Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6dd00134c

Novelty-aware evolutionary Bayesian optimisation for multi-objective discovery science

Maytham Aqeeli, Thatchathon Leelawat and David Shorthouse *

Efficient optimisation of complex experimental systems is a central challenge in modern discovery science, particularly in settings characterised by high-dimensional design spaces, expensive evaluations, and multiple competing objectives. Multi-objective Bayesian optimisation (MOBO) has emerged as a leading approach for such problems due to its sample efficiency, but can suffer from limited exploration and reduced diversity, especially in many-objective, multimodal, and constrained settings. Evolutionary algorithms, by contrast, excel at maintaining diversity across the Pareto front but typically require large evaluation budgets. Here, we systematically investigate hybrid evolutionary-Bayesian optimisation strategies that combine the strengths of both approaches. Building on the Evolutionary Guided Bayesian Optimisation (EGBO) framework, we benchmark multiple evolutionary generators within a unified acquisition-driven pipeline across ten synthetic test problems spanning multimodal, many-objective, and constrained regimes. We further introduce a novelty-aware batch selection strategy that explicitly promotes diversity within candidate batches while retaining model-guided prioritisation. Across benchmarks, hybrid methods consistently outperform acquisition-only MOBO in challenging optimisation regimes, achieving improved hypervolume, lower inverted generational distance, and more reliable convergence. Gains are most pronounced in many-objective and multimodal problems, as well as in feasibility-limited search spaces. However, performance advantages diminish in very high-dimensional feature spaces, where evolutionary exploration reduces sample efficiency. The proposed novelty-aware selection further improves performance by reducing redundancy within batches and mitigating optimisation stagnation. Importantly, these trends translate to real-world experimental datasets spanning reaction optimisation, pharmaceutical formulation, materials design, and drug screening. Together, these results demonstrate that hybrid evolutionary-Bayesian optimisation provides a robust and practical strategy for improving optimisation performance in autonomous and data-driven discovery workflows.

Received 23rd March 2026
Accepted 22nd May 2026

DOI: 10.1039/d6dd00134c

rsc.li/digitaldiscovery

1 Introduction

Modern discovery science across chemistry, pharmaceuticals, and materials science is increasingly adopting closed-loop optimisation, in which machine learning models iteratively propose new experiments to efficiently navigate complex design spaces. These approaches are being integrated into automated platforms such as self-driving laboratories (SDLs), which combine robotics, data acquisition, and adaptive decision-making to accelerate the identification of new molecules, materials, and processes.[1–7] In such systems, the optimisation algorithm forms the decision-making core of the experimental loop, determining which experiments are performed and, ultimately, how efficiently the design space is explored. Despite advances in automation and modelling, experimental evaluations remain inherently expensive, noisy, and limited in number, while the associated design spaces are often high-dimensional and governed by multiple competing objectives. Efficiently navigating such spaces therefore requires optimisation strategies that are both sample-efficient and capable of balancing trade-offs across objectives.

Bayesian optimisation (BO) has emerged as a dominant paradigm for this setting, owing to its ability to guide data-efficient exploration using probabilistic surrogate models.[8–10] In particular, multi-objective Bayesian optimisation (MOBO) methods based on hypervolume improvement, such as expected hypervolume improvement (EHVI) and its batch variants (*e.g.*, qLogNEHVI), are widely used in SDL workflows.[11] These approaches typically employ Gaussian process models (or model ensembles) to estimate uncertainty and select candidate experiments that maximise expected improvement of the Pareto front, representing optimal trade-offs between competing objectives. As a result, MOBO has been successfully applied across a wide range of discovery problems, including reaction optimisation,[12,13] materials design,[14,15] and pharmaceutical formulation.[16]

UCL School of Pharmacy, 29-32 Brunswick Square, London, WC1N 1AX, UK. E-mail: d.shorthouse@ucl.ac.uk

However, despite these successes, MOBO methods rely fundamentally on acquisition function maximisation, which prioritises regions of high expected improvement at each iteration. While effective in low-dimensional or well-behaved settings, this strategy introduces an inherent bias toward greedy, model-driven exploitation, rather than exploration of the experimental state space. In more complex scenarios such as many-objective optimisation, multimodal landscapes, or problems with complex feasibility constraints, this can lead to insufficient exploration of the Pareto front and reduced diversity in proposed solutions. In particular, hypervolume-based acquisition functions may struggle to adequately represent trade-offs across many objectives, resulting in premature convergence to limited regions of the objective space.

This limitation is especially critical in real-world discovery workflows, where identifying a diverse set of viable candidates is often more valuable than locating a single optimum. Motivated by this limitation, recent work by Low *et al.*[17] explored incorporating evolutionary search into acquisition-driven optimisation within autonomous laboratory workflows to improve diversity. These hybrid approaches have since been adopted in emerging experimental optimisation settings.[18,19] More broadly, evolutionary multi-objective optimisation (EMO) methods, such as AGE-MOEA-II,[20] U-NSGA-III,[21,22] and SMS-EMOA,[22] are explicitly designed to maintain diversity across the Pareto front through population-based search and non-dominated sorting. These approaches are well suited to many-objective problems and complex constraint landscapes, as they explore multiple regions of the design space simultaneously.[23] However, evolutionary methods typically require large numbers of function evaluations, making them less suitable for settings where experiments are expensive or time-consuming.

This creates a fundamental trade-off: Bayesian optimisation is sample-efficient but diversity-limited, whereas evolutionary algorithms are diversity-rich but evaluation-intensive. Bridging this gap represents a key opportunity for advancing optimisation in autonomous discovery. Low *et al.* recently introduced Evolutionary Guided Bayesian Optimisation (EGBO),[17] a hybrid framework coupling evolutionary candidate generation with acquisition-driven Bayesian optimisation. In this architecture, candidate solutions are generated from two complementary sources: (i) acquisition function optimisation using qLogNEHVI, targeting regions of high expected hypervolume improvement, and (ii) an evolutionary search, promoting diversity across the objective space. These candidates are then jointly evaluated under a unified acquisition-based ranking, and only the most promising subset is selected for evaluation. This competitive candidate generation approach allows exploitation- and exploration-driven proposals to coexist and be assessed on equal footing, enabling the optimisation process to balance local refinement with broader Pareto-front exploration.

However, even within hybrid frameworks such as EGBO, the candidate selection step remains vulnerable to redundancy: when the acquisition function consistently favours dense regions of the current Pareto front approximation, both evolutionary and BO-derived candidates can converge toward similar regions of the decision space, leading to stagnation and inefficient use of the experimental budget – a limitation that has motivated several recent efforts to incorporate diversity and novelty signals into BO frameworks. ROBOT[24] introduced rank-ordered trust regions to discover high-performing solutions satisfying a user-specified diversity constraint, demonstrating that explicitly promoting solution diversity can improve robustness to post-hoc feasibility constraints in single-objective settings. BEACON[25] proposed a sample-efficient novelty search algorithm built on multi-output Gaussian processes, selecting candidates by maximising a novelty metric derived from posterior samples to systematically uncover diverse system behaviours in expensive black-box systems. SANE[26] developed a cost-driven probabilistic acquisition function to navigate multimodal, non-differentiable single-objective landscapes, integrating a domain knowledge gate to distinguish true from spurious optima; this approach has since been extended to autonomous microscopy applications[27] illustrating the growing relevance of diversity-aware active learning across physical sciences SDL platforms. Collectively, these works highlight that standard acquisition-driven BO can suffer from insufficient exploration of the broader solution space, and that incorporating diversity or novelty signals meaningfully improves campaign outcomes.

However, these approaches are largely designed for either pure novelty search or single-objective multimodal optimisation, and do not directly address the problem of Pareto front stagnation in many-objective constrained settings – a challenge that becomes particularly acute in closed-loop SDL campaigns where experimental budgets are limited and redundant candidate selection represents a direct cost. How best to integrate novelty-aware selection into a hybrid evolutionary-Bayesian optimisation framework, without sacrificing the acquisition quality that drives convergence toward the Pareto front, remains an open question.

In this work we build upon the EGBO framework in two ways. First, we provide a systematic evaluation of hybrid evolutionary-Bayesian optimisation strategies across a wide range of optimisation regimes, including many-objective, multimodal, and constrained problems. Second, we introduce a simple novelty-aware batch-selection strategy designed to improve exploration efficiency within the EGBO framework by reducing redundancy within selected batches. We find that by simply weighting the final sample acquisition by novelty improves optimisation efficiency in low sample regimes typical of self driving labs.

Our results demonstrate that hybrid evolutionary-Bayesian strategies provide substantial benefits in challenging optimisation regimes common to self-driving laboratories, particularly in many-objective, multimodal, and feasibility-limited problems where maintaining diversity is critical, while remaining competitive on simpler tasks. We also identify settings in which these gains diminish, particularly in very high-dimensional feature spaces. Taken together, this work provides a unified perspective on hybrid optimisation in data-driven discovery, clarifying when and why such approaches are effective and offering practical guidance for their use in self-driving laboratories and related experimental workflows.

2 Results

2.1 Hybrid evolutionary-Bayesian optimisation improves performance on challenging benchmark problems

Inspired by previous work by Low *et al.*[17], which presented Evolutionary Guided Bayesian Optimisation (EGBO), we sought to systematically evaluate how consistently such hybrid approaches improve optimisation efficiency and reliability.

In a conventional multi-objective Bayesian optimisation campaign, candidate experiments are selected by maximising an acquisition function, such as normalised expected hypervolume improvement (qLogNEHVI), over the design space (Fig. 1A). To introduce diversity in candidate generation, EGBO augments this framework with an evolutionary search that independently proposes additional candidate solutions (Fig. 1B). At each optimisation iteration, candidate points are therefore generated from two sources: (i) acquisition-function optimisation using qLogNEHVI, and (ii) evolutionary search. These candidate sets are then combined into a single pool and evaluated again using the qLogNEHVI acquisition function, which ranks the candidates according to their expected contribution to Pareto front improvement. The top-ranked candidates are then selected for evaluation. In this architecture, qLogNEHVI acts as a common selection criterion, choosing among candidates proposed by both acquisition-driven and evolutionary search strategies.

We surmised that different evolutionary algorithms may perform differently within this framework on different problem types. We established a computational benchmarking framework to evaluate the performance of these hybrid evolutionary-Bayesian optimisation strategies, comparing the effects of different evolutionary algorithm performance to a baseline. All optimisation algorithms were initialised using identical multi-objective test problems and the same seed set of initial samples, allowing comparison of performance across equivalent settings. Optimisation campaigns were then executed using identical batch sizes, numbers of optimisation iterations, and surrogate modelling architectures. This design ensured that each algorithm operated under identical evaluation budgets, enabling fair comparison of optimisation efficiency and reliability. This framework was designed to mimic experimental optimisation campaigns, in which a fixed number of experiments are performed sequentially in batches while the surrogate model is iteratively updated.

To evaluate algorithm performance across a range of optimisation regimes, we selected ten commonly used benchmark problems drawn from the DTLZ,[28] ZDT,[29] and MW[30] test suites used for benchmarking optimisation algorithms. These problems span a range of characteristics representative of real

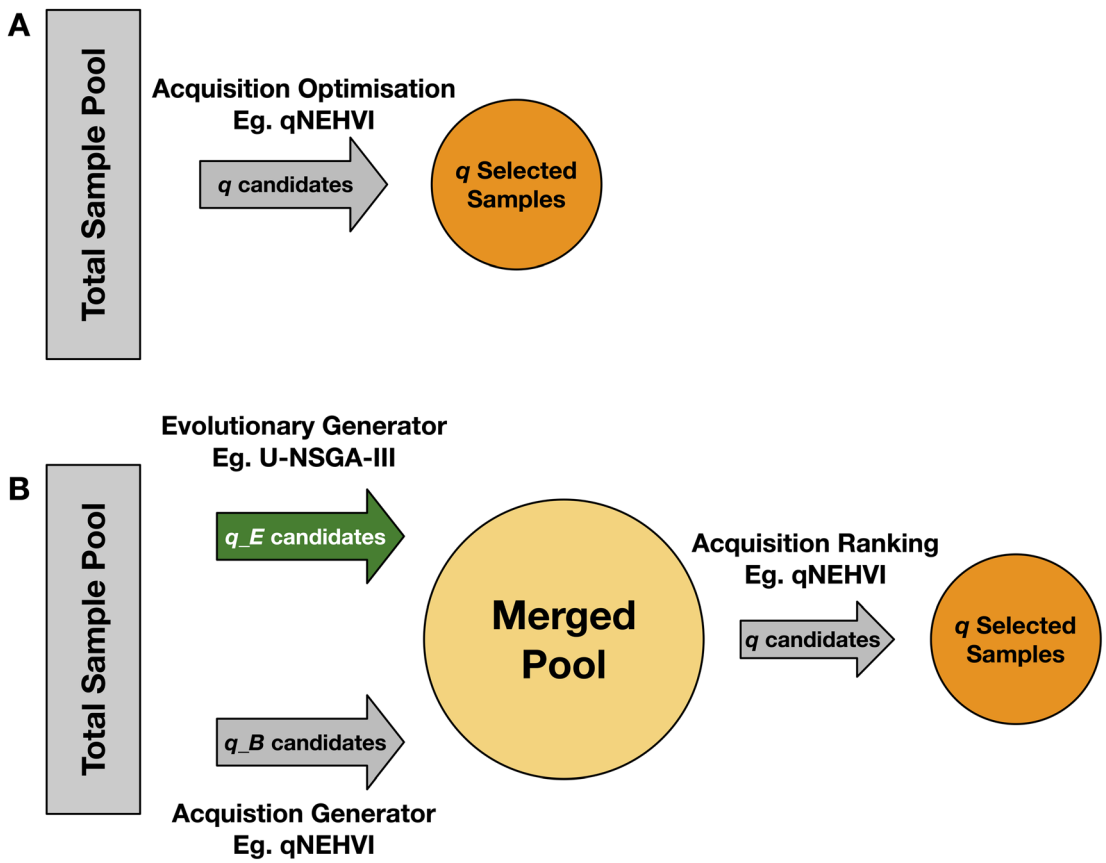

Fig. 1 (A) Traditional BO sample selection method, (B) Low *et al.*[17] method incorporating an evolutionary generator, making a merged pool, and then selecting from that pool using an acquisition ranking.

experimental optimisation challenges, including differing dimensionalities, Pareto front geometries, and constraint structures. Specifically, the benchmark set includes standard multi-objective problems (ZDT1, ZDT2, ZDT3, DTLZ1), a many-objective problem (DTLZ2 with five objectives), multimodal problems containing numerous local optima (ZDT4 and DTLZ3), and constrained optimisation problems (MW3, MW5, and MW7).

To assess the impact of incorporating evolutionary candidate generation, we compared traditional batch multi-objective Bayesian optimisation driven solely by the noisy expected hypervolume improvement (qLogNEHVI) acquisition function with hybrid architectures in which evolutionary algorithms were used to generate additional candidate solutions.

U-NSGA-III[21] is an extension of the widely-used NSGA-III[31,32] framework that incorporates a unified selection mechanism based on structured reference directions distributed across the objective space. At each generation, candidate solutions are assigned to reference directions and survival selection prioritises individuals that provide coverage of under-represented directions, explicitly encouraging a well-spread approximation of the Pareto front. This reference-direction approach makes U-NSGA-III particularly well-suited to many-objective problems, where hypervolume-based diversity metrics become computationally intractable, and it was the evolutionary generator used in the original EGBO study.

SMS-EMOA[22] is a steady-state evolutionary algorithm that uses hypervolume contribution as its primary selection criterion. At each generation, the individual contributing least to the hypervolume of the current population is removed, iteratively refining the population toward a front that maximises dominated volume relative to a reference point. This hypervolume-based survival mechanism directly optimises the same criterion used to assess optimisation quality.

AGE-MOEA-II[20] employs a geometry-aware diversity preservation strategy in which the shape of the Pareto front is estimated adaptively during the search and used to construct a problem-specific set of reference vectors. Survival selection then promotes solutions that provide uniform coverage relative to this estimated geometry, allowing the algorithm to adapt its

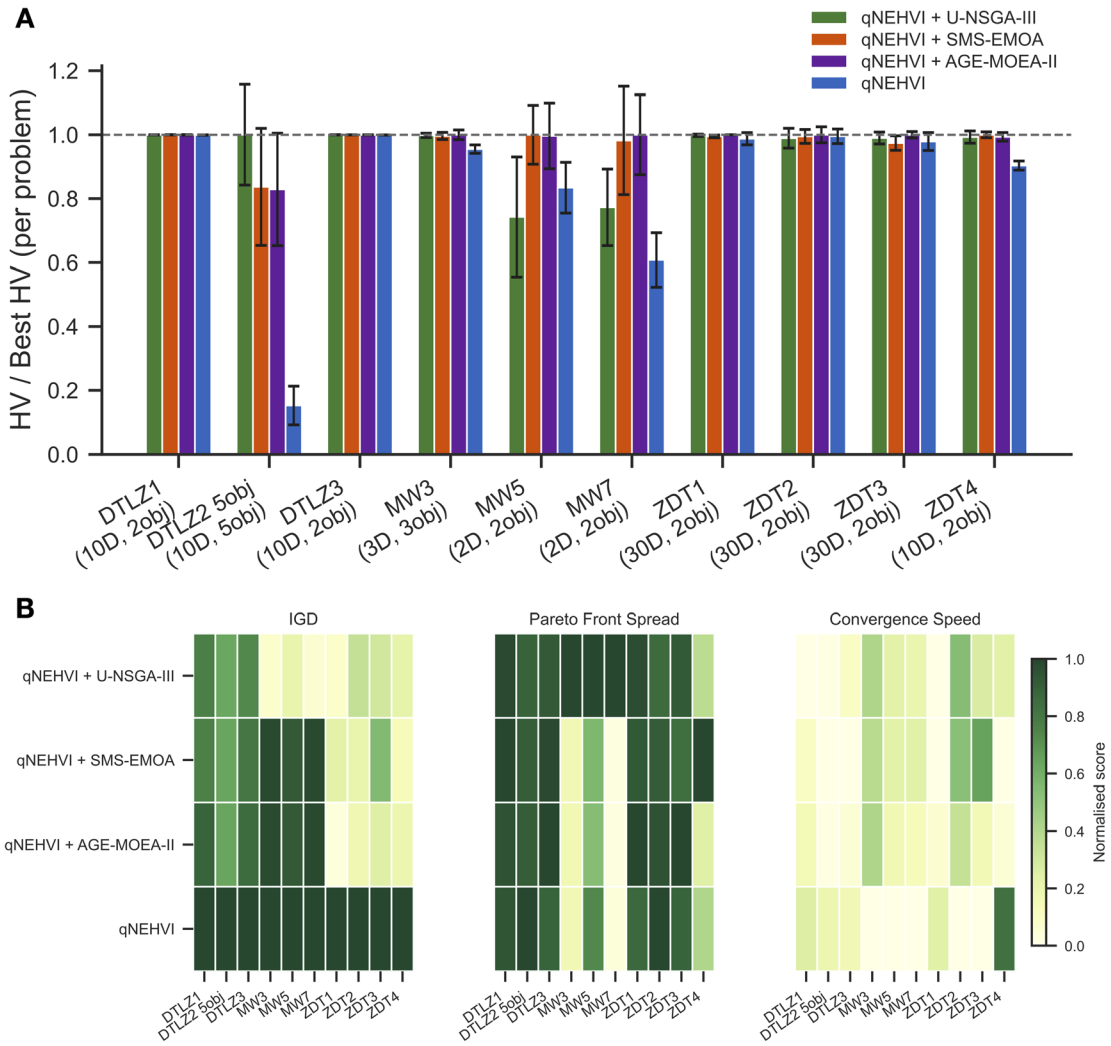

Fig. 2 Performance metrics of different optimisation architectures across the same benchmark datasets. (A) Proportional hypervolume of final campaigns, (B) (left to right) IGD, Pareto front spread, and convergence speed of each architecture.

diversity mechanism to the actual trade-off structure of the problem rather than assuming a fixed front shape.

These algorithms represent complementary diversity-preservation mechanisms; reference-direction-based, hypervolume-based, and geometry-adaptive. This enables systematic evaluation of how different evolutionary strategies interact with acquisition-driven optimisation within the EGBO framework. We compared numerous metrics for the optimisations, including final hypervolume (HV), Pareto front properties, and importantly – Inverted Generational Distance (IGD),[33] which measures average distances from each point on the true Pareto front to the closest point on the discovered front. Smaller values indicate a closer front to the real one, and this provides a measure of diversity as well as convergence.

Across the benchmark suite performed, hybrid optimisation approaches consistently outperformed qLogNEHVI on the most challenging problems (Fig. 2A and B). The largest gain was observed on the many-objective DTLZ2 task with five objectives, where the NEHVI + U-NSGA-III hybrid achieved more than five-fold higher hypervolume than the acquisition-only baseline. This advantage was also reflected in lower IGD (closer approximation to the reference Pareto front), broader Pareto-front spread (better trade-off coverage), and faster convergence (earlier attainment of high-quality fronts) (Fig. S1). Together, these results highlight a key limitation of acquisition-only optimisation in many-objective settings: the acquisition landscape becomes difficult to optimise, increasing the risk of premature concentration in narrow objective-space regions. Evolutionary candidate generation mitigates this by explicitly promoting exploration and front coverage.

A similar pattern was observed on multimodal problems, particularly ZDT4, where hybrid methods again achieved better HV and IGD, with wider front spread and more reliable convergence. This supports the hypothesis that evolutionary generation improves robustness in rugged landscapes where acquisition optimisation alone can become trapped in local optima. In contrast, differences between strategies were small on simpler two-objective problems with smooth fronts (ZDT1, ZDT2, ZDT3, and DTLZ1). Here, HV and IGD were generally close across methods, indicating limited practical benefit from additional evolutionary diversity under easier geometry.

Statistical testing remained consistent with this interpretation. Friedman tests confirmed strong overall algorithm effects for both HV and IGD metrics (HV $p = 5.24 \times 10^{-19}$, IGD $p = 2.62 \times 10^{-33}$). Post-hoc Wilcoxon tests with Holm correction showed that all three hybrid variants significantly outperformed pure acquisition-only qLogNEHVI on both HV (all Holm-adjusted p-values $< 2.1 \times 10^{-6}$) and IGD (Holm p for all $< 2.2 \times 10^{-13}$). Within-hybrid differences were metric-dependent: for HV, no pairwise hybrid comparison was significant whereas for IGD, the U-NSGA-III coupled (EGBO) algorithm was significantly better than SMS-EMOA (Holm $p = 6.42 \times 10^{-5}$) and AGE-MOEA-II (Holm $p = 3.47 \times 10^{-4}$).

To understand how each component contributed to batch selection, we quantified the proportion of selected points originating from the acquisition optimiser *versus* the evolutionary generator. For more complex problems with higher dimensions and tighter trade-off structure, the selected set remains dominated by evolutionary proposals, while acquisition contributes a smaller but consistently non-zero share that likely helps retain local refinement (Fig. 3A). We also calculated the number of Pareto points discovered by each generator and show that selection share alone is not a proxy for impact. The number of Pareto-optimal points contributed by each generator did not always mirror simple selection share, indicating that some generators converted selected proposals into front-quality solutions more efficiently than others (Fig. 3B). Together, these results support the view that hybrid performance is driven by complementary roles with evolutionary generators providing broad frontier coverage and acquisition targeting exploitation, rather than by either component winning in isolation. Overall however, we find U-NSGA-III provides better and more

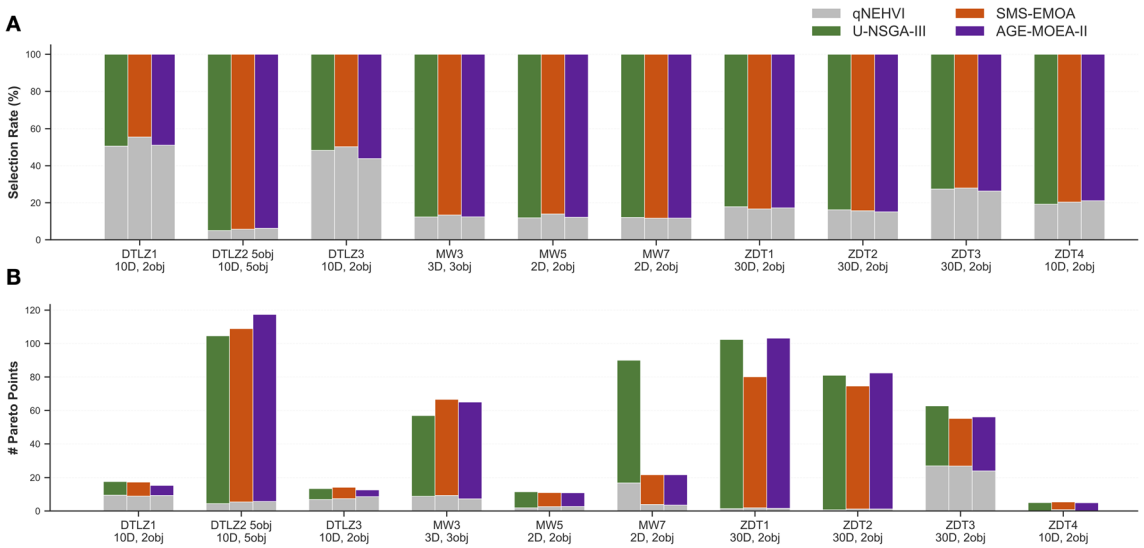

Fig. 3 Generator contribution to optimisation across 10 benchmark problems showing (A) selection rate for each component of each model, (B) number of Pareto front points discovered by each contributor.

consistent coverage of the Pareto front and is competitive with the other evolutionary algorithms in HV expansion.

Among the evolutionary algorithms evaluated, U-NSGA-III consistently produced the most reliable Pareto-front coverage across diverse benchmark problems. This behaviour likely reflects the use of reference directions, which explicitly guide population diversity across the objective space. In many-objective optimisation problems, maintaining diversity becomes increasingly challenging due to the exponential growth of possible trade-offs between objectives. The reference-direction strategy used in NSGA-III helps stabilise search behaviour in such settings, ensuring that candidate solutions remain distributed across the Pareto front. Within the hybrid optimisation framework, this diversity complements acquisition-driven exploitation, allowing evolutionary search to provide broad coverage while the acquisition function refines promising regions.

To assess whether incorporating exploration directly at the acquisition function level achieves comparable benefits to hybrid evolutionary search, we additionally benchmarked qParEGO,[34] a well-established multi-objective BO method that diversifies acquisition through random Chebyshev scalarisation, sampling a new weight vector from the unit simplex at each batch (Fig. S2). Across a representative subset of benchmarks spanning increasing problem complexity, qParEGO performed comparably to EGBO on the simplest two-objective unconstrained problem (ZDT1), but degraded substantially as complexity increased, achieving approximately seven-fold lower hypervolume than EGBO on the five-objective DTLZ2 problem and less than half the hypervolume on the constrained MW5 benchmark. These results suggest that exploration embedded at the acquisition level alone is insufficient as problem complexity scales, and that the complementary diversity provided by evolutionary candidate generation is not replicated by scalarisation-based acquisition diversification alone.

2.2 Combining multiple generators gives marginal performance increases at computational cost

We next sought to uncover whether combining multiple generators led to better performance of an evolutionary hybrid. To test this, we compared the single-generator EGBO configuration against multi-generator variants that combine additional evolutionary operators within the same BO loop. We designed 3 new models combining multiple evolutionary generators with qLogNEHVI based acquisition and tested them on 8 of the

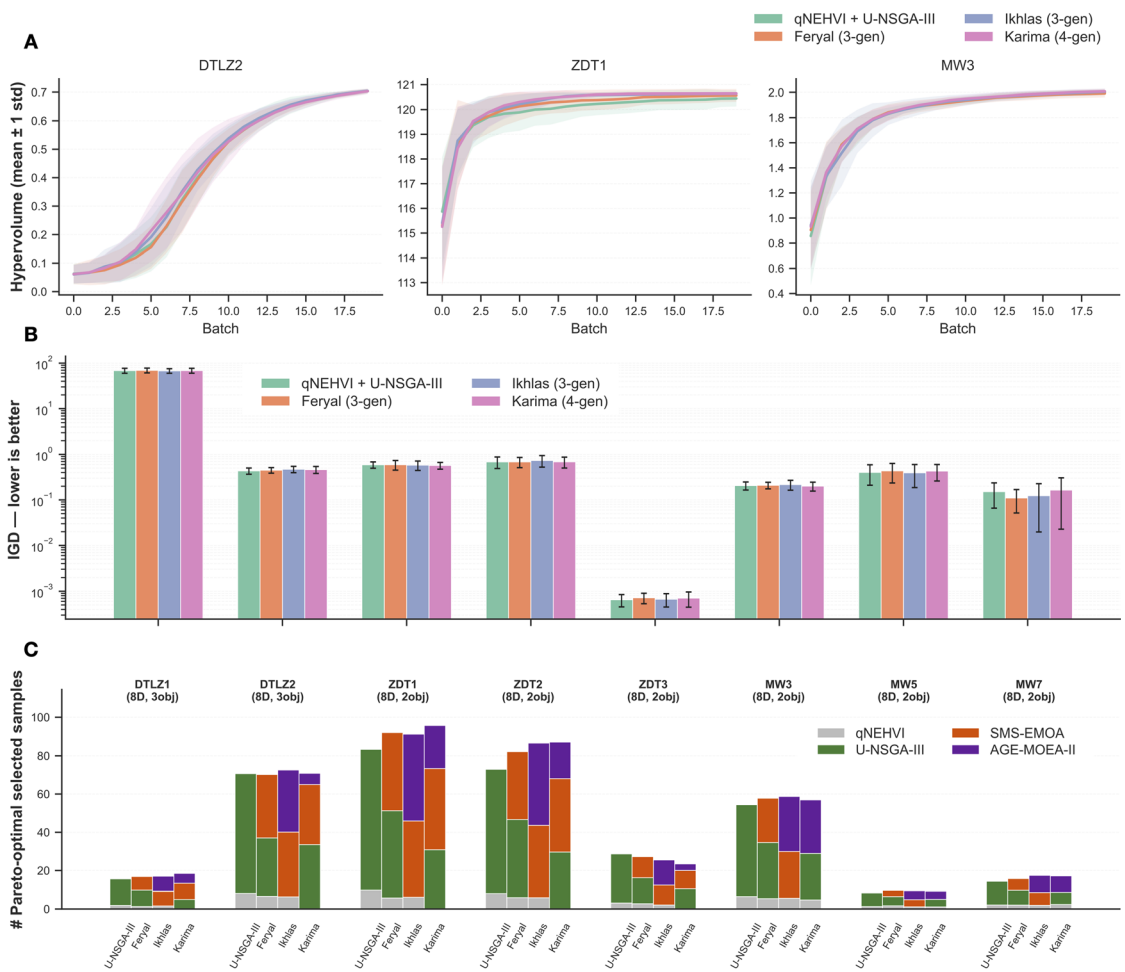

Fig. 4 Performance of multi-generator models across 8 test datasets showing (A) optimisation curves for DTLZ2, ZDT1, and MW3 (B) normalised IGD across 8 test problems, (C) number of Pareto optimal samples suggested by each component of each model.

original 10 problems in the same format as our previous tests, keeping the number of samples each evolutionary generator chooses constant:

- Feryal: Combining U-NSGA-III and SMS-EMOA.
- Ikhlas: Combining AGE-MOEA-II and U-NSGA-III.
- Karima: Combining U-NSGA-III, SMS-EMOA, and AGE-MOEA-II.

Across benchmark problems, multi-generator variants did not deliver large or practically transformative gains in final hypervolume over U-NSGA-III (EGBO) coupled optimisation (Fig. 4A and S3), though a significant Friedman statistic ($p = 5.67 \times 10^{-6}$) and post hoc (Holm p EGBO *vs.* Ikhlas = 0.00255, Karima = 0.00614) results show that two of the generators are consistently higher HV than EGBO alone, these gains were small relative to the added computational complexity. For IGD, we saw no difference between any of the generators (Friedman $p = 0.235$) (Fig. 4B). Overall improvements were small, suggesting that the additional compute cost of coupling multiple generators is unlikely to be universally beneficial.

Studying generator contributions also demonstrates that multiple generators do contribute to Pareto optimal results (Fig. 4C), but from HV and IGD results they converge to nearly identical solution distributions. U-NSGA-III alone achieves roughly the same Pareto front size and diversity as variants using multiple generators combined, indicating that the additional architectural complexity doesn't necessarily provide meaningful improvement in solution quality or exploration coverage, but does increase compute times significantly through both increasing the number of samples generated, and increasing the pool size qLogNEHVI optimises over to select the final samples.

2.3 Evolutionary generators 0069mprove robustness to noise and feasibility-limited search, but degrade in high-dimensional feature spaces

Having established that incorporation of evolutionary generators increases Pareto discovery and convergence speed of optimisation across a range of problem types, we next sought to study the effect of evolutionary diversity on noise robustness, constraint handling, and scalability.

We extended our analysis, keeping only the qLogNEHVI evolutionary variant coupled to U-NSGA-III (the original EGBO) to compare to traditional qLogNEHVI, as it consistently showed a better ability to map the Pareto front (demonstrated by a lower IGD) across problems compared to the other generators, and as additional generators did not meaningfully impact performance. We first explored how the addition of Gaussian noise impacted the ability of the algorithms to optimise, given that many experimental setups include unavoidable high noise levels, we sought to see how robust each method is to different variance. We tested the addition of gaussian noise to 4 test functions (ZDT2, ZDT3, DTLZ2 with 5 objectives, and MW5) at different levels, 0% (baseline), 1%, 5%, 10%, and 20% of the objective range (Fig. 5), keeping other parameters of the optimisation consistent with the previous studies (12 batches of size 8, with 10 repeats on the same random starting points).

As feature dimensionality increased, we found both EGBO and acquisition only driven optimisation maintained their ability to advance the Pareto front, though this generally decreased as noise levels increased. In particular – for the DTLZ2 problem with 5 objectives, we found that qLogNEHVI failed to increase the HV at all, but EGBOs ability to optimise degraded to a similar level as noise increased.

Next, we studied the ability of the models to handle constraints – U-NSGA-III and other evolutionary algorithms inherently contain constraint aware features, and so we studied the ability of EGBO and qLogNEHVI to meet constraints for MW3, MW5, and MW7 constrained optimisation problems. Across all 3 problems, EGBO generated more feasible points (Fig. 6A), and upon shifting the constraint boundaries of MW5 to make finding feasible points more difficult qLogNEHVI + U-NSGA-III maintained its ability to find more feasible samples than qLogNEHVI alone (Fig. 6B). We also tested the algorithms ability to optimise under high numbers of features – as is sometimes common in chemical optimisations where

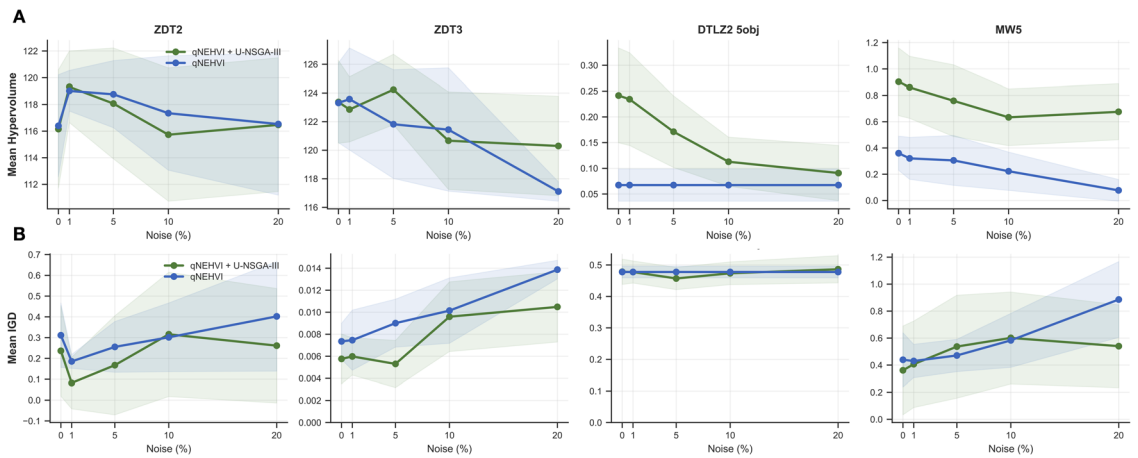

Fig. 5 Impact of Gaussian noise on optimisation performed by qLogNEHVI + U-NSGA-III (EGBO) and qLogNEHVI alone. (A) Final hypervolume discovered over 4 problems. (B) Final IGD score discovered over 4 problems.

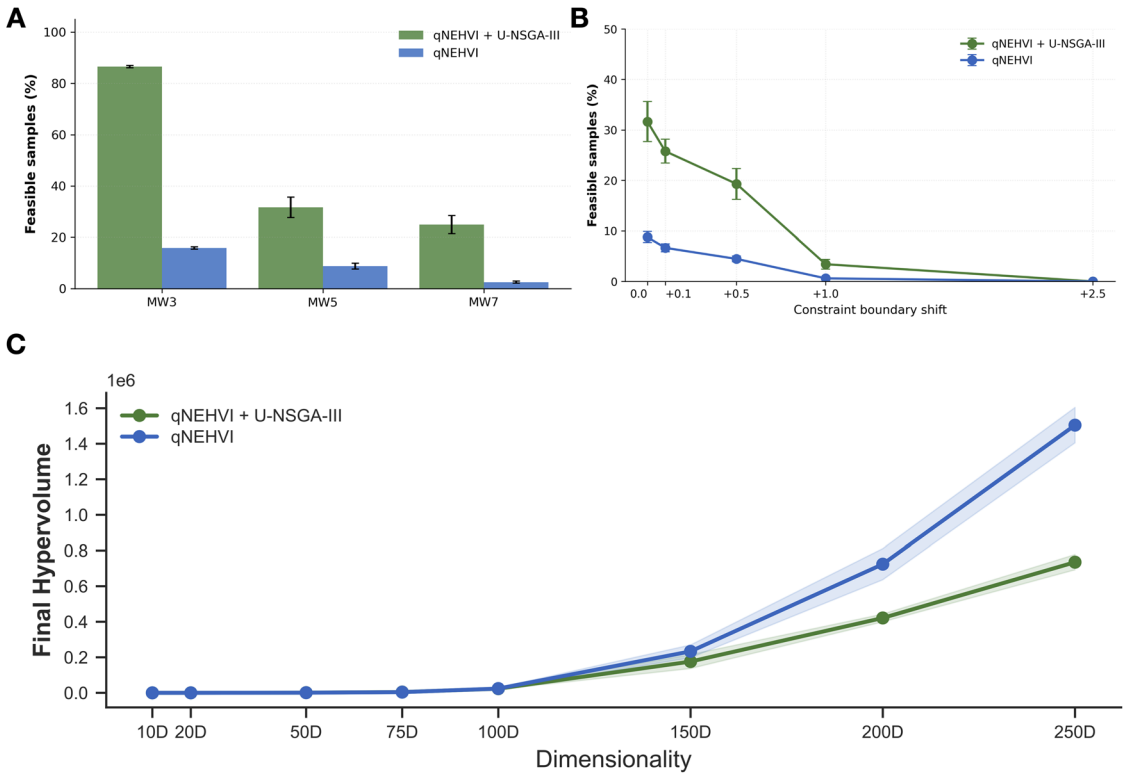

Fig. 6 Model ability to adhere to constraints and handle high dimensionality of features. (A) Number of feasible samples suggested by qLogNEHVI + U-NSGA-III (EGBO) and qLogNEHVI alone on three test problems. (B) Ability of model frameworks to handle increasing restriction of constraints. (C) Model framework final hypervolume over MW5 with increasing numbers of input features.

molecules are described by hundreds of descriptors. As features scale we find that EGBO showed progressively reduced ability to recover Pareto-optimal solutions as feature dimensionality increased, and was ultimately outperformed by qLogNEHVI alone (Fig. 6C), suggesting that for high-dimensional feature systems, using an evolutionary generator can dilute sample efficiency and hinder convergence by spreading evaluations too broadly across a sparse search space.

2.4 Incorporating novelty awareness improves hybrid model optimisation efficiency

Finally, we sought to assess whether evolutionary-Bayesian hybrid optimisation algorithms studied here could be further improved. In the standard evolutionary-coupled workflow, candidate solutions generated by evolutionary search (*e.g.* NSGA-III) and those proposed by qLogNEHVI are merged into a single pool and ranked according to their predicted hypervolume improvement by qLogNEHVI. While this strategy effectively combines model-driven exploitation with evolutionary exploration, it treats all candidates primarily through the lens of predicted improvement and does not explicitly account for redundancy or diversity within the selected batch.

We hypothesised that the final batch selection stage could therefore represent an opportunity for improvement. In particular, when large candidate pools are generated (*e.g.* hundreds of evolutionary candidates alongside a small set of qLogNEHVI proposals), multiple high-scoring candidates may occupy very similar regions of decision. Selecting several such candidates within a single batch may limit the effective exploration of the design space and reduce the information gained from each optimisation round, particularly in campaigns with small batch sizes – such as a batch size of 4 used in the original EGBO project.

To address this, we introduced a modified merge-selection strategy in which candidates from the combined pool are selected sequentially using a hybrid score incorporating both predicted merit and novelty. Specifically, candidates were first evaluated according to their predicted optimisation score (as in the previously studied evolutionary-coupled methods), and final batch members were then selected sequentially with an additional novelty term that favours candidates that are distant from previously selected points (Fig. 7). This approach encourages diversity within each batch while retaining the model-guided prioritisation of promising regions.

Our novelty term includes a weight parameter which controls the trade-off between acquisition merit and novelty, with a higher weight favouring more acquisition driven sample selection, and lower weight more novelty-based selection. We first performed a sensitivity analysis of the weight parameter by assessing the influence of different weights (between 0.3, 0.5, 0.7, and 0.9) on 4 of the test problems (Fig. S4). We found a systematic reduction in IGD as the weight increased across these problems. These results indicate that the balance should favour acquisition-driven selection, with the novelty component

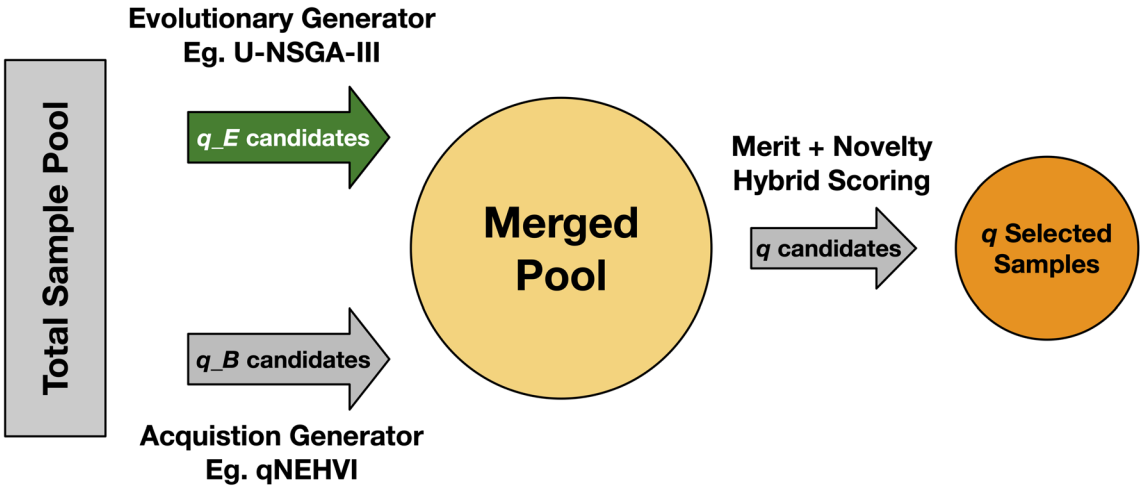

Fig. 7 Novelty-aware evolution guided Bayesian optimisation framework.

playing a supporting role. In particular though, we note that these problems have smooth objective spaces and we expect in fully experimental systems for the novelty term to become more influential, where noisier measurements and more complex objective landscapes reduce the reliability of the acquisition signal and increase the risk of premature convergence to a narrow region of the front. The decreasing returns at lower weights on synthetic problems likely reflect over-penalisation of high-merit candidates in settings where the acquisition signal is already reliable and the evolutionary candidate pool provides sufficient geometric diversity across the front. Having identified that a large novelty contribution degrades performance on these test function, we set our weight term to a value of 0.7 for the following analyses, keeping some contribution from novelty but allowing most of the weight to be driven by acquisition.

We re-ran our initial 10 problem sets, comparing U-NSGA-III coupled evolutionary optimisation without (EGBO) and with (Novelty-aware EGBO) our selection metric, as well as qLogNEHVI. Across the 10 benchmark problems, Novelty-aware EGBO produced modest but consistent gains over standard EGBO and marked gains over acquisition-only qLogNEHVI for both HV (Fig. 8A) and IGD (Fig. 8B). Friedman tests indicated significant overall differences for both metrics (HV $p = 6.93\ 10^{-12}$; IGD $p = 2.04 \times 10^{-12}$), although direct pairwise differences between EGBO and Novelty-aware EGBO were not uniformly significant across all benchmarks. To better characterise selection behaviour, we quantified an exploration score defined as one minus the average percentile rank of selected points within the merged candidate pool, such that values near 1 indicate more exploratory selection and values near 0 indicate more exploitative selection. A paired sign test showed that Novelty-aware EGBO was more exploratory than EGBO in all matched comparisons ($p = 1.58$ e−30; 100/100 pairs) (Fig. S5).

In addition, we performed a sensitivity sweep of the weight parameter of the Novelty-aware EGBO to study the impact of novelty on the efficiency (Fig. S6). We found no clear trend towards one weight value, suggesting a robustness to the inclusion of a novelty aware term, but that further improvements could be made by dynamically adjusting this value for each dataset, or within runs to maximise efficiency.

This difference in selection behaviour was most consequential on constrained problems (Fig. S7). Studying the MW benchmark problems (MW3, MW5, and MW7), which feature narrow or disconnected feasible Pareto regions analogous to real experimental systems, Novelty-aware EGBO reduced the average number of stagnant batches by ∼30% (1.67 *vs.* 2.37) and increased final hypervolume by 12.9% relative to EGBO. This is particularly relevant in the self-driving lab context, where a stagnant optimisation round corresponds to a batch of physical experiments that yields no Pareto-front improvement – consuming reagents, instrument time, and researcher effort with no improvement to the Pareto front. The ability of the novelty-augmented selection to escape locally dense regions of candidate space may therefore be of direct practical value in SDL campaigns targeting multi-objective formulation problems, where feasibility constraints partition the design space in ways that are not known *a priori* and must be discovered through experimentation.

We next evaluated the hybrid optimisation strategies on real-world experimental datasets, in order to assess performance across a range of practical discovery settings with differing design-space structures and objective relationships. These datasets represent problems from reaction optimisation, pharmaceutical formulation, industrial materials development, and drug screening, allowing the behaviour of the optimisation algorithms to be assessed across diverse experimental design spaces and objective structures. We chose to test 4 experimental datasets in this analysis – a Suzuki–Miyaura cross-coupling reaction originally reported by Reizman and Jensen;[35] a microparticle formulation campaign derived from experimental data reported from an automated lab generating long-acting injectable formulations;[16] an industrial coating formulation optimisation dataset from the ADA database, representing a typical multi-objective materials optimisation problem with competing performance criteria;[36] and a drug screening dataset from the Genomics of Drug Sensitivity in Cancer (GDSC) project[37] where

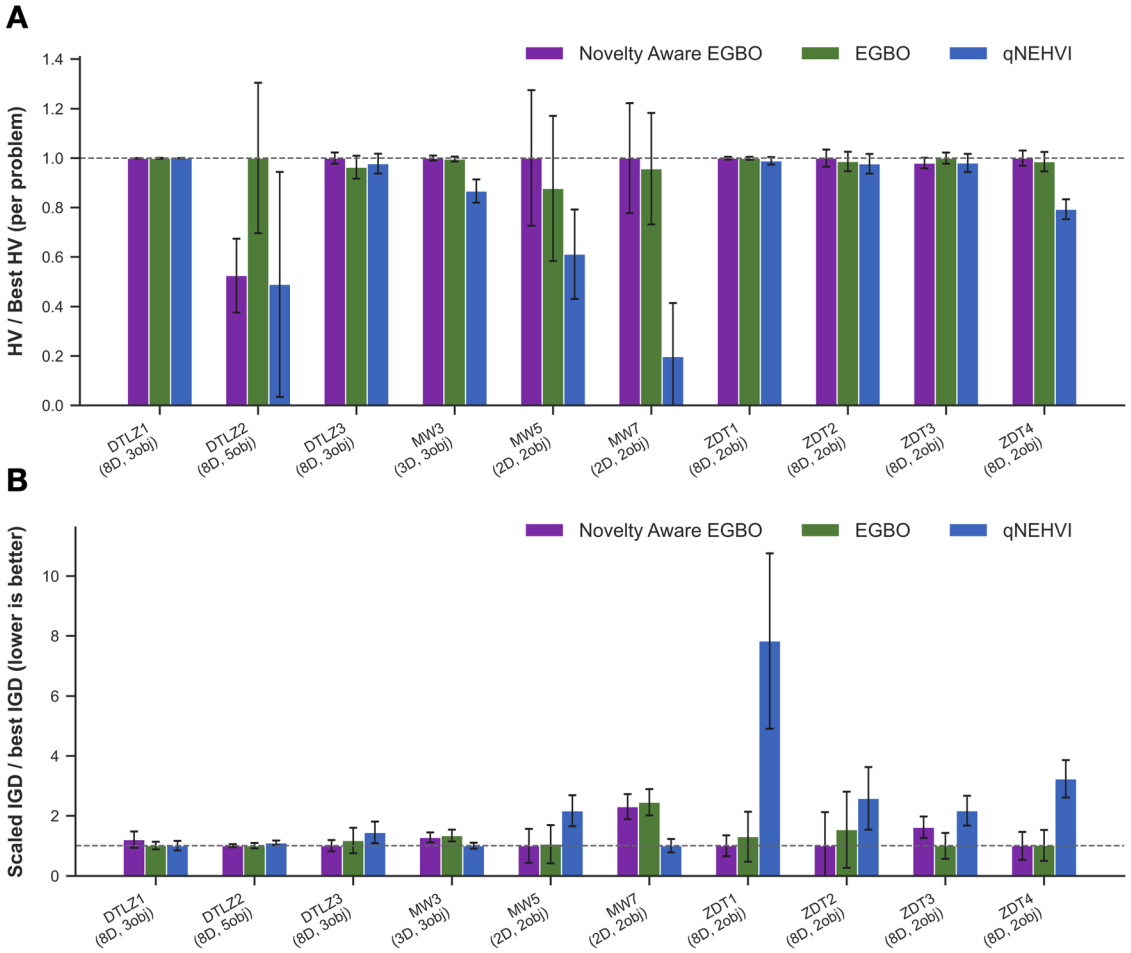

Fig. 8 Novelty-aware EGBO compared to EGBO alone and qLogNEHVI for 10 test problems. Showing (A) Hypervolume, (B) normalised IGD score.

we have extracted drug response for 5 genetically diverse colorectal cancer cell lines, and each cell line is treated as an independent optimisation objective. These datasets collectively represent a range of optimisation regimes encountered in experimental discovery workflows, including varying dimensionalities, objective trade-offs, and noise characteristics.

For each dataset we ran a post-hoc optimisation campaign in which optimisation algorithms sequentially selected candidate experiments from the existing dataset as if they were conducting a real experimental campaign. At each iteration, the selected sample was revealed from the dataset and used to update the surrogate model, allowing optimisation performance to be evaluated under realistic experimental budgets without performing additional laboratory experiments. We compared the performance of qLogNEHVI, EGBO, and Novelty-aware EGBO based optimisations. We ran 10 repeats of campaigns using 12 batches of 4 samples (in line with the original EGBO publication), where starting samples were shared between algorithms to ensure fairness.

Across the four experimental datasets Novelty-aware EGBO consistently achieved the highest HV values, outperforming both original EGBO and qLogNEHVI (Fig. 9A). Novelty-aware EGBO achieved the highest mean HV on three of the four datasets and remained competitive on the fourth, and overall is significantly superior to qLogNEHVI (Holm p *vs.* EGBO $= 0.00022$, Holm p *vs.* qLogNEHVI $= 0.0035$). This is also reflected in significantly lower IGD values for Novelty-aware EGBO compared to original EGBO and qLogNEHVI across all datasets (Holm p *vs.* EGBO $= 0.00032$, Holm p *vs.* qLogNEHVI $= 0.01$) (Fig. 9B).

These results suggest that the novelty-aware allocation strategy improves optimisation performance in realistic discovery settings where experimental noise, heterogeneous objective landscapes, and limited evaluation budgets are common. By incorporating information on the novelty of each potential samples alongside acquisition-driven exploitation and evolutionary exploration, Novelty-aware EGBO appears better able to maintain diversity while still prioritising promising regions of the design space.

Importantly, these observations are consistent with the trends observed in the synthetic benchmark experiments. In both settings, hybrid optimisation approaches provide the greatest benefit in complex optimisation landscapes, where maintaining diversity in candidate generation helps prevent

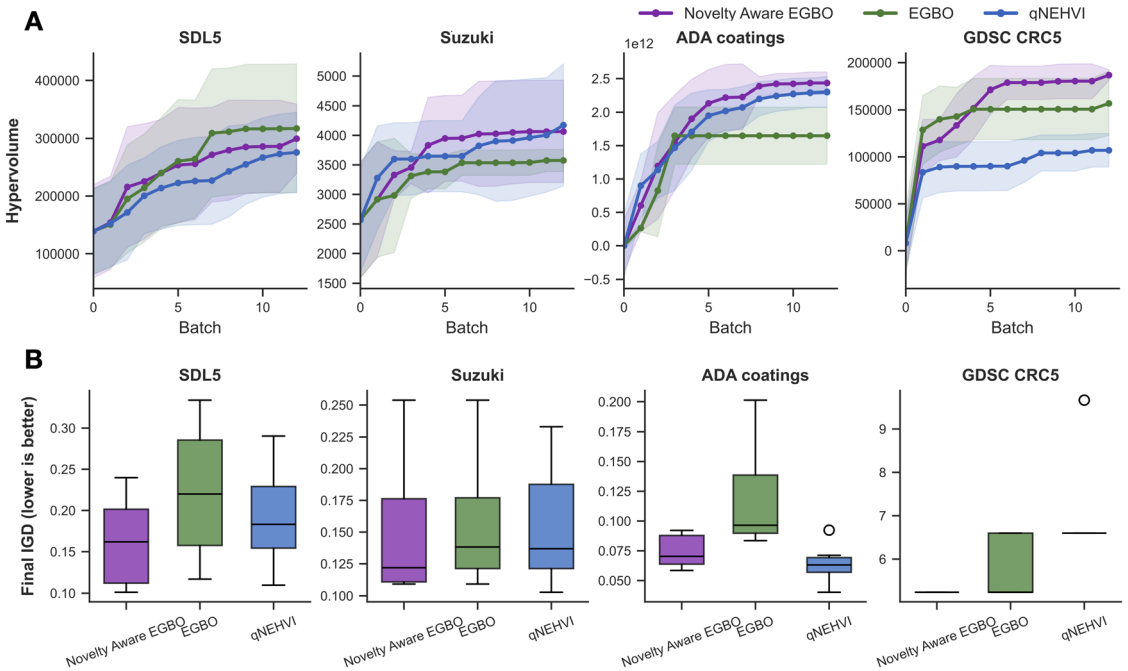

Fig. 9 Novelty-aware EGBO compared to EGBO alone and qLogNEHVI for 4 post-hoc real world problems showing (A) mean hypervolume traces, (B) IGD.

premature convergence and improves exploration of the objective space. The novelty-aware variant further enhances this behaviour by penalising candidates that are near to others in the normalised decision space, ensuring that samples selected concurrently are distributed across distinct regions of the input space, promoting broader exploration of the Pareto front.

Taken together, these results demonstrate that hybrid evolutionary-Bayesian optimisation strategies not only perform well on synthetic benchmarks but also translate effectively to real-world discovery problems, including reaction optimisation, formulation development, materials design, and multi-objective drug screening. This suggests that such approaches

Table 1 Optimisation problem choices

Problem	Features	Objectives	Reason chosen
ZDT1	8	2	Smooth, convex Pareto front; baseline test of standard multi-objective convergence
ZDT2	8	2	Non-convex Pareto front; tests whether methods can recover curved/non-convex trade-offs
ZDT3	8	2	Disconnected Pareto front; tests front coverage across separated regions
ZDT4	8	2	Highly multimodal landscape; stresses robustness to local optima and deceptive structure
DTLZ1	8	3	3-Objective benchmark with broad trade-off structure; evaluates extension beyond 2-objective settings
DTLZ3	8	3	Multimodal many-local-front variant; harder convergence challenge in 3-objective optimisation
DTLZ2 5obj	14	5	Many-objective setting; tests scalability of surrogate + selection under higher objective dimensionality
MW3	8	2	Constrained benchmark with nonlinear feasibility boundaries; tests constrained BO behavior
MW5	8	2	Constrained benchmark with multiple nonlinear constraints and reduced feasible area; tests feasibility search under tighter constraints
MW7	8	2	Constrained benchmark with challenging feasible geometry; tests stability of constrained exploration/exploitation

may provide a practical framework for improving optimisation performance in autonomous laboratories and data-driven experimental workflows, where limited experimental budgets and complex objective landscapes are common.

3 Conclusion

In this work, we systematically evaluated hybrid evolutionary-Bayesian optimisation strategies for multi-objective discovery problems. By incorporating evolutionary candidate generation into acquisition-driven optimisation workflows, hybrid approaches consistently improved optimisation performance in challenging regimes, particularly for many-objective and multimodal problems where traditional acquisition-only optimisation can become overly exploitative. Across a range of benchmark landscapes, evolutionary generators improved Pareto-front coverage, convergence reliability, and hypervolume expansion while maintaining comparable performance on simpler optimisation tasks.

Our results further show that these benefits arise from the complementary roles of acquisition-driven and evolutionary search. Evolutionary algorithms provide broad exploration of the objective space, while acquisition optimisation refines promising regions of the design landscape. Among the evolutionary algorithms evaluated, U-NSGA-III provided the most consistent Pareto-front coverage, while combining multiple evolutionary generators produced only marginal improvements relative to the additional computational cost. In contrast, these hybrid benefits diminished in very high-dimensional feature spaces, where broad evolutionary exploration significantly reduced sample efficiency. We also demonstrate that introducing a novelty-aware batch-selection strategy improves optimisation efficiency by promoting diversity within candidate batches, reducing optimisation stagnation and improving Pareto-front approximation. Whilst these improvements are modest – showing only slight improvements in hypervolume and IGD, they open an avenue for adjusting the sample selection method in a new way which could lead to further improvements with more sophisticated methods.

Importantly, these findings extend beyond synthetic benchmarks and translate effectively to real-world experimental optimisation problems, where hybrid optimisation strategies consistently outperformed traditional Bayesian optimisation approaches. These results are particularly relevant in the context of self-driving laboratories, where each optimisation batch corresponds to a set of physical experiments. In such settings, optimisation stagnation represents wasted experimental resources, including reagents, instrument time, and researcher effort. Hybrid optimisation strategies that maintain diversity in candidate generation therefore have practical advantages beyond purely computational metrics, as they reduce the likelihood of repeatedly sampling similar regions of the design space and improve the probability of discovering diverse high-performing solutions.

More broadly, this work suggests that diversity should be treated as a design principle in batched multi-objective optimisation, rather than as a by-product of acquisition maximisation alone. In practical terms, hybrid evolutionary-Bayesian approaches appear most valuable for discovery campaigns with complex trade-offs, constrained feasible regions, or rugged optimisation landscapes, whereas simpler acquisition-only methods may remain preferable in very high-dimensional settings where sample efficiency is paramount.

Future work should focus on extending these frameworks to higher-dimensional experimental design spaces, developing more adaptive selection rules that adjust exploration pressure during the campaign, and validating these methods prospectively in live self-driving laboratory workflows. Together, these results demonstrate that evolutionary-assisted Bayesian optimisation provides a robust and practically useful strategy for navigating complex multi-objective design spaces in autonomous discovery.

4 Methods

To enable reproducibility, we have added code to run all the frameworks described, as well as cumulative data from all experimental run into a Github repository available at: https://github.com/shorthouse-lab/Aqeeli_NoveltyAware_EGBO and a Zenodo repository available at: https://zenodo.org/records/20321979.

4.1 Optimisation framework

Our framework performed closed-loop, batched multi-objective optimisation. In each cycle, a surrogate model was fitted to all data collected to date, a batch of candidate experiments proposed, those experiments evaluated, and the model updated before the next cycle is initiated. Each system incorporated either a traditional qLogNEHVI acquisition strategy or an evolutionary hybrid. To ensure fair comparison and reproducibility, the following properties were enforced across all campaigns:

(i) All systems shared identical randomly-selected initialisation sets (*e.g.*, the same 10 initial sample sets were used for every algorithm tested on MW5).

(ii) Batch size was held constant across systems and problems;

(iii) Total experimental budget (number of batches) was equal for all systems and problems.

For all synthetic test problem assessments, we used 18 initial samples, 10 repeats of each campaign using consistent initial samples, with a batch size of 8, and collecting 12 batches for each campaign. For the post-hoc analysis on real world datasets, in line with experimental protocols in the original EGBO project, we performed 10 repeats of 18 initial samples, with 12 batches of a batch size 4.

4.2 Surrogate models

Each objective, and each constraint where applicable, was modelled using an independent Gaussian process (GP) refit after every batch on the full accumulated training set. Input variables were normalised to the unit hypercube prior to modelling. Each output was represented using a separate

SingleTaskGP in BoTorch, and individual models were combined in a ModelListGP. A standardising outcome transform was used to centre and scale each output. GP hyperparameters, including kernel lengthscales and observation noise, were estimated by maximising the summed marginal log-likelihood using L-BFGS-B *via* fit_gpytorch_mll. The fitted posterior was then used to construct a qLogNoisyExpectedHypervolumeImprovement (qLogNEHVI) acquisition function, evaluated using Sobol quasi-Monte Carlo sampling. All GP models used the Matern 5/2 kernel fit using default settings.

4.3 Evolutionary generators

Evolutionary candidate proposals were generated using population-based multi-objective optimisers applied in the normalised design space. At the start of each batch, the non-dominated observed points were extracted from the current training set and used to seed the initial evolutionary population; when fewer non-dominated points were available than required, the remainder of the population was filled by random samples within the feasible bounds. Each generator was then run for a fixed number of generations to produce a candidate pool for the current batch. We evaluated three algorithms: U-NSGA-III, which uses reference directions to preserve diversity across the objective space; SMS-EMOA, which uses hypervolume contribution for survival; and AGE-MOEA-II, which uses geometry-aware diversity preservation. All generators were run under the same variable bounds and comparable population sizes, and their outputs were used only as candidate proposals rather than as direct replacements for surrogate-guided acquisition optimisation. We have included details of the model implementation including the number of samples generated, pooled, and selected for each experiment, including which datasets were studied in Table S1.

4.4 Hybrid optimisation frameworks

The hybrid frameworks fuse surrogate-driven acquisition with population-based evolutionary generation to combine exploitation and exploration within each batch. At each iteration, the fitted GP surrogates are used to simultaneously generate two candidate pools: a small set of exploitation proposals obtained by directly optimising the qLogNEHVI acquisition function and a larger exploration pool produced by the evolutionary generator seeded from the current Pareto set. The two pools are concatenated, every candidate is scored by the same qLogNEHVI acquisition function, and the top B candidates by acquisition value are selected to form the evaluation batch; this shared scoring criterion ensures the two sources are compared on a common basis. Four framework variants were evaluated:

(i) Traditional qLogNEHVI, which generates candidates by acquisition optimisation only (no evolutionary generator).

(ii) Evolutionary generator coupled, which merges a fixed-size qLogNEHVI pool with a U-NSGA-III pool and ranks by raw acquisition.

(iii) Multi-generator coupled, which combines qLogNEHVI with two or more EA engines (*e.g.*, U-NSGA-III and SMS-EMOA) to further broaden coverage; and.

(iv) Novelty aware evolutionary generator coupled, which used the same merged pool as EGBO but selected the final batch using a novelty-aware greedy downselection procedure rather than pure acquisition ranking.

All variants used the same surrogate modelling procedure, batch size, and evaluation budget, so differences in performance arose solely from differences in candidate generation and batch selection.

4.5 Novelty aware EGBO

For the novelty-aware EGBO variant, we retained the same merged candidate-generation framework as standard EGBO but replaced pure acquisition-based top-B selection with a novelty-aware greedy downselection procedure. At each batch, GP surrogates were refit on all accumulated observations and a qLogNEHVI acquisition function was constructed as above. Two candidate pools were then generated in normalised decision space: (i) a qLogNEHVI pool of size B, obtained by joint acquisition optimisation, and (ii) an exploratory pool produced by one generation of U-NSGA-III with population size 256, seeded from the current non-dominated observed set. The two pools were merged, duplicate candidates were removed, and all remaining candidates were rescored using qLogNEHVI.

Final batch construction was then performed greedily. At each selection step, each candidate i in the merged pool was assigned a combined score:

$$\mathrm{Score}_i = w\tilde{a}_i + (1 - w)\tilde{n}_i$$

where $\tilde{a}_i$ is the min–max normalised qLogNEHVI acquisition value of candidate i, $\tilde{n}_i$ is the min–max normalised novelty score, and w controls the trade-off between acquisition merit and novelty, and was fixed at 0.7 for all experiments. w was fixed *a priori* and not tuned. Novelty was defined as the minimum Euclidean distance in normalised decision space from candidate i to any previously evaluated point or any point already selected for the current batch. After each batch member was selected, novelty scores were recomputed for the remaining candidates before the next selection step. This source-blind scoring rule allowed candidates from either generator to be selected while discouraging within-batch redundancy and resampling of near-duplicate designs.

4.6 Benchmark problems

We used 10 problems as benchmarks across the study – chosen from a range of problems available in pymoo. Table 1 details each problem, its number of features and objectives, and a note on why it was chosen.

To assess robustness beyond nominal settings, we performed three targeted stress tests. First, we evaluated noise robustness by adding zero-mean Gaussian perturbations to observed objective/constraint values at controlled levels (0–20% relative noise), while keeping initial seeds, batch size, and evaluation budget fixed across algorithms. Second, we evaluated constraint handling on constrained MW benchmarks (MW3, MW5, MW7), including modified MW5 variants with adjusted

constraint tightness (*e.g.*, tighter and looser feasible regions), to test each method's ability to discover and improve within limited feasible domains. Third, we evaluated high-dimensional scaling by increasing decision-space dimensionality (*e.g.*, from standard 8D settings to higher-dimensional variants such as 50D/100D), while maintaining the same closed-loop protocol, to quantify how performance and stability change as the feature space grows. Together, these analyses isolate sensitivity to measurement noise, feasibility geometry, and dimensionality-driven search complexity.

4.7 Real-world datasets and post-hoc optimisation

Real-world validation was conducted on four datasets:

- Suzuki: a Suzuki–Miyaura cross-coupling reaction originally reported by Reizman and Jensen[35] and included in the Summit python package.[38]

- SDL5: a microparticle formulation campaign derived from experimental data reported from an automated lab generating long-acting injectable formulations.[16]

- ADA coatings: an industrial coating formulation optimisation dataset from the ADA database.[36]

- GDSC CRC5: a drug screening dataset from the Genomics of Drug Sensitivity in Cancer (GDSC) project.[37]

The Suzuki, SDL5, and ADA coatings databases were used as-is, with every sample in the dataset used as input for modellig. For the CDSC CRC5 dataset, the Genomics of Drug Sensitivity in cancer dataset was downloaded and subset into only cells from colorectal adenocarcinoma. We then selected the 5 cell lines with the most coverage: SNU-C1, LS-1034, LS-513, LS-123, NCI-H747. For all cells lines we obtained the IC50 (the concentration required to kill 50% of the cells) of 349 drugs, and the optimisation task was set to uncover the drugs with minimal IC50 (lowest concentration needed to kill 50% of cells), treating each of the 5 cell lines as a separate objective. Drug response was represented using IC_{50} values across these five cell lines, with each cell line treated as a separate optimisation objective. Because of the limited sample size, drugs were represented using one-hot encoded ontology features derived from annotated putative targets and pathway labels rather than higher-dimensional SMILES-derived descriptor sets.

Because these datasets are retrospective, we used a post-hoc closed-loop protocol in which each algorithm sequentially proposes batches from the pool of unqueried experiments; selected samples are then "revealed" from the historical dataset and appended to the training set for the next iteration. This emulates practical batched Bayesian optimisation while preserving full comparability across methods. The same core protocol used for synthetic benchmarks was maintained in post-hoc studies, including matched initial seeds, fixed batch size and iteration budget, and multiple repeated runs, enabling direct, controlled performance comparisons between algorithms across synthetic and real-world settings. Due to these datasets being tabular by nature (rather than a continuous state space), for our post-hoc analysis we optimised in the normalised design space and then mapped each proposed offspring to the discrete experimental set through nearest-neighbour oracles at evaluation, with candidate selection preferring points that map to previously unseen dataset rows.

5 Evaluation metrics

We evaluated algorithm quality using four complementary criteria.

5.1 Hypervolume (HV)

HV was computed as the dominated volume of the approximation set relative to a fixed reference point in objective space. Larger HV indicates better joint convergence and diversity. Hypervolume was calculated related to a reference point representing a sample in the non-optimal area of state space. This reference point was shared across algorithms, and kept consistent across campaigns.

5.2 Inverted generational distance (IGD)

IGD was used to quantify distance to a high-quality reference Pareto front calculated as either the true Pareto front for existing problems, or the best samples for post-hoc analysis. Lower IGD indicates better convergence to the target front. Distances were computed in normalized objective space to ensure comparability across problems. For synthetic problems, the reference set was derived from the true Pareto front; for retrospective datasets, the reference set was constructed from the non-dominated subset of the full dataset. Lower IGD indicates closer approximation to the target front.

5.3 Pareto spread and convergence

We separately characterized:

- Convergence: trajectory of HV (and IGD) across batches, plus summary statistics over the optimization horizon.
- Spread/diversity: uniformity of solutions along the non-dominated front (*e.g.*, spacing/dispersion from nearest-neighbor distances in objective space), where lower dispersion indicates more even coverage.

5.4 Feasible sample count (constrained settings)

For constrained benchmarks, we tracked the number (and proportion) of evaluated points satisfying all constraints, this metric captures practical constraint-handling effectiveness independent of objective quality.

Additionally, we calculated an exploration score to assess how much algorithms were exploring new samples. This score was based on the acquisition-rank percentile of the candidates chosen for evaluation. For each trial, the percentile ranks of the selected points within the acquisition-ordered candidate pool were averaged to obtain a mean acquisition percentage for the selection (mean_selected_acq_percentile), and this was converted to an exploration metric as:

$$\text{Exploration score} = 1 - \text{mean\_selected\_acquisition\_percentile}$$

Thus, higher values indicate that the algorithm more often selected lower-ranked acquisition candidates, consistent with greater exploratory behaviour, whereas lower values indicate stronger exploitation of top-ranked acquisition candidates. Reported values correspond to the mean and standard deviation of this per-trial score across repeated runs.

5.5 Statistical analysis

Statistical comparisons were performed using a nonparametric matched design implemented in SciPy.[39] For each metric, algorithms were first ranked within matched experimental blocks and compared using a Friedman test to detect overall differences without assuming normality. Matched blocks were defined as problem x replicate for synthetic benchmarks and dataset x replicate for retrospective studies. When the Friedman test was significant, pairwise two-sided Wilcoxon signed-rank tests were applied to matched observations from the same block. Holm step-down correction was used to control the family-wise error rate across pairwise comparisons. Unless otherwise stated, statistical tests were performed on final campaign metrics rather than intermediate batch values.

5.6 Implementation

All experiments were implemented in Python, with Bayesian optimization components built on PyTorch[40] using BoTorch[41] and GPyTorch,[42] and evolutionary search implemented with pymoo.[43] To ensure reproducibility and fair cross-method comparison, runs used fixed random seeds and matched initial conditions across algorithms (*i.e.*, identical trial-level initialization per replicate). Core outputs (*e.g.*, per-trial performance traces and summary tables) were generated programmatically from the same pipeline. Code for reproduction and core outputs (tables of per-run algorithm performance) are available at a Github repository for this project: https://github.com/shorthouse-lab/Aqeeli_NoveltyAware_EGBO.

Author contributions

M. A.: methodology, software, investigation, writing – review and editing. T. L.: software, investigation. D. S.: conceptualization, methodology, investigation, writing – original draft, writing – review and editing, supervision.

Conflicts of interest

David Shorthouse is a cofounder of Implexis ltd. No other conflicts of interest are declared.

Data availability

Code and data generated as part of this manuscript are available at a Zenodo repository: https://doi.org/10.5281/zenodo.20321978. This includes aggregate data from optimisations runs suitable for reproducing analysis and plots, as well as code and datasets for running the analysis.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6dd00134c.

References

1 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid and A. Aspuru-Guzik, Self-Driving Laboratories for Chemistry and Materials Science, *Chem. Rev.*, 2024, **124**, 9633–9732, DOI: 10.1021/acs.chemrev.4c00055.

2 J. A. Bennett and M. Abolhasani, Autonomous chemical science and engineering enabled by self-driving laboratories, *Curr. Opin. Chem. Eng.*, 2022, **36**, 100831, DOI: 10.1016/j.coche.2022.100831.

3 M. Abolhasani and E. Kumacheva, The rise of self-driving labs in chemical and materials sciences, *Nat. Synth.*, 2023, **2**, 483–492, DOI: 10.1038/s44160-022-00231-0.

4 B. P. MacLeod, F. G. L. Parlane, C. C. Rupnow, K. E. Dettelbach, M. S. Elliott, T. D. Morrissey, T. H. Haley, O. Proskurin, M. B. Rooney, N. Taherimakhsousi, D. J. Dvorak, H. N. Chiu, C. E. B. Waizenegger, K. Ocean, M. Mokhtari and C. P. Berlinguette, A self-driving laboratory advances the Pareto front for material properties, *Nat. Commun.*, 2022, **13**, 995, DOI: 10.1038/s41467-022-28580-6.

5 H. Hysmith, E. Foadian, S. P. Padhy, S. V. Kalinin, R. G. Moore, O. S. Ovchinnikova and M. Ahmadi, The future of self-driving laboratories: from human in the loop interactive AI to gamification, *Digit. Discov.*, 2024, **3**, 621–636, DOI: 10.1039/D4DD00040D.

6 A. A. Volk and M. Abolhasani, Performance metrics to unleash the power of self-driving labs in chemistry and materials science, *Nat. Commun.*, 2024, **15**, 1378, DOI: 10.1038/s41467-024-45569-5.

7 H. Ros, Y. Abdalla, M. T. Cook and D. Shorthouse, Efficient discovery of new medicine formulations using a semi-self-driven robotic formulator, *Digit. Discov.*, 2025, **4**, 2263–2272, DOI: 10.1039/D5DD00171D.

8 S. Greenhill, S. Rana, S. Gupta, P. Vellanki and S. Venkatesh, Bayesian Optimization for Adaptive Experimental Design: A Review, *IEEE Access*, 2020, **8**, 13937–13948, DOI: 10.1109/ACCESS.2020.2966228.

9 P. I. Frazier and J. Wang, *Bayesian optimization for materials design*, in Springer Series in Materials Science, 2015, DOI: 10.1007/978-3-319-23871-5_3.

10 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, *Proc. IEEE*, 2016, **104**, 148–175, DOI: 10.1109/JPROC.2015.2494218.

11 A. D. Adesiji, J. Wang, C.-S. Kuo and K. A. Brown, Benchmarking self-driving labs, *Digit. Discov.*, 2026, **5**, 14–27, DOI: 10.1039/D5DD00337G.

12 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, Bayesian reaction optimization as a tool for chemical

synthesis, *Nature*, 2021, **590**, 89–96, DOI: 10.1038/s41586-021-03213-y.
13 J. A. Bennett, N. Orouji, M. Khan, S. Sadeghi, J. Rodgers and M. Abolhasani, Autonomous reaction Pareto-front mapping with a self-driving catalysis laboratory, *Nat. Chem. Eng.*, 2024, **1**, 240–250, DOI: 10.1038/s44286-024-00033-5.
14 T. Lookman, P. V. Balachandran, D. Xue and R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design, *npj Comput. Mater.*, 2019, **5**, 21, DOI: 10.1038/s41524-019-0153-8.
15 M. J. Tamasi, R. A. Patel, C. H. Borca, S. Kosuri, H. Mugnier, R. Upadhya, N. S. Murthy, M. A. Webb and A. J. Gormley, Machine Learning on a Robotic Platform for the Design of Polymer–Protein Hybrids, *Adv. Mater.*, 2022, **34**, 2201809, DOI: 10.1002/adma.202201809.
16 Z. Bao, F. Le Devedec and S. Huynh, A Data-Driven Workflow for Nanomedicine Optimization Using Active Learning and Automated Experimentation, *Mol. Pharm.*, 2025, **22**, 7465–7477, DOI: 10.1021/acs.molpharmaceut.5c00933.
17 A. K. Y. Low, F. Mekki-Berrada, A. Gupta, A. Ostudin, J. Xie, E. Vissol-Gaudin, Y.-F. Lim, Q. Li, Y. S. Ong, S. A. Khan and K. Hippalgaonkar, Evolution-guided Bayesian optimization for constrained multi-objective optimization in self-driving labs, *npj Comput. Mater.*, 2024, **10**, 104, DOI: 10.1038/s41524-024-01274-x.
18 S. T. Knox, S. J. Parkinson, C. Y. P. Wilding, R. A. Bourne and N. J. Warren, Autonomous polymer synthesis delivered by multi-objective closed-loop optimisation, *Polym. Chem.*, 2022, **13**, 1576–1585, DOI: 10.1039/D2PY00040G.
19 M. R. Athavale, R. Al-Abri, S. Church, W. W. Wong, A. K. Low, H. H. Tan, K. Hippalgaonkar and P. Parkinson, Accelerated Design of Microring Lasers with Multi-Objective Bayesian Optimization, *arXiv*, 2024, preprint arXiv:2411.04487, DOI: 10.48550/arXiv.2411.04487.
20 A. Panichella, An improved Pareto front modeling algorithm for large-scale many-objective optimization, in, *GECCO 2022 - Proceedings of the 2022 Genetic and Evolutionary Computation Conference*, 2022, DOI: 10.1145/3512290.3528732.
21 H. Seada and K. Deb, U-NSGA-III: A Unified Evolutionary Algorithm for Single, Multiple, and Many-Objective Optimization, in, *International Conference on Evolutionary Multi-Criterion Optimization*, 2015.
22 N. Beume, B. Naujoks and M. Emmerich, SMS-EMOA: Multiobjective selection based on dominated hypervolume, *Eur. J. Oper. Res.*, 2007, **181**, 1653–1669, DOI: 10.1016/j.ejor.2006.08.008.
23 J. Liu, R. Sarker, S. Elsayed, D. Essam and N. Siswanto, Large-scale evolutionary optimization: A review and comparative study, *Swarm Evol. Comput.*, 2024, **85**, 729–745, DOI: 10.1016/j.swevo.2023.101466.
24 N. Maus, K. Wu, D. Eriksson and J. Gardner, Discovering Many Diverse Solutions with Bayesian Optimization Proc. Mach. Learn. Res., *arXiv*, 2023. **206**, preprint arXiv:2210.10953, DOI: 10.48550/arXiv.2210.10953.
25 W.-T. Tang, A. Chakrabarty and J. A. Paulson, *BEACON: A Bayesian Optimization Strategy for Novelty Search in Expensive Black-Box Systems*, (2025).
26 A. Biswas, R. Vasudevan, R. Pant, I. Takeuchi, H. Funakubo and Y. Liu, SANE: strategic autonomous non-smooth exploration for multiple optima discovery in multi-modal and non-differentiable black-box functions, *Digit. Discov.*, 2025, **4**, 853–867, DOI: 10.1039/d4dd00299g.
27 R. Bulanadi, J. Chowdhury, H. Funakubo, M. Ziatdinov, R. Vasudevan, A. Biswas and Y. Liu, Beyond Optimization: Exploring Novelty Discovery in Autonomous Experiments, *ACS Nanosci. Au*, 2026, **6**(1), 86–94, DOI: 10.1021/acsnanoscienceau.5c00106.
28 K. Deb, L. Thiele, M. Laumanns and E. Zitzler, Scalable Test Problems for Evolutionary Multiobjective Optimization, in, *Evolutionary Multiobjective Optimization*, 2005, DOI: 10.1007/1-84628-137-7_6.
29 E. Zitzler, K. Deb and L. Thiele, Comparison of multiobjective evolutionary algorithms: empirical results, *Evol. Comput.*, 2000, **8**, 173–195, DOI: 10.1162/106365600568202.
30 Z. Ma and Y. Wang, Evolutionary constrained multiobjective optimization: Test suite construction and performance comparisons, *IEEE Trans. Evol. Comput.*, 2019, **23**, 972–986, DOI: 10.1109/TEVC.2019.2896967.
31 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 Trans. Evol. Comput.*, 2014, **18**, 577–601, DOI: 10.1109/TEVC.2013.2281535.
32 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 Trans. Evol. Comput.*, 2014, **18**, 602–622, DOI: 10.1109/TEVC.2013.2281534.
33 C. A. C. Coello and M. R. Sierra, A study of the parallelization of a coevolutionary multi-objective evolutionary algorithm, in, *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2004, DOI: 10.1007/978-3-540-24694-7_71.
34 J. Knowles, ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems, *IEEE Trans. Evol. Comput.*, 2006, **10**, 50–66, DOI: 10.1109/TEVC.2005.851274.
35 B. J. Reizman, Y. M. Wang, S. L. Buchwald and K. F. Jensen, Suzuki-Miyaura cross-coupling optimization enabled by automated feedback, *React. Chem. Eng.*, 2016, **1**, 658–666, DOI: 10.1039/c6re00153j.
36 C. C. Rupnow, B. P. MacLeod, M. Mokhtari, K. Ocean, K. E. Dettelbach, D. Lin, F. G. L. Parlane, H. N. Chiu, M. B. Rooney, C. E. B. Waizenegger, E. I. de Hoog, A. Soni and C. P. Berlinguette, A self-driving laboratory optimizes a scalable process for making functional coatings, *Cell Rep. Phys. Sci.*, 2023, **4**(5), DOI: 10.1016/j.xcrp.2023.101411.
37 W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson,

S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott and M. J. Garnett, Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells, *Nucleic Acids Res.*, 2013, **41**, D955–D961, DOI: 10.1093/nar/gks1111.

38 K. C. Felton, J. G. Rittig and A. A. Lapkin, Summit: Benchmarking Machine Learning Methods for Reaction Optimisation, *Chem. Methods.*, 2021, **1**, 116–122, DOI: 10.1002/cmtd.202000051.

39 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. Pietro Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko and Y. Vázquez-Baeza, SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nat. Methods*, 2020, **17**, 261–272, DOI: 10.1038/s41592-019-0686-2.

40 S. Imambi, K. B. Prakash and G. R. Kanagachidambaresan, PyTorch, in, *EAI/Springer Innovations in Communication and Computing*, 2021, DOI: 10.1007/978-3-030-57077-4_10.

41 M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson and E. Bakshy, BOTORCH: A framework for efficient Monte-Carlo Bayesian optimization, in, *Adv. Neural Inf. Process. Syst.*, 2020.

42 J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger and A. G. Wilson, Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration, in, *Adv. Neural Inf. Process. Syst.*, 2018.

43 J. Blank and K. Deb, Pymoo: Multi-Objective Optimization in Python, *IEEE Access*, 2020, **8**, 89497–89509, DOI: 10.1109/ACCESS.2020.2990567.