nature communications

Article

https://doi.org/10.1038/s41467-025-58503-0

# Uncertainty quantification with graph neural networks for efficient molecular design

Received: 17 July 2024

Accepted: 21 March 2025

Published online: 05 April 2025

Check for updates

Lung-Yi Chen [1] & Yi-Pei Li [1,2]

Optimizing molecular design across expansive chemical spaces presents unique challenges, especially in maintaining predictive accuracy under domain shifts. This study integrates uncertainty quantification (UQ), directed message passing neural networks (D-MPNNs), and genetic algorithms (GAs) to address these challenges. We systematically evaluate whether UQ-enhanced D-MPNNs can effectively optimize broad, open-ended chemical spaces and identify the most effective implementation strategies. Using benchmarks from the Tartarus and GuacaMol platforms, our results show that UQ integration via probabilistic improvement optimization (PIO) enhances optimization success in most cases, supporting more reliable exploration of chemically diverse regions. In multi-objective tasks, PIO proves especially advantageous, balancing competing objectives and outperforming uncertainty-agnostic approaches. This work provides practical guidelines for integrating UQ in computational-aided molecular design (CAMD).

The exploration of novel chemical materials is a pivotal scientific endeavor with the potential to significantly advance both the economy and society[1–4]. Historically, the discovery of innovative molecules has led to major breakthroughs in various fields, including the development of enhanced medical therapies[5], innovative catalysts for chemical reactions[6], and more efficient carbon capture technologies[7]. These discoveries have traditionally resulted from labor-intensive experimental processes characterized by extensive trial and error.

In response to the limitations of these traditional experimental approaches, computational-aided molecular design (CAMD) has emerged as a crucial innovation. By conceptualizing material design as an optimization problem, where molecular structures and their properties are treated as variables and objectives, CAMD harnesses computational power to efficiently predict and identify promising materials. The advent of sophisticated machine learning algorithms has marked a paradigm shift from conventional knowledge-based methods, such as the group contribution method[8–10], to advanced learning-based strategies[11]. Among these, deep learning has demonstrated exceptional accuracy and flexibility, modeling complex interrelations between chemical structures and properties that challenge traditional theoretical approaches[12]. For example, graph neural networks (GNNs) have emerged as powerful tools for representing molecular structures[13]. Unlike traditional models that rely on fixed molecular descriptors, GNNs operate directly on molecular graphs, capturing detailed connectivity and spatial relationships between atoms. This graph-based approach enables GNNs to model molecular interactions with high fidelity[14,15], making them particularly well-suited for applications in molecular design, where accurate structural representation is critical. Furthermore, GNNs offer scalability[16], enabling efficient processing of large datasets, which is essential for exploring the expansive chemical spaces required in CAMD.

As CAMD has evolved, it has incorporated various generative models and sophisticated optimization strategies that employ surrogate models as objective functions to enhance molecular design. For example, variational autoencoders (VAEs) have been widely used for molecular generation by encoding molecules into a latent space where new structures can be sampled and decoded, facilitating exploration of chemical space[17,18]. These VAEs are often coordinated with optimization techniques, such as evolutionary algorithms[19–21], Bayesian optimization (BO)[18,22], or Monte Carlo tree search (MCTS)[23] to guide the search toward novel molecules with desired properties. Similarly, SMILES-based recurrent neural networks (RNNs) have been employed to generate molecular structures[24–26]. After a pretraining phase, RNNs are often fine-tuned using reinforcement learning, enhancing the

[1]Department of Chemical Engineering, National Taiwan University, Taipei, Taiwan, ROC. [2]Taiwan International Graduate Program on Sustainable Chemical Science and Technology (TIGP-SCST), Taipei, Taiwan, ROC. e-mail: yipeili@ntu.edu.tw

model's ability to achieve goal-directed optimization. However, a significant challenge for generative models can be ensuring diversity in the generated molecules, especially if training data is limited or narrowly focused, which may limit their utility for exploring diverse chemical spaces[27]. Beyond generative models, some approaches apply optimization algorithms directly to molecular representations without requiring latent spaces. For instance, genetic algorithms (GAs) operate on molecular graphs[28,29] or SMILES strings[30,31], iteratively generating improved candidates through mutation and crossover operations. This approach bypasses the need for a pretrained generative model, making GAs adaptable and accessible for a variety of CAMD tasks. Compared with generative models, GAs can work well even with smaller datasets and may have lower initial computational demands, which can be beneficial for direct exploration and optimization of molecular properties. Additionally, their evolutionary principles naturally maintain diversity, supporting a broad exploration of chemical space and adaptability to specific property objectives[32].

Despite the promise of these optimization approaches, a major challenge with data-driven models in CAMD is their tendency to fail in accurately predicting properties for molecules outside their training scope. This limitation underscores the importance of integrating uncertainty quantification (UQ) into CAMD to assess prediction reliability. Previous studies have commonly addressed this challenge through BO frameworks, often using Gaussian process regression (GPR), including Kriging models[33]. These non-parametric models make predictions with uncertainty estimates based on the posterior distribution, leveraging a kernel function to define the covariance between training data points. However, the matrix inversion required for non-parametric methods can become time-consuming, particularly for large datasets, as the computational complexity scales $O(n^3)$ with the number of training data and $O(n)$ with the dimension of molecular features[34,35]. As a result, GPR models are typically constrained to smaller training datasets[36], limiting the chemical space that can be explored in CAMD[37]. To alleviate this computational bottleneck and enable the use of larger datasets, several approximation strategies have been proposed for GPR[38]. Low-rank or sparse methods (e.g., inducing-point approaches) address the $O(n^3)$ scaling by selecting a small subset of points (inducing points), reducing the effective size of the Gram matrix, and leading to a more manageable $O(nm^2)$ complexity. Random feature expansions—such as random Fourier features—approximate the kernel function by mapping data into a lower-dimensional feature space, converting GPR into an approximately linear model that can be trained in $O(nD)$ or $O(nD^2)$, where $D \ll n$. Distributed or parallel GPs divide the dataset across multiple machines or computational nodes, combining local posteriors to maintain predictive accuracy while managing larger data volumes. These techniques collectively address the high computational burden of Gaussian process models and may expand the applicability of BO-driven CAMD to larger chemical search spaces.

In contrast, parametric models like GNNs offer a scalable alternative, as they maintain a fixed number of parameters regardless of dataset size, allowing efficient handling of larger datasets. UQ has been successfully integrated with parametric models for active learning and virtual screening, enhancing workflow efficiency[39,40]. However, optimizing over expansive chemical spaces presents distinct challenges, as accurate UQ under domain shifts remains notoriously difficult[41,42]. To the best of our knowledge, whether UQ integration within parametric models can enable effective optimization across broad, open-ended chemical spaces—and how best to implement this—remains an open question. Such an approach is particularly valuable for CAMD, as it enables exploration across vast and less-characterized chemical spaces essential for discovering novel compounds.

In this work, we address this issue by combining GNNs with GAs for molecular optimization, allowing direct exploration of chemical space without reliance on predefined libraries or generative models. To mitigate errors associated with surrogate model predictions in extrapolated regions, we integrate UQ into our GNN framework[43–45]. Inspired by acquisition functions used in BO[46], we systematically investigate different ways to incorporate UQ into CAMD, including probabilistic improvement and expected improvement methods. Our experiments show that the probabilistic improvement optimization (PIO) approach, which uses probabilistic assessments to guide the optimization process, is particularly effective in facilitating exploration of chemical space with GNNs. Given that practical applications often require molecular properties to meet specific thresholds rather than extreme values[47,48], the PIO method quantifies the likelihood that a candidate molecule will exceed predefined property thresholds, reducing the selection of molecules outside the model's reliable range and promoting candidates with superior properties.

Our study includes a comprehensive evaluation of uncertainty-agnostic and uncertainty-aware optimization approaches using the Tartarus[49] and GuacaMol[48] platforms, both open-source molecular design tools addressing a range of design challenges. Tartarus utilizes physical modeling across various software packages to estimate target properties, effectively simulating the experimental evaluations required in molecular design processes, while GuacaMol focuses on drug discovery tasks such as similarity searches and physicochemical property optimization. The benchmarking workflow, illustrated in Fig. 1, starts with datasets from these platforms to develop GNN-based surrogate models using the directed message passing neural network (D-MPNN) implemented in Chemprop[50]. These models predict molecular properties and their uncertainties, which, when coupled with a GA, optimize molecular structures based on the PIO and other selected fitness functions. Our results indicate that the PIO method substantially improves the likelihood of meeting threshold requirements, especially in multi-objective optimization tasks.

In summary, this integration of UQ with GNNs for CAMD represents a pioneering approach, offering a more reliable and scalable strategy for discovering novel chemical materials. Through extensive benchmarking and validation, our work demonstrates the potential of uncertainty-aware GNN algorithms in molecular design, with promising applications across domains such as organic electronics, biochemistry, and materials science.

## Results

### Molecular design benchmarks

To effectively evaluate molecular design strategies, tasks must be complex enough to reflect the challenges encountered in real-world applications. Our study provides a comprehensive assessment of different optimization approaches across 19 molecular property datasets, encompassing 10 single-objective and 6 multi-objective tasks (Table 1), derived from the Tartarus[49] and GuacaMol[48] platforms.

The first platform, Tartarus[49], offers a sophisticated suite of benchmark tasks tailored to address practical molecular design challenges within the realms of materials science, pharmaceuticals, and chemical reactions. Utilizing well-established computational chemistry techniques, including force fields and density functional theory (DFT), Tartarus models complex molecular systems with high computational efficiency. The benchmarks encompass a wide array of applications, ranging from optimizing organic photovoltaics and discovering novel organic light-emitting diodes (OLEDs) to designing protein ligands and pioneering new chemical reactions. This breadth enables a comprehensive evaluation of various molecular design algorithms across multiple real-world simulation scenarios.

Three molecular design categories from Tartarus, comprising seven single-objective and two multi-objective tasks, are listed in Table 1. Each task employs specific computational methods: organic emitter design involves conformer sampling[51], semi-empirical quantum mechanical methods for geometry optimization[52,53], and time-dependent DFT for single-point energy calculations[54]. Protein ligand

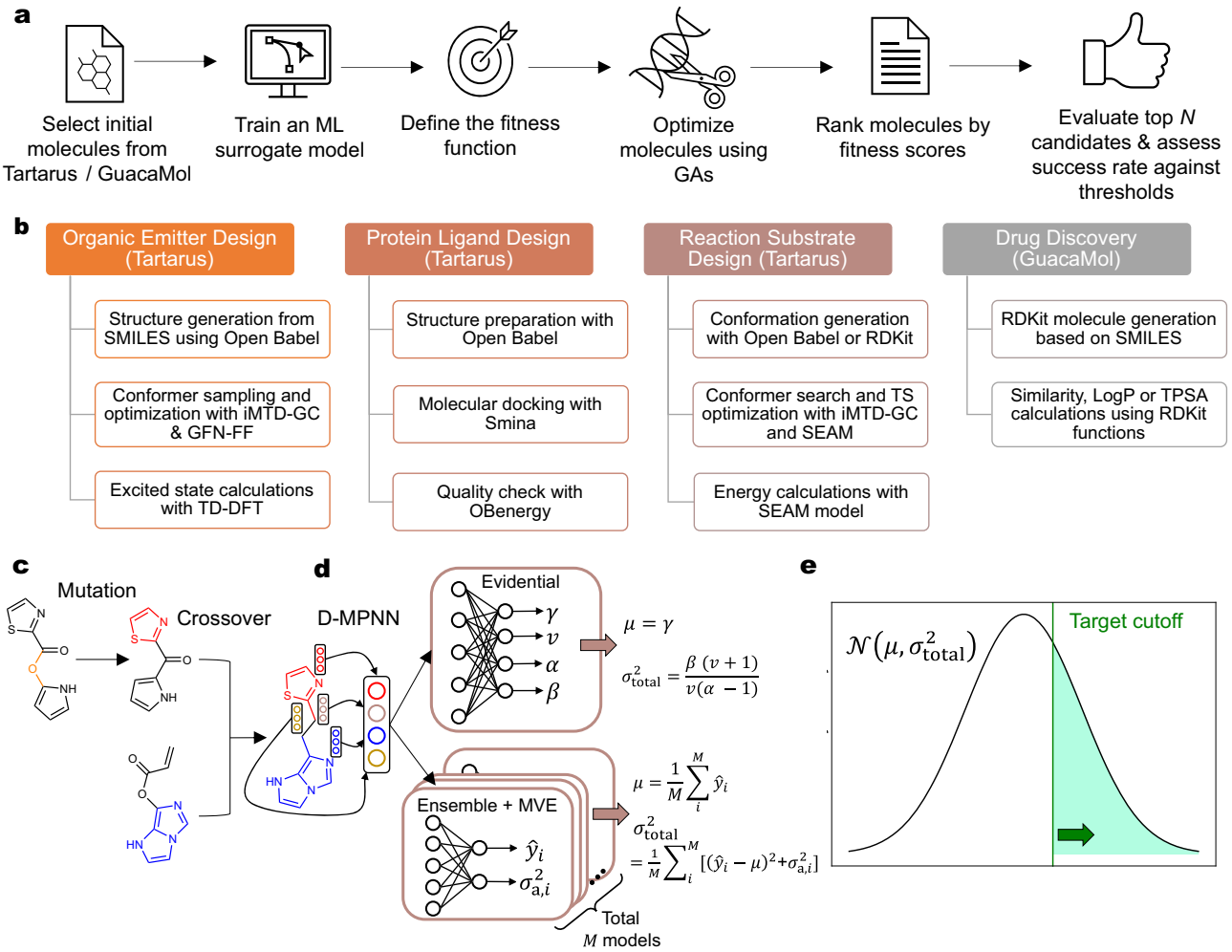


**Fig. 1 | Workflow and methodology for illustrating probabilistic improvement optimization (PIO) strategy. a** Schematic diagram illustrating the overall workflow used in this study to evaluate the optimization strategy. **b** Description of the benchmark tasks comprising three molecular design challenges from the Tartarus suite, which utilize physical modeling across different software packages to estimate target properties, circumventing the need for actual experimental assessments. Seven additional drug discovery tasks were selected from GuacaMol, using similarity metrics and physicochemical descriptor calculations as oracle functions to evaluate molecular properties. **c** Schematic representation of the genetic algorithm (GA), where the mutation operator randomly modifies molecular structures, and crossover operations generate new molecular structures through recombination. **d** Construction of a machine learning (ML) surrogate model employing the directed message passing neural network (D-MPNN) architecture, designed to predict molecular properties and their associated uncertainties via either the evidential method or the ensemble with mean-variance estimation (MVE) method. **e** The PIO fitness function, calculated using probability improvement, generally enhances the likelihood of meeting threshold requirements. TS transition state, log*P* octanol-water partition coefficient, TPSA topological polar surface area.

design utilizes docking pose searches to determine stable binding energies[55], supplemented by empirical functions for final score calculations[56]. Reaction substrate design tasks employ force fields for optimizing reactant and product structures[57], with transition state structures further refined using the SEAM method[58]. These methods include stochastic elements such as conformer search and docking site sampling, introducing variability in simulation outcomes due to the inherent randomness of geometry optimization. For multi-objective tasks, a typical approach might involve aggregating multiple objectives into a single composite score. However, this can lead to suboptimal compromises where certain objectives are sacrificed to maximize the overall score. In practical applications, molecules often need to satisfy multiple objectives simultaneously, which can be particularly challenging when these objectives are mutually constraining. To evaluate the efficacy of molecular design strategies under these conditions, we analyzed each objective within multi-objective tasks, choosing scenarios where objectives could potentially conflict. For example, the task of simultaneously minimizing both activation energy and reaction energy was excluded due to their positive correlation, as explained by the Bell–Evans–Polanyi principle[59,60]. Conversely, we included the task of simultaneously maximizing activation energy while minimizing reaction energy, as it poses a significant challenge by deviating from conventional expectations and thus aligns more closely with the aims of our study. These choices are detailed in Table 1, illustrating the structured approach to assessing molecular design algorithms against complex, real-world criteria.

The second molecular design platform, GuacaMol[48], serves as a widely recognized benchmark in drug discovery and is extensively utilized in various molecular optimization studies. The design tasks include marketed drug rediscovery, similarity assessment, median molecule generation, and isomer generation. From these, we selected tasks suitable for molecular property optimization, comprising three single-objective tasks aimed at identifying structures similar to a specific drug and four multi-objective tasks focused on finding median molecules between two drugs or achieving multi-property optimization (MPO), as detailed in Table 1. Unlike the physical simulations in

**Table 1 | Summary of the molecular design tasks investigated in this study**

| Benchmark platform | Design task | Objective | No. of reference data |
|---|---|---|---|
| Tartarus | Organic emitters | Singlet-triplet gap (↓) | 403,947 |
| | Organic emitters | Oscillator strength (↑) | |
| | Organic emitters | Singlet-triplet gap (↓) + Oscillator strength (↑) + Absolute difference of vertical excitation energy (VEE) (↓) | |
| | Protein ligands | 1SYH score (↓) | 152,296 |
| | Protein ligands | 6Y2F score (↓) | |
| | Protein ligands | 4LDE score (↓) | |
| | Reaction substrates | Activation energy (↓) | 60,828 |
| | Reaction substrates | Reaction energy (↓) | |
| | Reaction substrates | Activation energy (↑) + Reaction energy (↓) | |
| GuacaMol | Aripiprazole similarity | Similarity to aripiprazole (↑) | 22,000 (downsampled from 1.2 million GuacaMol entries) |
| | Albuterol similarity | Similarity to albuterol (↑) | |
| | Mestranol similarity | Similarity to mestranol (↑) | |
| | Median molecules 1 | Similarity to tadalafil (↑) + Similarity to sildenafil (↑) | |
| | Median molecules 2 | Similarity to camphor (↑) + Similarity to menthol (↑) | |
| | Fexofenadine MPO | Similarity to fexofenadine (↑) + TPSA (↑) + log*P* (↓) | |
| | Ranolazine MPO | Similarity to ranolazine (↑) + TPSA (↑) + log*P* (↑) | |

Tartarus, GuacaMol uses deterministic functions implemented in RDKit to compute property values, thereby eliminating data randomness. To simulate real-world scenarios where machine learning (ML) surrogate models are rarely perfect, we downsample the GuacaMol dataset to build ML surrogate models for fitness prediction during the GA process. In this setup, the molecular design process initially relies on a potentially imperfect surrogate model to propose molecular structures, which are subsequently validated using the RDKit-based oracle functions.

### Uncertainty-aware and uncertainty-agnostic fitness functions

In conventional molecular design, the typical single-objective optimization approach focuses on maximizing a specific fitness function $F_{\mathrm{DOM}}(m)$ without consideration of uncertainty. This naïve approach, referred to as the direct objective maximization (DOM), or greedy method[61], is defined as

$$F_{\mathrm{DOM}}(m) = \eta\mu(m) \tag{1}$$

where $\mu(m)$ represents the predicted property value of molecule $m$ by the surrogate model, and $\eta$ is the sign factor taking the values of +1 or -1. This factor is assigned a value of +1 when a higher $\mu$ is desired, and −1 when a lower $\mu$ is preferred. However, practical applications often do not necessitate driving the property values to their extremes. Instead, it is usually sufficient for the property to meet a certain threshold $\delta$ that is deemed acceptable for a given application[47,48].

In such scenarios, the goal should shift from merely optimizing the property value to ensuring that the property of the molecule $m$ exceeds this threshold $\delta$. Assuming the property predicted by the surrogate model follows a Gaussian distribution with mean $\mu(m)$ and variance $\sigma^2(m)$, the PIO fitness function can be defined as

$$F_{\mathrm{PIO}}(m;\delta) = \eta\int_{\delta}^{\eta\infty}\frac{1}{\sigma(m)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu(m)}{\sigma(m)}\right)^2\right)\mathrm{d}x \tag{2}$$

where $F_{\mathrm{PIO}}(m;\delta)$ ranges between 0 and 1. In this expression, $F_{\mathrm{PIO}}(m;\delta)$ quantifies the probability that the property value of molecule $m$ will exceed the threshold $\delta$. The PIO approach inherently incorporates the uncertainty (variance) of the prediction and mitigates the risk of extrapolating the surrogate model beyond its reliable range and has been recently adopted in other works utilizing active learning for drug discovery[61], co-cured polycyanurates[62], organic semiconductors[63] and boron–carbon–nitrogen crystal structure design[64]. By establishing a realistic threshold $\delta$, this method significantly enhances the practicality and applicability of molecular design optimization in real-world settings.

An alternative approach to incorporating uncertainty into the fitness function is the expected improvement (EI) method, which evaluates the expected magnitude of the improvement[46]. Assuming the property predicted by the surrogate model follows a Gaussian distribution with mean $\mu(m)$ and variance $\sigma^2(m)$, the fitness function for EI can be defined as

$$F_{\mathrm{EI}}(m;\delta) = \eta\int_{\delta}^{\eta\infty}\frac{x-\delta}{\sigma(m)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu(m)}{\sigma(m)}\right)^2\right)\mathrm{d}x \tag{3}$$

The primary difference between PIO (Eq. 2) and EI (Eq. 3) lies in whether the integration over possible improvements at a given $x$ is considered. The PIO method focuses solely on the likelihood of improvement at a specific threshold, making it a unitless probability value, whereas the EI method considers both the probability and magnitude of potential gains, yielding a value with the same unit as the target variable. Both approaches are commonly used as acquisition functions in BO[46]. In this work, we compare the performance of DOM with both PIO and EI for single-objective optimization tasks. This comparative analysis aims to highlight the advantages of incorporating UQ with GNN in molecular design and to determine the most effective optimization strategy for various practical applications.

When considering multiple properties in molecular design, a common method, known as scalarization, aggregates all objectives into a single value using their corresponding weights. The weights should be carefully chosen to balance the contributions of each property in the objective function. One approach is to use the reciprocal of the standard deviation $\tilde{\sigma}$ of each target's distribution in the dataset as weights[65], scaling the contribution of each property according to the potential variability of its values:

$$F_{\mathrm{WS}}^{\mathrm{multi}}(m) = \sum_{i=1}^{k}\frac{F_{\mathrm{DOM}}^{i}(m)}{\tilde{\sigma}_i} \tag{4}$$

where $k$ is the number of properties considered. However, this weighted sum (WS) approach can still lead to suboptimal compromises where certain objectives are sacrificed to enhance the overall score[65]. In practical scenarios, molecules often need to meet multiple objectives simultaneously, typically represented by thresholds. To

address this, we propose calculating the product of individual probabilities that each property surpasses its respective threshold, representing the overall probability of meeting all specified targets:

$$F_{\mathrm{PIO}}^{\mathrm{multi}}(m;\delta_1,\delta_2,\ldots,\delta_k)=\prod_{i=1}^{k}F_{\mathrm{PIO}}^{i}(m;\delta_i) \tag{5}$$

If any single probability approaches zero, the overall fitness score will also approach zero, regardless of high scores in other targets. This method emphasizes balancing trade-offs and aligns more closely with the complex demands of real-world applications[66–68].

One limitation of the WS method in Eq. 4 is that it does not account for specific optimization thresholds, which may lead to unbalanced solutions in multi-objective optimization. Inspired by the $\varepsilon$-constraint method[69], which reformulates additional objectives as constraints with threshold values to ensure feasible trade-offs, this study explores alternative formulations, such as the normalized Manhattan distance (NMD) to the ideal threshold values $(\delta_1,\delta_2,\ldots,\delta_k)$ as the objective function[70]. This approach treats objective values that meet or exceed the thresholds as equally favorable, potentially reducing the risk of overemphasizing certain properties at the expense of others

$$F_{\mathrm{NMD}}^{\mathrm{multi}}(m;\delta_1,\delta_2,\ldots,\delta_k)=\sum_{i=1}^{k}\frac{\min\left(\eta_i(\mu(m)-\delta_i),0\right)}{\widetilde{\sigma}_i} \tag{6}$$

where $F_{\mathrm{NMD}}^{\mathrm{multi}}\le 0$. Both the NMD and $\varepsilon$-constraint methods aim to achieve balanced solutions by incorporating objective-specific limits. However, while NMD minimizes cumulative deviations to treat all objectives meeting thresholds as equally favorable, the $\varepsilon$-constraint method enforces strict feasibility by converting secondary objectives into constraints, resulting in a more rigid adherence to specified bounds. A key limitation of NMD, though, is that it restricts further optimization once all thresholds are met. To overcome this, we propose a hybrid fitness function that combines the NMD and the simple WS approach, transitioning form $F_{\mathrm{NMD}}^{\mathrm{multi}}$ to $F_{\mathrm{WS}}^{\mathrm{multi}}$ once all property values meet their respective thresholds

$$F_{\mathrm{NMD-WS}}^{\mathrm{multi}}(m;\delta_1,\delta_2,\ldots,\delta_k)=\begin{cases}F_{\mathrm{NMD}}^{\mathrm{multi}}(m;\delta_1,\delta_2,\ldots,\delta_k),\ \mathrm{if}\ F_{\mathrm{NMD}}^{\mathrm{multi}}(m;\delta_1,\delta_2,\ldots,\delta_k)<0\\ F_{\mathrm{WS}}^{\mathrm{multi}}(m;\delta_1,\delta_2,\ldots,\delta_k),\ \mathrm{if}\ F_{\mathrm{NMD}}^{\mathrm{multi}}(m;\delta_1,\delta_2,\ldots,\delta_k)=0\end{cases} \tag{7}$$

This hybrid approach (NMD-WS), similar to methods combining $\varepsilon$-constraint and weighted sum techniques[69], aims to combine the strengths of both methods, achieving a balanced optimization that respects the thresholds while allowing further improvements once the initial conditions are met.

## Surrogate model and UQ performance

For effective CAMD, it is crucial that the surrogate model accurately represents the molecular properties of interest. To this end, we first assessed the performance of the D-MPNN model along with two UQ methods—deep ensemble combined with mean-variance estimation (MVE)[71] and evidential learning[72]—on the target properties of the design tasks specified in Table 1. Our evaluations revealed that neither UQ method delivered consistent performance across all datasets. Notably, the MVE loss function exhibited a tendency to diverge when training models on the reactivity dataset, which occasionally led to the premature termination of training sessions. In contrast, the evidential loss function faced convergence issues during the training of models for the organic emitter dataset, resulting in reduced accuracy. These challenges in model training may be partly attributed to data noise inherent in the property values, a consequence of the non-deterministic computational procedures used to generate these data, as detailed in the method section and illustrated in Supplementary Figs. S1, S2, and S3. These observations highlight the critical need for further development and refinement of these UQ methods to enhance their robustness. In response to these findings, we selected the deep ensemble and MVE approach for the organic emitter dataset, while applying evidential regression for the other datasets in our molecular design experiments. The efficacy of these approaches was visually assessed using parity plots and confidence-based calibration curves, displayed in Figs. 2 and 3, respectively. These figures show that the D-MPNN effectively captures the trends in property values, with the estimated uncertainties generally well-calibrated against the test set.

It is important to recognize that prediction uncertainties may arise from multiple sources, such as data noise and model uncertainty, meaning that the residuals between predicted and reference values may not always follow a Gaussian distribution. Therefore, we validated the Gaussian distribution assumption by examining the actual distribution of residuals. This was achieved by using confidence-based calibration curves (Fig. 3)[73,74]. These curves assess the proportion of test data points that fall within a confidence interval around their predicted values. The intervals are calculated based on predicted variance under the Gaussian assumption, and the observed proportions are then compared to the expected confidence levels. Ideally, a perfect calibration curve would follow a diagonal line, where predicted probabilities align with observed proportions across various confidence levels. To quantify deviations from this ideal calibration, we calculated the area under the calibration error curve (AUCE), with higher AUCE values reflecting greater deviations from perfect calibration. As shown in Fig. 3, the calibration curves closely follow the diagonal line across all test sets, with AUCE values remaining below 0.1, suggesting that the residual distribution for the test data does not significantly deviate from Gaussian assumptions and aligns well with estimated variance. Nonetheless, further improvement in uncertainty estimation may be achieved through additional recalibration steps[75,76] or by employing alternative UQ methods[77,78] that do not rely on strong distributional assumptions. Incorporating these enhanced UQ methods with uncertainty-aware optimization presents a promising direction for future research.

## Optimization results of single-objective task

This section evaluates the optimization results obtained using DOM (Eq. 1), PIO (Eq. 2), and EI (Eq. 3) fitness functions across ten single-objective tasks, focusing on the hit rate of molecules—i.e., their ability to exceed predetermined threshold values. This metric assesses whether the integration of uncertainty into the fitness function could improve the success rate of generating molecules that surpass these thresholds. As shown in Table 2, the PIO method consistently achieved the highest hit rates for most tasks. Additionally, Fig. 4 illustrates that, under the PIO approach, the top-100 molecules for these tasks exhibit a greater proportion of candidates meeting or exceeding the threshold, compared to those generated by DOM, further demonstrating the benefit of incorporating uncertainty in the optimization process. However, despite integrating uncertainty, the EI method does not consistently outperform the uncertainty-agnostic DOM method.

To understand why only the PIO method outperformed its uncertainty-agnostic counterpart while the EI method did not, we generated parity plots of the molecules optimized by each method (Fig. 5). These plots show that the leading molecules selected by EI tend to exhibit the highest uncertainties in the surrogate model compared to those identified by DOM and PIO. Conversely, DOM-selected molecules often display extreme predicted mean values—either lowest or highest for minimization or maximization tasks, respectively. This outcome is expected, as DOM focuses solely on optimizing the predicted mean without considering uncertainty, often pushing optimization toward extrapolative regions where predictions are less reliable. While EI does incorporate uncertainty, its performance in most single-objective optimization tasks was not particularly robust. This outcome likely stems from EI's tendency to favor candidates with high uncertainty

https://doi.org/10.1038/s41467-025-58503-0

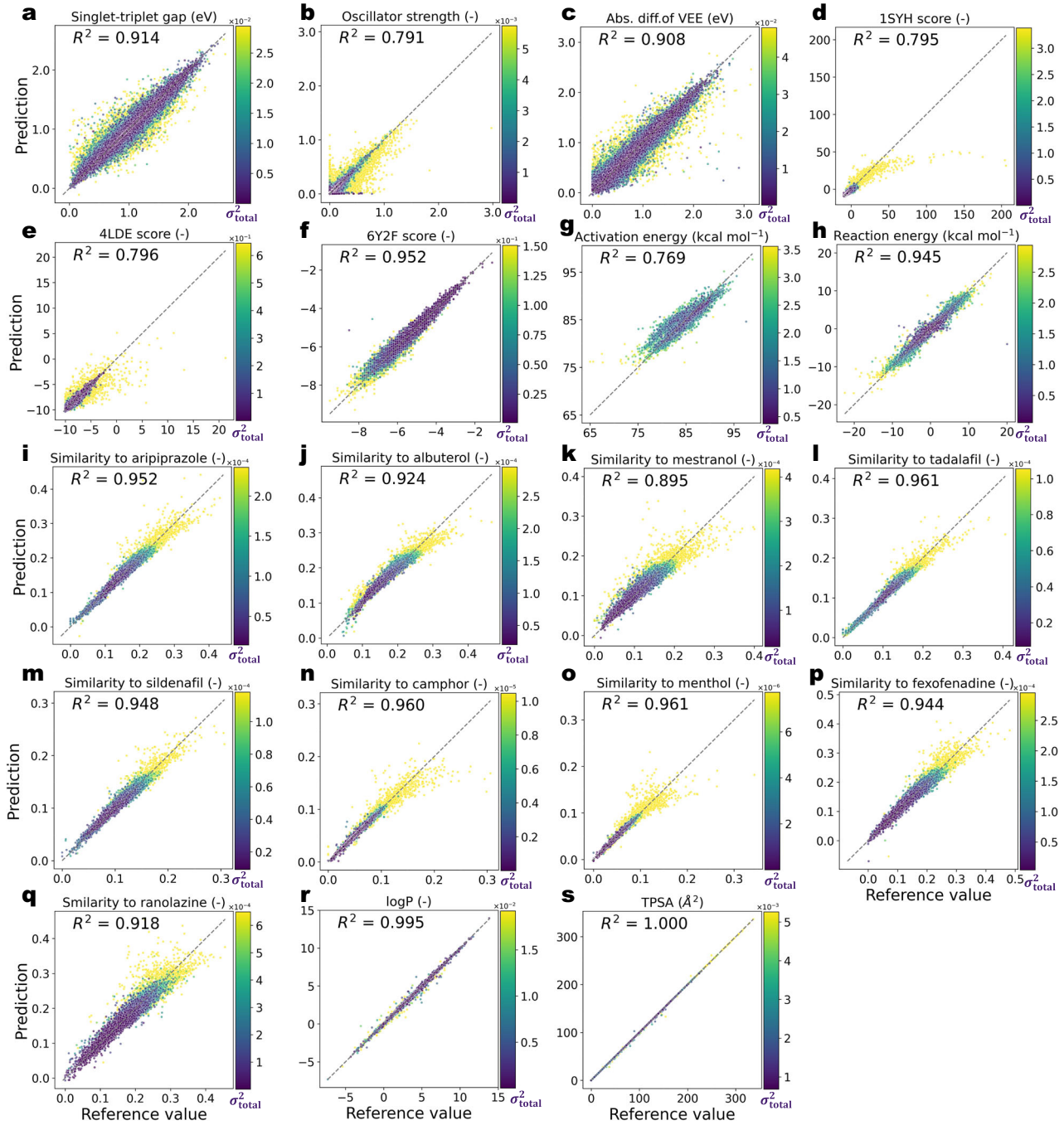


**Fig. 2 | Parity plots comparing reference values with predictions from the directed message passing neural network (D-MPNN) surrogate models on the test set.** The color coding of the data points indicates the level of total uncertainty ($\sigma^2_{total}$) in the model predictions. Uncertainty quantification (UQ) across the panels varies: **a**–**c** ensemble and mean-variance estimation (MVE) methods were utilized; **d**–**s** the evidential method was applied in panels. The molecular structure similarity is calculated using the Tanimoto similarity metric. Abs. diff. of VEE absolute difference of vertical excitation energy, $R^2$ (coefficient of determination), log*P* octanol-water partition coefficient, TPSA topological polar surface area. Source data is provided as a Source Data file.

when their predicted mean values are similar, as it calculates expected improvements as the fitness function. Such a preference can lead to the selection of molecules with significant prediction uncertainties, which often causes discrepancies between predicted and actual properties, contributing to EI's relatively unstable performance across tasks. It is worth noting that EI is widely used as an acquisition function in BO with Gaussian processes[79–83], where it effectively identifies optimal solutions within smaller, more confined search spaces over fewer iterations. However, molecular design requires navigating a much larger chemical space, where D-MPNN surrogate models can assign considerable uncertainty to numerous candidate structures, inflating expected improvements and diminishing EI's effectiveness in our test cases. In contrast, PIO focuses exclusively on the probability of improvement, yielding a bounded fitness value between 0 and 1, which makes it less susceptible to the issues of extreme variance. By emphasizing candidates with a higher probability of exceeding the threshold without overemphasizing uncertain regions, PIO achieves more stable and reliable performance. This balance enables PIO to identify candidates that meet cutoff criteria while maintaining lower uncertainties, leading to more reliable predictions.

https://doi.org/10.1038/s41467-025-58503-0

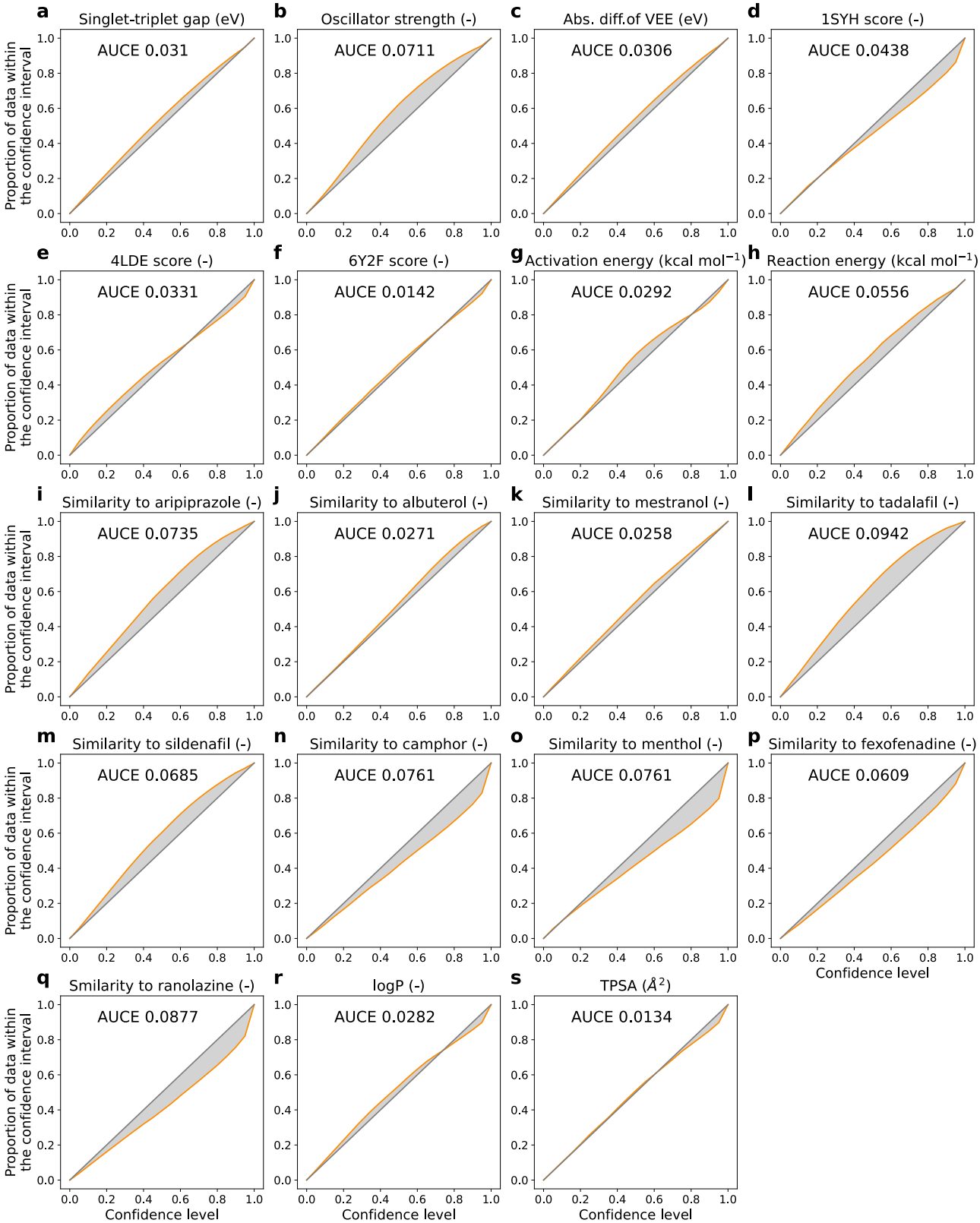


**Fig. 3 | Confidence-based calibration curves (orange) for various models assessed using testing data.** The area under the calibration error curve (AUCE), or miscalibration area[106] (gray area in this figure), is calculated, with perfect calibration indicated by an AUCE of 0. Uncertainty quantification (UQ) across the panels varies: **a**–**c** ensemble and mean-variance estimation (MVE) methods were utilized; **d**–**s** the evidential method was applied. The molecular structure similarity is calculated using the Tanimoto similarity metric. Abs. diff. of VEE absolute difference of vertical excitation energy, log*P* octanol-water partition coefficient, TPSA topological polar surface area. Source data is provided as a Source Data file.

**Table 2 | Comparison of top-k hit rates for single-objective optimization results across various methods**

| Design task | Objective | Method | Top-10 hit rate | Top-50 hit rate | Top-100 hit rate |
|---|---|---|---|---|---|
| Organic emitters | Singlet-triplet gap (↓) | DOM | 0 | 0 | 0 |
| | | EI | 0 | 0 | 0 |
| | | PIO (ours) | 0 | 0 | **0.02** |
| Organic emitters | Oscillator strength (↑) | DOM | 0.20 | 0.12 | 0.16 |
| | | EI | 0 | 0.14 | 0.12 |
| | | PIO (ours) | **0.30** | **0.28** | **0.21** |
| Protein ligands | 1syh score (↓) | DOM | 0 | 0 | 0 |
| | | EI | 0 | 0 | 0 |
| | | PIO (ours) | 0 | **0.02** | **0.02** |
| Protein ligands | 4lde score (↓) | DOM | 0.40 | 0.48 | 0.49 |
| | | EI | 0 | 0.02 | 0.03 |
| | | PIO (ours) | **0.90** | **0.94** | **0.92** |
| Protein ligands | 6y2f score (↓) | DOM | 0.40 | 0.54 | **0.50** |
| | | EI | 0 | 0 | 0 |
| | | PIO (ours) | **0.80** | **0.56** | 0.49 |
| Reaction substrates | Activation energy (↓) | DOM | 0.20 | 0.48 | 0.56 |
| | | EI | **0.50** | 0.46 | 0.58 |
| | | PIO (ours) | **0.50** | **0.56** | **0.67** |
| Reaction substrates | Reaction energy (↓) | DOM | 0.50 | 0.36 | 0.40 |
| | | EI | 0.70 | 0.68 | 0.57 |
| | | PIO (ours) | **0.90** | **0.82** | **0.76** |
| Aripiprazole similarity | Similarity to aripiprazole (↑) | DOM | 0.50 | 0.52 | 0.53 |
| | | EI | 0 | 0 | 0.06 |
| | | PIO (ours) | **1.00** | **0.72** | **0.58** |
| Albuterol similarity | Similarity to albuterol (↑) | DOM | 0 | 0.04 | 0.03 |
| | | EI | 0 | 0 | 0 |
| | | PIO (ours) | **1.00** | **0.76** | **0.62** |
| Mestranol similarity | Similarity to mestranol (↑) | DOM | 0 | 0 | 0 |
| | | EI | 0 | 0 | 0 |
| | | PIO (ours) | 0 | 0 | 0 |

The highest hit rate is highlighted in bold font.

However, certain challenging tasks reveal limitations across all methods—DOM, EI, and PIO—in identifying candidates that surpass thresholds. For example, tasks involving singlet-triplet gap, 1SYH score, and similarity to mestranol demonstrate cases where all of these approaches struggle to find candidates meeting the set criteria, highlighting areas for further improvement. The first limitation arises from the dependency of both PIO and EI on threshold-based guidance during the search process. When thresholds are set too stringently, far beyond the performance range of the current population, the fitness score remains zero regardless of optimization direction, which can impede the search for high-performing molecules. In our study, most thresholds were set near the performance of top molecules within each task's original dataset (as shown in Supplementary Table S5). Consequently, the difficulty of exceeding these thresholds varies depending on the structural diversity in each dataset. The second limitation involves decreased model accuracy in predicting property mean and variance in extrapolated regions. This issue is evident in the mestranol similarity task, where the objective is to identify molecules resembling mestranol's complex polycyclic structure, featuring four fused rings (Fig. 4j). Accurately capturing these complex ring structures remains challenging for D-MPNN, which would benefit from additional structural features—such as ring size indicators—to improve prediction accuracy for complex species in highly extrapolated regions[84]. Therefore, although D-MPNN performed reasonably well on the test set for this task (Fig. 2k), it struggled to identify similar molecules within the broader chemical space, consistently yielding similarity predictions for recommended candidates that deviated significantly from the true reference values (Fig. 5j). An additional concern is that these predictions frequently showed small uncertainty estimates, suggesting that D-MPNN may have inaccurately assessed uncertainty in these cases. This finding underscores a critical limitation: even well-calibrated models may struggle to generalize accurately during molecular optimization over an extensive chemical space, leading to unreliable predictions not only for mean values but also for variance estimates with current UQ methods. One approach to address these challenges is adaptive modeling, which iteratively incorporates newly validated molecules to refine predictions and improve uncertainty estimates. However, improving the reliability of UQ methods is essential to address these challenges and strengthen the robustness of molecular design workflows.

### Optimization results of multi-objective tasks

In this subsection, we evaluate the impact of various fitness function designs on the performance of molecule generation for multi-objective tasks. These designs included uncertainty-agnostic methods such as the WS (Eq. 4), NMD (Eq. 6), and the hybrid approach NMD-WS (Eq. 7), as well as the uncertainty-aware PIO method (Eq. 5), which calculates the product of single-objective probabilities where each indicator exceeds its corresponding cutoff. A molecule was considered a hit in multi-objective tasks if it met all specified property cutoffs.

As detailed in Table 3, the PIO approach emerged as the most effective in identifying molecules that satisfied criteria for multi-objective criteria, achieving the highest hit rates for most tasks. Among the uncertainty-agnostic methods, no single approach demonstrated consistent superiority across all tasks. The NMD method showed higher success rates in generating viable molecules for organic emitter designs and the fexofenadine MPO task, while the hybrid NMD-WS method outperformed other uncertainty-agnostic approaches in the remaining multi-objective tasks. In contrast, the WS method consistently struggled, failing to identify molecules that met all required thresholds in any multi-objective task.

A primary challenge in multi-objective tasks, as opposed to single-objective tasks, lies in balancing the contributions of different properties. For instance, in the organic emitter design task, there is a moderate positive correlation between the singlet-triplet gap and oscillator strength (Supplementary Fig. S24), complicating the task, which demands minimizing the singlet-triplet gap while maximizing oscillator strength, thereby creating conflicting optimization directions. This complexity was exacerbated by the disproportionate emphasis on oscillator strength, whose unbounded maximum value could lead the WS method to overly prioritize this trait, neglecting the other (Fig. 6). Similar challenges were observed in the fexofenadine and ranolazine MPO tasks, which involve maximizing similarity to target molecules while optimizing octanol-water partition coefficient (log*P*) and topological polar surface area (TPSA). Here, the WS method tended to prioritize log*P* and TPSA optimization at the expense of similarity scores (Fig. 6). This observation aligns with previous research[85], emphasizing the challenge of balancing each target's contribution in the fitness function to prevent bias in multi-objective optimization scenarios. Methods such as NMD and NMD-WS, which incorporate cutoff values into fitness functions, better address this balancing challenge. However, these uncertainty-agnostic methods can still lead to over-optimization in regions beyond the model's predictive range, potentially resulting in discrepancies between predicted and actual outcomes. Consequently, the PIO method generally demonstrates a higher hit rate by incorporating uncertainty information and thus achieving a better balance across all targets.

https://doi.org/10.1038/s41467-025-58503-0

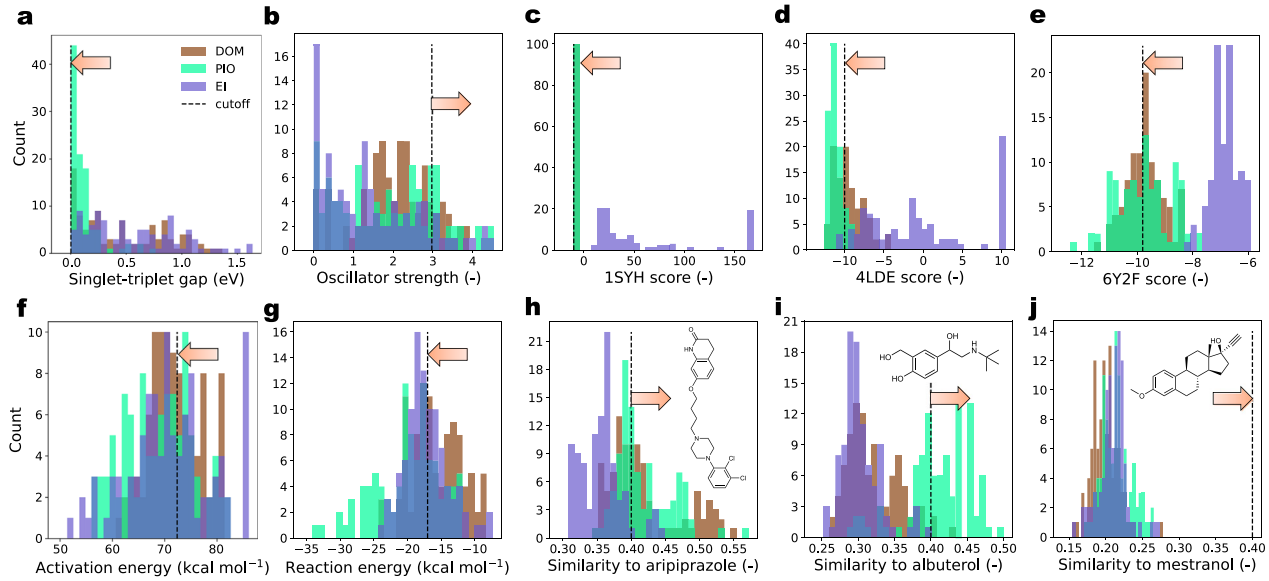


**Fig. 4 | Comparative distribution of true property values for the top-100 molecules generated by different methods. a–j** These plots show direct objective maximization (DOM, brown), expected improvement (EI, purple), and probabilistic improvement optimization (PIO, green) results. The black dotted line represents the cutoff values, while orange arrows illustrate the desired optimization direction. For the final three similarity optimization tasks, the structures of the target molecules are displayed within their respective figures. The molecular structure similarity is calculated using the Tanimoto similarity metric. Source data are provided as a Source Data file.

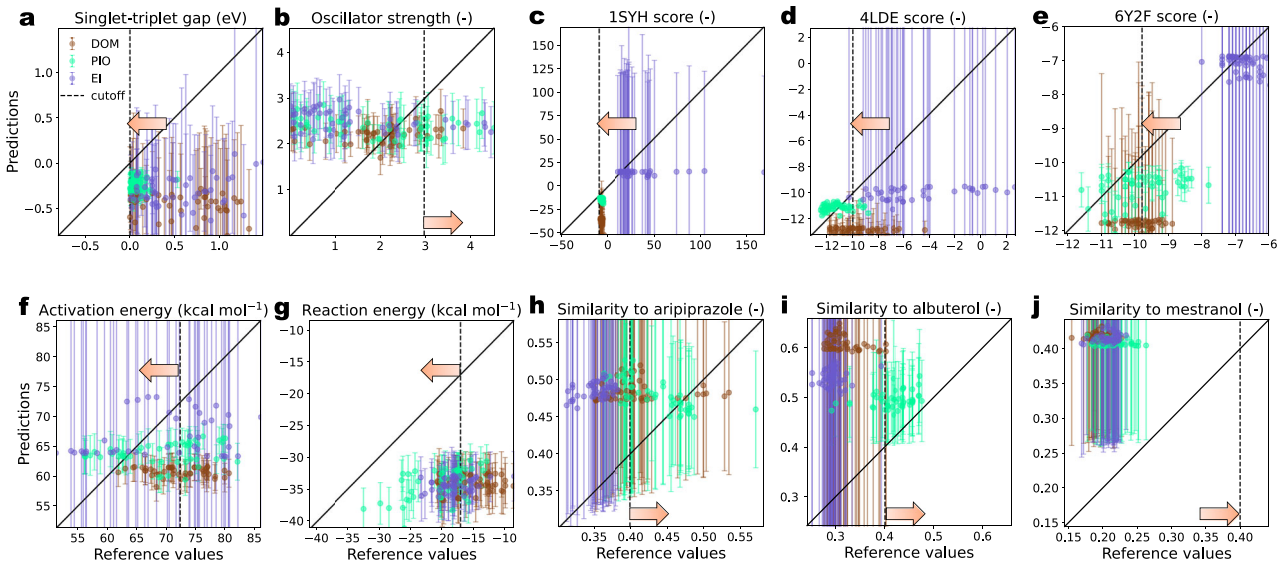


**Fig. 5 | Parity plots comparing reference values with predictions as well as uncertainties from the directed message passing neural network (D-MPNN) models. a–j** These plots show top-50 candidate molecules generated based on the fitness values from direct objective maximization (DOM, brown), expected improvement (EI, purple), and probabilistic improvement optimization (PIO, green). Predictions are presented as means with standard deviations (error bars), capturing both aleatoric and epistemic uncertainty, as estimated by the D-MPNN model. The black dotted line represents the cutoffs, while the orange arrows illustrate the desired direction for optimization. The molecular structure similarity is calculated using the Tanimoto similarity metric. Source data is provided as a Source Data file.

Despite the overall success of the PIO method, one task presented challenges for all optimizers: the median molecules 2 task, which aimed to find molecules similar to both camphor and menthol. In this case, none of the optimization methods succeeded in identifying molecules with similarity scores exceeding the cutoff of 0.2. This difficulty is likely due to the low similarity scores in the original dataset, where the majority of scores fall below 0.1 for these target molecules (Supplementary Figs. S17 and S18). This task proved more challenging compared to the median molecules 1 task, where similarity scores with the target molecules (tadalafil and sildenafil) in the original data generally ranged between 0.1 and 0.2, closer to the target value of 0.2 (Supplementary Figs. S15 and S16).

Multi-objective optimization problems are prevalent in fields such as chemical, drug, and material design, where property cutoffs are often required to meet specific commercial objectives. The PIO method achieves the highest hit rates across most multi-objective tasks by integrating uncertainty information and balancing optimization across all targets. In the PIO fitness function (Eq. 5), any molecule

**Table 3 | Comparison of top-k hit rates for multi-objective optimization results across various methods**

| Design task | Objective | Method | Top-10 hit rate | Top-50 hit rate | Top-100 hit rate |
|---|---|---|---|---|---|
| Organic emitters | Singlet-triplet gap (↓) + Oscillator strength (↑) + Absolute difference of VEE (↓) | WS | 0 | 0 | 0 |
| | | NMD | 0.60 | 0.36 | 0.35 |
| | | NMD-WS | 0 | 0.20 | 0.29 |
| | | PIO (ours) | **0.80** | **0.56** | **0.36** |
| Reaction substrates | Activation energy (↑) + Reaction energy (↓) | WS | 0.20 | 0.08 | 0.07 |
| | | NMD | 0.10 | 0.10 | 0.10 |
| | | NMD-WS | **0.40** | 0.16 | 0.11 |
| | | PIO (ours) | **0.40** | **0.22** | **0.22** |
| Median molecules 1 | Similarity to tadalafil (↑) + Similarity to sildenafil (↑) | WS | 0 | 0 | 0 |
| | | NMD | 0.60 | 0.62 | 0.59 |
| | | NMD-WS | **0.90** | 0.86 | **0.84** |
| | | PIO (ours) | **0.90** | **0.90** | 0.83 |
| Median molecules 2 | Similarity to camphor (↑) + Similarity to menthol (↑) | WS | 0 | 0 | 0 |
| | | NMD | 0 | 0 | 0 |
| | | NMD-WS | 0 | 0 | 0 |
| | | PIO (ours) | 0 | 0 | 0 |
| Fexofenadine MPO | Similarity to fexofenadine (↑) + TPSA (↑) + log*P* (↓) | WS | 0 | 0 | 0 |
| | | NMD | **0.40** | **0.32** | 0.31 |
| | | NMD-WS | 0.10 | 0.08 | 0.12 |
| | | PIO (ours) | 0.30 | **0.32** | **0.38** |
| Ranolazine MPO | Similarity to ranolazine (↑) + TPSA (↑) + log*P* (↑) | WS | 0 | 0 | 0 |
| | | NMD | 0 | **0.12** | 0.12 |
| | | NMD-WS | 0 | 0.04 | 0.10 |
| | | PIO (ours) | **0.20** | **0.12** | **0.20** |

The highest hit rate is highlighted in bold font.

deviating significantly from a target threshold receives a lower overall score, guiding the optimization process to consider all objectives equally. When certain objectives are of lower priority, cutoff values for these properties can be relaxed to minimum acceptable levels, reducing their impact on the overall fitness score as long as the values remain within acceptable ranges. Conversely, if a property approaches its minimum acceptable threshold, it appropriately impacts the fitness score, signaling the need for further optimization in that direction.

## Discussions

This study addresses a central challenge in molecular design: optimizing across expansive chemical spaces, where maintaining predictive accuracy is difficult, especially under domain shifts. The PIO method introduced here integrates UQ within molecular optimization frameworks, combining D-MPNNs with GAs to enhance reliability in exploring broad chemical spaces. Our systematic analysis evaluates the strengths and limitations of PIO in comparison to another UQ-integrated method, EI, providing insights into each method's ability to adapt to domain shifts and effectively guide exploration. Previous research has indicated that in virtual screening settings, uncertainty-agnostic acquisition functions can exhibit surprisingly equivalent or even superior performance compared to uncertainty-aware active learning approaches[39,86], suggesting that purely exploitative methods can be highly efficient in the well-defined chemical library. In contrast, our experimental setup explores an open-ended chemical space and continuously updates the optimization trajectory using fitness values. Under these conditions, PIO outperforms uncertainty-agnostic methods in most instances, whereas the EI approach proves less effective.

Benchmarking results on the Tartarus and GuacaMol platforms indicate that PIO generally improves optimization success compared to traditional uncertainty-agnostic methods. In single-objective tasks, PIO balances the search between well-understood regions and less-explored areas, reducing the risk of selecting candidates where predictions may be unreliable. This approach contrasts with EI, which often focuses on high-variance areas, leading to inconsistent performance. However, it is important to note that PIO's performance may diminish in tasks where the required properties differ significantly from those represented in the available data. This highlights an area for further methodological improvement.

In multi-objective optimization scenarios, PIO consistently proves advantageous, balancing competing objectives more effectively than weighted scalarization methods, which can skew optimization toward particular properties at the expense of others. By incorporating UQ directly into the fitness function, PIO supports a more balanced approach, generally achieving higher hit rates across multiple objectives. This is particularly relevant in CAMD, where real-world applications often require that multiple property thresholds be met concurrently. The ability of PIO to adapt to varying objectives without overemphasizing any single goal enhances its practical utility in discovering compounds suitable for complex applications.

This study's comparative analysis of UQ-integrated methods also reveals the critical role of UQ calibration in determining optimization outcomes. Our results show that robust UQ calibration is fundamental to the success of UQ-driven methods. When UQ calibration is poor, PIO's advantages are reduced, underscoring the need for more accurate and robust UQ techniques in molecular optimization. This finding suggests a direction for future research, where advancements in UQ methodologies, such as those that dynamically adapt to domain shifts, could further enhance the reliability of PIO and similar approaches in broad chemical spaces. In conclusion, this research provides valuable insights into the role of UQ in optimizing molecular design across diverse chemical spaces, demonstrating that the integration of UQ can mitigate some of the limitations posed by domain shifts. The PIO method presents a promising pathway for exploring large chemical spaces with enhanced reliability, paving the way for uncertainty-informed optimization strategies in CAMD.

## Methods

### Surrogate models and uncertainty quantification

The choice of surrogate method is crucial in molecular design, as it directly impacts predictive accuracy and computational efficiency. For this study, we selected D-MPNN, a type of GNN architecture, due to its scalability, computational efficiency, and established performance in predicting both mean properties and associated uncertainties in molecular datasets. Although Bayesian inference-based[87] methods offer theoretical advantages, their adoption in molecular property prediction has been limited by challenges such as computational costs, intractability in deep neural networks, and complex implementation requirements[41]. This has restricted their scalability in large datasets, which is a key requirement for molecular design. To assess the effectiveness of molecular design strategies, we utilized the D-MPNN model, as implemented in Chemprop[50], which facilitates the automatic extraction and learning of significant structural features of molecules by leveraging atom and bond information. It updates hidden atom states based on molecular connectivity, ultimately deriving a molecular fingerprint from the summation or averaging of all hidden atomic vectors[88]. This fingerprint is then utilized as input for subsequent feed-forward neural networks. The D-MPNN model has demonstrated robust performance in various studies focused on the prediction of chemical properties[89–91].

Various methods have been proposed to quantify uncertainty, such as Bayesian neural networks[92], Monte Carlo dropout[93], ensemble learning[94], MVE[71], and evidential learning[72,95]. Chemprop also incorporates techniques to quantify uncertainty from various sources,

categorizing it into aleatoric and epistemic types[96]. Aleatoric uncertainty, arising from inherent data randomness due to experimental or computational errors, poses challenges for mitigation as it requires enhancements in data accuracy. Conversely, epistemic uncertainty, stemming from model ignorance, can be addressed by enriching the training dataset or improving molecular feature encoding.

One of the UQ methods implemented in Chemprop is the combination of deep ensemble and MVE[71]. The deep ensemble method estimates epistemic uncertainty, $\sigma_e^2$, by training multiple models and evaluating the variance among their predictions[94]. Specifically, for $M$ models within Chemprop, with each model's prediction denoted as $\hat{y}_i$, the final prediction $\tilde{y}$ is the average of these individual predictions, and the epistemic uncertainty is calculated as:

$$\sigma_e^2 = \frac{1}{M}\sum_i^M (\tilde{y} - \hat{y}_i)^2 \tag{8}$$

In this study, we prepared ten models with different initialization seeds to form the deep ensemble. On the other hand, MVE is used to calculate aleatoric uncertainty, $\sigma_a^2$, by introducing an additional output neuron that predicts the data-dependent uncertainty, ensuring positivity via the softplus activation. In MVE, the residuals between the predicted value and the reference value are assumed to follow a Gaussian distribution with mean 0 and variance $\sigma_a^2$. This assumption justifies the use of the negative log likelihood (NLL) of a Gaussian distribution as the loss function:

$$\mathrm{NLL}(y, \hat{y}_i, \sigma_a^2) = \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(\sigma_a^2) + \frac{(y-\hat{y}_i)^2}{2\sigma_a^2} \tag{9}$$

where $y \in \mathbb{R}$ is the reference property value. When using the ensemble approach, the aleatoric uncertainties of each model are averaged to derive a composite aleatoric uncertainty value, with each model trained using the NLL as the loss function[44].

Additionally, evidential learning, another UQ method in Chemprop, avoids the need for multiple model training by directly predicting the parameters of an evidential distribution[72]. This approach involves imposing a prior Gaussian distribution on the unknown mean $\mu \sim \mathcal{N}\left(\gamma, \frac{\sigma^2}{\upsilon}\right)$ and an inverse-Gamma prior on the unknown variance $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$. The joint posterior distribution $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$ takes the form of Normal Inverse-Gamma (NIG) distribution:

$$\begin{aligned} p(\mu, \sigma^2|\gamma, \upsilon, \alpha, \beta) &= p(\mu|\sigma^2, \gamma, \upsilon) \cdot p(\sigma^2|\alpha, \beta) \\ &= \frac{\sqrt{\upsilon}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\upsilon(\gamma-\mu)^2}{2\sigma^2}\right) \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{\beta}{\sigma^2}\right) \\ &= \frac{\beta^\alpha\sqrt{\upsilon}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{2\beta + \upsilon(\gamma-\mu)^2}{2\sigma^2}\right) \end{aligned} \tag{10}$$

where $\Gamma(\cdot)$ represents the gamma function, and the NIG parameters $\gamma \in \mathbb{R}$, $\upsilon > 0$, $\alpha > 1$, $\beta > 0$ determine the mean ($\mathbb{E}[\mu] = \gamma$) and uncertainty associated with the likelihood function. In Chemprop, four neuron outputs are used to predict the NIG parameters. The softplus activation function is applied to $\upsilon$, $\alpha$ and $\beta$, ensuring their outputs are always greater than zero. The aleatoric uncertainty and epistemic uncertainty can be separately derived by:

$$\sigma_a^2 = \mathbb{E}\left[\sigma^2\right] = \frac{\beta}{\alpha - 1} \tag{11}$$

$$\sigma_e^2 = \mathrm{Var}[\mu] = \frac{\beta}{\upsilon(\alpha - 1)} \tag{12}$$

Ultimately, both deep ensemble with MVE and evidential learning can estimate total uncertainties by combining aleatoric and epistemic uncertainties: $\sigma_a^2 + \sigma_e^2 = \sigma_{\mathrm{total}}^2$ [41]. The PIO and EI algorithm for molecular design then both use total uncertainty to calculate the probability that a molecule's properties will meet the specified cutoff.

### Genetic algorithm for molecular optimization

In this study, we employed the GA for molecular optimization. GA is a population-based metaheuristic designed to iteratively refine a pool of candidate solutions, aiming to discover the optimal configuration for complex problems characterized by large search spaces[97]. Our method utilizes an advanced version of GA, known as Janus[98], which specifically manipulates SELFIES[99] representations of molecular structures. In contrast to traditional SMILES[100] representations, which are limited by stringent syntax rules[101], SELFIES ensures that any textual modifications maintain chemical validity, thus preserving the structural integrity of molecules even after random modifications. For further insights into the operational principles and efficiency of the Janus algorithm, readers are encouraged to refer to the foundational work by Nigam et al.[98]. Within our experiments, all hyperparameters were set according to the default specifications of the Janus package unless otherwise noted.

### Computational details

The data volumes used to develop the D-MPNN model for each prediction task are summarized in Table 1. For the Tartarus dataset, each of the three design tasks was divided into training, validation, and testing subsets using an 8:1:1 random split. Within each task, a multi-task learning strategy was employed, enabling the model to predict all designated targets simultaneously for the given dataset. In the case of GuacaMol, all design tasks utilized the same training, validation, and testing subsets, consisting of 10,000, 2000, and 10,000 data points, respectively. These subsets were randomly downsampled from the platform's original dataset. The distribution of molecular properties for each dataset is illustrated in Supplementary Figs. S4–S22.

In this study, we systematically examine the performance of each fitness function formulation for single (Eqs. 1–3) and multi-objective (Eqs. 4–7) optimization tasks in molecular design. Each design task incorporates a penalty term $P(m)$ to ensure that the molecules adhere to specific structural constraints required for each task. The specific definitions of these penalty terms can be found in Supplementary Tables S1–S4, and closely align with the original definitions used in the Tartarus and GuacaMol platforms. In the EI method, $\sigma(m)$ is capped at 100 to avoid the situation where a very large uncertainty value could make the EI fitness value infinite. For the median molecules tasks in the GuacaMol dataset, the objective is to identify molecules with high similarity scores to two target molecules simultaneously. Because similarity scores range from 0 to 1 and there is minimal variation across the training dataset, no additional weighting was applied between indicators in the WS, NMD, and NMD-WS methods for these tasks. Specific thresholds for each design task are listed in Supplementary Table S5.

For each molecular optimization experiment, we initiated with a pool of the top 10,000 molecules from the datasets, selected based on their performance under DOM or WS fitness functions (Eqs. 1 and 4). This pool underwent 15 independent optimization runs using the GA and the D-MPNN surrogate model, with hyperparameters detailed in the Supplementary Table S6. Each optimization run was structured to update the candidate pool across 10 iterations, introducing 500 new molecules per iteration through mutation and crossover processes. The final candidate molecules from these runs were then amalgamated to minimize variability inherent in the stochastic nature of the GA[102].

https://doi.org/10.1038/s41467-025-58503-0

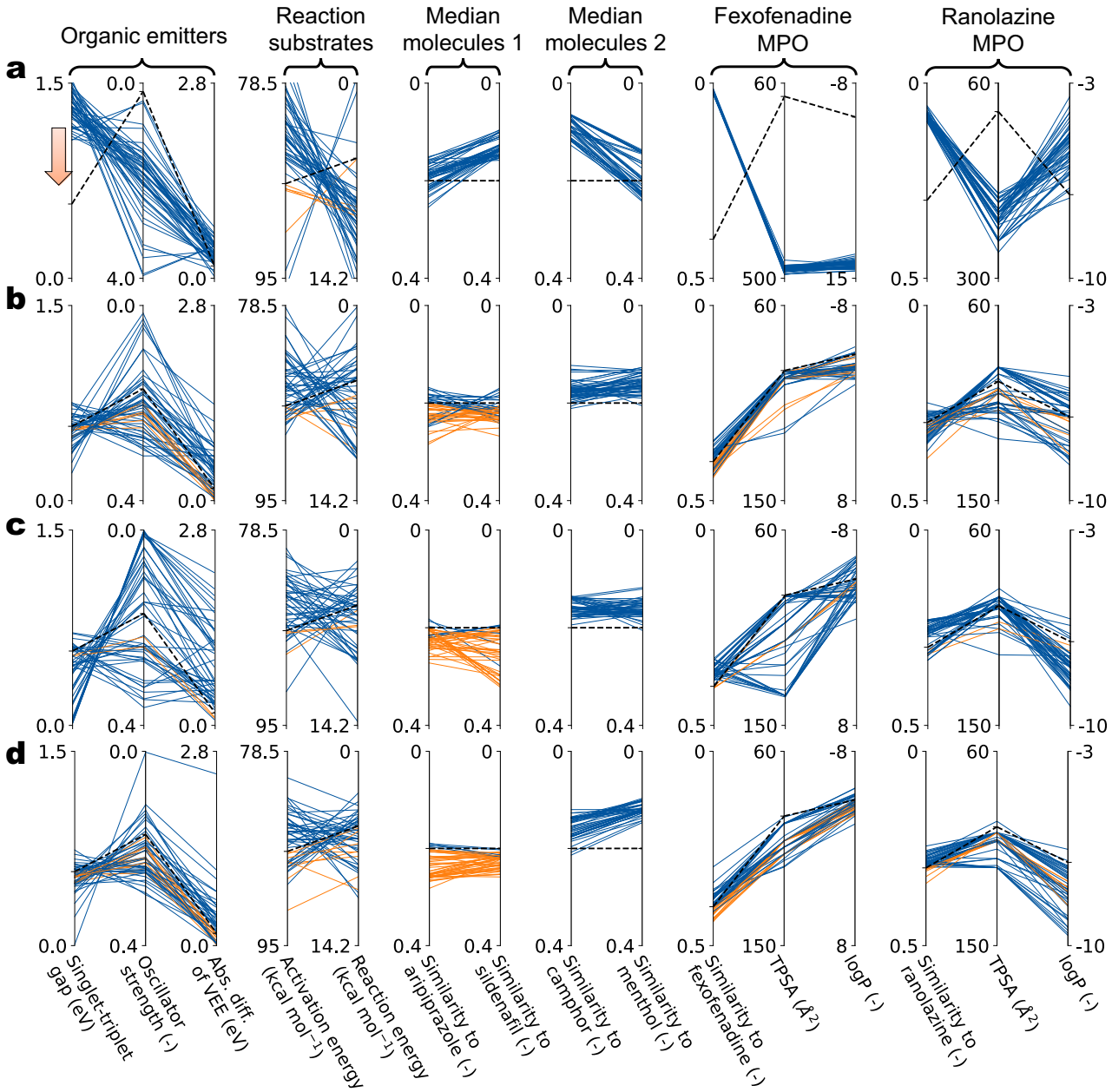


**Fig. 6 | Parallel coordinate plots illustrating the true property values for the top-50 molecules derived from various optimization methods in multi-objective design tasks.** Each subplot displays molecules generated by the methods: **a** weighted sum (WS), **b** normalized Manhattan distance (NMD), **c** hybrid approach (NMD-WS), and **d** probabilistic improvement optimization (PIO), organized into six sections arranged from left to right, corresponding to the design tasks for organic emitters, reaction substrates, median molecules 1, median molecules 2, fexofenadine multi-property optimization (MPO), and ranolazine MPO. Blue lines represent molecules that failed to meet all established cutoffs, while orange lines signify those that met all criteria. Black dotted lines across the plots denote the cutoffs. Orange arrows indicate the desired direction for optimization. The molecular structure similarity is calculated using the Tanimoto similarity metric. Abs. diff. of VEE absolute difference of vertical excitation energy, $R^2$ (coefficient of determination) log*P* octanol-water partition coefficient, TPSA topological polar surface area. Source data is provided as a Source Data file.

The consolidated list of molecules was subsequently ranked based on their fitness scores.

This procedure was carried out for each fitness function formulation. The top-performing molecules, derived using each fitness function, were then subjected to validation simulations within the Tartarus or GuacaMol frameworks to verify their actual properties. Our primary metric for evaluation was the hit rate of these molecules, particularly their ability to exceed the predefined threshold values. This analysis provides key insights into the optimization strategies that best improve success rates in generating molecules meeting targeted criteria.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The datasets for the docking, organic emitter, and reactivity designs within the Tartarus task are available at the Zenodo repository (https://doi.org/10.5281/zenodo.8072249)[103]. The GuacaMol dataset for the drug discovery task is accessible on Figshare (https://doi.org/10.6084/m9.figshare.7322252.v2)[104]. The molecules and their properties

generated in this study are provided in the Source Data file. Source data are provided with this paper.

## Code availability

The code described in this manuscript is publicly available at the Zenodo repository (https://doi.org/10.5281/zenodo.14729022)[105].

## References


1. Cheng, Y., Gong, Y., Liu, Y., Song, B. & Zou, Q. Molecular design in drug discovery: a comprehensive review of deep generative models. *Brief. Bioinform.* **22**, bbab344 (2021).
2. Pollice, R. et al. Data-driven strategies for accelerated materials design. *Acc. Chem. Res.* **54**, 849–860 (2021).
3. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
4. Schifferstein H. N., Wastiels L. Sensing materials: exploring the building blocks for experiential design. In: *Materials Experience* (Elsevier, 2014).
5. Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **10**, 787–797 (2003).
6. Zhang, X., Guo, S.-X., Gandionco, K. A., Bond, A. M. & Zhang, J. Electrocatalytic carbon dioxide reduction: from fundamental principles to catalyst design. *Mater. Today Adv.* **7**, 100074 (2020).
7. Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **1**, 37–46 (2009).
8. Chen, Y., Kontogeorgis, G. M. & Woodley, J. M. Group contribution based estimation method for properties of ionic liquids. *Ind. Eng. Chem. Res.* **58**, 4277–4292 (2019).
9. Gani, R., Nielsen, B. & Fredenslund, A. A group contribution approach to computer-aided molecular design. *AIChE J.* **37**, 1318–1332 (1991).
10. Struebing, H. et al. Computer-aided molecular design of solvents for accelerated reaction kinetics. *Nat. Chem.* **5**, 952–957 (2013).
11. Alshehri, A. S., Gani, R. & You, F. Deep learning and knowledge-based methods for computer-aided molecular design—toward a unified approach: state-of-the-art and future directions. *Comput. Chem. Eng.* **141**, 107005 (2020).
12. Chen L.-Y., Li Y.-P. Machine learning applications in chemical kinetics and thermochemistry. In *Proc. Machine Learning in Molecular Sciences* (Springer, 2023).
13. Wieder, O. et al. A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today. Technol.* **37**, 1–12 (2020).
14. Garg, V. Generative AI for graph-based drug design: recent advances and the way forward. *Curr. Opin. Struct. Biol.* **84**, 102769 (2024).
15. Xiong, J., Xiong, Z., Chen, K., Jiang, H. & Zheng, M. Graph neural networks for automated de novo drug design. *Drug Discov. Today* **26**, 1382–1393 (2021).
16. Corso, G., Stark, H., Jegelka, S., Jaakkola, T. & Barzilay, R. Graph neural networks. *Nat. Rev. Methods Prim.* **4**, 17 (2024).
17. Cheng, A. H. et al. Group SELFIES: a robust fragment-based molecular string representation. *Digit. Discov.* **2**, 748–758 (2023).
18. Jin W., Barzilay R. & Jaakkola T. Junction tree variational auto-encoder for molecular graph generation. In *Proc. International Conference on Machine Learning* (PMLR, 2018).
19. Grantham, K. et al. Deep evolutionary learning for molecular design. *IEEE Comput. Intell. Mag.* **17**, 14–28 (2022).
20. Matsukiyo, Y., Yamanaka, C. & Yamanishi, Y. De novo generation of chemical structures of inhibitor and activator candidates for therapeutic target proteins by a transformer-based variational autoencoder and Bayesian optimization. *J. Chem. Inf. Model.* **64**, 2345–2355 (2023).
21. Winter, R. et al. Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **10**, 8016–8024 (2019).
22. Griffiths, R.-R. & Hernández-Lobato, J. M. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chem. Sci.* **11**, 577–586 (2020).
23. Iwata, H. et al. VGAE-MCTS: a new molecular generative model combining the variational graph Auto-Encoder and Monte Carlo Tree Search. *J. Chem. Inf. Model.* **63**, 7392–7400 (2023).
24. Blaschke, T. et al. REINVENT 2.0: an AI tool for de novo drug design. *J. Chem. Inf. Model.* **60**, 5918–5922 (2020).
25. Guo, J., Schwaller, P. Augmented memory: sample-efficient generative molecular design with reinforcement learning. *JACS Au* **4**, 2160–2172 (2024).
26. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 1–14 (2017).
27. Anstine, D. M. & Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *J. Am. Chem. Soc.* **145**, 8736–8750 (2023).
28. Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **10**, 3567–3572 (2019).
29. Verhellen, J. Graph-based molecular Pareto optimisation. *Chem. Sci.* **13**, 7526–7535 (2022).
30. Alberga, D. et al. DeLA-DrugSelf: empowering multi-objective de novo design through SELFIES molecular representation. *Comput. Biol. Med.* **175**, 108486 (2024).
31. Yoshikawa, N. et al. Population-based de novo molecule generation, using grammatical evolution. *Chem. Lett.* **47**, 1431–1434 (2018).
32. Yu, Q. et al. A survey on evolutionary computation based drug discovery. In *Proc. IEEE Transactions on Evolutionary Computation* (IEEE, 2024).
33. Teixeira, A. L. & Falcao, A. O. Structural similarity based kriging for quantitative structure activity and property relationship modeling. *J. Chem. Inf. Model.* **54**, 1833–1849 (2014).
34. Jiang, M., Pedrielli, G. & Ng, S. H. Gaussian processes for high-dimensional, large data sets: a review. In *Proc. Winter Simulation Conference (WSC)* (IEEE, 2022).
35. Deringer, V. L. et al. Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
36. Moss, H. B. & Griffiths, R.-R. Gaussian process molecule property prediction with flowmo. *arXiv preprint arXiv:201001118*, (2020).
37. Korovina, K. et al. Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations. In *Proc. International Conference on Artificial Intelligence and Statistics* (PMLR, 2020).
38. Williams, C. K. & Rasmussen, C. E. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
39. Fromer, J. C. & Graff, D. E., Coley, C. W. Pareto optimization to accelerate multi-objective virtual screening. *Digit. Discov.* **3**, 467–481 (2024).
40. van Tilborg, D. & Grisoni, F. Traversing chemical space with active deep learning for low-data drug discovery. *Nat. Comput. Sci.* **4**, 1–11 (2024).
41. Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P. & Green, W. H. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J. Chem. Inf. Model.* **60**, 2697–2717 (2020).
42. Yin, T., Panapitiya, G., Coda, E. D. & Saldanha, E. G. Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. *J. Cheminform.* **15**, 105 (2023).
43. Abdar, M. et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021).

44. Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R. & Coley, C. W. Uncertainty quantification using neural networks for molecular property prediction. *J. Chem. Inf. Model.* **60**, 3770–3780 (2020).
45. Yang, C.-I. & Li, Y.-P. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *J. Cheminform.* **15**, 13 (2023).
46. Frazier P. I. A tutorial on Bayesian optimization. *arXiv preprint arXiv:180702811* (2018).
47. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **4**, 828–849 (2019).
48. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
49. Nigam, A. K. et al. Tartarus: a benchmarking platform for realistic and practical inverse molecular design. *Adv. Neural Inf. Process. Syst.* **36**, 3263–3306 (2023).
50. Heid, E. et al. Chemprop: a machine learning package for chemical property prediction. *J. Chem. Inform. Model.* **64**, 9–17 (2023).
51. Grimme, S. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *J. Chem. Theory Comput.* **15**, 2847–2862 (2019).
52. Bannwarth, C., Ehlert, S. & Grimme, S. 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**, 1652–1671 (2019).
53. Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements ($Z = 1$–86). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).
54. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098 (1988).
55. Alhossary, A., Handoko, S. D. & Mu, Y. Kwoh C-K. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics* **31**, 2214–2216 (2015).
56. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
57. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).
58. Jensen, F. Locating minima on seams of intersecting potential energy surface. An application to transition structure modeling. *J. Am. Chem. Soc.* **114**, 1596–1603 (1992).
59. Bell, R. P. The theory of reactions involving proton transfers. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **154**, 414–429 (1936).
60. Evans, M. & Polanyi, M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Trans. Faraday Soc.* **32**, 1333–1360 (1936).
61. Crivelli-Decker, J. E. et al. Machine learning guided aqfep: a fast and efficient absolute free energy perturbation solution for virtual screening. *J. Chem. Theory Comput.* **20**, 7188–7198 (2024).
62. Xu, X., Zhao, W., Wang, L., Lin, J. & Du, L. Efficient exploration of compositional space for high-performance copolymers via Bayesian optimization. *Chem. Sci.* **14**, 10203–10211 (2023).
63. Kawagoe, R., Ando, T., Matsuzawa, N. N., Maeshima, H., Kaneko, H. Exploring molecular descriptors and acquisition functions in Bayesian optimization for designing molecules with low hole reorganization energy. *ACS Omega* **9**, 49 (2024).
64. Li, C.-N., Liang, H.-P., Zhang, X., Lin, Z. & Wei, S.-H. Graph deep learning accelerated efficient crystal structure search and feature extraction. *NPJ Comput. Mater.* **9**, 176 (2023).
65. Fromer, J. C., Coley, C. W. Computer-aided multi-objective optimization in small molecule discovery. *Patterns* **4**, 100678 (2023).
66. Bigman, L. S. & Levy, Y. Proteins: molecules defined by their trade-offs. *Curr. Opin. Struct. Biol.* **60**, 50–56 (2020).
67. Chen, Y.-H. et al. Vacuum-deposited small-molecule organic solar cells with high power conversion efficiencies by judicious molecular design and device optimization. *J. Am. Chem. Soc.* **134**, 13616–13623 (2012).
68. Zhang, Q., Khetan, A., Sorkun, E. & Er, S. Discovery of aza-aromatic anolytes for aqueous redox flow batteries via high-throughput screening. *J. Mater. Chem. A* **10**, 22214–22227 (2022).
69. Miettinen, K. *Nonlinear Multiobjective Optimization* (Springer Science & Business Media, 1999).
70. Chiu, W.-Y., Yen, G. G. & Juan, T.-K. Minimum Manhattan distance approach to multiple criteria decision making in multiobjective optimization problems. *IEEE Trans. Evolut. Comput.* **20**, 972–985 (2016).
71. Nix D. A., Weigend A. S. Estimating the mean and variance of the target probability distribution. In *Proc. IEEE International Conference on Neural Networks (ICNN'94)* (IEEE, 1994).
72. Soleimany, A. P. et al. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).
73. Gustafsson F. K., Danelljan M., Schon T. B. Evaluating scalable Bayesian deep learning methods for robust computer vision. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*) (IEEE, 2020).
74. Heid, E., McGill, C. J., Vermeire, F. H., Green, W. H. Characterizing uncertainty in machine learning for chemistry. *J. Chem. Inform. Model.* **13**, 63 (2023).
75. Laves, M.-H., Ihler, S., Fast, J. F., Kahrs, L. A. & Ortmaier, T. Recalibration of aleatoric and epistemic regression uncertainty in medical imaging. *arXiv preprint arXiv:210412376* (2021).
76. Jiang, S., Qin, S., Van Lehn, R. C., Balaprakash, P. & Zavala, V. M. Uncertainty quantification for molecular property predictions with graph neural architecture search. *Digit. Discov.* **3**, 1534–1553 (2024).
77. Huang, K., Jin, Y., Candes, E. & Leskovec, J. Uncertainty quantification over graph with conformalized graph neural networks. *Adv. Neural Inform. Process. Syst.* **36**, 26699–26721 (2024).
78. Romano, Y., Patterson, E. & Candes, E. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, **32** (NIPS, 2019).
79. Ando, T. et al. Design of molecules with low hole and electron reorganization energy using DFT calculations and Bayesian optimization. *J. Phys. Chem. A* **126**, 6336–6347 (2022).
80. Hickman, R. J., Aldeghi, M., Häse, F. & Aspuru-Guzik, A. Bayesian optimization with known experimental and design constraints for chemistry applications. *Digit. Discov.* **1**, 732–744 (2022).
81. Nambiar, A. M. et al. Bayesian optimization of computer-proposed multistep synthetic routes on an automated robotic flow platform. *ACS Cent. Sci.* **8**, 825–836 (2022).
82. Shields, B. J. et al. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
83. Wang, K. & Dowling, A. W. Bayesian optimization for chemical products and functional materials. *Curr. Opin. Chem. Eng.* **36**, 100728 (2022).
84. Chen, L.-Y., Hsu, T.-W., Hsiung, T.-C. & Li, Y.-P. Deep learning-based increment theory for formation enthalpy predictions. *J. Phys. Chem. A* **126**, 7548–7556 (2022).
85. Nigam, A., Pollice, R., Krenn, M., dos Passos Gomes, G. & Aspuru-Guzik, A. Beyond generative models: superfast traversal,

optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chem. Sci.* **12**, 7079–7090 (2021).
86. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **12**, 7866–7881 (2021).
87. Ryu, S., Kwon, Y. & Kim, W. Y. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chem. Sci.* **10**, 8438–8446 (2019).
88. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
89. Lim, M. A., Yang, S., Mai, H. & Cheng, A. C. Exploring deep learning of quantum chemical properties for absorption, distribution, metabolism, and excretion predictions. *J. Chem. Inf. Model.* **62**, 6336–6341 (2022).
90. Liu, G., et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nat. Chem. Biol.* **19**, 1–9 (2023).
91. Muthiah, B., Li, S.-C. & Li, Y.-P. Developing machine learning models for accurate prediction of radiative efficiency of greenhouse gases. *J. Taiwan Inst. Chem. Eng.* **151**, 105123 (2023).
92. Zhang, Y. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10**, 8154–8163 (2019).
93. Gal, Y. & Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proc. International Conference on Machine Learning* (PMLR, 2016).
94. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, **30** (NIPS, 2017).
95. Sensoy, M., Kaplan, L. & Kandemir, M. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, **31** (NIPS, 2018).
96. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **110**, 457–506 (2021).
97. Lambora, A., Gupta, K. & Chopra, K. Genetic algorithm—a literature review. In *Proc. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (IEEE, 2019).
98. Nigam, A., Pollice, R. & Aspuru-Guzik, A. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *Digit. Discov.* **1**, 390–404 (2022).
99. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **1**, 045024 (2020).
100. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
101. Krenn, M. et al. SELFIES and the future of molecular string representations. *Patterns* **3**, 10 (2022).
102. Gao, W., Fu, T., Sun, J. & Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. *Adv. Neural Inf. Process. Syst.* **35**, 21342–21357 (2022).
103. Nigam, A., Tom, G. & Wille, J. aspuru-guzik-group/Tartarus: v0.1.0. https://doi.org/10.5281/zenodo.8072249 (2023).
104. Fiscato, M., Vaucher, A. C. & Segler, M. GuacaMol All SMILES. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.7322252.v2 (2018).
105. Chen, L.-Y. & Li, Y.-P. Uncertainty quantification with graph neural networks for efficient molecular design. Lung-Yi/uncmoo. https://doi.org/10.5281/zenodo.14729022 (2025).
106. Varivoda, D., Dong, R., Omee, S. S. & Hu, J. Materials property prediction with uncertainty quantification: a benchmark study. *Appl. Phys. Rev.* **10**, 021409 (2023).

## Acknowledgements

We are grateful to the National Center for High-performance Computing (NCHC) and the Computer and Information Networking Center at NTU for the support of computing facilities. L.Y.C. is supported by the Graduate Students Study Abroad Program (113-2917-I-002-018) sponsored by National Science and Technology Council in Taiwan. Y.P.L. is supported by Taiwan NSTC (113-2628-E-002-017-MY3 and 113-2622-8-002-015-SB) and the Higher Education Sprout Project by the Ministry of Education in Taiwan (114L7763). During the preparation of this work, the authors used ChatGPT to correct grammatical mistakes and enhance the fluency of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## Author contributions

L.Y.C. designed the methodology, performed the formal analysis, and wrote the initial draft of the manuscript. Y.P.L. acquired funding, supervised the project, and reviewed and edited the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58503-0.

**Correspondence** and requests for materials should be addressed to Yi-Pei Li.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025