A Electronic Supporting Information: Bayesian optimisation for additive screening and yield improvements - beyond one-hot encoding

A.1 Bayesian optimisation algorithm

The BO algorithm consists of two core components, a surrogate model and an acquisition function. The surrogate is typically a probabilistic model such as a Gaussian process[1] that captures the prior belief about the nature of $f(\mathbf{x})$. The uncertainty estimates afforded by the surrogate are crucial in representing knowledge about the unobserved values of the black-box function $f\mathbf{x}$ and act to inform further data collection via a policy known as an acquisition function. The acquisition function, $\alpha(f(\mathbf{x}),\mathcal{D})$, is responsible for selecting the next data point on a given iteration of the BO algorithm. The acquisition function achieves this by leveraging the uncertainty estimates of the surrogate to trade off between exploration and exploitation in the black-box objective $f(\mathbf{x})$. The acquisition function should be cheaper to evaluate relative to the black-box and easy to optimise[2–4]. Modifications to classical BO surrogates, which typically assume design spaces, $\mathcal{X}$ that are compact subsets of $\mathbb{R}^d$, are necessary to operate on molecular spaces. We utilise such models in this work[5–7]. The pseudocode for BO is given in Algorithm 1.

Algorithm 1 Bayesian optimisation (BO)

input: initial dataset $\mathcal{D}$ ▷ possibly empty
repeat
select $\mathbf{x}$ by optimising the acquisition function α

$$\mathbf{x} \leftarrow \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha(\mathbf{x};\mathcal{D})$$

$y \leftarrow$ Evaluate($\mathbf{x}$) ▷ evaluate the black-box at the selected input
$\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{x},y)\}$ ▷ update dataset and surrogate
until termination condition ▷ e.g. evaluation budget exhausted
return $\mathcal{D}$

A.2 Data preprocessing

The four reactions used in this paper are as follows:
Reaction plate 1: Informer X2[8] with cyclohexanoic acid - This experiment explores the role of small-molecule additives on the decarboxylative C-C coupling of Informer X2 and cyclohexanoic acid, a non-activated secondary carboxylic acid. Different additives can have significant effects on the outcome of this reaction, with the potential for improved efficiency and yield.
Reaction plate 2: 3-bromo-5-phenylpyridine with cyclohexanoic acid shows the influence of diverse additives on the decarboxylative C-C coupling between 3-bromo-5-phenylpyridine, a substrate prone to dehalogenation and Minisci radical addition side-products, and cyclohexanoic acid, a non-activated secondary carboxylic acid. The goal is to identify additives that can mitigate side reactions and improve overall reactivity.
Reaction plate 3: 7-bromo-3,4-dihydroquinolin-2(1H)-one with cyclohexanoic acid - This experiment examined the role of small-molecule additives in the decarboxylative C-C coupling of 7-bromo-3,4-dihydroquinolin-2(1H)-one, an electron-rich and slowly reactive substrate, with cyclohexanoic acid. Identifying the right additives can significantly enhance the reaction rate and yield.
Reaction plate 4: Informer X2 with hexanoic acid - This reaction assessed the impact of diverse small-molecule additives on the decarboxylative C-C coupling between Informer X2, a typically highly reactive substrate, and hexanoic acid, a non-activated primary carboxylic acid.

We extracted reactions from the four plates used in the Ni-catalysed photoredox decarboxylative arylations study[9]. We combined all columns containing molecular SMILES to form a reaction SMILES. The reactants and reagents were added on the left side of the "»" and the products on the right side. We excluded the solution SMILES while building the reaction smiles and kept only the reactant, reagent and additives. We canonicalized the molecular SMILES with RDKit[10]. When a reaction with the same additive was run multiple times within a plate we averaged the yield values. Prieto Kullmer *et al.*[9] approximated the yields using "UV210_Prod AreaAbs" and estimated the yield improvements by dividing through a baseline reaction without the additive.

A.3 Reproducibility

We implemented the code in Python. We used PyTorch[11], PyTorch Lightning[12], Gauche[6] and BoTorch[13]. Different trials were set using a set of seeds ranging from 1 to 20 with PyTorch Lightning's seed_everything function. The surrogate model used is a Gaussian process implemented with SingleTaskGP from Botorch. We used the Tanimoto kernel[7,14] from GAUCHE and Linear and Matern kernels from GPyTorch[15]. We encoded reaction representations with DRFP[16] (512 dimensional with bond radius equaling 7), RXNFP[17], CDDD[18], XTB[19], ChemBERTa[20], fingerprints and fragprints (512-dimensional with bond radius equaling 3) from RDKit under GauChe interface[6]. We tried acquisition functions: UpperConfidenceBound and ExpectedImprovement from BoTorch. We used 0.1 for UCB beta value.

A.4 Extended analysis of parameter impact on BO performance and model fit metrics

As demonstrated in the main text, various parameters significantly influence the performance of Bayesian optimisation in additive screening. These parameters include data representation methods, kernels, initialisation strategies, and acquisition functions. In the main manuscript, we have focused on the performance of different data representations in terms of top-5 discovery, averaged across various kernels, acquisition functions, and initialisation strategies. We have also analysed the influence of combining specific kernels with particular representations for top-5 additive discovery. Here we extend this analysis by visualising the results presented in Table 2 of the main paper, along with additional performance metrics in Figure 1. We also consider two additional metrics and parameters: 'average similarity' and 'noise,' which provide more nuanced insights into the model's behaviour. Noise represents the noise associated with the observations in the data, modelled by the surrogate model. Understanding this metric can provide valuable information on the reliability of the model's predictions. The average similarity metric quantifies the mean similarity between all pairs of data points in the feature space as defined by the kernel function of the Gaussian process surrogate model. This metric is normalised between 0 and 1, allowing us to understand the general level of similarity or dissimilarity among observations. A higher average similarity value indicates that the data points are, on average, more similar to each other, which may influence the performance of the Bayesian optimisation algorithm. Consequently and aligned with our expectation, OHE representation results in average similarity of 0, proving the orthogonality of the design space built by this representation in our dataset. We show averaged results on all these metrics and parameters in Table 2.

We notice several emerging patterns from the extended analysis. DRFP consistently outperforms other representations in identifying the top 1 and top 5 additives. We attribute this result to its effective feature extraction capabilities that can guide toward the promising regions of the chemical reactions. Interestingly FRAGPRINTS excel in identifying the top 10 additives, potentially maximising the benefits of clustering initialisation strategies. The Matern kernel is particularly effective in BO performance.

Concerning acquisition functions, UCB emerges as a better performer across diverse top-n criteria, reinforcing its capacity to strike an optimal balance between exploratory and exploitative behaviours within the search landscape.

Lastly, in the realm of initialisation, clustering slightly edges out other techniques in terms of BO efficiency. On the other hand, maxmin, although viable in some scenarios, results in especially poor R^2 validation scores in comparison to other initialisation strategies.

These observations sharpen our understanding of how varying parameters interact to influence the overall performance of BO in chemical reaction screening. On top of the single parameter influence, we have analysed their interplay and presented results in Figure 2. On the x-axis we can observe how different kernels influence the BO performance (percentage of the top 10 discovered additives) per data representation. With different colours denoting the initialisation strategies, we can clearly see that clustering is the right choice across the majority of better performing representations. Interestingly, FINGERPRINTS work better with random initialisation, again strongly emphasising how important are the added molecular fragments to this representation in order to find valuable clusters at the initial stage of BO search. Mirrored y-axis shows the difference between the two employed acquisition functions. Even though often similar in performance, we can see clear asymmetry especially in DRFP and RXNFP representations.

Additionally, we noticed that the performance of the underlying surrogate model does not always correlate with the efficacy of the Bayesian optimisation. For instance, the representation that achieves the highest R^2 score in validation does not necessarily excel in identifying the most promising top-n additives. This observation raises important questions about the interplay between model fit and optimisation performance, challenging the conventional wisdom that a better-fitting model always leads to better optimisation outcomes.

To illustrate this point, we include a figure (Figure 3) that contrasts the best-performing representation in terms of validation R^2 (CDDD) score against the representation most effective at discovering important BO values (DRFP). The figure reveals a divergence in the areas of the chemical space that each representation considers promising, further underscoring the complexity of the relationship between model fit and BO performance. To dive deeper into the problem, we evaluated regression scores and negative log probability density (NLPD) on the regions of the dataset where it contains highest and lowest objective values. More precisely, we evaluated the surrogate model on the top 5% of the dataset and similarly, bottom 5% of the data containing the lowest target values. Lower NLPD values suggest that the model's predicted distribution aligns well with the true distribution of the data, providing a robust measure of the model's performance. In the context of BO, NLPD serves as a valuable evaluation criterion for the surrogate model, complementing traditional metrics like MAE score. Our goal was to gain insights into whether the surrogate model demonstrates signs of 'targeted inference' where it does not fit the entire input space but compensates its shortcomings by modelling well the promising regions of the space. We present these results in the Table 3. While we do observe noticeable difference in the performance between the bottom and the top regions (especially for the DRFP) of the input space we cannot conclude that the targeted inference takes place. However, NLPD values tend to show us that a better predicted distribution on the higher regions of the space correlates with better BO performance.

Repr.	Dimension	Type
DRFP	512	binary
FRAGPRINTS	597	binary
FINGERPRINTS	512	mixed
CDDD	512	continuous
RXNFP	256	continuous
XTB	11	continuous
MQN	42	continuous
OHE	722	binary
CHEMBERTA	768	continuous

Table 1 An overview of the different representations used in the Bayesian optimisation process along with their respective dimensions and types. The 'Dimension' column indicates the number of features in each representation, while the 'Type' column specifies the nature of the data — binary, mixed (for FRAGPRINTS since they include encoded fragments on top of the FINGERPRINT representation) or continuous. The table presents the diversity in the data representations explored in this study, illustrating the range of complexity and information encapsulated in each.

Understanding this relationship can be crucial for the effective application of Bayesian optimisation in complex chemical spaces. It suggests that focusing solely on the fit of the surrogate model might be an oversimplification, and that a more holistic view that takes into account various performance metrics could be more insightful.

In the main paper, we also show the clustering results using FRAGPRINTS representation on the first reaction plate. The result of clustering with this representation would be the same for the remaining plates, as we use the same set of additives across the four reactions. However, with different reaction representations, namely DRFP, the results of clustering also differ across the plates. Additionally, we present the clustering outcomes on DRFP design space in Figure 4. As shown, different reactions uncover diverse clusters and the initial set of selected data points (additives) varies across reactions, while noticeably, certain cluster centres repeat on multiple reaction plates.

A.5 Bayesian optimisation on Buchwald Hartwig dataset

The Buchwald-Hartwig dataset[21] is a benchmark dataset commonly used in organic chemistry for evaluating and comparing reaction prediction models. It consists of a large number of reactions that involve the formation of C-N bonds commonly using palladium catalysts, which is known as the Buchwald-Hartwig amination reaction. Buchwald and co-workers compiled the dataset in 2002. It has since become indispensable in developing and evaluating reaction prediction models based on machine learning and artificial intelligence. The dataset contains 3955 data points describing Pd-catalysed Buchwald–Hartwig C–N cross-couplings each with the respective yield. We split these data points by the product making a distinction between five unique reactions to optimise the yield for. Each of these reactions has a unique combination of 3 bases, 22 additives, 3 aryl halides and 4 ligands making the OHE

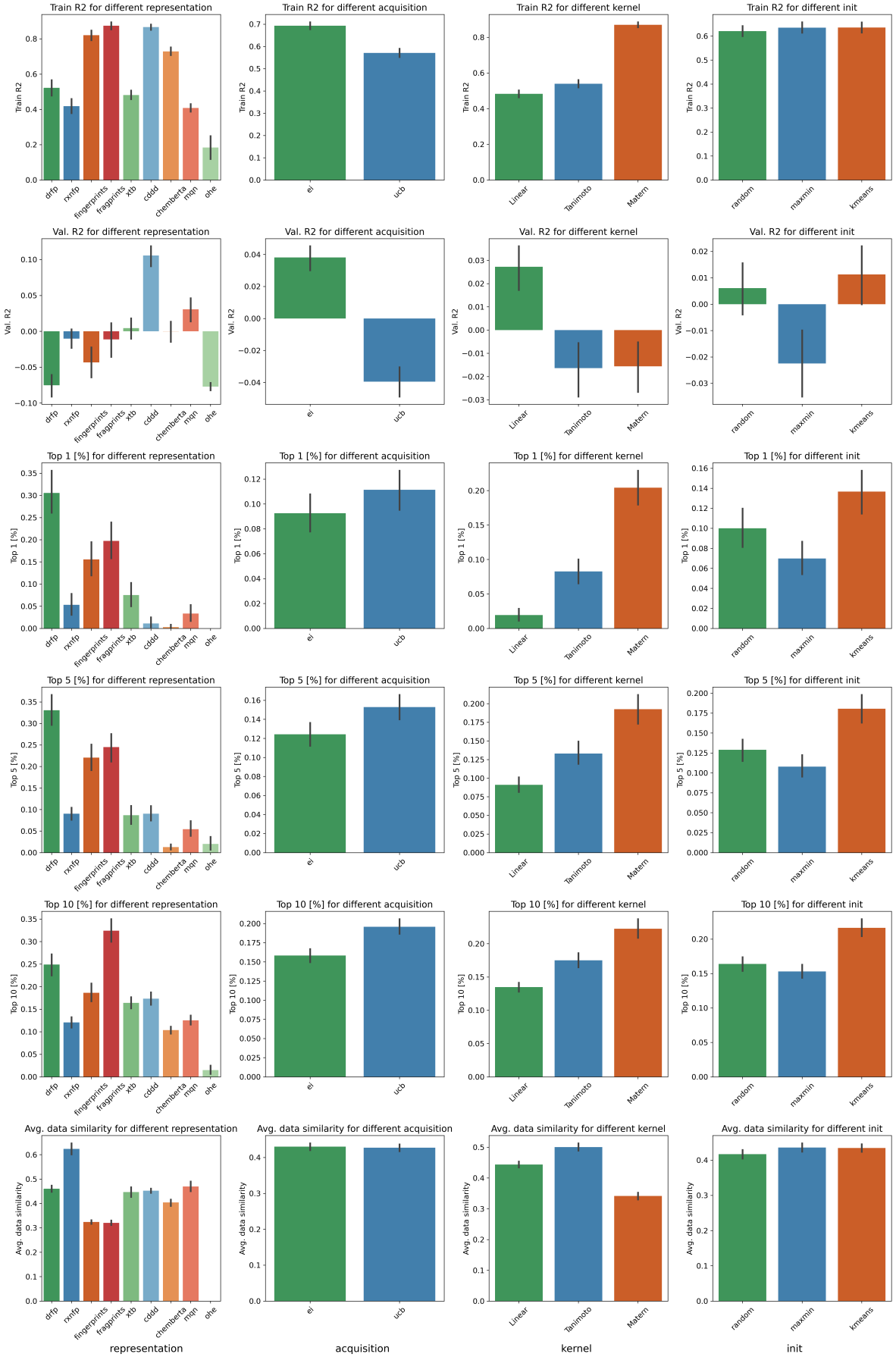

Figure 1 Single parameter influence on various metrics. For each of the subplots, the data is averaged across the remaining parameters and 20 different seed-runs. We calculated these metrics on the reaction plate 1 to uncover the optimal configuration to use for the remaining plates.

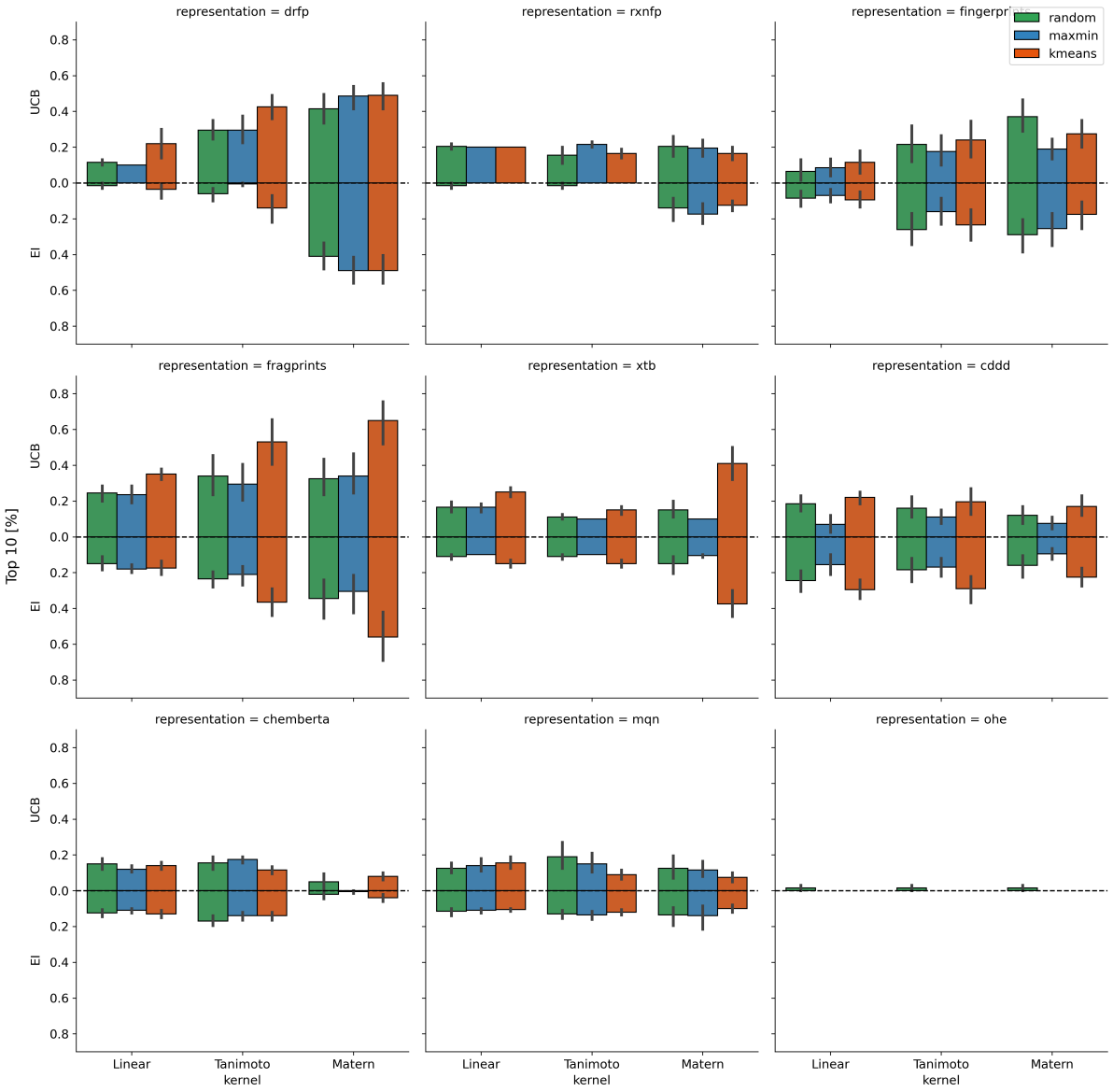

Figure 2 Combined parameter influence on Top 10 count [%] metric. Each subplot represents a different reaction representation, while the x-axis spans across different kernels. The colours denote initialisation methods and the y-axis measures the percentage of discovered additives from the top 10 set, while the mirroring of the axis allows to show the performance of UCB and EI acquisition functions.

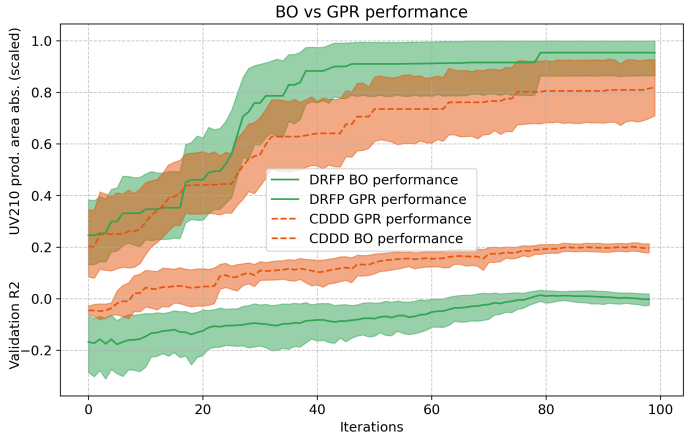

Figure 3 Visual comparison of BO paths and validation R^2 scores for DRFP with Matern kernel and UCB acquisition function and CDDD with Linear kernel and EI acquisition function representations. Results from the first reaction plate. Despite CDDD's higher validation scores it trails behind DRFP in reaching the top regions of the search space.

representation 32-dimensional.

We conducted additional experiments on this dataset to extend the findings from our primary study. The Buchwald-Hartwig dataset is particularly informative because it has a more manageable data-to-dimension ratio, especially when using one-hot encoding. Specifically, the dataset contains approximately 790 data points per reaction, and OHE results in a 32-dimensional feature space. This is contrasting our primary dataset, where the feature dimensionality essentially equals the number of data points, leading to a $d = n$ scenario.

We evaluated the performance of three primary representations: OHE, RXNFP and DRFP across different kernels: Linear, Matern, and Tanimoto, using the three defined initialisation strategies. Our findings confirm previous studies showing that OHE is remarkably effective for this dataset. However, we also discovered that DRFP has a matching performance, therefore reinforcing its robustness across different datasets for BO-related tasks. We present the results in the Figure 5.

A.6 Comparison of random search and Bayesian optimisation selected yields

We compared random search (where the following point is chosen at random by permuting the heldout set) to Bayesian optimisation to test whether BO makes better guided decisions.

We recorded the yields obtained by these methods (BO with DRFP using Matern kernel and kmeans initialisation with UCB acquisition function versus random search) across four reactions over twenty different trials with 100 selected points in each trial. The results allow us to obtain a comparative distribution of selected yields for BO and RS.

We conducted a further comparative analysis using the two-sample bootstrapped Kolmogorov-Smirnov (K-S) test, a non-parametric test that determines whether two samples come from the same distribution. The resulting p-values substantially below the 0.05 threshold firmly reject the null hypothesis that the BO and RS distributions are identical. We visualised the distributions in the box plots in Figure 6. Our findings confirm that the intelligent search strategy adopted by BO outperforms the random search methodology. This underscores the power of advanced optimisation techniques for navigating complex experimental landscapes and maximising desirable outcomes.

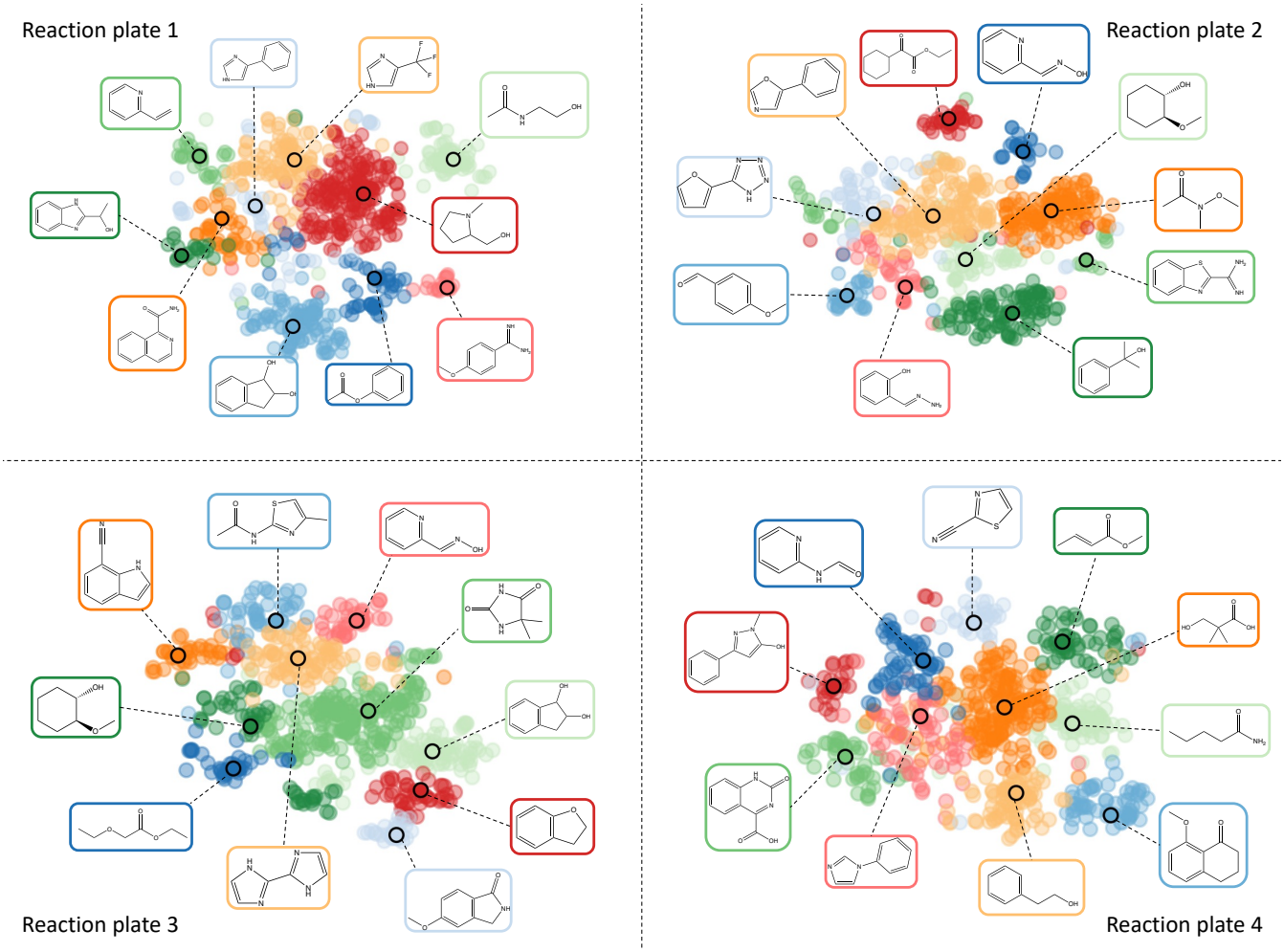

Figure 4 t-sne visualisation of the clustering results on DRFP reaction space for the four reaction plates. We first perform the PCA to reduce the dimensionality of the vectors to 10 principal components and cluster using the k-means method. While uncovering different clusters for each of the separate reaction plates, some of the reaction centres repeat across multiple reactions.

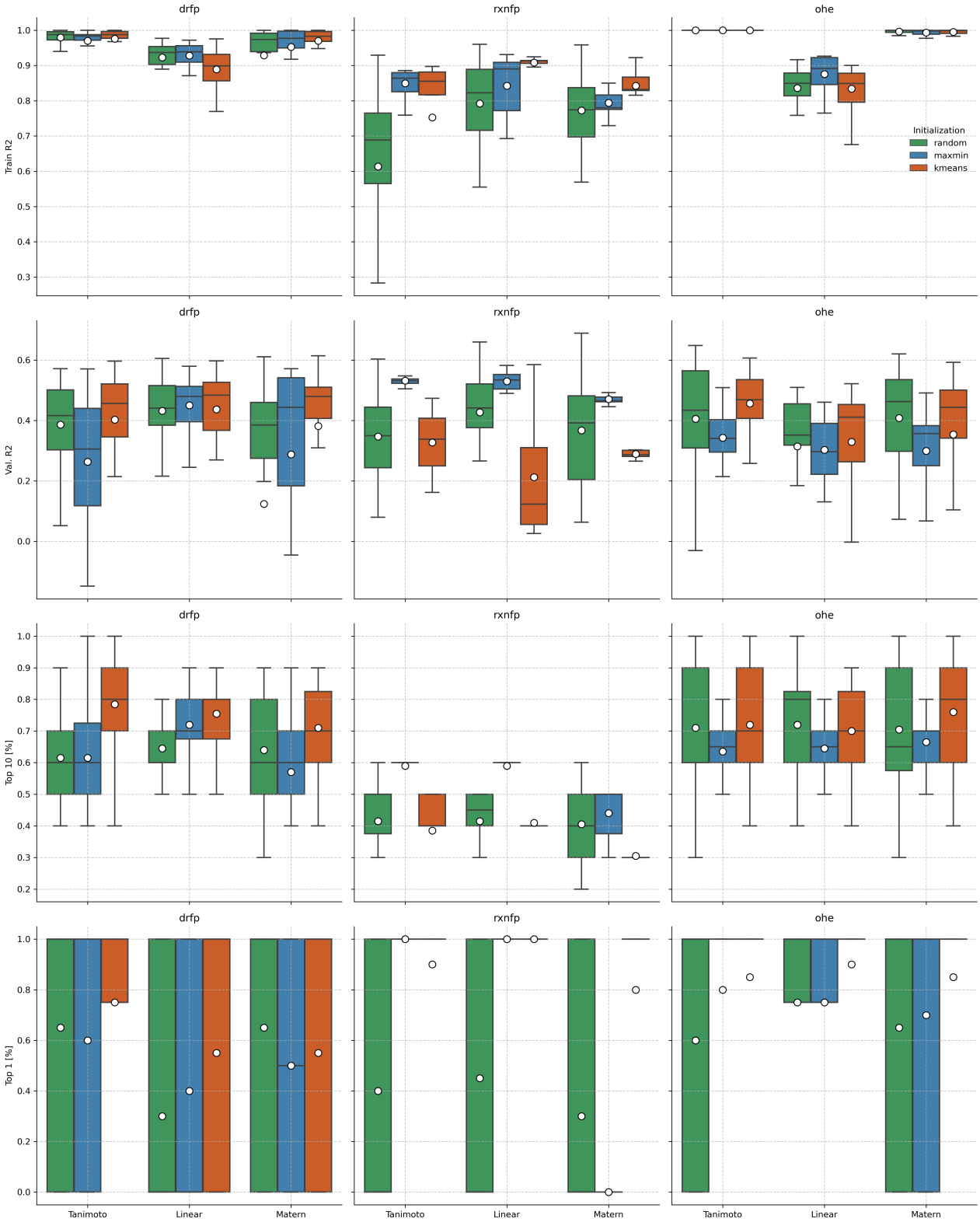

Figure 5 Visualisation of representation, kernel, and initialisation interplay on the Buchwald-Hartwig dataset. A comprehensive facet plot offering insights into the combined effects of representation techniques (OHE, DRFP, RXNFP), kernel choices (Linear, Matern, Tanimoto), and initialisation strategies (clustering, maxmin, random) for Bayesian optimisation. Each panel provides a snapshot of the interaction between these parameters, emphasising the optimal configurations for navigating the Buchwald-Hartwig reaction space.

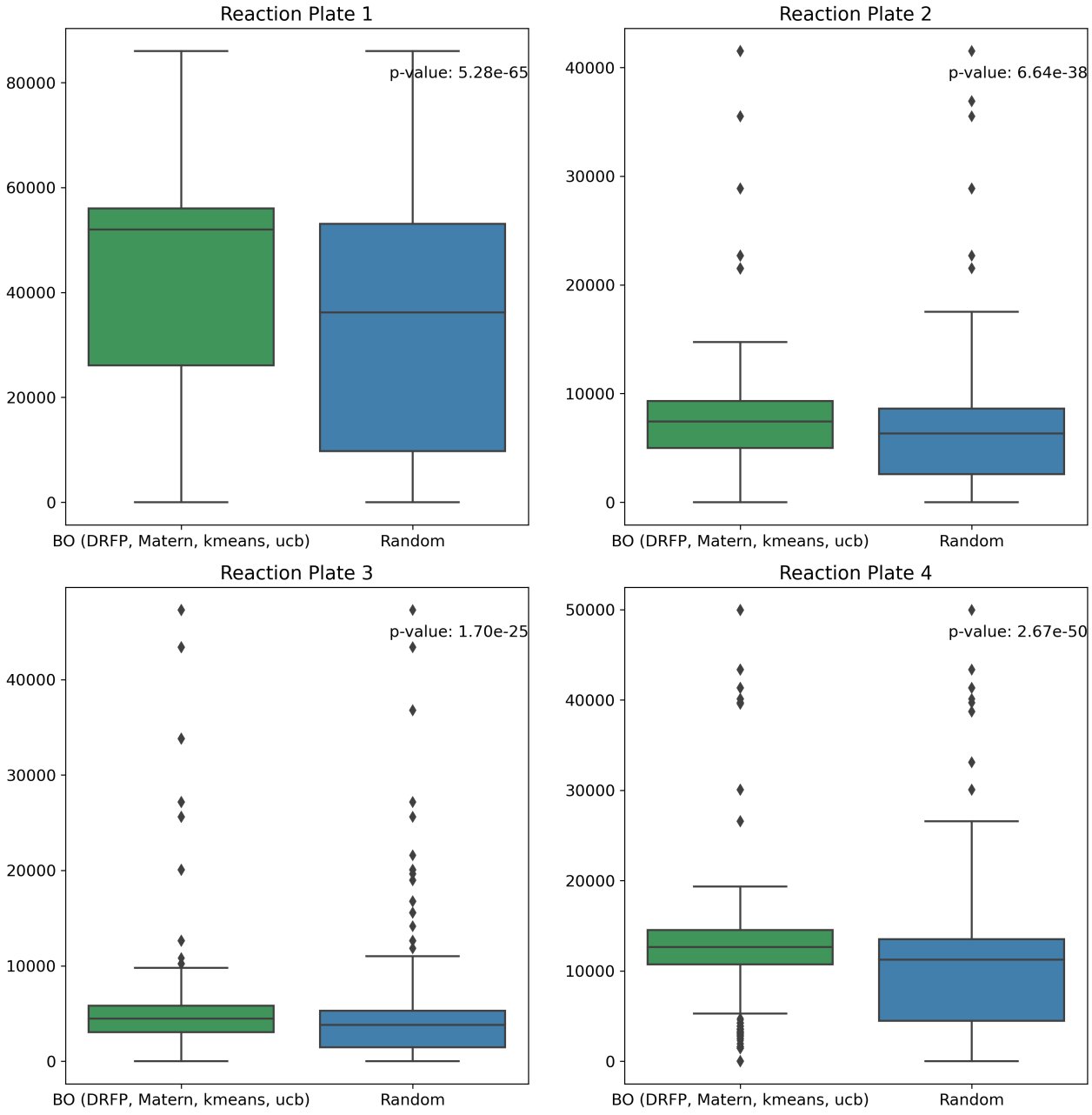

Figure 6 Comparison of Bayesian optimisation BO and random search RS selected yields across four reactions. For BO we use DRFP representations with Matern kernel, kmeans clustering initialisation strategy and UCB acquisition function. Each boxplot displays the distribution of selected yields over 100 iterations and 20 seed-runs for each reaction. We calculated the statistical difference between the BO and RS selected yields using a two-sample bootstrapped Kolmogorov-Smirnov (K-S) test. The results indicate that BO consistently outperforms random search across the studied reactions.

Repr.	Kernel	Init.	Acq.	T. R2	Val. R2	Top 1 [%]	Top 5 [%]	Top 10 [%]	Avg. sim.	Noise
cddd	Lin.	kmeans	ei	0.80 ± 0.07	0.20 ± 0.04	0.00 ± 0.00	0.21 ± 0.18	0.30 ± 0.13	0.55 ± 0.02	0.34 ± 0.08
			ucb	0.74 ± 0.13	0.15 ± 0.05	0.00 ± 0.00	0.07 ± 0.13	0.22 ± 0.08	0.49 ± 0.01	0.39 ± 0.13
		maxmin	ei	0.85 ± 0.09	0.07 ± 0.21	0.00 ± 0.00	0.11 ± 0.10	0.15 ± 0.12	0.56 ± 0.02	0.27 ± 0.12
			ucb	0.87 ± 0.09	0.06 ± 0.11	0.00 ± 0.00	0.02 ± 0.06	0.07 ± 0.11	0.52 ± 0.03	0.24 ± 0.12
		random	ei	0.79 ± 0.10	0.17 ± 0.09	0.00 ± 0.00	0.19 ± 0.23	0.25 ± 0.13	0.55 ± 0.02	0.33 ± 0.13
			ucb	0.73 ± 0.16	0.12 ± 0.09	0.05 ± 0.22	0.07 ± 0.18	0.18 ± 0.10	0.50 ± 0.02	0.39 ± 0.17
	Mat.	kmeans	ei	0.90 ± 0.11	0.10 ± 0.19	0.00 ± 0.00	0.13 ± 0.18	0.23 ± 0.13	0.41 ± 0.12	0.20 ± 0.19
			ucb	0.93 ± 0.11	0.11 ± 0.08	0.05 ± 0.22	0.08 ± 0.21	0.17 ± 0.13	0.35 ± 0.07	0.14 ± 0.19
		maxmin	ei	0.98 ± 0.05	0.12 ± 0.09	0.00 ± 0.00	0.03 ± 0.07	0.10 ± 0.08	0.41 ± 0.07	0.04 ± 0.10
			ucb	0.96 ± 0.05	0.10 ± 0.09	0.00 ± 0.00	0.00 ± 0.00	0.07 ± 0.08	0.40 ± 0.05	0.10 ± 0.12
		random	ei	0.96 ± 0.10	0.04 ± 0.12	0.05 ± 0.22	0.09 ± 0.15	0.16 ± 0.15	0.40 ± 0.10	0.08 ± 0.16
			ucb	0.95 ± 0.10	0.03 ± 0.12	0.00 ± 0.00	0.03 ± 0.10	0.12 ± 0.11	0.39 ± 0.08	0.09 ± 0.17
	Tan.	kmeans	ei	0.83 ± 0.10	0.18 ± 0.05	0.00 ± 0.00	0.21 ± 0.19	0.29 ± 0.16	0.46 ± 0.02	0.34 ± 0.12
			ucb	0.81 ± 0.18	0.08 ± 0.11	0.00 ± 0.00	0.09 ± 0.19	0.20 ± 0.18	0.40 ± 0.01	0.33 ± 0.21
		maxmin	ei	0.92 ± 0.10	0.13 ± 0.09	0.00 ± 0.00	0.12 ± 0.12	0.17 ± 0.12	0.46 ± 0.04	0.19 ± 0.17
			ucb	0.88 ± 0.22	0.07 ± 0.25	0.00 ± 0.00	0.03 ± 0.07	0.11 ± 0.09	0.40 ± 0.02	0.23 ± 0.22
		random	ei	0.88 ± 0.10	0.13 ± 0.08	0.05 ± 0.22	0.10 ± 0.15	0.18 ± 0.15	0.47 ± 0.03	0.25 ± 0.14
			ucb	0.80 ± 0.25	0.02 ± 0.25	0.00 ± 0.00	0.05 ± 0.16	0.16 ± 0.14	0.42 ± 0.03	0.31 ± 0.27
chemberta	Lin.	kmeans	ei	0.67 ± 0.07	-0.02 ± 0.09	0.00 ± 0.00	0.01 ± 0.04	0.13 ± 0.05	0.47 ± 0.02	0.43 ± 0.07
			ucb	0.54 ± 0.06	0.04 ± 0.05	0.00 ± 0.00	0.01 ± 0.04	0.14 ± 0.05	0.46 ± 0.01	0.55 ± 0.05
		maxmin	ei	0.64 ± 0.04	0.03 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.11 ± 0.03	0.47 ± 0.01	0.47 ± 0.03
			ucb	0.48 ± 0.07	0.07 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.12 ± 0.04	0.46 ± 0.02	0.61 ± 0.06
		random	ei	0.61 ± 0.07	0.04 ± 0.08	0.00 ± 0.00	0.02 ± 0.06	0.12 ± 0.04	0.47 ± 0.02	0.49 ± 0.07
			ucb	0.48 ± 0.15	-0.02 ± 0.34	0.00 ± 0.00	0.02 ± 0.06	0.15 ± 0.07	0.45 ± 0.02	0.60 ± 0.13
	Mat.	kmeans	ei	0.89 ± 0.13	0.09 ± 0.03	0.00 ± 0.00	0.01 ± 0.04	0.04 ± 0.05	0.30 ± 0.13	0.23 ± 0.17
			ucb	0.94 ± 0.11	0.02 ± 0.07	0.00 ± 0.00	0.01 ± 0.04	0.08 ± 0.05	0.27 ± 0.14	0.11 ± 0.15
		maxmin	ei	0.93 ± 0.09	0.10 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.02	0.23 ± 0.11	0.19 ± 0.12
			ucb	0.85 ± 0.12	0.07 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.39 ± 0.21	0.27 ± 0.13
		random	ei	0.96 ± 0.05	0.08 ± 0.04	0.00 ± 0.00	0.02 ± 0.06	0.02 ± 0.05	0.23 ± 0.08	0.14 ± 0.08
			ucb	0.97 ± 0.04	0.00 ± 0.10	0.05 ± 0.22	0.05 ± 0.14	0.05 ± 0.09	0.22 ± 0.05	0.09 ± 0.10
	Tan.	kmeans	ei	0.87 ± 0.09	-0.13 ± 0.13	0.00 ± 0.00	0.01 ± 0.04	0.14 ± 0.05	0.46 ± 0.01	0.26 ± 0.13
			ucb	0.58 ± 0.13	-0.02 ± 0.10	0.00 ± 0.00	0.01 ± 0.04	0.12 ± 0.05	0.49 ± 0.02	0.56 ± 0.11
		maxmin	ei	0.90 ± 0.08	-0.17 ± 0.10	0.00 ± 0.00	0.00 ± 0.00	0.14 ± 0.05	0.47 ± 0.02	0.21 ± 0.14
			ucb	0.43 ± 0.12	-0.01 ± 0.07	0.00 ± 0.00	0.00 ± 0.00	0.17 ± 0.04	0.47 ± 0.02	0.70 ± 0.09
		random	ei	0.82 ± 0.13	-0.06 ± 0.18	0.00 ± 0.00	0.02 ± 0.06	0.17 ± 0.07	0.47 ± 0.02	0.31 ± 0.16
			ucb	0.53 ± 0.24	-0.08 ± 0.18	0.00 ± 0.00	0.04 ± 0.10	0.15 ± 0.09	0.50 ± 0.04	0.59 ± 0.22
drfp	Lin.	kmeans	ei	0.16 ± 0.26	-0.04 ± 0.05	0.10 ± 0.31	0.05 ± 0.14	0.04 ± 0.09	0.30 ± 0.01	0.89 ± 0.20
			ucb	0.10 ± 0.18	-0.20 ± 0.12	0.25 ± 0.44	0.34 ± 0.24	0.22 ± 0.19	0.36 ± 0.04	0.93 ± 0.12
		maxmin	ei	0.12 ± 0.16	-0.08 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.32 ± 0.02	0.92 ± 0.09
			ucb	0.00 ± 0.00	-0.32 ± 0.02	0.00 ± 0.00	0.20 ± 0.00	0.10 ± 0.00	0.37 ± 0.00	0.99 ± 0.00
		random	ei	0.19 ± 0.24	-0.08 ± 0.03	0.00 ± 0.00	0.02 ± 0.06	0.02 ± 0.04	0.31 ± 0.02	0.88 ± 0.16
			ucb	0.00 ± 0.00	-0.31 ± 0.04	0.00 ± 0.00	0.22 ± 0.06	0.12 ± 0.04	0.38 ± 0.01	0.99 ± 0.00
	Mat.	kmeans	ei	0.99 ± 0.06	-0.00 ± 0.07	0.80 ± 0.41	0.66 ± 0.27	0.49 ± 0.18	0.39 ± 0.10	0.02 ± 0.11
			ucb	0.99 ± 0.06	-0.00 ± 0.06	0.75 ± 0.44	0.66 ± 0.26	0.49 ± 0.17	0.42 ± 0.10	0.02 ± 0.11
		maxmin	ei	1.00 ± 0.00	0.04 ± 0.03	0.60 ± 0.50	0.58 ± 0.27	0.49 ± 0.18	0.46 ± 0.04	0.00 ± 0.00
			ucb	1.00 ± 0.00	0.04 ± 0.01	0.75 ± 0.44	0.65 ± 0.27	0.48 ± 0.15	0.49 ± 0.04	0.00 ± 0.00
		random	ei	1.00 ± 0.00	0.02 ± 0.05	0.50 ± 0.51	0.50 ± 0.31	0.41 ± 0.17	0.42 ± 0.05	0.00 ± 0.00
			ucb	1.00 ± 0.00	0.01 ± 0.05	0.55 ± 0.51	0.53 ± 0.31	0.42 ± 0.18	0.46 ± 0.07	0.00 ± 0.00
	Tan.	kmeans	ei	0.53 ± 0.34	0.01 ± 0.06	0.25 ± 0.44	0.21 ± 0.30	0.14 ± 0.18	0.55 ± 0.03	0.61 ± 0.28
			ucb	0.39 ± 0.40	-0.22 ± 0.25	0.55 ± 0.51	0.53 ± 0.31	0.42 ± 0.16	0.62 ± 0.06	0.69 ± 0.34
		maxmin	ei	0.38 ± 0.33	-0.05 ± 0.04	0.00 ± 0.00	0.01 ± 0.04	0.01 ± 0.02	0.55 ± 0.02	0.73 ± 0.24
			ucb	0.42 ± 0.18	-0.03 ± 0.09	0.20 ± 0.41	0.34 ± 0.29	0.30 ± 0.18	0.70 ± 0.02	0.72 ± 0.12
		random	ei	0.54 ± 0.33	-0.04 ± 0.07	0.00 ± 0.00	0.10 ± 0.14	0.06 ± 0.08	0.56 ± 0.04	0.60 ± 0.27
			ucb	0.60 ± 0.30	-0.10 ± 0.19	0.20 ± 0.41	0.35 ± 0.24	0.30 ± 0.14	0.61 ± 0.04	0.56 ± 0.26
fingerprints	Lin.	kmeans	ei	0.73 ± 0.12	0.13 ± 0.05	0.00 ± 0.00	0.14 ± 0.18	0.10 ± 0.10	0.37 ± 0.02	0.44 ± 0.12
			ucb	0.77 ± 0.10	0.08 ± 0.06	0.00 ± 0.00	0.13 ± 0.18	0.12 ± 0.15	0.37 ± 0.03	0.39 ± 0.11
		maxmin	ei	0.78 ± 0.12	0.09 ± 0.06	0.00 ± 0.00	0.14 ± 0.17	0.07 ± 0.09	0.38 ± 0.02	0.39 ± 0.13
			ucb	0.67 ± 0.17	0.07 ± 0.07	0.05 ± 0.22	0.12 ± 0.20	0.09 ± 0.11	0.36 ± 0.02	0.50 ± 0.17
		random	ei	0.82 ± 0.13	0.05 ± 0.10	0.10 ± 0.31	0.17 ± 0.19	0.09 ± 0.09	0.38 ± 0.01	0.34 ± 0.17
			ucb	0.79 ± 0.16	0.09 ± 0.07	0.20 ± 0.41	0.12 ± 0.25	0.07 ± 0.14	0.37 ± 0.02	0.36 ± 0.17
	Mat.	kmeans	ei	1.00 ± 0.00	-0.07 ± 0.26	0.05 ± 0.22	0.15 ± 0.27	0.17 ± 0.17	0.32 ± 0.11	0.00 ± 0.00
			ucb	1.00 ± 0.00	-0.17 ± 0.24	0.10 ± 0.31	0.20 ± 0.28	0.28 ± 0.18	0.29 ± 0.12	0.00 ± 0.00
		maxmin	ei	1.00 ± 0.00	-0.19 ± 0.28	0.15 ± 0.37	0.30 ± 0.29	0.26 ± 0.20	0.28 ± 0.12	0.00 ± 0.01
			ucb	1.00 ± 0.00	-0.21 ± 0.27	0.00 ± 0.00	0.12 ± 0.18	0.19 ± 0.13	0.29 ± 0.12	0.00 ± 0.00
		random	ei	0.99 ± 0.02	-0.10 ± 0.21	0.35 ± 0.49	0.38 ± 0.32	0.29 ± 0.21	0.34 ± 0.10	0.03 ± 0.08
			ucb	1.00 ± 0.00	-0.10 ± 0.18	0.45 ± 0.51	0.43 ± 0.35	0.37 ± 0.22	0.33 ± 0.09	0.00 ± 0.00
	Tan.	kmeans	ei	0.88 ± 0.21	0.03 ± 0.09	0.15 ± 0.37	0.26 ± 0.32	0.24 ± 0.20	0.28 ± 0.02	0.21 ± 0.27
			ucb	0.53 ± 0.48	-0.18 ± 0.21	0.20 ± 0.41	0.22 ± 0.37	0.24 ± 0.25	0.29 ± 0.03	0.50 ± 0.47
		maxmin	ei	0.89 ± 0.18	0.03 ± 0.06	0.10 ± 0.31	0.23 ± 0.27	0.16 ± 0.18	0.29 ± 0.02	0.22 ± 0.23
			ucb	0.39 ± 0.38	-0.25 ± 0.24	0.15 ± 0.37	0.20 ± 0.29	0.17 ± 0.20	0.31 ± 0.03	0.71 ± 0.30
		random	ei	0.91 ± 0.12	0.03 ± 0.06	0.45 ± 0.51	0.38 ± 0.34	0.26 ± 0.22	0.29 ± 0.02	0.20 ± 0.20

Continued on next page

Repr.	Kernel	Init.	Acq.	T. R2	Val. R2	Top 1 [%]	Top 5 [%]	Top 10 [%]	Avg. sim.	Noise
			ucb	0.61 ± 0.34	-0.09 ± 0.21	0.30 ± 0.47	0.28 ± 0.34	0.21 ± 0.24	0.30 ± 0.02	0.54 ± 0.29
fragprints	Lin.	kmeans	ei	0.85 ± 0.11	0.19 ± 0.07	0.00 ± 0.00	0.13 ± 0.18	0.17 ± 0.09	0.30 ± 0.02	0.29 ± 0.14
			ucb	0.77 ± 0.14	0.11 ± 0.05	0.00 ± 0.00	0.28 ± 0.12	0.35 ± 0.07	0.31 ± 0.01	0.37 ± 0.16
		maxmin	ei	0.88 ± 0.07	0.12 ± 0.06	0.00 ± 0.00	0.09 ± 0.15	0.18 ± 0.05	0.30 ± 0.02	0.26 ± 0.11
			ucb	0.81 ± 0.16	0.05 ± 0.10	0.00 ± 0.00	0.12 ± 0.18	0.24 ± 0.11	0.29 ± 0.02	0.33 ± 0.20
		random	ei	0.89 ± 0.10	0.11 ± 0.13	0.10 ± 0.31	0.12 ± 0.16	0.15 ± 0.09	0.31 ± 0.02	0.23 ± 0.15
			ucb	0.91 ± 0.14	0.05 ± 0.08	0.10 ± 0.31	0.15 ± 0.14	0.25 ± 0.10	0.31 ± 0.01	0.18 ± 0.17
	Mat.	kmeans	ei	1.00 ± 0.00	-0.04 ± 0.26	0.60 ± 0.50	0.51 ± 0.39	0.56 ± 0.32	0.37 ± 0.11	0.00 ± 0.00
			ucb	1.00 ± 0.00	-0.20 ± 0.25	0.70 ± 0.47	0.60 ± 0.33	0.65 ± 0.26	0.34 ± 0.08	0.00 ± 0.00
		maxmin	ei	1.00 ± 0.00	-0.09 ± 0.30	0.20 ± 0.41	0.19 ± 0.32	0.30 ± 0.24	0.41 ± 0.12	0.00 ± 0.00
			ucb	0.99 ± 0.06	-0.23 ± 0.38	0.25 ± 0.44	0.20 ± 0.34	0.34 ± 0.25	0.36 ± 0.14	0.02 ± 0.11
		random	ei	1.00 ± 0.00	0.06 ± 0.11	0.25 ± 0.44	0.26 ± 0.35	0.35 ± 0.25	0.46 ± 0.10	0.00 ± 0.00
			ucb	0.99 ± 0.05	0.03 ± 0.13	0.20 ± 0.41	0.19 ± 0.32	0.33 ± 0.25	0.44 ± 0.06	0.02 ± 0.10
	Tan.	kmeans	ei	0.87 ± 0.13	0.05 ± 0.08	0.20 ± 0.41	0.33 ± 0.29	0.36 ± 0.17	0.25 ± 0.01	0.26 ± 0.22
			ucb	0.44 ± 0.32	-0.29 ± 0.27	0.45 ± 0.51	0.53 ± 0.42	0.53 ± 0.30	0.28 ± 0.02	0.68 ± 0.29
		maxmin	ei	0.89 ± 0.19	0.03 ± 0.13	0.00 ± 0.00	0.14 ± 0.18	0.21 ± 0.12	0.25 ± 0.01	0.19 ± 0.26
			ucb	0.74 ± 0.37	-0.23 ± 0.34	0.15 ± 0.37	0.17 ± 0.31	0.30 ± 0.23	0.26 ± 0.01	0.32 ± 0.39
		random	ei	0.90 ± 0.12	0.09 ± 0.08	0.10 ± 0.31	0.17 ± 0.16	0.24 ± 0.10	0.26 ± 0.02	0.21 ± 0.20
			ucb	0.81 ± 0.19	-0.02 ± 0.14	0.25 ± 0.44	0.23 ± 0.36	0.34 ± 0.27	0.27 ± 0.02	0.32 ± 0.27
mqn	Lin.	kmeans	ei	0.44 ± 0.08	0.11 ± 0.04	0.00 ± 0.00	0.01 ± 0.04	0.11 ± 0.02	0.38 ± 0.02	0.62 ± 0.07
			ucb	0.41 ± 0.11	0.11 ± 0.06	0.00 ± 0.00	0.11 ± 0.15	0.15 ± 0.08	0.42 ± 0.07	0.63 ± 0.11
		maxmin	ei	0.37 ± 0.08	0.11 ± 0.04	0.00 ± 0.00	0.02 ± 0.06	0.11 ± 0.03	0.37 ± 0.01	0.69 ± 0.07
			ucb	0.26 ± 0.10	0.05 ± 0.06	0.00 ± 0.00	0.08 ± 0.18	0.14 ± 0.09	0.44 ± 0.06	0.78 ± 0.09
		random	ei	0.32 ± 0.05	0.13 ± 0.02	0.00 ± 0.00	0.02 ± 0.06	0.12 ± 0.05	0.38 ± 0.02	0.73 ± 0.05
			ucb	0.36 ± 0.15	0.04 ± 0.17	0.00 ± 0.00	0.07 ± 0.13	0.12 ± 0.06	0.41 ± 0.07	0.68 ± 0.14
	Mat.	kmeans	ei	0.52 ± 0.15	0.13 ± 0.08	0.00 ± 0.00	0.00 ± 0.00	0.10 ± 0.05	0.60 ± 0.15	0.58 ± 0.16
			ucb	0.53 ± 0.27	-0.01 ± 0.16	0.00 ± 0.00	0.00 ± 0.00	0.07 ± 0.06	0.47 ± 0.25	0.55 ± 0.29
		maxmin	ei	0.67 ± 0.26	-0.16 ± 0.24	0.10 ± 0.31	0.08 ± 0.25	0.14 ± 0.15	0.24 ± 0.28	0.43 ± 0.33
			ucb	0.55 ± 0.30	-0.09 ± 0.18	0.05 ± 0.22	0.04 ± 0.18	0.12 ± 0.10	0.34 ± 0.31	0.54 ± 0.33
		random	ei	0.47 ± 0.18	0.04 ± 0.18	0.05 ± 0.22	0.06 ± 0.23	0.14 ± 0.12	0.48 ± 0.27	0.64 ± 0.18
			ucb	0.41 ± 0.27	-0.15 ± 0.20	0.10 ± 0.31	0.09 ± 0.28	0.12 ± 0.16	0.26 ± 0.28	0.70 ± 0.25
	Tan.	kmeans	ei	0.46 ± 0.06	0.15 ± 0.02	0.00 ± 0.00	0.04 ± 0.08	0.12 ± 0.04	0.57 ± 0.01	0.64 ± 0.06
			ucb	0.36 ± 0.07	-0.01 ± 0.11	0.00 ± 0.00	0.00 ± 0.00	0.09 ± 0.06	0.64 ± 0.06	0.72 ± 0.06
		maxmin	ei	0.37 ± 0.11	0.10 ± 0.12	0.00 ± 0.00	0.07 ± 0.12	0.14 ± 0.06	0.56 ± 0.01	0.72 ± 0.08
			ucb	0.19 ± 0.18	-0.14 ± 0.22	0.10 ± 0.31	0.07 ± 0.22	0.15 ± 0.12	0.66 ± 0.06	0.85 ± 0.14
		random	ei	0.41 ± 0.06	0.13 ± 0.03	0.00 ± 0.00	0.05 ± 0.11	0.13 ± 0.06	0.57 ± 0.02	0.68 ± 0.05
			ucb	0.24 ± 0.16	-0.00 ± 0.12	0.20 ± 0.41	0.17 ± 0.35	0.19 ± 0.18	0.65 ± 0.07	0.82 ± 0.12
ohe	Lin.	random	ucb	0.00 ± 0.00	-0.08 ± 0.02	0.00 ± 0.00	0.02 ± 0.06	0.02 ± 0.04	0.00 ± 0.00	0.99 ± 0.00
	Mat.	random	ucb	0.55 ± 0.01	-0.08 ± 0.02	0.00 ± 0.00	0.02 ± 0.06	0.02 ± 0.04	0.00 ± 0.00	0.66 ± 0.01
	Tan.	random	ucb	0.00 ± 0.00	-0.08 ± 0.02	0.00 ± 0.00	0.02 ± 0.06	0.02 ± 0.04	0.00 ± 0.00	0.99 ± 0.00
rxnfp	Lin.	kmeans	ei	0.29 ± 0.02	0.07 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.77 ± 0.01	0.77 ± 0.02
			ucb	0.00 ± 0.00	-0.11 ± 0.01	0.00 ± 0.00	0.20 ± 0.00	0.20 ± 0.00	0.78 ± 0.01	0.99 ± 0.00
		maxmin	ei	0.30 ± 0.02	0.05 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.75 ± 0.01	0.76 ± 0.02
			ucb	0.00 ± 0.00	-0.10 ± 0.01	0.00 ± 0.00	0.20 ± 0.00	0.20 ± 0.00	0.78 ± 0.00	0.99 ± 0.00
		random	ei	0.31 ± 0.04	0.08 ± 0.02	0.00 ± 0.00	0.01 ± 0.04	0.02 ± 0.04	0.76 ± 0.01	0.76 ± 0.03
			ucb	0.00 ± 0.02	-0.10 ± 0.02	0.00 ± 0.00	0.21 ± 0.04	0.21 ± 0.04	0.78 ± 0.00	0.99 ± 0.01
	Mat.	kmeans	ei	0.95 ± 0.07	0.02 ± 0.12	0.15 ± 0.37	0.04 ± 0.10	0.12 ± 0.06	0.21 ± 0.05	0.17 ± 0.13
			ucb	0.87 ± 0.13	0.03 ± 0.10	0.05 ± 0.22	0.02 ± 0.09	0.17 ± 0.08	0.48 ± 0.11	0.26 ± 0.17
		maxmin	ei	0.98 ± 0.07	0.02 ± 0.12	0.37 ± 0.50	0.17 ± 0.20	0.17 ± 0.12	0.17 ± 0.05	0.03 ± 0.12
			ucb	0.91 ± 0.14	0.06 ± 0.10	0.15 ± 0.37	0.11 ± 0.23	0.20 ± 0.11	0.45 ± 0.08	0.17 ± 0.21
		random	ei	0.97 ± 0.09	0.08 ± 0.12	0.15 ± 0.37	0.08 ± 0.18	0.14 ± 0.15	0.19 ± 0.06	0.08 ± 0.15
			ucb	0.93 ± 0.10	0.07 ± 0.08	0.10 ± 0.31	0.07 ± 0.13	0.21 ± 0.13	0.47 ± 0.10	0.15 ± 0.16
	Tan.	kmeans	ei	0.34 ± 0.12	0.09 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.77 ± 0.03	0.73 ± 0.10
			ucb	0.00 ± 0.00	-0.18 ± 0.02	0.00 ± 0.00	0.13 ± 0.12	0.17 ± 0.06	0.75 ± 0.00	0.99 ± 0.00
		maxmin	ei	0.29 ± 0.08	0.05 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.78 ± 0.02	0.77 ± 0.07
			ucb	0.00 ± 0.00	-0.20 ± 0.01	0.00 ± 0.00	0.23 ± 0.07	0.21 ± 0.04	0.76 ± 0.00	0.99 ± 0.00
		random	ei	0.31 ± 0.09	0.08 ± 0.02	0.00 ± 0.00	0.01 ± 0.04	0.02 ± 0.04	0.78 ± 0.04	0.75 ± 0.08
			ucb	0.10 ± 0.24	-0.20 ± 0.11	0.00 ± 0.00	0.15 ± 0.14	0.15 ± 0.11	0.78 ± 0.06	0.91 ± 0.20
xtb	Lin.	kmeans	ei	0.36 ± 0.07	0.11 ± 0.04	0.00 ± 0.00	0.10 ± 0.10	0.15 ± 0.05	0.46 ± 0.02	0.67 ± 0.06
			ucb	0.29 ± 0.03	0.05 ± 0.04	0.00 ± 0.00	0.11 ± 0.10	0.25 ± 0.06	0.45 ± 0.02	0.74 ± 0.03
		maxmin	ei	0.45 ± 0.02	-0.12 ± 0.11	0.00 ± 0.00	0.00 ± 0.00	0.10 ± 0.00	0.43 ± 0.01	0.58 ± 0.02
			ucb	0.30 ± 0.02	-0.08 ± 0.09	0.00 ± 0.00	0.00 ± 0.00	0.17 ± 0.05	0.43 ± 0.01	0.73 ± 0.02
		random	ei	0.41 ± 0.06	0.04 ± 0.09	0.00 ± 0.00	0.01 ± 0.04	0.11 ± 0.03	0.45 ± 0.02	0.62 ± 0.06
			ucb	0.31 ± 0.05	-0.04 ± 0.14	0.00 ± 0.00	0.02 ± 0.06	0.17 ± 0.07	0.45 ± 0.02	0.72 ± 0.05
	Mat.	kmeans	ei	0.72 ± 0.28	-0.14 ± 0.14	0.60 ± 0.50	0.48 ± 0.40	0.38 ± 0.17	0.16 ± 0.15	0.36 ± 0.35
			ucb	0.77 ± 0.23	-0.13 ± 0.10	0.55 ± 0.51	0.44 ± 0.41	0.41 ± 0.22	0.14 ± 0.11	0.32 ± 0.31
		maxmin	ei	0.67 ± 0.29	-0.01 ± 0.11	0.00 ± 0.00	0.00 ± 0.00	0.11 ± 0.02	0.35 ± 0.24	0.38 ± 0.32
			ucb	0.72 ± 0.34	-0.11 ± 0.20	0.00 ± 0.00	0.00 ± 0.00	0.10 ± 0.00	0.20 ± 0.23	0.33 ± 0.39
		random	ei	0.65 ± 0.27	-0.01 ± 0.16	0.10 ± 0.31	0.09 ± 0.25	0.15 ± 0.11	0.35 ± 0.23	0.42 ± 0.30
			ucb	0.62 ± 0.27	-0.01 ± 0.11	0.10 ± 0.31	0.09 ± 0.25	0.15 ± 0.11	0.32 ± 0.19	0.48 ± 0.28
	Tan.	kmeans	ei	0.49 ± 0.03	0.15 ± 0.05	0.00 ± 0.00	0.10 ± 0.10	0.15 ± 0.05	0.62 ± 0.01	0.55 ± 0.03

Continued on next page

Repr.	Kernel	Init.	Acq.	T. R2	Val. R2	Top 1 [%]	Top 5 [%]	Top 10 [%]	Avg. sim.	Noise
			ucb	0.26 ± 0.05	0.12 ± 0.03	0.00 ± 0.00	0.10 ± 0.10	0.15 ± 0.05	0.65 ± 0.01	0.79 ± 0.05
		maxmin	ei	0.48 ± 0.01	-0.02 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.10 ± 0.00	0.62 ± 0.01	0.56 ± 0.01
			ucb	0.31 ± 0.03	0.10 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.10 ± 0.00	0.67 ± 0.01	0.75 ± 0.03
		random	ei	0.51 ± 0.05	0.09 ± 0.09	0.00 ± 0.00	0.01 ± 0.04	0.11 ± 0.03	0.62 ± 0.01	0.53 ± 0.05
			ucb	0.35 ± 0.06	0.09 ± 0.11	0.00 ± 0.00	0.01 ± 0.04	0.11 ± 0.03	0.67 ± 0.02	0.70 ± 0.06

Table 2 Aggregated performance metrics for the evaluated representations, kernels, initialisation strategies and acquisition functions over 20 different seed-runs on the first reaction plate. We present the training and validation R^2 scores alongside top-n BO performance metrics. Average similarity is a measure of how similar the data points are to each other, within the context of the kernel function used in the Gaussian process surrogate model. The noise represents the noise associated with the observations in the data, modeled by the surrogate model.

Repr.	NLPD top ↓	NLPD bottom ↓	NLPD all ↓	MAE top ↓	MAE bottom ↓	MAE all ↓	Top 10 [%] ↑	Top 1 [%] ↑
fingerprints	11.25 ± 0.16	11.84 ± 0.70	11.49 ± 0.24	0.30 ± 0.04	0.42 ± 0.05	0.23 ± 0.01	0.05 ± 0.08	0.00 ± 0.00
drfp	11.32 ± 0.21	12.41 ± 1.02	11.48 ± 0.16	0.28 ± 0.05	0.45 ± 0.05	0.23 ± 0.01	0.11 ± 0.11	0.10 ± 0.31
fragprints	11.38 ± 0.08	11.49 ± 0.31	11.33 ± 0.07	0.35 ± 0.06	0.36 ± 0.06	0.22 ± 0.01	0.29 ± 0.23	0.40 ± 0.51
rxnfp	11.57 ± 0.29	11.87 ± 0.56	11.58 ± 0.63	0.34 ± 0.05	0.41 ± 0.07	0.23 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
xtb	11.64 ± 0.26	12.32 ± 0.80	15.22 ± 11.61	0.31 ± 0.05	0.43 ± 0.06	0.23 ± 0.01	0.14 ± 0.12	0.10 ± 0.31
cddd	13.46 ± 1.23	13.35 ± 1.34	11.81 ± 0.34	0.38 ± 0.05	0.38 ± 0.05	0.24 ± 0.00	0.03 ± 0.04	0.00 ± 0.00
mqn	14.13 ± 1.67	12.17 ± 0.82	12.73 ± 4.32	0.46 ± 0.04	0.30 ± 0.05	0.24 ± 0.01	0.00 ± 0.00	0.00 ± 0.00
ohe	15.27 ± 1.30	12.13 ± 0.49	11.95 ± 0.27	0.48 ± 0.05	0.29 ± 0.05	0.24 ± 0.01	0.02 ± 0.04	0.00 ± 0.00
chemberta	15.49 ± 0.82	11.69 ± 0.18	11.95 ± 0.18	0.51 ± 0.02	0.26 ± 0.02	0.25 ± 0.00	0.01 ± 0.02	0.00 ± 0.00

Table 3 Performance metrics showing negative log probability density, normalised MAE scores and top-*n* BO scores calculated on the top and bottom 5% of the heldout set (in terms of the target values), alongside evaluation on the whole heldout set after 10 BO iterations starting from 10 initial points.

Notes and references

[1] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, MIT press Cambridge, MA, 2006.

[2] J. Wilson, F. Hutter and M. Deisenroth, *Advances in Neural Information Processing Systems*, 2018, **31**, 9884–9895.

[3] A. M. Schweidtmann, D. Bongartz, D. Grothe, T. Kerkenhoff, X. Lin, J. Najman and A. Mitsos, *arXiv preprint arXiv:2005.10902*, 2020.

[4] A. Grosnit, A. I. Cowen-Rivers, R. Tutunov, R.-R. Griffiths, J. Wang and H. Bou-Ammar, *Journal of Machine Learning Research*, 2021, **22**, 160–1.

[5] H. Moss, D. Leslie, D. Beck, J. Gonzalez and P. Rayson, *Advances in Neural Information Processing Systems*, 2020, **33**, 15476–15486.

[6] R.-R. Griffiths, L. Klarner, H. Moss, A. Ravuri, S. T. Truong, B. Rankovic, Y. Du, A. R. Jamasb, J. Schwartz, A. Tripp, G. Kell, A. Bourached, A. Chan, J. Moss, C. Guo, A. Lee, P. Schwaller and J. Tang, ICML 2022 2nd AI for Science Workshop, 2022.

[7] H. B. Moss and R.-R. Griffiths, *arXiv e-prints*, 2020, arXiv:2010.01118.

[8] P. S. Kutchukian, J. F. Dropinski, K. D. Dykstra, B. Li, D. A. DiRocco, E. C. Streckfuss, L.-C. Campeau, T. Cernak, P. Vachal, I. W. Davies *et al.*, *Chemical Science*, 2016, **7**, 2604–2613.

[9] C. N. Prieto Kullmer, J. A. Kautzky, S. W. Krska, T. Nowak, S. D. Dreher and D. W. MacMillan, *Science*, 2022, **376**, 532–539.

[10] G. Landrum, P. Tosco, B. Kelley, Ric, sriniker, gedeck, R. Vianello, NadineSchneider, D. Cosgrove, E. Kawashima, A. Dalke, D. N, G. Jones, B. Cole, M. Swain, S. Turk, AlexanderSavelyev, A. Vaucher, M. Wójcikowski, I. Take, D. Probst, K. Ujihara, V. F. Scalfani, guillaume godin, A. Pahl, F. Berenger, JLVarjo, strets123, JP and DoliathGavid, 2022.

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, *Advances in Neural Information Processing Systems*, 2019, **32**, 8024–8035.

[12] W. Falcon and The PyTorch Lightning team, *PyTorch Lightning*, 2019, `https://github.com/Lightning-AI/lightning`.

[13] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson and E. Bakshy, Advances in Neural Information Processing Systems 33, 2020.

[14] R.-R. Griffiths, J. L. Greenfield, A. R. Thawani, A. R. Jamasb, H. B. Moss, A. Bourached, P. Jones, W. McCorkindale, A. A. Aldrick, M. J. Fuchter *et al.*, *Chemical Science*, 2022, **13**, 13541–13551.

[15] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger and A. G. Wilson, Advances in Neural Information Processing Systems, 2018.

[16] D. Probst, P. Schwaller and J.-L. Reymond, *Digital Discovery*, 2022, **1**, 91–97.

[17] P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino and J.-L. Reymond, *Nature Machine Intelligence*, 2021, **3**, 144–152.

[18] R. Winter, F. Montanari, F. Noé and D.-A. Clevert, *Chemical Science*, 2019, **10**, 1692–1701.

[19] *digital-chemistry-laboratory/morfeus: A Python package for calculating molecular features*, `https://github.com/digital-chemistry-laboratory/morfeus`, (Accessed on 05/17/2023).

[20] *seyonec/ChemBERTa-zinc-base-v1 · Hugging Face*, `https://huggingface.co/seyonec/ChemBERTa-zinc-base-v1`, (Accessed on 05/17/2023).

[21] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.