Bayesian illumination: Inference and quality-diversity accelerate generative molecular models

Jonas Verhellen

Data Science for Drug Design Group, Center for Pharmaceutical Data Science Education, Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, 2100 Copenhagen, Denmark
Centre for Integrative Neuroplasticity, Department of Biosciences, University of Oslo, Blindernveien 31, 0371 Oslo, Norway

jverhell@gmail.com

Abstract

In recent years, there have been considerable academic and industrial research efforts to develop novel generative models for high-performing, small molecules. Traditional, rules-based algorithms such as genetic algorithms have, however, been shown to rival deep learning approaches in terms of both efficiency and potency [Jensen, *Chem. Sci.*, 2019, 12, 3567-3572]. In previous work, we showed that the addition of a quality-diversity archive to a genetic algorithm resolves stagnation issues and substantially increases search efficiency [Verhellen, *Chem. Sci.*, 2020, 42, 11485-11491]. In this work, we expand on these insights and leverage the availability of bespoke kernels for small molecules [Griffiths, *Adv. Neural. Inf. Process. Syst.*, 2024, 36] to integrate Bayesian optimisation into the quality-diversity process. This novel generative model, which we call Bayesian Illumination, produces a larger diversity of high-performing molecules than standard quality-diversity optimisation methods. In addition, we show that Bayesian Illumination further improves search efficiency compared to previous generative models for small molecules, including deep learning approaches, genetic algorithms, and standard quality-diversity methods.

Received 2025-02-05
Accepted 2025-05-12
Published 2025-05-28
doi:10.21468/SciPostChem.4.1.001

Check for updates

Contents

1 Introduction

Despite a surge [1] of deep learning papers focused on generative models for small molecules, it remains difficult to out-compete more traditional, rules-based approaches [2–4] such as a genetic algorithms (GA). In prior research, we introduced quality-diversity methods to the de novo design of small molecules and showed that these methods, which explicitly balance exploitation and exploration, resolve stagnation issues and are more efficient in exploring chemical space than both deep learning models and genetic algorithms. In this work, we extend this approach by leveraging and integrating Bayesian optimisation methods to present a novel generative molecular model, which we call Bayesian Illumination. This algorithm produces a larger diversity of high-performing molecules and further improves search efficiency compared to genetic algorithms, deep learning approaches, and standard quality-diversity methods.

We present the technical details of the Bayesian Illumination algorithm alongside a comprehensive hyperparameter scan (considering different molecular representations) on a standardised fingerprint-based rediscovery benchmark. In addition, we introduce a novel type of benchmark, where molecules are rediscovered on the basis of a sample of conformers and associated USRCAT or Zernike descriptors. This descriptor-based rediscovery of small molecules is both challenging for generative models and computationally affordable. To facilitate this benchmark, we provide a novel and efficient implementation of Zernike descriptors for small molecules. Finally, we also apply the Bayesian Illumination algorithm to docking based tasks. To avoid pure exploitation of docking scores and to avoid unrealistic structures for the predicted binders, we apply stringent structural and physicochemical filters on the candidate molecules and modulate the docking scores with a factor based on the synthetic accessibility of the candidate molecule.

2 Algorithmic methodology

Genetic algorithms [5, 6] offer a powerful approach to molecular optimisation, particularly in scenarios where the exact mathematical form of the evaluation function is inaccessible. They generate molecules by iteratively modifying molecules from a database or those previously obtained by the algorithm. The optimisation process of genetic algorithms involves two fundamental operations: mutations and crossovers. Mutations involve randomly changing molecules from the current population, whereas crossovers stochastically combine parts of molecules from the population. Selection pressure is applied in each generation of the optimisation process, where only the most fit molecules are retained, based on a given and external evaluation function. This process mimics natural selection, promoting the survival and spread of the most relevant motives from high-scoring molecules.

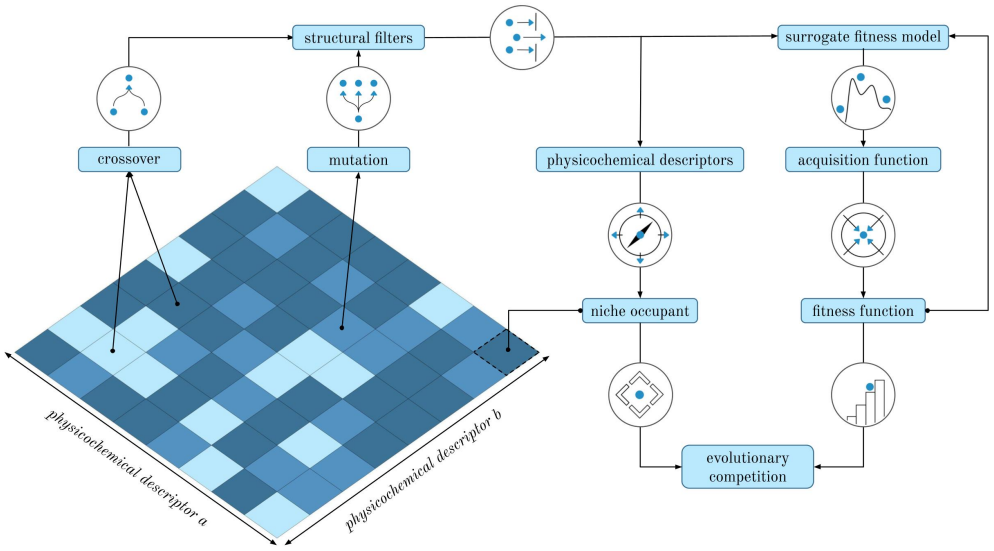

Figure 1: Illustration of the Bayesian Illumination algorithm, with its individual steps in the molecular optimisation process, proceeding clockwise from the bottom left. Random selection from the physicochemical archive, mutations, and crossovers initiate exploration of the search space. Structural filtering, physicochemical descriptors and niche determination refine molecule selection. Surrogate fitness and acquisition function calculations inform the selection of a single molecule to be compared in direct evolutionary competition with the current occupant of the niche. This algorithm is applied iteratively, in batches known as generations, until a predetermined fitness function budget is exhausted or the algorithm has converged.

The most effective genetic algorithm currently available for molecular optimisation is the graph-based genetic algorithm [3] (GB-GA), which encodes candidate molecules by their molecular graphs. This encoding allows for the generation of novel molecules through mutations or crossovers acting directly on the molecular graphs from within the existing population. The initial set of candidate molecules for GB-GA is typically sourced from publicly available molecular datasets such as ZINC [7] or ChEMBL [8]. GB-GA is highly effective in straightforward optimisation problems. However, like other genetic algorithms, GB-GA faces difficulties when encountering low-performing valleys or local optima in chemical space due to collapse of molecular diversity in its evolutionary population [9, 10].

Quality-diversity algorithms address the challenges of navigating low-performing valleys and escaping local optima by incorporating novelty search and behavioural diversity into the optimisation process of genetic algorithms. The multi-dimensional archive of phenotypic elites [11] (MAP-Elites) is the premier quality-diversity algorithm. It rewards novelty and enforces diversity by replacing the evolutionary population of a genetic algorithm with a multi-dimensional archive containing the best solution found so far in each niche of a pre-defined search space. MAP-Elites produces more diverse and higher performing populations throughout its optimisation run compared to genetic algorithms, providing the algorithm with crucial stepping stones [12] that allow it to traverse low-performing valleys and bypass local optima.

In a previous paper, we introduced MAP-Elites to molecular optimisation by dividing small-molecule chemical space into a grid of niches based on physicochemical properties. In that paper, we combined MAP-Elites with the graph-based molecular representation, mutations, and crossovers from GB-GA. This novel algorithm, which we call graph-based elite patch illumination (GB-EPI) [13], solved the stagnation issues seen in GB-GA while simultaneously

improving search efficiency and producing a larger diversity of high-performing, yet qualitatively different molecules. The resulting multi-dimensional archive is filled with locally elite molecules, providing a comprehensive overview or illumination of the fitness potential of the explored parts of chemical space.

The other major limitation of using genetic algorithms in molecular optimisation stems from the inherent randomness of the molecules generated by mutations and crossovers. A genetic algorithm does not explicitly make use of information about the fitness of molecules encountered during its optimisation run. This limitation hinders the algorithm's ability to optimally make use of its exploration of chemical space and slows down its convergence towards the desired molecules. The information generated by genetic algorithms could be exploited to prioritise the evaluation of promising molecules. A potential approach to use this information is to integrate Bayesian optimisation [14–21] (based on a surrogate fitness model and an appropriate acquisition function) with the genetic algorithm to determine which new molecules are selected for fitness evaluation.

In this paper we combine all of the above, see Figure 1, by applying state-of-the-art Gaussian processes for molecular representations to create surrogate fitness models for GB-EPI, the previously described quality-diversity algorithm for molecular optimisation based on GB-GA. We test a range of modern acquisition functions on the surrogate fitness models to determine their ability to solve the ensemble problem of identifying the best molecule for every niche. The resulting algorithm, which we call the graph-based Bayesian illumination (GB-BI), significantly improves efficiency compared to previous approaches. The upcoming paragraphs provide an overview and technical details of how genetic algorithms, quality-diversity algorithms, and Bayesian optimisation are applied, adapted, and combined to formulate GB-BI. A lightweight, open-source version of GB-BI is available for download on GitHub.

2.1 Graph-based genetic algorithm (GB-GA)

For its fundamental operations, GB-BI relies on the molecular representations, mutations and crossovers and the core logic of GB-GA. Specifically, GB-GA provides an optimisation cycle ensuring the refinement and exploration of molecular structures upon which we apply extensions from quality-diversity algorithms and Bayesian optimisation. This optimisation cycle consist of three steps:

1. mutations and crossovers act on molecules, randomly selected from the population, to introduce variability,

2. each newly generated molecule undergoes evaluation based on a predefined fitness or evaluation function,

3. as a form of selection pressure, only the highest-scoring molecules present in the population are retained.

These steps are iterated until the fitness call budget is exhausted or a predefined maximum of generations is reached. In GB-GA, and hence also in GB-BI and GB-EPI, molecules are encoded by their molecular graphs, and the mutations and the crossovers act on these graphs. In practice, the mutations and the crossovers of GB-GA are implemented using the chemical reaction capabilities of the open-source cheminformatics package RDKit [22]. To rule out unwanted and potentially toxic molecules, GB-GA discards molecules containing macrocycles, allene centers in rings, fewer than five heavy atoms, or incorrect valences. GB-EPI also applies functional group filters from the ChEMBL database [8,23] in combination with restrictions on absorption, distribution, metabolism and excretion (ADME) properties [24–26] on candidate molecules before they enter the evaluation step.

2.2 Graph-based elite patch illumination (GB-EPI)

For its quality-diversity capabilities, GB-BI relies on the data-architecture and selection policies of GB-EPI, which in turn used MAP-Elites to solve the stagnation issues of GB-GA. To reliably outperform deep generative models for small molecule design, GB-EPI mimics diversity in biological evolution by assigning candidate solutions from a GB-GA to different niches depending on their characterising features. In each generation, the best performing candidate molecule in each of the individual niches is retained, creating a population of locally elite and diverse solutions that can be used as a resource to escape evolutionary stagnation. In GB-EPI, and hence also in GB-BI, users can choose which physicochemical properties of interest to use in spanning chemical space [27, 28], and select the boundaries within which a grid of niches is created.

Limiting the number of niches in MAP-Elites is important to avoid dilution of the evolutionary pressure that guides the algorithm. To counter the exponential growth of the amount of niches, advanced implementations of MAP-Elites, including GB-EPI and GP-BI, make use of a centroidal Voronoi tessellation [29–32] (CVT) instead of a regular grid, because it can cover a high-dimensional space with a fixed and predefined number of niches regardless of the amount of properties used to span it. GB-EPI and GB-BI also make use of positional analogue scanning [33] (structure modifications are applied in systematic batches by the mutation operator) and fitness function memoisation [34] (keeping an on-the-fly record of obtained results to ensures that an algorithm does not unnecessarily repeat expensive calculations) to increase their efficiency.

Algorithm: Graph-Based Bayesian Illumination (GB-BI)

Input: G – the number of generations, $\mathcal{M}_0$ – the initial population, $\mathcal{N}$ – the collection of niches

$\mathcal{F}_0 \leftarrow \mathsf{fitness}(\mathcal{M}_0)$;
for $i = 1 \rightarrow G$ **do**
  $\mathcal{M}_i \leftarrow \mathcal{M}_{i-1}, \mathcal{F}_i \leftarrow \mathcal{F}_{i-1}$;
  $\mathcal{M}' \leftarrow \mathsf{mutation}(\mathcal{M}_i) + \mathsf{crossover}(\mathcal{M}_i)$;
  for *molecule* **in** $\mathcal{M}'$ **do**
    niche $\leftarrow \mathsf{features}($*molecule*$)$;
    performance $\leftarrow \mathsf{surrogate}($*molecule*$)$;
    improvement $\leftarrow \mathsf{acquisition}($*performance, niche*$)$;
  end
  for *niche* **in** $\mathcal{N}$ **do**
    molecule $\leftarrow \mathsf{argmax}($*improvement*$[$*niche*$])$;
    fitness $\leftarrow \mathsf{evaluate}($*molecule*$)$;
    if *fitness* $> \mathcal{F}_{i-1}$*[niche]* **then**
      $\mathcal{M}_i$*[niche]* $\leftarrow$ *molecule*;
      $\mathcal{F}_i$*[niche]* $\leftarrow$ *fitness*;
  end
  surrogate $\leftarrow \mathsf{update}(\mathcal{M}_i, \mathcal{F}_i)$;
end
Result: $\mathcal{M}_N$ – molecules, $\mathcal{F}_N$ – fitnesses

Figure 2: Pseudocode description of the Bayesian Illumination algorithm as applied to the optimisation of small molecules.

2.3 Gaussian processes in chemistry (GAUCHE)

To address the limited exploitation of fitness information in GB-GA and GB-EPI, GB-BI makes use of a surrogate model to guide the selection of new molecules, coming from mutations and crossovers, for fitness evaluation. A good surrogate model [35, 36] accurately approximates the true fitness landscape based on the data it observed and estimates its own uncertainties. In molecular machine learning, Bayesian neural networks [37, 38] and ensembles [39, 40] are sometimes used for property predictions with uncertainty estimation. However, in scenarios where the dataset is small, deep learning models face challenges in achieving out-of-distribution generalisation. Gaussian processes are often a preferred alternative [41] for the small data regime, because of their ability to perform exact Bayesian inference and their minimal need for manual determination of hyper-parameters [41].

At their core, Gaussian processes are non-parametric probabilistic models that define a distribution over functions, where any finite set of function values follows a joint Gaussian distribution. The kernel function [42], which defines the degree of similarity between pairs of input points, plays a central role in Gaussian processes. This allows Gaussian processes to capture complex relationships between inputs and outputs while providing uncertainty estimates for the predictions. Training Gaussian processes on molecules does, however, introduce new challenges, particularly given the unfavourable properties of common molecular representations such as SMILES [43] and SELFIES [44] strings (variable lengths), fingerprints [45–47] (high-dimensional and sparse), and molecular graphs [48, 49] (non-continuous).

Extending Gaussian processes to efficiently work with these representations is nontrivial, but recent work, under the name GAUCHE [50], has provided the cheminformatics community with bespoke kernels designed to deal with these challenges. In this paper, we will specifically make use of the Tanimoto kernel provided by GAUCHE, which is defined as

$$K_{\text{Tanimoto}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \frac{< \mathbf{x}, \mathbf{x}' >}{||\mathbf{x}||^2 + ||\mathbf{x}'||^2 - < \mathbf{x}, \mathbf{x}' >}, \tag{1}$$

where is $\mathbf{x}$ and $\mathbf{x}'$ are binary fingerprint vectors, $< \cdot, \cdot >$ is the Euclidean inner product, $|| \cdot ||^2$ is the Euclidean norm and σ_f is the scalar kernel signal amplitude hyperparameter. In accordance with the GAUCHE documentation, we initially applied the Tanimoto kernel to extended connectivity fingerprints [51, 52] (ECFP4, ECFP6) and bag-of-characters [53, 54] representations of SMILES and SELFIES strings. Note that the bag-of-characters representation was previously used in the SMILES string kernel [55].

In this paper, we aimed to enhance the exploration of molecular representations for the surrogate fitness models, going beyond the recommendations and documentation of GAUCHE. Our investigation expanded to include the application of the Tanimoto kernel to a diverse set of fingerprints [56, 57]. In line with the extended connectivity fingerprints, we make use of feature-based connective fingerprints (FCFP4, FCFP6) which add internal labels, denoting whether an atom is basic or acidic, aromatic, halogen, or a hydrogen bond donor or acceptor. We also use the fingerprints included with RDKit (RDFP), atom pair fingerprints [58] (APFP) which focus on local substructures, and topological torsion fingerprints [59] (TTFP) which aim to predominantly capture long-range substructures. By incorporating this varied set of fingerprints into our study of GB-BI, we intend to cover a wider range of molecular characteristics and aim to evaluate the effectiveness of these different fingerprints in aligning with underlying structure-fitness relationships.

2.4 Bayesian optimisation of elites (BOP-Elites)

In GB-BI, we use Bayesian optimisation with GB-EPI to select candidate molecules for evaluation. Bayesian optimisation employs the combination of a surrogate model (including un-

certainties) and an acquisition function to balance exploration and exploitation, guiding the selection of candidate solutions. In quality-diversity algorithms, this translates into an ensemble problem where the population must balance novelty search, behavioural diversity, and efficient exploitation of the fitness function. The recently developed Bayesian optimisation for the MAP-Elites algorithm [60] (BOP-Elites) addresses this issue by constructing surrogate models for the objective function and descriptor functions simultaneously and select candidate solutions to maximise a global acquisition function across niches.

In the typical use-cases of GB-BI, the descriptor functions are physicochemical properties which can be efficiently calculated, so that we can make use of a simplified version of BOP-Elites in which niche occupancy is determined exactly and a surrogate model is only constructed for the fitness function. As acquisition function, BOP-Elites uses the expected improvement [61–63] (EI) which considers both the probability of improving on the current solution in a niche and the magnitude of the predicted improvement. In the case where descriptor functions are calculated exactly, this acquisition function boils down to

$$\mathrm{EI}(x) = \sigma(x)\, h\left(\frac{\mu(x) - y}{\sigma(x)}\right), \tag{2}$$

where x denotes the candidate solution, $\mu(\cdot)$ and $\sigma(\cdot)$ are respectively the posterior mean and variance of the surrogate fitness model, and y is the best function value observed so far in the niche, also referred to as the *incumbent* value. In the above equation, the helper function $h(\cdot)$ is defined as

$$h(z) = \phi(z) + z\Phi(z), \tag{3}$$

where ϕ and Φ are respectively the probability density function and the cumulative density function of the normal distribution.

In GB-BI, we follow the approach of BOP-Elites: every generation the algorithm only evaluates the candidate solution with the highest EI in each niche. In addition, we also explore the effectiveness of other acquisition functions as alternatives for EI. The most straightforward acquisition function we consider is the posterior mean, $\mu(x)$. We also consider the upper confidence bound [14] (UCB) acquisition function, which is commonly used in multi-armed bandit problems [64, 65] and is defined as

$$\mathrm{UCB}(x) = \mu(x) + \beta\sigma(x), \tag{4}$$

where β is a trade-off hyperparameter (sometimes denoted as the *confidence* parameter), which controls the balance between exploitation and exploration. For the purposes of this paper, we fix this hyperparameter to 0.2. The final acquisition function considered in this paper, is a numerically stable variant of the logarithm of the expected improvement [66] (logEI), which was recently introduced to alleviate the vanishing gradient problems sometimes encountered in the classical version of EI and is defined as

$$\mathrm{logEI}(x) = \log_h\left(\frac{\mu(x) - y}{\sigma(x)}\right) + \log(\sigma(x)), \tag{5}$$

in which the helper function $\log_h(\cdot)$ is a numerically stable implementation [67, 68] of the composite function $\log \circ h$.

Before BOP-Elites, the Surrogate-Assisted Illumination (SAIL) algorithm [69] introduced the idea of using surrogate models to efficiently explore and map a design space based on user-defined features. SAIL differs from BOP-Elites in several key ways, including reliance on the UCB acquisition function, use of a pre-defined computational budget, and the exclusive application of surrogate models to fitness functions. The latter aspect is reminiscent of Bayesian optimisation as applied in this paper, however, it is important to emphasise that SAIL

was designed specifically for continuous optimisation problems. A major distinction between molecular optimisation and continuous optimisation lies in how fitness is improved: molecular optimisation often requires specific mutations or crossovers to enhance fitness, while continuous spaces allow for smoother transitions between solutions. BOP-Elites, in contrast, showed promise in the discrete optimisation setting by making use of EI as an acquisition function. Hence, in this paper, we will refer to BOP-Elites as the basis for our exploration of chemical space with a surrogate-assisted quality-diversity algorithm.

2.5 Graph-based bayesian illumination (GB-BI)

To recapitulate, GB-BI is an illumination algorithm for efficiently generating optimised small molecules combining ideas from genetic algorithms (GB-GA), quality-diversity algorithms (GB-EPI) and Gaussian processes for small molecule representations (GAUCHE) and Bayesian optimisation (BOP-Elites). In GB-BI, molecules are acted upon as molecular graphs for mutations and crossovers, and represented by either fingerprints or as a bag-of characters based on SMILES or SELFIES for use in Gaussian processes. Mutations and crossovers are used to generate new molecules which are added to the evolutionary population, based on their comparative fitness with the current occupier of their physicochemical niche. Gaussian processes are used to create a surrogate fitness function and only those molecules with the highest acquisition function value within a single niche[1] receive a fitness evaluation.

In this manner, GB-BI aims to combine the stepping stones of quality-diversity algorithms with the functional call efficiency of Bayesian optimisation. Note that in GB-BI, exploration-vs-exploitation is controlled by a combination of the specifics of the physicochemical archive, the accuracy of the Gaussian processes, and the chosen acquisition function. For near-optimal fitness values, GB-BI might suffer from numerically vanishing or otherwise uninformative acquisition function values and – as an unfortunate consequence – the reintroduction of stagnation issues. To avoid this in this paper, we will explore the combinations of nine different molecular representations (ECFP4, ECFP6, FCFP4, FCFP6, RDFP, APFP, TTFP, SMILES and SELFIES) and four different acquisition functions (EI, Posterior Mean, UCB, and logEI) for their efficiency in rediscovering existing molecules.

3 Results and benchmarks

Several open-source benchmarking suites for the de novo design of small molecules have been developed in recent years. Most notable among these are GuacaMol [2], Tartarus [71], and the Therapeutics Data Commons [72]. GuacaMol offers lightweight tasks focused on molecule rediscovery where the fitness of a generated molecule is assessed using the Tanimoto similarity [73,74] between the generated molecule and the target molecule, based on their respective extended-connectivity fingerprints. Tartarus and the Therapeutics Data Commons, conversely, present more challenging and computationally intensive benchmark tests utilising docking methods [75,76], which evaluate the theoretical affinity between a small molecule and a target protein. These benchmarking suites have been extensively employed, facilitating a fair and open comparison with other published methods for generating small molecules.

GuacaMol has rediscovery tasks for three known drugs: Celecoxib (an anti-inflammatory), Troglitazone (an antidiabetic), and Thiothixene (an antipsychotic). Note that these three rediscovery tasks have been previously solved by several different methods [2], while exhibiting

[1] In contrast with trust regions in Bayesian optimisation, which dynamically adjust the search space to balance exploration and exploitation [70], niches in GB-BI serve as fixed subspaces that promote diversity in solution discovery. While trust regions focus on high-performing candidates, niches ensure that solutions are distributed across feature space, preserving a range of high-quality yet diverse results.

varying levels of reliability and efficiency. In previous work [13], we used the hardest of these tasks, the rediscovery of Troglitazone, to quantify the efficiency between GB-EPI and GB-GA. In this paper, we apply the Troglitazone rediscovery task to assess the effectiveness of GB-BI. Unlike computationally demanding docking tasks, the evaluation of Tanimoto similarities incurs limited computational costs. This enables us to efficiently gather a substantial amount of statistical data on the performance of GB-BI with regard to the use of different molecular representations and acquisition functions. A drawback of the simple rediscovery tasks of GuacaMol is their relatively lack of discriminative capabilities [77].

The commonly employed alternative to fingerprint-based rediscovery are docking-based tasks, as seen in Tartarus and the Therapeutics Data Commons. However, these tasks come with a relatively high computational cost [78–81]. To strike a balance, our paper introduces a new type of benchmark: descriptor-based rediscovery of small molecules. Instead of relying on fingerprints, these tasks encode a randomly selected conformer of the target molecule using either a Ultrafast Shape Recognition with CREDO Atom Types [82] (USRCAT) or Zernike [83] descriptor. We evaluate the fitness of generated molecules using customized similarity metrics that broadly align with Tanimoto similarity trends. These tasks prove to be more challenging and discriminative than fingerprint-based rediscovery, yet significantly less resource-intensive than docking tasks. This approach allows for statistical calculations, facilitating the thorough comparison of different generative models.

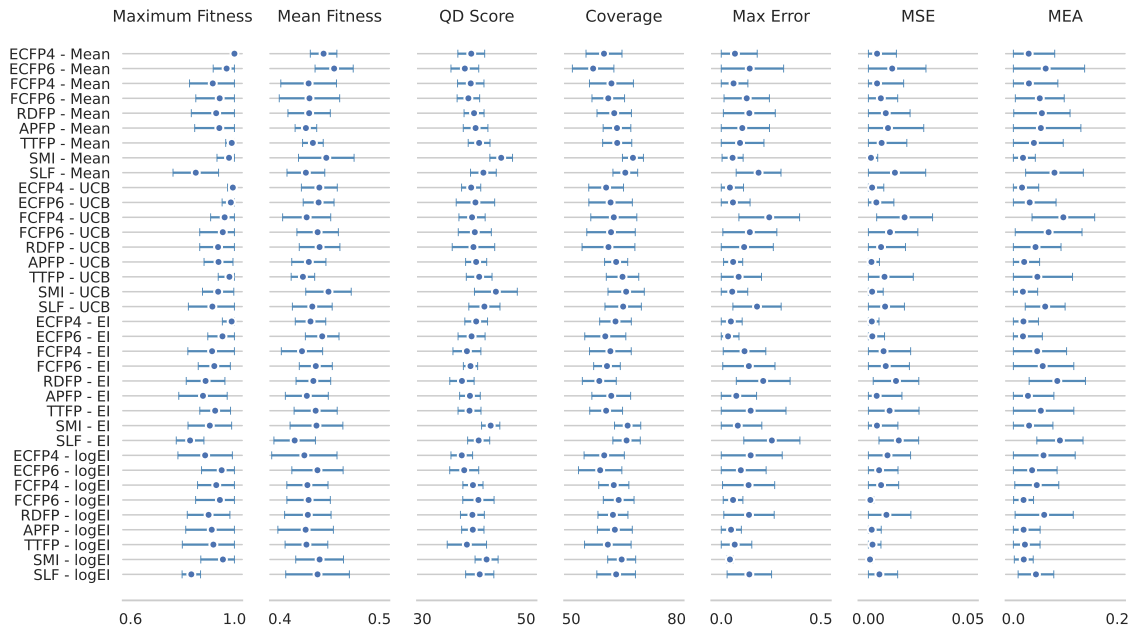

Figure 3: A comprehensive dot plot with error bars (one standard deviation) for all combinations of all representations (ECFP4, ECFP6, FCFP4, FCFP6, RDFP, APFP, TTFP, SMILES and SELFIES) and acquisition functions (EI, Posterior Mean, UCB, and logEI) the following metrics: the maximum and mean fitness within the evolutionary population, the QD-score, and the archive coverage the maximum prediction error, the mean squared error (MSE), and the mean absolute error (MAE). All values were determined based on ten independent, fingerprint-based Troglitazone rediscovery runs of GB-BI for each combination of acquisition function and representation, and were calculated on the final evolutionary population for each of those runs. For brevity, SMILES and SELFIES are abbreviated to SMI and SLF, respectively.

To complete the investigation into the efficiency of GB-BI, we conduct a comparison with GB-BA and GB-EPI using the standard docking task within the Therapeutics Data Commons. This task focuses on the dopamine receptor DRD3 [84, 85] as the protein target. Due to the computational expense of docking scores, the Therapeutics Data Commons restricts the number of function calls, enhancing the discriminative power of the task and demanding quick adaptation from generative models. In this paper, we also apply this approach to the molecular representation and acquisition function scans in both the fingerprint-based and descriptor-based rediscovery tasks. Additionally, since docking tasks are theoretically open-ended without a distinct optimal fitness value, the Therapeutics Data Commons introduces supplementary metrics to gauge the realism of the generated molecules.

Note that, while the Practical Molecular Optimisation (PMO) benchmark [86] is a more recent framework for molecular optimisation, and has a specific focus on function call efficiency, we opted not to use it in our study due to several critical flaws that make it unsuitable for fair comparisons. One major issue is the inconsistency in initial populations: deep generative models are pretrained on the entire initial database, while genetic algorithms in PMO are restricted to a small subset of the first 1000 molecules. This discrepancy skews results and alters the optimisation landscape [87], making it difficult to draw fair conclusions. Additionally, PMO does not standardise molecules across different tasks, which introduces bias, particularly when charged and neutral molecules are compared by the same pretrained oracle. Finally, PMO includes deep learning-based oracles which are easily exploitable [77] and have faced reproducibility issues, as documented in the public code repository [88, 89].

3.1 Fingerprint-based rediscovery of small molecules

In GuacaMol, molecules highly similar to the target (bit-vector Tanimoto similarity above 0.323) are removed from the initial population of molecules to increase the effectiveness of the benchmarks. That initial population is randomly selected from ChEMBL, which exclusively consists of molecules that have both been synthesised in a lab and tested against biological targets. To set up GB-BI for the GuacaMol rediscovery benchmark of Troglitazone, we chose the feature space to be spanned by molecular mass, 225 u. to 555 u., a lipophilicity, $\log \mathrm{P} = -0.4$ to $\log \mathrm{P} = 5.6$, a topological polar surface area (TPSA) between 0 Å^2 and 140 Å^2, and a Wildman-Crippen molar refractivity value between 40 and 130. The ranges were chosen to roughly correspond to properties of orally active drugs, and in accordance with our previous work the archive was set up with 150 niches.

To enhance the discriminative power of these benchmarks, we limit the maximum number of fitness function calls to 5000 per run and gather statistics of ten runs for each possible combination of molecular representation and acquisition function. In addition, to increase real-world relevance, we filter out molecules that fail at Veber's rule [26]. In quality-diversity algorithms, like GB-BI, there are several relevant metrics to track the performance of an algorithm. These metrics include the maximum and mean fitness within the evolutionary population, the QD-score [90, 91] (sum of fitness values in the evolutionary population), and the coverage of the archive (percentage of niches containing a molecule). Additionally, we monitor metrics assessing the accuracy of the surrogate fitness model, including the maximum prediction error, mean squared error (MSE), and mean absolute error (MAE). Each of these metrics is calculated for every generation across all GB-BI runs analysed in this paper.

Based on these metrics, a thorough analysis of quality-diversity and surrogate fitness model metrics is conducted during the parameter scan of GB-BI on the GuacaMol rediscovery task of Troglitazone. A comprehensive overview of the results is presented in Figure 3 which displays the mean and standard deviation for all six metrics, calculated for the final generations of ten independent runs per combination of molecular representations and acquisition functions. An

Table 1: Efficiency of GB-BI, GB-EPI and GB-GA in the rediscovery of Troglitazone, in terms of the average number of required score evaluations and the success ratio. We also include the publication year of the algorithm. For GB-BI, we present the results for ten independent, randomly seeded runs for the ECFP4 representation in combination with the posterior mean acquisition function. For GB-EPI and GB-GA, we use the results of a previous study in which 100 independent, randomly seeded runs of both those algorithms were analysed.

Rediscovery of Troglitazone			
Algorithm	**Fitness Evaluations (↓)**	**Success Ratio (↑)**	**Year**
GB-BI	**629**	**100 %**	2024
GB-EPI	14,258	**100 %**	2020
GB-GA	24,216	81 %	2019

initial examination of these results highlights both the importance and challenges of tracking multiple metrics while studying the effectiveness of quality-diversity methods. The QD-Score, for instance, largely displays the same trends as the archive coverage, but is less reflective of the actual mean fitness value of the evolutionary populations considered here. At the same time, the three metrics assessing the accuracy of the surrogate fitness model, display nearly identical trends but lack strong correlation with either the maximum or mean fitness values.

To gain a more comprehensive understanding of how various molecular representations and acquisition functions influence the performance of GB-BI, we present a detailed analysis incorporating maximum fitness value, maximum error, rediscovery rate, and average fitness calls required for rediscovery. These insights are depicted in a series of heatplots showcased in Figure 4. Notably, these figures clearly reveal that only the posterior mean coupled to the ECFP4 fingerprint representation achieves a perfect rediscovery rate within the allocated budget of 5000 fitness function calls. Surprisingly, neither EI nor logEI outperform simpler acquisition functions like UCB or the posterior mean in terms of both maximum fitness or rediscovery rate. Moreover, a high maximum fitness score or a low maximum error does not necessarily correlate with a high rediscovery rate, underscoring the significance of evaluating multiple complimentary metrics when analysing the effectiveness of different quality-diversity methods for small molecule generation.

GB-BI, utilising the ECFP4 representation for the surrogate model and the posterior mean as an acquisition function, achieves a perfect rediscovery rate with, on average, 629 fitness function calls required. This marks a substantial improvement compared to GB-EPI, which is approximately 23 times less efficient than GB-BI, and GB-GA, which is roughly 38 times less efficient than GB-BI, as indicated in Table 1. Notably, GB-GA encounters stagnation issues, necessitating a minimum of 3 searches for successful rediscovery with at least 99% certainty. Factoring in this requirement would escalate the necessary number of fitness function calls for GB-GA to about 72,000, making it roughly 115 times less efficient than GB-BI. Recently, it has been asserted [92] that novel generative molecular models must demonstrate a clear advantage over genetic algorithms to be considered impactful in advancing the research field. GB-BI and GP-EPI unequivocally meet this GA criterion.

Furthermore, despite the vastness of chemical space, estimated at 10^{60} molecules, there is an argument [87] suggesting that an ideal, all-powerful search algorithm could pinpoint small drug-like molecules within a few hundred fitness function evaluations. Considering the efficiency demonstrated by GB-BI and keeping this idealized benchmark in mind, it becomes evident that the long-term efficacy of restricting the maximum allowed fitness calls to enhance the discriminative power of fingerprint rediscovery tasks for small molecules is at risk. To pre-

vent future benchmarking issues and to further challenge and engage the research field, this paper introduces two novel classes of benchmarks. These benchmarks involve encoding a randomly selected conformer of a target molecule into either a USRCAT or Zernike descriptor, along with tailored similarity metrics. Although both these benchmarks closely resemble fingerprint rediscovery tasks, they present significantly increased optimisation challenges while simultaneously remaining computationally affordable in comparison to docking-based tasks.

3.2 Descriptor-based rediscovery of small molecules

In response to the computational challenges posed by conventional fingerprint-based rediscovery and resource-intensive docking tasks, this paper pioneers an alternative approach: descriptor-based rediscovery of small molecules. Departing from the conventional method of determining a target molecule and calculating a fingerprint as the basis for similarity, our alternative involves sampling a random conformer and replacing the fingerprint with either a USRCAT or Zernike descriptor. To evaluate the similarity of a candidate molecule to the target, we sample multiple conformers, calculate corresponding descriptors, and apply custom, aggregating similarity metrics to this collection of descriptors. These metrics are expressly designed to align with the broad trends of the corresponding fingerprint-based Tanimoto similarities, while creating a more challenging pathway for optimisation algorithms to reach the target molecule.

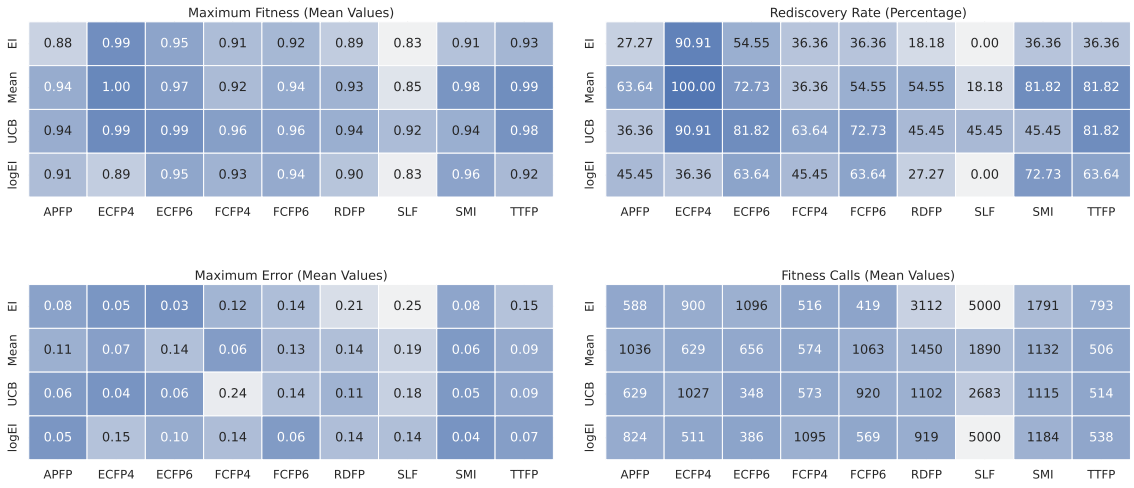

Figure 4: Heatmaps for the mean values for the maximum fitness (upper, left) and the maximum error (bottom, left) are shown. Additional calculations were made to determine and show heatplots for the rate of rediscovery of Troglitazone (upper, right) and the average amount of fitness calls it requires (bottom, right). These four heatmaps reflect the performance of the various combinations of representations and acquisition functions for GB-BI based on ten independent runs for each combination. Darker hue's correspond to a better performance. Note that only successful Troglitazone rediscoveries were considered in computing the mean fitness call value, i.e. displayed values are excluding failed runs, but in the two cases were there was no successful rediscovery of Troglitazone at all, the mean fitness calls were fixed to the overall maximum (5000). For compactness and improved readability, SMILES and SELFIES are abbreviated to SMI and SLF, respectively.

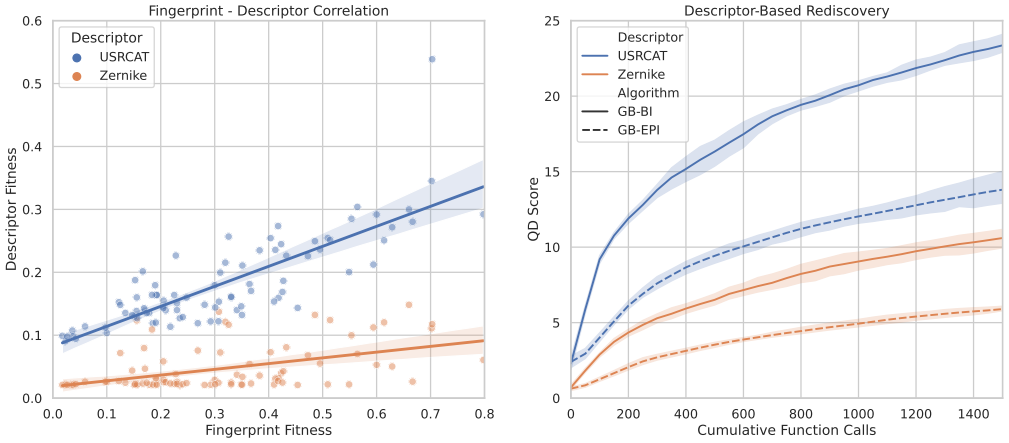

Figure 5: Left: A comparison of the correlation between USRCAT (blue) and Zernike (orange) fitness values and the ECFP4 fingerprint fitness for the rediscovery of Troglitazone. One molecule for each occupied niche was sampled from the entire evolutionary history of a single randomly chosen but successful GB-BI run for this task. Both distributions are shown on the same scale. Right: The QD-score for the descriptor-based rediscovery of Troglitazone in function of the cumulative function calls (capped at a maximum of 1500), making use of either the USRCAT descriptor (blue) or the Zernike descriptor (orange), for 5 independent runs of the GB-BI (full line) and GB-EPI (dashed line) algorithms. Results for both rediscovery tasks are shown on the same scale.

Ultrafast Shape Recognition [93] (USR) descriptors are characterised by a set of statistical moments of distance distributions created by measuring distances between atoms and reference points. USRCAT and Zernike descriptors are extension on these USR descriptors. USRCAT descriptors [82] incorporate additional information about the presence of hydrophobic, aromatic, hydrogen bond donor, and acceptor atoms while Zernike descriptors [83] replace statistical moments with projections on orthogonal basis functions. Both USRCAT and Zernike descriptors have been shown to outperform traditional USR descriptors in virtual screening benchmarks [82,83]. In practical terms, we rely on the RDKit implementation of USRCAT and to calculate the Zernike descriptors efficiently, we make use of the just-in-time (JIT) compilation capabilities of the open-source LLVM package Numba [94].

The similarity between individual pairs of USRCAT descriptors is canonically calculated by the USRscore [82], a special variant of the Manhattan distance. In line with the use of USRscore for USRCAT descriptors, we use the Canberra distance [95,96], which is a weighted version of the Manhattan distance, to calculate the similarity between individual pairs of Zernike descriptors. For use in the descriptor-based rediscovery benchmarks presented in this paper, we propose using a conformer-aggregated similarity metric S which, with respect to the target molecule, is defined [97] as

$$S = \max_{i=1}^{k} (\text{similarity}(c_i, t)) \,, \tag{6}$$

where t is a given, fixed descriptor for the target molecule and $(c_1, c_2, ..., c_k)$ is a collection of descriptors for sampled conformers of the candidate molecule. We employ the stochastic conformer generator provided by RDKit (ETKDG.v3) to sample and generate the necessary conformers for the candidate molecules. Throughout this paper, we will sample 15 conformers for each candidate molecule. This approach results in a set of similarity measures which were explicitly designed to follow the overall trends (but not the numerical values) of fingerprint-based Tanimoto similarities.

Table 2: Correlations of the target similarities of the USRCAT and Zernike descriptors with respect to the fingerprint-based similarity for the rediscovery of Troglitazone, in terms of the Pearson correlation, Spearman's ρ and Kendall's τ. These correlations were computed based on molecules uniformly sampled across archive niches, originating from a single, randomly selected, successful GB-BI run for the Troglitazone rediscovery task.

Correlation with Fingerprint Similarity			
Descriptor	**Pearson**	**Spearman's ρ**	**Kendall's τ**
USRCAT	0.81	0.79	0.61
Zernike	0.46	0.44	0.30

To explicitly demonstrate the relationship between descriptor-based and fingerprint-based rediscovery, we randomly selected a successful GB-BI run for fingerprint-based rediscovery of Troglitazone. From this specific run, we systematically sampled molecules, ensuring the inclusion of a single molecule (that had already been selected for fitness evaluation) from each occupied niche. Subsequently, we applied the two novel similarity metrics to these molecules, and compared those with the previously recorded fingerprint fitness values, as shown on the left-hand side of Figure 5. To quantitatively evaluate the correlation between descriptor-based similarities and fingerprint-based similarities, we calculated the Pearson correlation [98–100], which gauges the linear relationship between two variables, alongside Spearman's ρ [101] and Kendall's τ [102]. Both of these latter correlation measures are based on the rank of the data rather than the raw values. The results, presented in Table 2, demonstrate a stronger correlation between USRCAT similarity and fingerprint-based similarity than the correlation between Zernike similarity and fingerprint-based similarity. Note that values coming from a conformer-aggregated similarity metric, as defined here, are strictly non-negative and remain bound between zero and unity (e.g. self-similarity).

Based on these new conformer-aggregated similarity metrics, we conducted a comparative analysis involving five distinct runs, each allocated a budget of 1500 fitness calls, of both the GB-BI and GB-EPI algorithms for the USRCAT and Zernike-based rediscovery tasks of Troglitazone. To facilitate this evaluation, we repurposed the initial dataset of molecules, along with the structural filters and ADMET requirements, originally employed in the GuacaMol rediscovery benchmarks. Similarly, both algorithms were initialised with the archive configuration previously used by GB-BI in the GuacaMol rediscovery benchmark. The archive thus encompasses 150 niches and is spanned along four physicochemical properties: molecular mass (ranging from 225 u. to 555 u.), lipophilicity (log P ranging from -0.4 to 5.6), TPSA (ranging from 0 Å^2 to 140 Å^2), and Wildman-Crippen molar refractivity (ranging between 40 and 130). For the surrogate fitness model, we choose to make use of the ECFP4 fingerprint as molecular representation and decided on the posterior mean as acquisition function.

To assess the efficacy of GB-BI and GB-EPI algorithms in the context of USRCAT and Zernike-based Troglitazone rediscovery, we monitored the maximum and mean fitness within the evolutionary population, as well as the QD-score. The statistical summaries of these metrics across the final population of each run, are presented in Table 3. From these results, we can make the rather remarkable observation that neither GB-BI nor GB-EPI have managed to rediscover Troglitazone based on either USRCAT or Zernike descriptors. Clearly, the descriptor-based rediscovery tasks are substantially more challenging than their fingerprint equivalent. This might be due to the presence of activity cliffs – pairs of molecules with highly similar molecular graphs but displaying large differences in fitness – in descriptor-based rediscovery. Another striking observation, seen in both benchmarks, is the lack of significant difference in

performance between GB-BI and GB-EPI in terms of either the maximum or mean fitness in the population. GB-BI does perform significantly better in terms of QD-score, see the right-hand side of Figure 5, which indicates that it manages to generate a larger and more diverse population of (comparatively) high-scoring molecules.

3.3 Efficient organic photovoltaics

The Tartarus benchmarking suite includes tasks aimed at enhancing the efficiency of organic solar cells by evaluating single point GFN2-xTB calculations [103] of candidate molecules. These tasks focus on identifying small organic donor molecules with optimal power conversion efficiency for use in bulk heterojunction devices. Specifically, we selected four tasks for this study: maximising HOMO-LUMO gap values, minimising LUMO energy values, maximising the molecular dipole moment and maximising a combined score defined as the molecular dipole moment plus the HOMO-LUMO gap minus the LUMO energy. This combined score approximates the power conversion efficiency of the proposed molecules. For ease of evaluation, we make use of a surrogate deep learning model, delivered with Tartarus, which was trained on a subset of approximately 25,000 molecules sampled from the Harvard Clean Energy Project Database [104] to predict the optimisation values for these tasks.

The results, summarised at the top of Table 4, demonstrate the superior performance of GB-BI compared to previous generative models. Notably, the JANUS model performs well in these tasks due to its explicit use of molecular fragments. To further refine these tasks, we imposed a maximum limit of 2500 fitness function evaluations. The deep learning model within Tartarus uses ECFP4 fingerprints for molecular representation and was trained to predict optimisation values for specific tasks. For GB-BI and GB-EPI, we employed an archive of 150 niches, covering molecular masses from 200 u to 700 u, lipophilicity values from $\log P = -0.5$ to $\log P = 5.5$, topological polar surface areas (TPSA) from 0 $Å^2$ to 300 $Å^2$, and Wildman-Crippen molar refractivity values between 0 and 300. No structural filters were applied.

Table 3: Performance evaluation of GB-BI and GB-EPI in the descriptor-based rediscovery tasks for Troglitazone, in terms of the mean and standard deviation of the maximum fitness and mean fitness of the evolutionary population and the QD-score. The presented statistics were obtained based on five independent runs of GB-BI and GB-EPI, each with a maximum budget of 1500 fitness calls, for both benchmarks.

USRCAT Rediscovery			
Algorithm	**Max Fitness (↑)**	**Mean Fitness (↑)**	**QD Score (↑)**
GB-BI	0.39 ± 0.03	0.26 ± 0.01	23.36 ± 0.80
GB-EPI	0.42 ± 0.02	0.26 ± 0.02	13.80 ± 1.29
Zernike Rediscovery			
Algorithm	**Max Fitness (↑)**	**Mean Fitness (↑)**	**QD Score (↑)**
GB-BI	0.24 ± 0.01	0.13 ± 0.01	10.60 ± 0.83
GB-EPI	0.24 ± 0.03	0.12 ± 0.02	5.90 ± 0.21

Table 4: Top: Optimisation results obtained in five independent runs of four tasks (maximising HOMO-LUMO gap values, minimising LUMO energy values, maximising the molecular dipole moment and maximising a combined score) related to optimal power conversion efficiency of organic photovoltaics for GB-BI, GB-EPI, GB-GA, JANUS and REINVENT. Each optimisation run is limited to 2500 fitness function calls. Bottom: Optimisation results obtained in three independent runs of the DRD3, ABL1, and EGFR docking tasks (including SAS modulation) for the GB-BI, GB-EPI and GB-GA algorithms, limited to 1000 fitness function calls and subject to structural filters from ChEMBL and Veber's rule of druglikeness. In addition to the minimum and mean docking score, we also report the QD-score and archive coverage for both quality-diversity algorithms.

Efficient Organic Photovoltaics				
Algorithm	**Humo-Lumo Gap (↑)**	**Lumo Energy (↓)**	**Molecular Dipole Moment (↑)**	**Power Conversion Efficiency (↑)**
GB-BI	**2.76 ± 0.00**	**-9.44 ± 0.01**	**8.22 ± 0.21**	**18.18 ± 0.13**
GB-EPI	**2.76 ± 0.00**	-9.40 ± 0.01	8.04 ± 0.10	18.17 ± 0.10
GB-GA	2.73 ± 0.00	-9.29 ± 0.05	7.68 ± 0.45	17.46 ± 0.16
JANUS	2.75 ± 0.00	-9.42 ± 0.02	7.74 ± 0.38	18.11 ± 0.21
REINVENT	2.59 ± 0.03	-9.18 ± 0.04	6.73 ± 0.11	16.91 ± 0.36

Small Molecule Protein Binders				
Algorithm	**Minimum Docking (↓)**	**Mean Docking (↓)**	**Quality-Diversity Score (↓)**	**Archive Coverage (↑)**
Target Protein: Dopamine D3 Receptor (DRD3)				
GB-BI	**-12.05 ± 0.25**	**-10.77 ± 0.17**	**-638.76 ± 21.36**	**45.11 % ± 1.39 %**
GB-EPI	-11.10 ± 0.30	-9.98 ± 0.18	-471.89 ± 19.47	34.89 % ± 3.15 %
GB-GA	-10.81 ± 0.18	-9.64 ± 0.20	N/A	N/A
Target Protein: Tyrosine-Protein Kinase ABL (ABL1)				
GB-BI	**-11.99 ± 0.44**	**-10.97 ± 0.37**	**-652.82 ± 4.39**	**45.11 % ± 1.54 %**
GB-EPI	-11.10 ± 0.34	-9.93 ± 0.06	-443.82 ± 20.74	33.78 % ± 2.14 %
GB-GA	-10.72 ± 0.24	-9.53 ± 0.23	N/A	N/A
Target Protein: Epidermal Growth Factor Receptor (EGFR)				
GB-BI	**-12.22 ± 0.08**	**-11.17 ± 0.11**	**-674.63 ± 19.32**	**46.67 % ± 4.16 %**
GB-EPI	-11.06 ± 0.07	-10.01 ± 0.15	-461.80 ± 20.24	35.11 % ± 1.68 %
GB-GA	-10.85 ± 0.32	-9.69 ± 0.23	N/A	N/A

3.4 Small molecule protein binders

The Therapeutics Data Commons (TDC) provides docking molecule generation benchmarks[2] which evaluate the theoretical binding affinity between small molecules and target proteins. Docking is widely used for virtual screening of compounds, as molecules with higher theoretical binding affinities are statistically more likely to have a higher bioactivity [106]. To increase real-life relevance, we apply stringent structural and ADME filters to candidate molecules and modulate [81] the docking results with a synthethic accessibility score (SAS), as suggested in the documentation, for the proposed small molecule. We select three different targets from the TDC benchmarking suite: a dopamine receptor (DRD3) implicated in schizophrenia [85] and essential tremor syndrome [107], a tyrosine-protein kinase (ABL1) implicated in chronic myelogenous leukemia [108], and the epidermal growth factor receptor (EGFR) which has been strongly associated with a number of cancers [109], including lung cancer [110], glioblastoma [111] and epithelial tumours of the head and neck [112].

[2] It is important to note that the objective functions employed in this study cannot be compared with those on the public leaderboard due to broken backwards compatibility [105] for docking tasks, TDC versioning has been updated to 1.0.0 to reflect this.

To set up the archives for GB-BI and GB-EPI, we created 150 niches by defining a feature space that aligns with the properties of orally active drugs. The chosen ranges for these features were: molecular mass from 225 u to 555 u, lipophilicity (expressed as log P) from -0.4 to 5.6, topological polar surface area (TPSA) from 0 Å^2 to 140 Å^2, and Wildman-Crippen molar refractivity from 40 to 130. These specific ranges were selected to reflect characteristics commonly found in orally active drugs, ensuring that the molecules evaluated would be relevant for potential pharmaceutical applications. For all three algorithms – GB-GA, GB-EPI, and GB-BI – we utilized a batch size of 40 molecules, maintaining this size for the initial selection as well. This consistent batch size ensures that each algorithm evaluates an equivalent set of molecules, providing a fair comparison of their performance. Additionally, we applied stringent criteria to filter out unsuitable molecules. Specifically, molecules that exhibited structural alerts, which are indicative of potential toxicity or other undesirable properties, were excluded from the archives. Furthermore, we removed any molecules that failed to meet Veber's rule of drug-likeness.

The results of the protein binding tasks for three independent runs of GB-BI, GB-EPI, and GB-GA are presented at the bottom of Table 4. GB-BI consistently outperforms both GB-EPI and GB-GA across all three tasks, achieving superior results in terms of both the minimum obtained docking score and the mean docking score for the 100 best compounds at the end of optimisation for each algorithm. To evaluate the quality-diversity effectiveness of GB-BI and GB-EPI, we calculate the QD-score [113] (the sum of all fitness values of the molecules present in the archive at the last generation) and the percentage of archive niches occupied by a molecule, known as archive coverage [114]. Since GB-GA is not a quality-diversity algorithm, we do not calculate the QD-score or archive coverage for it. The QD-score is a widely used quality-diversity measure that assesses an algorithm's ability to populate its archive with diverse yet high-performing solutions. In the context of this paper, the archive represents a section of chemical space, and our results indicate that GB-BI demonstrates a significantly enhanced capacity to generate a variety of optimised molecules compared to the standard quality-diversity approach used in GB-EPI.

4 Conclusion and outlook

Recently, it has been asserted [92] that to progress the research field, novel generative molecular models must demonstrate a clear advantage over genetic algorithms. Traditional deep generative models and genetic algorithms have struggled to consistently deliver optimised small molecules, either due to inefficiencies in information utilisation or evolutionary stagnation. In this paper, we introduce Bayesian Illumination, a novel approach that combines Gaussian processes with quality-diversity methods to address these shortcomings. Through an extensive series of molecular optimisation tasks, – ranging from drug rediscovery and multi-property optimisation to efficient power conversion and the design of protein binders – based on three independent benchmarking suites, we robustly show that Bayesian Illumination displays state-of-the-art efficiency in finding optimal molecular structures in chemical space. In addition, it is worth noting that Bayesian Illumination also generates a larger diversity of high-scoring molecules than a standard quality-diversity method without Bayesian optimisation.

In conclusion, by combining key aspects of genetic algorithms, quality-diversity methods, and Bayesian optimisation, Bayesian illumination sets a new baseline for the efficient and effective molecular optimisation of small molecules. Bayesian illumination's success is an important indication that there is plenty of opportunity left for improvement over current deep generative models and genetic algorithms. For instance, during numerical experiments, we noticed that the performance of the surrogate and acquisition functions can in some cases rely

strongly on the chosen molecular representation. This is a potential shortcoming of Bayesian Illumination and the integration of a data driven molecular representation is an interesting subject for future work. Bayesian Illumination also opens up new avenues for future research and applications regarding the optimisation of chemical reactions [115], particularly in the context of data-driven representations [116], the design of optimal protein and peptide structures [117, 118], and the efficient exploration of chemical databases [119].

Acknowledgments

The author wishes to acknowledge useful feedback on this manuscript by K. Beshkov and P. Coppin and to acknowledge J. Van den Abeele for interesting discussions during the early stages of the research presented here. The author would like to thank the community, and anonymous reviewers at the ICML24 ML4LMS workshop, for their valuable feedback on earlier versions of this manuscript, particularly regarding the SAIL algorithm, the PMO benchmarking suite, and the diversity of fingerprints used in this paper.

Funding information The work presented here is supported by the Carlsberg Foundation, grant CF23-0939 and by UiO:Life Science through the 4MENT convergence environment.

Data availability Full code for the implementation of GB-BI is available at https://github.com/Jonas-Verhellen/Bayesian-Illumination. To ensure reproducibility, a permanent GitHub release tagged as v1.0-paper-submission has been created, which captures the exact version of the code and data used in this manuscript. An easy-to-use, online tool for using GB-BI with a limited set of fitness functions can be found at https://huggingface.co/spaces/jonas-verhellen/Bayesian-Illumination.

Conflicts of interest There are no conflicts to declare.

References

[1] D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, *Deep learning for molecular design – A review of the state of the art*, Mol. Syst. Des. Eng. **4**, 828 (2019), doi:10.1039/c9me00039a.

[2] N. Brown, M. Fiscato, M. H. S. Segler and A. C. Vaucher, *GuacaMol: Benchmarking models for de novo molecular design*, J. Chem. Inf. Model. **59**, 1096 (2019), doi:10.1021/acs.jcim.8b00839.

[3] J. H. Jensen, *A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space*, Chem. Sci. **10**, 3567 (2019), doi:10.1039/c8sc05372c.

[4] M. Mukaidaisi, A. Vu, K. Grantham, A. Tchagang and Y. Li, *Multi-objective drug design based on graph-fragment molecular representation and deep evolutionary learning*, Front. Pharmacol. **13** (2022), doi:10.3389/fphar.2022.920747.

[5] J. H. Holland, *Adaptation in natural and artificial systems*, MIT Press, Cambridge, USA, ISBN 9780262275552 (1992), doi:10.7551/mitpress/1090.001.0001.

[6] D. E. Goldberg and J. H. Holland, *Genetic algorithms and machine learning*, Mach. Learn. **3**, 95 (1988), doi:10.1007/bf00113892.

[7] T. Sterling and J. J. Irwin, *ZINC 15 – Ligand discovery for everyone*, J. Chem. Inf. Model. **55**, 2324 (2015), doi:10.1021/acs.jcim.5b00559.

[8] D. Mendez et al., *ChEMBL: Towards direct deposition of bioassay data*, Nucleic Acids Res. **47**, D930 (2018), doi:10.1093/nar/gky1075.

[9] A. Nigam, P. Friederich, M. Krenn and A. Aspuru-Guzik, *Augmenting genetic algorithms with deep neural networks for exploring the chemical space*, (arXiv preprint) doi:10.48550/arXiv.1909.11655.

[10] Z. Zhou and K. D. M. Harris, *Counteracting stagnation in genetic algorithm calculations by implementation of a micro genetic algorithm strategy*, Phys. Chem. Chem. Phys. **10**, 7262 (2008), doi:10.1039/B807326K.

[11] J.-B. Mouret and J. Clune, *Illuminating search spaces by mapping elites*, (arXiv preprint) doi:10.48550/arXiv.1504.04909.

[12] J. Nordmoen, F. Veenstra, K. O. Ellefsen and K. Glette, *MAP-Elites enables powerful stepping stones and diversity for modular robotics*, Front. Robot. AI **8** (2021), doi:10.3389/frobt.2021.639173.

[13] J. Verhellen and J. Van den Abeele, *Illuminating elite patches of chemical space*, Chem. Sci. **11**, 11485 (2020), doi:10.1039/d0sc03544k.

[14] H. J. Kushner, *A versatile stochastic model of a function of unknown and time varying form*, J. Math. Anal. Appl. **5**, 150 (1962), doi:10.1016/0022-247x(62)90011-2.

[15] H. J. Kushner, *A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise*, J. Basic Eng. **86**, 97 (1964), doi:10.1115/1.3653121.

[16] M. Pelikan, D. E. Goldberg and E. Cantú-Paz, *BOA: The Bayesian optimization algorithm*, in *Proceedings of the 1st annual conference on genetic and evolutionary computation*, Morgan Kaufmann Publishers, San Francisco, USA, ISBN 9781558606111 (1999).

[17] 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 **104**, 148 (2016), doi:10.1109/JPROC.2015.2494218.

[18] J. P. Janet, S. Ramesh, C. Duan and H. J. Kulik, *Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization*, ACS Cent. Sci. **6**, 513 (2020), doi:10.1021/acscentsci.0c00026.

[19] R. Garnett, *Common Bayesian optimization policies*, in *Bayesian optimization*, Cambridge University Press, Cambridge, UK, ISBN 9781108348973 (2023), doi:10.1017/9781108348973.008.

[20] X. Wang, Y. Jin, S. Schmitt and M. Olhofer, *Recent advances in Bayesian optimization*, ACM Comput. Surv. **55**, 1 (2023), doi:10.1145/3582078.

[21] H. Kneiding and D. Balcells, *Augmenting genetic algorithms with machine learning for inverse molecular design*, Chem. Sci. **15**, 15522 (2024), doi:10.1039/D4SC02934H.

[22] G. Landrum, *Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling* (2013), https://www.rdkit.org.

[23] P. Walters, *Practical cheminformatics: Filtering chemical libraries* (2018), https://practicalcheminformatics.blogspot.com/2018/08/filtering-chemical-libraries.html.

[24] C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Adv. Drug Deliv. Rev. **23**, 3 (1997), doi:10.1016/S0169-409X(96)00423-1.

[25] W. J. Egan, K. M. Merz, and J. J. Baldwin, *Prediction of drug absorption using multivariate statistics*, J. Med. Chem. **43**, 3867 (2000), doi:10.1021/jm000292e.

[26] D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *Molecular properties that influence the oral bioavailability of drug candidates*, J. Med. Chem. **45**, 2615 (2002), doi:10.1021/jm020017n.

[27] C. M. Dobson, *Chemical space and biology*, Nature **432**, 824 (2004), doi:10.1038/nature03192.

[28] J.-L. Reymond, *The chemical space project*, Acc. Chem. Res. **48**, 722 (2015), doi:10.1021/ar500432k.

[29] Q. Du, V. Faber and M. Gunzburger, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev. **41**, 637 (1999), doi:10.1137/s0036144599352836.

[30] Q. Du and M. Gunzburger, *Grid generation and optimization based on centroidal Voronoi tessellations*, Appl. Math. Comput. **133**, 591 (2002), doi:10.1016/s0096-3003(01)00260-0.

[31] Y. Liu, W. Wang, B. Lévy, F. Sun, D.-M. Yan, L. Lu and C. Yang, *On centroidal Voronoi tessellation—energy smoothness and fast computation*, ACM Trans. Graph. **28**, 1 (2009), doi:10.1145/1559755.1559758.

[32] Q. Du, M. Emelianenko and L. Ju, *Convergence of the Lloyd algorithm for computing centroidal Voronoi tessellations*, SIAM J. Numer. Anal. **44**, 102 (2006), doi:10.1137/040617364.

[33] N. A. Meanwell, *Improving drug design: An update on recent applications of efficiency metrics, strategies for replacing problematic elements, and compounds in nontraditional drug space*, Chem. Res. Toxicol. **29**, 564 (2016), doi:10.1021/acs.chemrestox.6b00043.

[34] J. Hughes, *Lazy memo-functions*, in *Lecture notes in computer science*, Springer, Berlin, Heidelberg, Germany, ISBN 9783540159759 (1985), doi:10.1007/3-540-15975-4_34.

[35] P. I. Frazier, *A tutorial on Bayesian optimization*, (arXiv preprint) doi:10.48550/arXiv.1807.02811.

[36] B. Lei, T. Q. Kirk, A. Bhattacharya, D. Pati, X. Qian, R. Arroyave and B. K. Mallick, *Bayesian optimization with adaptive surrogate models for automated experimental design*, Npj Comput. Mater. **7**, 194 (2021), doi:10.1038/s41524-021-00662-x.

[37] Y. Zhang and A. A. Lee, *Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning*, Chem. Sci. **10**, 8154 (2019), doi:10.1039/c9sc00616h.

[38] G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li and W. H. Green, *Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction*, J. Chem. Inf. Model. **60**, 2697 (2020), doi:10.1021/acs.jcim.9b00975.

[39] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay and C. W. Coley, *Uncertainty quantification using neural networks for molecular property prediction*, J. Chem. Inf. Model. **60**, 3770 (2020), doi:10.1021/acs.jcim.0c00502.

[40] J. Busk, P. B. Jørgensen, A. Bhowmik, M. N. Schmidt, O. Winther and T. Vegge, *Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks*, Mach. Learn.: Sci. Technol. **3**, 015012 (2021), doi:10.1088/2632-2153/ac3eb3.

[41] C. E. Rasmussen, *Gaussian processes in machine learning*, in *Lecture notes in computer science*, Springer, Berlin, Heidelberg, Germany, ISBN 9783540231226 (2004), doi:10.1007/978-3-540-28650-9_4.

[42] M. Kanagawa, P. Hennig, D. Sejdinovic and B. K. Sriperumbudur, *Gaussian processes and kernel methods: A review on connections and equivalences*, (arXiv preprint) doi:10.48550/arXiv.1807.02582.

[43] E. Anderson, G. D. Veith and D. Weininger, *SMILES, a line notation and computerized interpreter for chemical structures*, Tech. Rep. 600M87021, Environmental Protection Agency, Washington D.C., USA (1987).

[44] M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation*, Mach. Learn.: Sci. Technol. **1**, 045024 (2020), doi:10.1088/2632-2153/aba947.

[45] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé and G. Pujadas, *Molecular fingerprint similarity search in virtual screening*, Methods **71**, 58 (2015), doi:10.1016/j.ymeth.2014.08.005.

[46] I. Muegge and P. Mukherjee, *An overview of molecular fingerprint similarity search in virtual screening*, Expert Opin. Drug Discov. **11**, 137 (2015), doi:10.1517/17460441.2016.1117070.

[47] S. Riniker and G. A. Landrum, *Similarity maps – A visualization strategy for molecular fingerprints and machine-learning methods*, J. Cheminformatics **5**, 43 (2013), doi:10.1186/1758-2946-5-43.

[48] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta and J. Leskovec, *Open graph benchmark: Datasets for machine learning on graphs*, in *Proceedings of the 34th international conference on neural information processing systems*, Curran Associates, Red Hook, USA, ISBN 9781713829546 (2020).

[49] S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, *Molecular graph convolutions: Moving beyond fingerprints*, J. Comput.-Aided Mol. Des. **30**, 595 (2016), doi:10.1007/s10822-016-9938-8.

[50] R.-R. Griffiths et al., *GAUCHE: A library for Gaussian processes in chemistry*, (arXiv preprint) doi:10.48550/arXiv.2212.04450.

[51] H. L. Morgan, *The generation of a unique machine description for chemical structures–a technique developed at chemical abstracts service*, J. Chem. Doc. **5**, 107 (1965), doi:10.1021/c160017a018.

[52] D. Rogers and M. Hahn, *Extended-connectivity fingerprints*, J. Chem. Inf. Model. **50**, 742 (2010), doi:10.1021/ci100050t.

[53] Z. S. Harris, *Distributional structure*, in *Papers on syntax*, Springer, Dordrecht, Netherlands, ISBN 9789027712677 (1981), doi:10.1007/978-94-009-8467-7_1.

[54] A. Z. Broder, S. C. Glassman, M. S. Manasse and G. Zweig, *Syntactic clustering of the Web*, Comput. Netw. ISDN Syst. **29**, 1157 (1997), doi:10.1016/s0169-7552(97)00031-7.

[55] D.-S. Cao, J.-C. Zhao, Y.-N. Yang, C.-X. Zhao, J. Yan, S. Liu, Q.-N. Hu, Q.-S. Xu and Y.-Z. Liang, *In silico toxicity prediction by support vector machine and SMILES representation-based string kernel*, SAR QSAR Environ. Res. **23**, 141 (2012), doi:10.1080/1062936x.2011.645874.

[56] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick and J. W. Davies, *How similar are similarity searching methods? A principal component analysis of molecular descriptor space*, J. Chem. Inf. Model. **49**, 108 (2009), doi:10.1021/ci800249s.

[57] D. Bajusz, A. Rácz and K. Héberger, *Chemical data formats, fingerprints, and other molecular descriptions for database analysis and searching*, in *Comprehensive medicinal chemistry III*, Elsevier, Amsterdam, Netherlands, ISBN 9780128032015 (2017), doi:10.1016/b978-0-12-409547-2.12345-5.

[58] R. E. Carhart, D. H. Smith and R. Venkataraghavan, *Atom pairs as molecular features in structure-activity studies: Definition and applications*, J. Chem. Inf. Comput. Sci. **25**, 64 (1985), doi:10.1021/ci00046a002.

[59] R. Nilakantan, N. Bauman, J. S. Dixon and R. Venkataraghavan, *Topological torsion: A new molecular descriptor for SAR applications. Comparison with other descriptors*, J. Chem. Inf. Comput. Sci. **27**, 82 (1987), doi:10.1021/ci00054a008.

[60] P. Kent, A. Gaier, J.-B. Mouret and J. Branke, *BOP-Elites, a Bayesian optimisation approach to quality diversity search with black-box descriptor functions*, (arXiv preprint) doi:10.48550/arXiv.2307.09326.

[61] V. R. Saltines, *One method of multiextremum optimization*, Avtom. Vychisl. Tekh. **5**, 33 (1971).

[62] J. Močkus, *On Bayesian methods for seeking the extremum*, in *Lecture notes in computer science*, Springer, Berlin, Heidelberg, Germany, ISBN 9783540071655 (1975), doi:10.1007/3-540-07165-2_55.

[63] D. R. Jones, M. Schonlau and W. J. Welch, *Efficient global optimization of expensive black-box functions*, J. Glob. Optim. **13**, 455 (1998), doi:10.1023/A:1008306431147.

[64] T. L. Lai and H. Robbins, *Asymptotically efficient adaptive allocation rules*, Adv. Appl. Math. **6**, 4 (1985), doi:10.1016/0196-8858(85)90002-8.

[65] T. Leung Lai, *Adaptive treatment allocation and the multi-armed bandit problem*, Ann. Stat. **15**, 1091 (1987), doi:10.1214/aos/1176350495.

[66] S. Ament, S. Daulton, D. Eriksson, M. Balandat and E. Bakshy, *Unexpected improvements to expected improvement for Bayesian optimization*, (arXiv preprint) doi:10.48550/arXiv.2310.20708.

[67] M. Mächler, *Accurately computing* $\log(1-\exp(-|a|))$ *assessed by the Rmpfr package* (2012), https://cran.r-project.org/web/packages/Rmpfr/vignettes/log1mexp-note.pdf.

[68] 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*, (arXiv preprint) doi:10.48550/arXiv.1910.06403.

[69] A. Gaier, A. Asteroth and J.-B. Mouret, *Data-efficient exploration, optimization, and modeling of diverse designs through surrogate-assisted illumination*, Proc. Genet. Evol. Comput. Conf. 99 (2017), doi:10.1145/3071178.3071282.

[70] D. Eriksson, M. Pearce, J. R. Gardner, R. Turner and M. Poloczek, *Scalable global optimization via local Bayesian optimization*, (arXiv preprint) doi:10.48550/arXiv.1910.01739.

[71] A. Nigam, R. Pollice, G. Tom, K. Jorner, J. Willes, L. A. Thiede, A. Kundaje and A. Aspuru-Guzik, *Tartarus: A benchmarking platform for realistic and practical inverse molecular design*, (arXiv preprint) doi:10.48550/arXiv.2209.12487.

[72] K. Huang et al., *Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development*, F1000Research **12** (2021), doi:10.12688/f1000research.channels.700.

[73] T. T. Tanimoto, *An elementary mathematical theory of classification and prediction*, Tech. Rep. 196401, International Business Machines, New York, USA (1958).

[74] D. Bajusz, A. Rácz and K. Héberger, *Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?*, J. Cheminformatics **7**, 20 (2015), doi:10.1186/s13321-015-0069-3.

[75] G. M. Morris and M. Lim-Wilby, *Molecular docking*, in *Methods in molecular biology*, Humana Press, Totowa, USA, ISBN 9781588298645 (2008), doi:10.1007/978-1-59745-177-2_19.

[76] N. S. Pagadala, K. Syed and J. Tuszynski, *Software for molecular docking: A review*, Biophys. Rev. **9**, 91 (2017), doi:10.1007/s12551-016-0247-1.

[77] P. Renz, D. Van Rompaey, J. K. Wegner, S. Hochreiter and G. Klambauer, *On failure modes in molecule generation and optimization*, Drug Discov. Today: Technol. **32**, 55 (2019), doi:10.1016/j.ddtec.2020.09.003.

[78] F. Stanzione, I. Giangreco and J. C. Cole, *Use of molecular docking computational tools in drug discovery*, in *Progress in medicinal chemistry*, Elsevier, Amsterdam, Netherlands, ISBN 9780323850568 (2021), doi:10.1016/bs.pmch.2021.01.004.

[79] D. E. Graff, E. I. Shakhnovich and C. W. Coley, *Accelerating high-throughput virtual screening through molecular pool-based active learning*, Chem. Sci. **12**, 7866 (2021), doi:10.1039/d0sc06805e.

[80] G. Corso, H. Stärk, B. Jing, R. Barzilay and T. Jaakkola, *DiffDock: Diffusion steps, twists, and turns for molecular docking*, (arXiv preprint) doi:10.48550/arXiv.2210.01776.

[81] C. Steinmann and J. H. Jensen, *Using a genetic algorithm to find molecules with good docking scores*, PeerJ Phys. Chem. **3**, e18 (2021), doi:10.7717/peerj-pchem.18.

[82] A. M. Schreyer and T. Blundell, *USRCAT: Real-time ultrafast shape recognition with pharmacophoric constraints*, J. Cheminformatics **4**, 27 (2012), doi:10.1186/1758-2946-4-27.

[83] V. Venkatraman, P. R. Chakravarthy and D. Kihara, *Application of 3D Zernike descriptors to shape-based ligand similarity searching*, J. Cheminformatics **1**, 19 (2009), doi:10.1186/1758-2946-1-19.

[84] B. Levant, *The D3 dopamine receptor: Neurobiology and potential clinical relevance*, Pharmacol. Rev. **49**, 231 (1997), doi:10.1016/s0031-6997(24)01327-9.

[85] R. Moraga-Amaro, H. Gonzalez, R. Pacheco and J. Stehberg, *Dopamine receptor D3 deficiency results in chronic depression and anxiety*, Behav. Brain Res. **274**, 186 (2014), doi:10.1016/j.bbr.2014.07.055.

[86] W. Gao, T. Fu, J. Sun and C. W. Coley, *Sample efficiency matters: A benchmark for practical molecular optimization*, (arXiv preprint) doi:10.48550/arXiv.2206.12411.

[87] E. S. Henault, M. H. Rasmussen and J. H. Jensen, *Chemical space exploration: How genetic algorithms find the needle in the haystack*, PeerJ Phys. Chem. **2**, e11 (2020), doi:10.7717/peerj-pchem.11.

[88] *Oracle unittests are failing. Issue #291*, https://github.com/mims-harvard/TDC/issues/291.

[89] *Oracles not satisfying public benchmarks. Issue #245*, https://github.com/mims-harvard/TDC/issues/245.

[90] J. K. Pugh, L. B. Soros, P. A. Szerlip and K. O. Stanley, *Confronting the challenge of quality diversity*, Proc. 2015 Annu. Conf. Genet. Evol. Comput. 967 (2015), doi:10.1145/2739480.2754664.

[91] B. Tjanaka, M. C. Fontaine and S. Nikolaidis, *Quantifying efficiency in quality diversity optimization*, in *Workshop on benchmarks for quality-diversity algorithms*, Boston, USA (2022), https://btjanaka.net/static/qd-auc/qd-auc-paper.pdf.

[92] A. Tripp and J. M. Hernández-Lobato, *Genetic algorithms are strong baselines for molecule generation*, (arXiv preprint) doi:10.48550/arXiv.2310.09267.

[93] P. J. Ballester and W. Graham Richards, *Ultrafast shape recognition to search compound databases for similar molecular shapes*, J. Comput. Chem. **28**, 1711 (2007), doi:10.1002/jcc.20681.

[94] S. K. Lam, A. Pitrou and S. Seibert, *Numba*, Proc. Second Workshop LLVM Compil. Infrastruct. HPC 1 (2015), doi:10.1145/2833157.2833162.

[95] G. N. Lance and W. T. Williams, *Computer programs for hierarchical polythetic classification ("similarity analyses")*, Comput. J. **9**, 60 (1966), doi:10.1093/comjnl/9.1.60.

[96] G. N. Lance and W. T. Williams, *Mixed-data classificatory programs I - Agglomerative systems*, Aust. Comput. J. **1**, 15 (1967).

[97] J. Robles, F. Sotelo, C. Rojas, J. Hurtado and J. Lopez, *Performance analysis of XGBoost models with ultrafast shape recognition descriptors in ligand-based virtual screening*, Proc. 2021 8th Int. Conf. Bioinform. Res. Appl. 8 (2021), doi:10.1145/3487027.3487029.

[98] K. Pearson, *VII. Note on regression and inheritance in the case of two parents*, Proc. R. Soc. Lond. **58**, 240 (1895), doi:10.1098/rspl.1895.0041.

[99] A. Bravais, *Analyse mathématique sur les probabilités des erreurs de situation d'un point*, Impr. Royale (1844).

[100] S. Wright, *Correlation and causation*, J. Agric. Res. **20**, 557 (1921).

[101] C. Spearman, *The proof and measurement of association between two things*, in *Studies in individual differences: The search for intelligence*, Appleton-Century-Crofts, New York, USA (1961), doi:10.1037/11491-005.

[102] M. G. Kendall, *A new measure of rank correlation*, Biometrika **30**, 81 (1938), doi:10.2307/2332226.

[103] C. Bannwarth, S. Ehlert and S. Grimme, *GFN2-xTB–An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions*, J. Chem. Theory Comput. **15**, 1652 (2019), doi:10.1021/acs.jctc.8b01176.

[104] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid*, J. Phys. Chem. Lett. **2**, 2241 (2011), doi:10.1021/jz200866s.

[105] *[Docking leaderboard DRD3] reproducibility issues. Issue #235* (2024), https://github.com/mims-harvard/TDC/issues/235.

[106] J. Lyu et al., *Ultra-large library docking for discovering new chemotypes*, Nature **566**, 224 (2019), doi:10.1038/s41586-019-0917-9.

[107] H. Deng, W. Le and J. Jankovic, *Genetics of essential tremor*, Brain **130**, 1456 (2007), doi:10.1093/brain/awm018.

[108] M. Baccarani et al., *The proportion of different BCR-ABL1 transcript types in chronic myeloid leukemia. An international overview*, Leukemia **33**, 1173 (2019), doi:10.1038/s41375-018-0341-4.

[109] N. Normanno et al., *Epidermal growth factor receptor (EGFR) signaling in cancer*, Gene **366**, 2 (2006), doi:10.1016/j.gene.2005.10.018.

[110] G. da Cunha Santos, F. A. Shepherd and M. S. Tsao, *EGFR mutations and lung cancer*, Annu. Rev. Pathol.: Mech. Dis. **6**, 49 (2011), doi:10.1146/annurev-pathol-011110-130206.

[111] M. Westphal, C. L. Maire and K. Lamszus, *EGFR as a target for glioblastoma treatment: An unfulfilled promise*, CNS Drugs **31**, 723 (2017), doi:10.1007/s40263-017-0456-6.

[112] M. Zimmermann, A. Zouhair, D. Azria and M. Ozsahin, *The epidermal growth factor receptor (EGFR) in head and neck cancer: Its role and treatment implications*, Radiat. Oncol. **1**, 11 (2006), doi:10.1186/1748-717x-1-11.

[113] J. K. Pugh, L. B. Soros and K. O. Stanley, *Quality diversity: A new frontier for evolutionary computation*, Front. Robot. AI **3** (2016), doi:10.3389/frobt.2016.00040.

[114] M. Flageat, B. Lim, L. Grillotti, M. Allard, S. C. Smith and A. Cully, *Benchmarking quality-diversity algorithms on neuroevolution for reinforcement learning*, (arXiv preprint) doi:10.48550/arXiv.2211.02193.

[115] J. Seumer, J. Kirschner Solberg Hansen, M. Brøndsted Nielsen and J. H. Jensen, *Computational evolution of new catalysts for the Morita-Baylis-Hillman reaction*, Angew. Chem. Int. Ed. **62**, e202218565 (2023), doi:10.1002/anie.202218565.

[116] B. Ranković, R.-R. Griffiths, H. B. Moss and P. Schwaller, *Bayesian optimisation for additive screening and yield improvements – Beyond one-hot encoding*, Digit. Discov. **3**, 654 (2024), doi:10.1039/D3DD00096F.

[117] B. Basanta, M. J. Bick, A. K. Bera, C. Norn, C. M. Chow, L. P. Carter, I. Goreshnik, F. Dimaio and D. Baker, *An enumerative algorithm for de novo design of proteins with diverse pocket structures*, Proc. Natl. Acad. Sci. **117**, 22135 (2020), doi:10.1073/pnas.2005412117.

[118] K. Boone, C. Wisdom, K. Camarda, P. Spencer and C. Tamerler, *Combining genetic algorithm with machine learning strategies for designing potent antimicrobial peptides*, BMC Bioinform. **22**, 239 (2021), doi:10.1186/s12859-021-04156-x.

[119] K. Klarich, B. Goldman, T. Kramer, P. Riley and W. P. Walters, *Thompson sampling–An efficient method for searching ultralarge synthesis on demand databases*, J. Chem. Inf. Model. **64**, 1158 (2024), doi:10.1021/acs.jcim.3c01790.