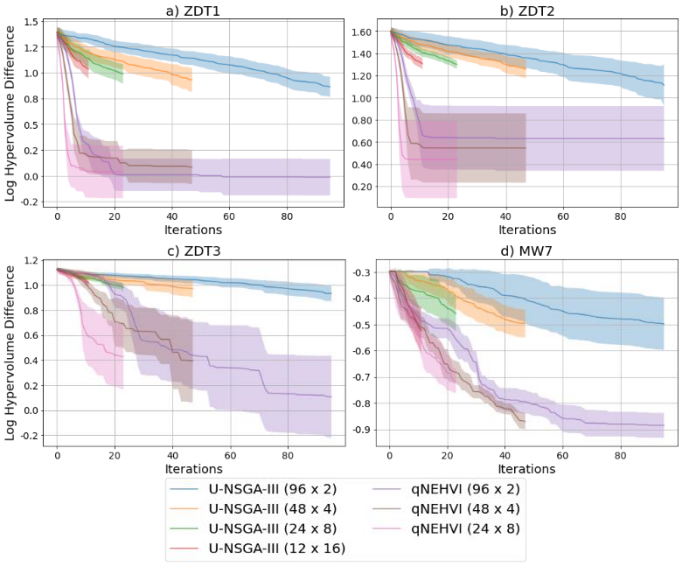

Fig. 6: Convergence at different batch sizes with the same total evaluation budget of 24 x 8. a) ZDT1. b) ZDT2. c) ZDT3. d) MW7. We omitted qNEHVI for batch of 16 due to prohibitively high computation cost when scaling up. Plots are taken with mean and 95% confidence interval of $\log_{10}(HV_{max} - HV_{current})$, with HV_{max} being computed from known PF in pymoo. We follow the same details as for Fig. 5. Results suggest that qNEHVI works better with low batching on disconnected PF.

B. Real-world Benchmarks

Based on results reported in Fig. 4, 5, 6, we formulate the hypothesis that qNEHVI as a MOBO strategy is very sample efficient, i.e. able to arrive at the PF rapidly with few evaluations and is superior in maximizing hypervolume as a performance metric. In comparison, we found that U-NSGA-III provides a more consistent search due to its heuristic evolution nature over that of stochastic QMC sampling in qNEHVI, and furthermore maintain a larger pool of near-Pareto samples that is not reflected by the HV performance metric. We also report that smaller batch sizes are generally better in both strategies over the two-objective jobs used.

To test this hypothesis, we repeated our experiments on real-world multi-objective datasets. An unavoidable issue of empirically benchmarking optimisation strategies on real-world problems is that some surrogate model must be used in-lieu of a black-box where new data is experimentally validated. Alternatively, a candidate selection problem can be used where optimisation is limited to only proposing new candidates from a pre-labelled dataset until eventually the 'pool' of samples is exhausted [65], [75], [99], [100]. The benefit of this method over surrogate-based methods is that only real data from the black-box is used, rather than data extrapolated from a model approximating its behaviour. However, the candidate selection approach assumes that the existing dataset contains all data points necessary to perfectly represent the search space and true PF. It is generally not possible to prove that this is the case, unless the exact function mapping input to output of the black box is known, or the dataset contains all possible combination of input/output pairs and is therefore a complete representation of the problem like that of inverse design.

Here, due to the relatively small size of the datasets (~10^2 data points), the candidate selection method was not implemented. Instead, we relied on training an appropriate regressor to model the dataset. The two real-world benchmarks used in this paper are presented in Table 2, and results are shown in Fig. 7, 8, 9. Materials datasets with constraints are hard to find from available HTE literature, asides from simple combinatorial setups that need to sum to 100%, [101]. Another example is Cao L. et al [70], which included complex constraints in the form of solubility, although we were unable to attain their full dataset and solubility classifier.

Similar to synthetic benchmark experiments, we compare both approaches based on 3 metrics:

1. Optimisation trajectory in objective space for 100 iterations x 8 points per batch
2. Probability density function in objective space for 24 iterations x 8 points per batch
3. Comparison of batch size for log hypervolume difference

Fig. 7 further supports our conclusions drawn from results reported in Fig. 4. As seen in Fig. 7 b) and d), qNEHVI is highly sample efficient, with points at or near the PF within the first 20 iterations or so, indicated by the darker points lying on the red line.

However, qNEHVI shows a large random distribution of non-optimal points away from PF across the entire optimisation as seen by both dark and bright points, which we attribute to the stochastic QMC sampling. U-NSGA-III performs a gradual evolution of points towards the PF as seen in Fig.7 a) and c), as well as maintaining a large pool of near-optimal solutions. This is reflected by the lower HV scores for U-NSGA-III compared to those of qNEHVI.

At a smaller evaluation budget, we observe that U-NSGA-III consistently maintains a large pool of near-optimal solutions, as the bright region is seen nearer to the PF, while reporting a lower mean HV compared to qNEHVI (Fig. 8 a) and e)). Fig. 8 b) for the Thin Film problem also corroborates our findings that qNEHVI proposes many non-optimal solutions, as seen by the bright region away from PF, which indicates a higher probability of occurrence.

TABLE II
LIST OF REAL-WORLD BENCHMARKS

Name	Optimization Problem	Modelling Technique[1]	n_var	n_obj	n_constr[2]	ref_pt
Thin film [70]	Minimize process temperature and maximize conductivity of spray coated palladium films	GP regressor	4	2	0	[1.019, -0.048]
Concrete Slump [102]	Maximize slump and compressive strength in concrete formulations	Neural network ensemble	7	2	0	[0, 0]

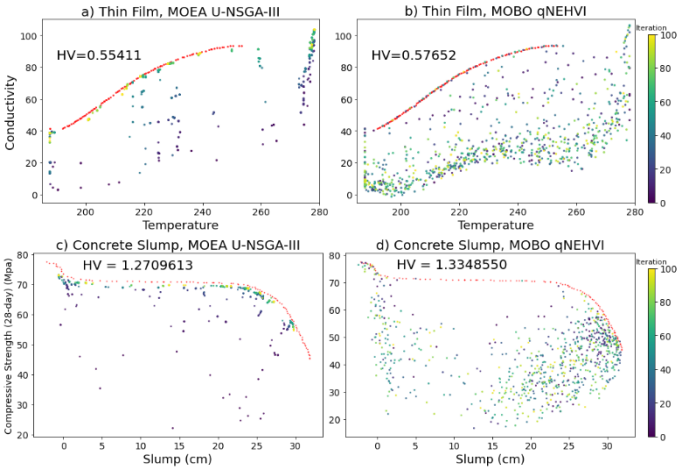

Fig. 7: Optimisation trajectory in objective space for a single optimisation run of 100 iterations x 8 points per batch. a-b) Thin Film. c-d) Concrete Slump. across objective space for a single run of 100 iterations x 8 points per batch. The red line represents the PF. PFs for real-world datasets were virtually generated using NSGA-II for 500 generations with population size of 100. The colour of each experiment refers to the number of iterations. The results here corroborate the 'wastage' of solutions in qNEHVI, although which algorithm is superior appears to be problem dependent.

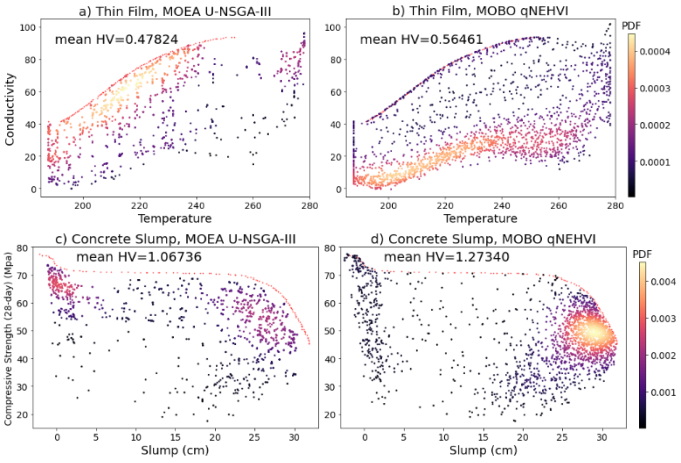

Fig. 8: Probability density maps in objective space for 10 optimisation runs of 24 iterations x 8 points per batch. a-b) Thin Film. c-d) Concrete Slump. The evaluated data points are plotted with a Gaussian kernel density estimate using SciPy to illustrate the distribution of points across objective space, with a colour bar to represent the numerical value of probability density. Results are averaged over 10 runs, taking a smaller evaluation budget of 24 iterations x 8 points = 192. The results here reinforce the finding that qNEHVI has a more random distribution of points, but still outperforms U-NSGA-III for a low evaluation budget.

Interestingly, in Fig. 8d) for Concrete Slump problem, we observe that qNEHVI is consistently converging to a specific region in objective space, while the U-NSGA-III search follows that of Fig. 8b) with concentration of solutions at the near-optimal region close to PF. We hypothesize that qNEHVI's performance for this problem is influenced by how the underlying GP surrogate model learns the function and strongly biases solutions to that specific region. We show further proof in SI 2, where we illustrate the expected PF given by the GP surrogate model.

In contrast, both problems here indicated that U-NSGA-III benefited more from larger batch sizes, as seen by the green line, which is different from what we observed in Fig. 6 for synthetic problems. Our hypothesis is that the modelled datasets present a more mathematically difficult optimisation problem, with various 'obstacles' that inhibit the evolution of solutions towards the PF. We support this by referring to our discussions for Fig. 7 c) and d) on Concrete Slump regarding local optima, as well as observing a notable blank region of objective space which U-NSGA-III fails to flesh out in Fig. 7 a) for Thin Film problem. Overall, results reported here suggest that given state-of-the-art implementations in HT experiments, a small batch-size with MOBO is the right strategy to converge rapidly.

Finally, we also studied the effect of batch size on convergence using the two optimisation approaches on the two real-world datasets. Results present both similarities and differences with what we observe for synthetic benchmarks as in Fig. 6. A lower batch size in qNEHVI was better for both problems, as seen by the purple line, which is consistent with our findings for Fig. 6.

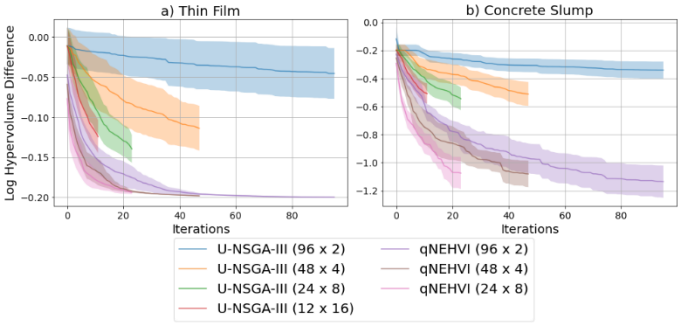

Fig. 9: Convergence at different batch sizes with the same total evaluation budget of 24 x 8. a) Thin Film. b) Concrete Slump. We omitted qNEHVI for batch of 16 due to prohibitively high computation cost when scaling up. Plots are taken with mean and 95% confidence interval of $\log_{10}(HV_{max} - HV_{current})$, with HV_{max} being computed from known PF in pymoo. The results shown here support our conclusions for qNEHVI in Fig. 6 but have marked differences for U-NSGA-III.

[1] Details of their implementation can be found in SI 4.
[2] Further elaboration on implementing constraints for real-world problems can be found in SI 5.

IV. CONCLUSION

We have compared qNEHVI and U-NSGA-III using both synthetic and real-world benchmarks, considering different experimental parameters such as dimensionality and batch size which materials scientists may face when implementing closed loop optimisation in HTE. Our results suggest that qNEHVI is extremely sample efficient in arriving at the PF to maximise HV gain but fails to exploit it. In contrast, we report that U-NSGA-III has a consistent optimisation trajectory, and better exploits the PF while maintaining more near-optimal solutions.

We thus make the case for MOEAs for materials experimentation besides computational design. We also argue that such implementations would be best when the objective space is mildly discontinuous (which can be the case for structural problems such as alloys) since small changes in inputs can cause the outputs to vary wildly in objective space, and an evolutionary-based strategy can navigate with better resolution. This is consistent with work by Liang Q. et al [100] on single-objective optimisation, which noted that having "multiple well-performing candidates allows one to not only observe regions in design space that frequently yield high-performing samples but also have backup options for further evaluation should the most optimal candidate fail in subsequent evaluations".

Furthermore, MOEAs also scale better in terms of computational cost for a high dimensional and high throughput context, where they have the means to converge while maintaining both diversity and feasibility. HV-based MOBOs such as qNEHVI scale poorly to high dimensionality and many-objective problems due to the cost of computing HV. Depending on the HTE set-up, the ML component may not be able to leverage on powerful cluster computing for computationally intensive problems/models. MOEAs with lower computation overhead such as U-NSGA-III would be a better choice in such scenarios. With advancements in HTE set ups allowing for automation and parallel sampling, we expect research groups to leverage on higher throughput systems with short turnarounds. This makes the implementation of MOEAs much more practical to explore complex search spaces when paired with larger evaluation budgets of 10^3 to 10^4 data points.

The choice of batch size to balance optimisation performance while minimising experimental cycles is also important. Empirically, our results obtained suggest that a smaller batch size of around 4 is ideal for the limited evaluation budget of 192 points, although larger batch sizes are preferred for more complex problems (with added difficulty from disconnected regions in objective space, or perhaps presence of local optima).

A caveat of our work here is that the synthetic problems we chose are a generalisation of bi-objective spaces with specific Pareto geometry that may not translate well for real-life experimentation especially for many-objective (M>3) problems. Newer benchmarks with higher difficulties and complex geometries/PFs [103] are tailored towards challenging MOEAs with massive evaluation budgets of up to 10^7 total observations. An example would be MW5 from the MW test suite, which has a narrow tunnel-like feasible regions that are practically impossible for GPs to model, resulting in MOBOs failing to converge. Indeed, R. W. Epps et al noted that it is "difficult to impose complex structure on the GPs, which often simply encode continuity, smoothness, or periodicity" [74]. We refer to other publications which study the differences between surrogate models in BO [100], [104], [105], as well as AI techniques that scale MOBOs to higher dimensional spaces [94], [95].

Furthermore, materials experimentation is usually afflicted with real-world imperfections and deviations during synthesis, or uncertainty due to characterization equipment resolution. For example, MacLeod et al noted that "the tendency of drop-casted samples to exhibit a wide range of downwards deviations in the apparent conductivity due to the poor sample morphology" [68]. The effect of noise causes deviations in objective values from the 'true' ground truth, and although unclear, is an unavoidable aspect of optimisation which should be tackled [106], [107]. In SI 6, we perform a comparison of qNEHVI and U-NSGA-III on varying amounts of white noise on outputs.

In conclusion, our results illustrate that existing performance metrics such as HV do not really reflect the goal of fleshing out the PF region, where HV-based methods like qNEHVI may not achieve satisfactorily. This reflects an aspect of optimisation which might be neglected in the purview of multi-objective materials discovery: which is to find a diverse set of optimal solutions that can adequately convey the trade-offs between conflicting objectives. We thus present alternative illustrative means such as probability density maps to better benchmark the performance of optimisation strategies for such purposes. Moving ahead, we hope that this can spur further improvement for MOBOs and a stronger consideration for the use of MOEAs for materials problems due to its heuristic nature in exploiting the PF.

REFERENCES

[1] T. Lookman *et al.*, "A perspective on materials informatics: state-of-the-art and challenges," in *Information science for materials discovery and design*, Springer, 2016, pp. 3–12.

[2] T. Lookman, F. J. Alexander, and K. Rajan, *Information science for materials discovery and design*, vol. 1. Springer, 2016.

[3] Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," *Journal of Materiomics*, vol. 3, no. 3, pp. 159–177, 2017.

[4] J.-P. Correa-Baena *et al.*, "Accelerating materials development via automation, machine learning, and high-performance computing," *Joule*, vol. 2, no. 8, pp. 1410–1420, 2018.

[5] G. Zhang and D. E. Block, "Using highly efficient nonlinear experimental design methods for optimization of Lactococcus lactis fermentation in chemically defined media," *Biotechnol Prog*, vol. 25, no. 6, pp. 1587–1597, 2009.

[6] S. M. Mennen *et al.*, "The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future," *Org Process Res Dev*, vol. 23, no. 6, pp. 1213–1242, 2019.

[7] S. Sun *et al.*, "Accelerating photovoltaic materials development via high-throughput experiments and machine-learning-assisted diagnosis," *arXiv preprint arXiv:1812.01025*, 2018.

[8] S. Sun *et al.*, "Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis," *Joule*, vol. 3, no. 6, pp. 1437–1451, 2019.

[9] B. Burger *et al.*, "A mobile robotic chemist," *Nature*, vol. 583, no. 7815, pp. 237–241, 2020.

[10] A. Dave *et al.*, "Autonomous discovery of battery electrolytes with robotic experimentation and machine learning," *Cell Rep Phys Sci*, vol. 1, no. 12, p. 100264, 2020.
[11] A. E. Gongora *et al.*, "A Bayesian experimental autonomous researcher for mechanical design," *Sci Adv*, vol. 6, no. 15, p. eaaz1708, 2020.
[12] S. Langner *et al.*, "Beyond ternary OPV: high-throughput experimentation and self-driving laboratories optimize multicomponent systems," *Advanced Materials*, vol. 32, no. 14, p. 1907801, 2020.
[13] J. Li *et al.*, "Autonomous discovery of optically active chiral inorganic perovskite nanocrystals through an intelligent cloud lab," *Nat Commun*, vol. 11, no. 1, pp. 1–10, 2020.
[14] R. Shimizu, S. Kobayashi, Y. Watanabe, Y. Ando, and T. Hitosugi, "Autonomous materials synthesis by machine learning and robotics," *APL Mater*, vol. 8, no. 11, p. 111110, 2020.
[15] L. Wang, L. R. Karadaghi, R. L. Brutchey, and N. Malmstadt, "Self-optimizing parallel millifluidic reactor for scaling nanoparticle synthesis," *Chemical Communications*, vol. 56, no. 26, pp. 3745–3748, 2020.
[16] D. Bash *et al.*, "Accelerated automated screening of viscous graphene suspensions with various surfactants for optimal electrical conductivity," *Digital Discovery*, vol. 1, no. 2, pp. 139–146, 2022.
[17] J. R. Deneault *et al.*, "Toward autonomous additive manufacturing: Bayesian optimization on a 3D printer," *MRS Bull*, vol. 46, no. 7, pp. 566–575, 2021.
[18] F. Mekki-Berrada *et al.*, "Two-step machine learning enables optimized nanoparticle synthesis," *NPJ Comput Mater*, vol. 7, no. 1, pp. 1–10, 2021.
[19] Z. Li, K. G. Pradeep, Y. Deng, D. Raabe, and C. C. Tasan, "Metastable high-entropy dual-phase alloys overcome the strength–ductility trade-off," *Nature*, vol. 534, no. 7606, pp. 227–230, 2016.
[20] I. Ramirez, M. Causa', Y. Zhong, N. Banerji, and M. Riede, "Key tradeoffs limiting the performance of organic photovoltaics," *Adv Energy Mater*, vol. 8, no. 28, p. 1703551, 2018.
[21] S. Ren *et al.*, "Molecular electrocatalysts can mediate fast, selective CO2 reduction in a flow cell," *Science (1979)*, vol. 365, no. 6451, pp. 367–369, 2019.
[22] N. Alsharif, J. R. Uzarski, T. J. Lawton, and K. A. Brown, "High-Throughput Multiobjective Optimization of Patterned Multifunctional Surfaces," *ACS Appl Mater Interfaces*, vol. 12, no. 28, pp. 32069–32077, 2020.
[23] D. Bash *et al.*, "Machine learning and high-throughput robust design of P3HT-CNT composite thin films for high electrical conductivity," *arXiv preprint arXiv:2011.10382*, 2020.
[24] J. Grizou, L. J. Points, A. Sharma, and L. Cronin, "A curious formulation robot enables the discovery of a novel protocell behavior," *Sci Adv*, vol. 6, no. 5, p. eaay4237, 2020.
[25] K. Abdel-Latif, R. W. Epps, F. Bateni, S. Han, K. G. Reyes, and M. Abolhasani, "Self-Driven Multistep Quantum Dot Synthesis Enabled by Autonomous Robotic Experimentation in Flow," *Advanced Intelligent Systems*, vol. 3, no. 2, p. 2000245, 2021.
[26] W. Yong, H. Zhang, H. Fu, Y. Zhu, J. He, and J. Xie, "Improving prediction accuracy of high-performance materials via modified machine learning strategy," *Comput Mater Sci*, vol. 204, p. 111181, 2022.
[27] Y.-F. Lim, C. K. Ng, U. S. Vaitesswar, and K. Hippalgaonkar, "Extrapolative Bayesian Optimization with Gaussian Process and Neural Network Ensemble Surrogate Models," *Advanced Intelligent Systems*, vol. 3, no. 11, p. 2100101, 2021.
[28] A. Sabharwal, H. Samulowitz, and G. Tesauro, "Selecting near-optimal learners via incremental data allocation," 2016.
[29] A. Klein, S. Bartels, S. Falkner, P. Hennig, and F. Hutter, "Towards efficient Bayesian optimization for big data," 2015.
[30] A. M. Gopakumar, P. v Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, "Multi-objective optimization for materials discovery via adaptive design," *Sci Rep*, vol. 8, no. 1, pp. 1–12, 2018.
[31] R. S. Niculescu, T. M. Mitchell, R. B. Rao, K. P. Bennett, and E. Parrado-Hernández, "Bayesian network learning with parameter constraints.," *Journal of machine learning research*, vol. 7, no. 7, 2006.
[32] V. Asvatourian, P. Leray, S. Michiels, and E. Lanoy, "Integrating expert's knowledge constraint of time dependent exposures in structure learning for Bayesian networks," *Artif Intell Med*, vol. 107, p. 101874, 2020.
[33] Z. Liu *et al.*, "Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing," *Joule*, vol. 6, no. 4, pp. 834–849, 2022.
[34] K. Deb, "Multi-objective optimisation using evolutionary algorithms: an introduction," in *Multi-objective evolutionary optimisation for product design and manufacturing*, Springer, 2011, pp. 3–34.
[35] M. Mitchell, *An introduction to genetic algorithms*. MIT press, 1998.
[36] Z. Fan *et al.*, "An improved epsilon constraint-handling method in MOEA/D for CMOPs with large infeasible regions," *Soft comput*, vol. 23, no. 23, pp. 12491–12510, 2019.
[37] B. Xu and Z. Zhang, "Constrained optimization based on ensemble differential evolution and two-level-based epsilon method," *IEEE Access*, vol. 8, pp. 213981–213997, 2020.
[38] Y. Tian, Y. Zhang, Y. Su, X. Zhang, K. C. Tan, and Y. Jin, "Balancing objective optimization and constraint satisfaction in constrained evolutionary multiobjective optimization," *IEEE Trans Cybern*, 2021.
[39] B. Li, J. Li, K. Tang, and X. Yao, "Many-objective evolutionary algorithms: A survey," *ACM Computing Surveys (CSUR)*, vol. 48, no. 1, pp. 1–35, 2015.
[40] P. Zhang, Y. Qian, and Q. Qian, "Multi-objective optimization for materials design with improved NSGA-II," *Mater Today Commun*, vol. 28, p. 102709, 2021.
[41] T. K. Patra, V. Meenakshisundaram, J.-H. Hung, and D. S. Simmons, "Neural-network-biased genetic algorithms for materials design: evolutionary algorithms that learn," *ACS Comb Sci*, vol. 19, no. 2, pp. 96–107, 2017.
[42] R. Jha, F. Pettersson, G. S. Dulikravich, H. Saxen, and N. Chakraborti, "Evolutionary design of nickel-based superalloys using data-driven genetic algorithms and related strategies," *Materials and Manufacturing Processes*, vol. 30, no. 4, pp. 488–510, 2015.
[43] C. A. Coello Coello and R. L. Becerra, "Evolutionary multiobjective optimization in materials science and engineering," *Materials and manufacturing processes*, vol. 24, no. 2, pp. 119–129, 2009.
[44] S. Ganguly, S. Datta, and N. Chakraborti, "Genetic algorithms in optimization of strength and ductility of low-carbon steels," *Materials and Manufacturing Processes*, vol. 22, no. 5, pp. 650–658, 2007.
[45] M. Mahfouf, M. Jamei, and D. A. Linkens, "Optimal design of alloy steels using multiobjective genetic algorithms," *Materials and Manufacturing processes*, vol. 20, no. 3, pp. 553–567, 2005.
[46] S. Wu, C. M. Hamel, Q. Ze, F. Yang, H. J. Qi, and R. Zhao, "Evolutionary Algorithm-Guided Voxel-Encoding Printing of Functional Hard-Magnetic Soft Active Materials," *Advanced Intelligent Systems*, vol. 2, no. 8, p. 2000060, Aug. 2020, doi: https://doi.org/10.1002/aisy.202000060.
[47] P. Avery, C. Toher, S. Curtarolo, and E. Zurek, "XtalOpt Version r12: An open-source evolutionary algorithm for crystal structure prediction," *Comput Phys Commun*, vol. 237, pp. 274–275, 2019, doi: https://doi.org/10.1016/j.cpc.2018.11.016.
[48] E. Berardo, L. Turcani, M. Miklitz, and K. E. Jelfs, "An evolutionary algorithm for the discovery of porous organic cages," *Chem Sci*, vol. 9, no. 45, pp. 8513–8527, 2018.
[49] M. Pakhnova, I. Kruglov, A. Yanilkin, and A. R. Oganov, "Search for stable cocrystals of energetic materials using the evolutionary algorithm USPEX," *Physical Chemistry Chemical Physics*, vol. 22, no. 29, pp. 16822–16830, 2020, doi: 10.1039/D0CP03042B.
[50] R. P. Carvalho, C. F. N. Marchiori, D. Brandell, and C. M. Araujo, "Tuning the Electrochemical Properties of Organic Battery Cathode Materials: Insights from Evolutionary Algorithm DFT Calculations," *ChemSusChem*, vol. 13, no. 9, pp. 2402–2409, May 2020, doi: https://doi.org/10.1002/cssc.201903450.
[51] P. C. Jennings, S. Lysgaard, J. S. Hummelshøj, T. Vegge, and T. Bligaard, "Genetic algorithms for computational materials discovery accelerated by machine learning," *NPJ Comput Mater*, vol. 5, no. 1, p. 46, 2019, doi: 10.1038/s41524-019-0181-4.
[52] D. Salley, G. Keenan, J. Grizou, A. Sharma, S. Martín, and L. Cronin, "A nanomaterials discovery robot for the Darwinian evolution of shape programmable gold nanoparticles," *Nat Commun*, vol. 11, no. 1, pp. 1–7, 2020.
[53] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian

Optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016, doi: 10.1109/JPROC.2015.2494218.

[54] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*, 2003, pp. 63–71.

[55] E. C. Garrido-Merchán and D. Hernández-Lobato, "Predictive entropy search for multi-objective bayesian optimization with constraints," *Neurocomputing*, vol. 361, pp. 50–68, 2019.

[56] S. Belakaria, A. Deshwal, and J. R. Doppa, "Max-value entropy search for multi-objective Bayesian optimization with constraints," *arXiv preprint arXiv:2009.01721*, 2020.

[57] D. Fernández-Sánchez, E. C. Garrido-Merchán, and D. Hernández-Lobato, "Max-value entropy search for multi-objective bayesian optimization with constraints," 2020.

[58] S. Daulton, M. Balandat, and E. Bakshy, "Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization," *Adv Neural Inf Process Syst*, vol. 33, pp. 9851–9864, 2020.

[59] S. Suzuki, S. Takeno, T. Tamura, K. Shitara, and M. Karasuyama, "Multi-objective Bayesian optimization using pareto-frontier entropy," in *International Conference on Machine Learning*, 2020, pp. 9279–9288.

[60] E. Bradford, A. M. Schweidtmann, and A. Lapkin, "Efficient multiobjective optimization employing Gaussian processes, spectral sampling and a genetic algorithm," *Journal of global optimization*, vol. 71, no. 2, pp. 407–438, 2018.

[61] Q. Zhang, W. Liu, E. Tsang, and B. Virginas, "Expensive multiobjective optimization by MOEA/D with Gaussian process model," *IEEE Transactions on Evolutionary Computation*, vol. 14, no. 3, pp. 456–474, 2009.

[62] A. Mannodi-Kanakkithodi, G. Pilania, R. Ramprasad, T. Lookman, and J. E. Gubernatis, "Multi-objective optimization techniques to design the Pareto front of organic dielectric polymers," *Comput Mater Sci*, vol. 125, pp. 92–99, 2016.

[63] A. Solomou *et al.*, "Multi-objective Bayesian materials discovery: Application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling," *Mater Des*, vol. 160, pp. 810–827, 2018.

[64] R. Yuan *et al.*, "Accelerated discovery of large electrostrains in BaTiO3-based piezoelectrics using active learning," *Advanced materials*, vol. 30, no. 7, p. 1702884, 2018.

[65] J. P. Janet, S. Ramesh, C. Duan, and H. J. Kulik, "Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization," *ACS Cent Sci*, vol. 6, no. 4, pp. 513–524, 2020.

[66] M. Karasuyama, H. Kasugai, T. Tamura, and K. Shitara, "Computational design of stable and highly ion-conductive materials using multi-objective bayesian optimization: Case studies on diffusion of oxygen and lithium," *Comput Mater Sci*, vol. 184, p. 109927, 2020.

[67] K. Hanaoka, "Bayesian optimization for goal-oriented multi-objective inverse material design," *iScience*, vol. 24, no. 7, p. 102781, 2021.

[68] B. P. MacLeod *et al.*, "A self-driving laboratory advances the Pareto front for material properties," *Nat Commun*, vol. 13, no. 1, pp. 1–10, 2022.

[69] B. P. MacLeod *et al.*, "Self-driving laboratory for accelerated discovery of thin-film materials," *Sci Adv*, vol. 6, no. 20, pp. eaaz8867–eaaz8867, 2020.

[70] L. Cao *et al.*, "Optimization of formulations using robotic experiments driven by machine learning DoE," *Cell Rep Phys Sci*, vol. 2, no. 1, p. 100295, 2021.

[71] T. Erps *et al.*, "Accelerated discovery of 3D printing materials using data-driven multiobjective optimization," *Sci Adv*, vol. 7, no. 42, pp. eabf7435–eabf7435, 2021.

[72] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, and A. A. Lapkin, "Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives," *Chemical Engineering Journal*, vol. 352, pp. 277–282, 2018.

[73] M. Christensen *et al.*, "Data-science driven autonomous process optimization," *Commun Chem*, vol. 4, no. 1, pp. 1–12, 2021.

[74] R. W. Epps *et al.*, "Artificial chemist: An autonomous quantum dot synthesis bot," *Advanced Materials*, vol. 32, no. 30, p. 2001626, 2020.

[75] K. Hanaoka, "Comparison of conceptually different multi-objective Bayesian optimization methods for material design problems," *Mater Today Commun*, vol. 31, p. 103440, 2022.

[76] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler, "Hypervolume-based multiobjective optimization: Theoretical foundations and practical implications," *Theor Comput Sci*, vol. 425, pp. 75–103, 2012.

[77] A. P. Guerreiro, C. M. Fonseca, and L. Paquete, "The hypervolume indicator: Problems and algorithms," *arXiv preprint arXiv:2005.00515*, 2020.

[78] F. Häse, L. M. Roch, and A. Aspuru-Guzik, "Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories," *Chem Sci*, vol. 9, no. 39, pp. 7642–7655, 2018.

[79] H. Zhang, H. Fu, S. Zhu, W. Yong, and J. Xie, "Machine learning assisted composition effective design for precipitation strengthened copper alloys," *Acta Mater*, vol. 215, p. 117118, 2021.

[80] S. Daulton, M. Balandat, and E. Bakshy, "Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement," *Adv Neural Inf Process Syst*, vol. 34, pp. 2187–2200, 2021.

[81] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[82] S. Daulton, M. Balandat, and E. Bakshy, "Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization," *Adv Neural Inf Process Syst*, vol. 33, pp. 9851–9864, 2020.

[83] M. Balandat *et al.*, "BoTorch: a framework for efficient Monte-Carlo Bayesian optimization," *Adv Neural Inf Process Syst*, vol. 33, pp. 21524–21538, 2020.

[84] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints," *IEEE transactions on evolutionary computation*, vol. 18, no. 4, pp. 577–601, 2013.

[85] H. Jain and K. Deb, "An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach, part II: Handling constraints and extending to an adaptive approach," *IEEE Transactions on evolutionary computation*, vol. 18, no. 4, pp. 602–622, 2013.

[86] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[87] J. Blank and K. Deb, "Pymoo: Multi-objective optimization in python," *IEEE Access*, vol. 8, pp. 89497–89509, 2020.

[88] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints," *IEEE transactions on evolutionary computation*, vol. 18, no. 4, pp. 577–601, 2013.

[89] H. Jain and K. Deb, "An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach, part II: Handling constraints and extending to an adaptive approach," *IEEE Transactions on evolutionary computation*, vol. 18, no. 4, pp. 602–622, 2013.

[90] H. Seada and K. Deb, "U-NSGA-III: A unified evolutionary algorithm for single, multiple, and many-objective optimization," *COIN report*, vol. 2014022, 2014.

[91] S. Daulton, M. Balandat, and E. Bakshy, "Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement," *Adv Neural Inf Process Syst*, vol. 34, pp. 2187–2200, 2021.

[92] E. Zitzler, K. Deb, and L. Thiele, "Comparison of multiobjective evolutionary algorithms: Empirical results," *Evol Comput*, vol. 8, no. 2, pp. 173–195, 2000.

[93] Z. Ma and Y. Wang, "Evolutionary constrained multiobjective optimization: Test suite construction and performance comparisons," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 6, pp. 972–986, 2019.

[94] R. Moriconi, M. P. Deisenroth, and K. S. Sesh Kumar, "High-dimensional Bayesian optimization using low-dimensional feature spaces," *Mach Learn*, vol. 109, no. 9, pp. 1925–1943, 2020.

[95] D. Eriksson and M. Jankowiak, "High-dimensional Bayesian optimization with sparse axis-aligned subspaces," in *Uncertainty in Artificial Intelligence*, 2021, pp. 493–503.
[96] Q. Wang, L. Wang, W. Huang, Z. Wang, S. Liu, and D. A. Savić, "Parameterization of NSGA-II for the optimal design of water distribution systems," *Water (Basel)*, vol. 11, no. 5, p. 971, 2019.
[97] M. Hort and F. Sarro, "The effect of offspring population size on NSGA-II: a preliminary study," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2021, pp. 179–180.
[98] R. Tanabe and A. Oyama, "The impact of population size, number of children, and number of reference points on the performance of NSGA-III," in *International Conference on Evolutionary Multi-Criterion Optimization*, 2017, pp. 606–621.
[99] A. M. Gopakumar, P. v Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, "Multi-objective optimization for materials discovery via adaptive design," *Sci Rep*, vol. 8, no. 1, pp. 1–12, 2018.
[100] Q. Liang *et al.*, "Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains," *NPJ Comput Mater*, vol. 7, no. 1, pp. 1–10, 2021.
[101] T. Erps *et al.*, "Accelerated discovery of 3D printing materials using data-driven multiobjective optimization," *Sci Adv*, vol. 7, no. 42, p. eabf7435, 2021.
[102] I.-C. Yeh, "Modeling slump of concrete with fly ash and superplasticizer," *Computers and Concrete, An International Journal*, vol. 5, no. 6, pp. 559–572, 2008.
[103] Z. Fan *et al.*, "Difficulty adjustable and scalable constrained multiobjective test problem toolkit," *Evol Comput*, vol. 28, no. 3, pp. 339–378, 2020.
[104] Y.-F. Lim, C. K. Ng, U. S. Vaitesswar, and K. Hippalgaonkar, "Extrapolative Bayesian Optimization with Gaussian Process and Neural Network Ensemble Surrogate Models," *Advanced Intelligent Systems*, vol. 3, no. 11, p. 2100101, 2021.
[105] Y. Yan, D. Lu, and K. Wang, "Accelerated discovery of single-phase refractory high entropy alloys assisted by machine learning," *Comput Mater Sci*, vol. 199, p. 110723, 2021.
[106] P. Koch, T. Wagner, M. T. M. Emmerich, T. Bäck, and W. Konen, "Efficient multi-criteria optimization on noisy machine learning problems," *Appl Soft Comput*, vol. 29, pp. 357–370, 2015.
[107] D. Horn, M. Dagge, X. Sun, and B. Bischl, "First investigations on noisy model-based multi-objective optimization," in *International Conference on Evolutionary Multi-Criterion Optimization*, 2017, pp. 298–313.