Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

A unified active learning framework for photosensitizer design†

Yizhe Chen,[a‡] Shomik Verma,[b‡] Kevin P. Greenman,[cde‡] Haoyu Yin,[a] Zhihao Wang,[a] Lanjing Wang,[f] Jiali Li,[*g] Rafael Gómez-Bombarelli,[*h] Aron Walsh,[*i] and Xiaonan Wang[*a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

The design of high-performance photosensitizers for next-generation photovoltaic and clean energy applications remains a formidable challenge due to the vast chemical space, competing photophysical trade-offs, and computational limitations of traditional quantum chemistry methods. While machine learning offers potential solutions, existing approaches suffer from data scarcity and inefficient exploration of molecular configurations. This work introduces a unified active learning framework that systematically integrates semi-empirical quantum calculations with adaptive molecular screening strategies to accelerate photosensitizer discovery. Our methodology combines three principal components: (1) A hybrid quantum mechanics/machine learning pipeline generating a chemically diverse molecular dataset while maintaining quantum chemical accuracy at significantly reduced computational costs; (2) A graph neural network architecture and uncertainty quantification; (3) Novel acquisition strategies that dynamically balance broad chemical space exploration with targeted optimization of photophysical objectives. The framework demonstrates superior performance in predicting critical energy levels (T_1/S_1) compared to conventional screening approaches, while effectively prioritizing synthetically feasible candidates. By open-sourcing both the curated molecular dataset and implementation tools, this work establishes an extensible platform for data-driven discovery of optoelectronic materials, with immediate applications in solar energy conversion and beyond.

1 Introduction

Photosensitizers (PSs) have emerged as critical functional materials in modern energy and biomedical technologies, driving innovations from solar energy harvesting to photodynamic therapy.[1 2 3 4] Their expanding applications in wearable devices, sterilization systems, and optical sensors demand precise control over photophysical properties such as triplet yields and excited-state lifetimes.[5 6] However, the rational design of high-performance PSs faces three fundamental challenges that hinder rapid progress.

First, previous work has shown that combinatorial D–A assembly yields libraries containing more than one million PS candidates, far exceeding the capacity of conventional trial-and-error approaches.[7 8 9] For example, subtle structural variations in porphyrin derivatives—such as peripheral substituent patterns—can shift absorption maxima by over 50 nm while altering quantum yields by orders of magnitude.[7 10 11] Second, the intricate balance between competing photophysical properties creates a complex optimization landscape. A PS optimized for strong visible light absorption may suffer from rapid triplet-triplet annihilation, while molecules with ideal S_1/T_1 energy ratios often exhibit poor solubility or photostability.[12 13] Third, computational screening methods such as time-dependent density-functional theory (TD-DFT), though theoretically rigorous, become prohibitively expensive for geometry optimization and large-scale exploration, requiring days of computation for a single medium-sized molecule (50+ atoms).[14 15]

Recent advances in machine learning (ML) offer promising solutions to these bottlenecks.[16 17] By establishing quantitative

[a] Department of Chemical Engineering, Tsinghua University, Beijing 100084, P. R. China. E-mail: wangxiaonan@tsinghua.edu.cn
[b] Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: skverma@mit.edu
[c] Department of Chemical Engineering, Catholic Institute of Technology, Cambridge, Massachusetts, United States of America.E-mail: kgreenman@catholic.tech
[d] Department of Chemistry, Catholic Institute of Technology, Cambridge, Massachusetts, United States of America.
[e] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America.
[f] Institute of Flexible Electronics (IFE) & Frontiers Science Center for Flexible Electronics, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China. E-mail: wanglanjing@mail.nwpu.edu.cn
[g] Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, P. R. China. E-mail: jlli@rcees.ac.cn
[h] Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: rafagb@mit.edu
[i] Department of Materials, Imperial College London, London SW7 2AZ, UK. E-mail: a.walsh@imperial.ac.uk

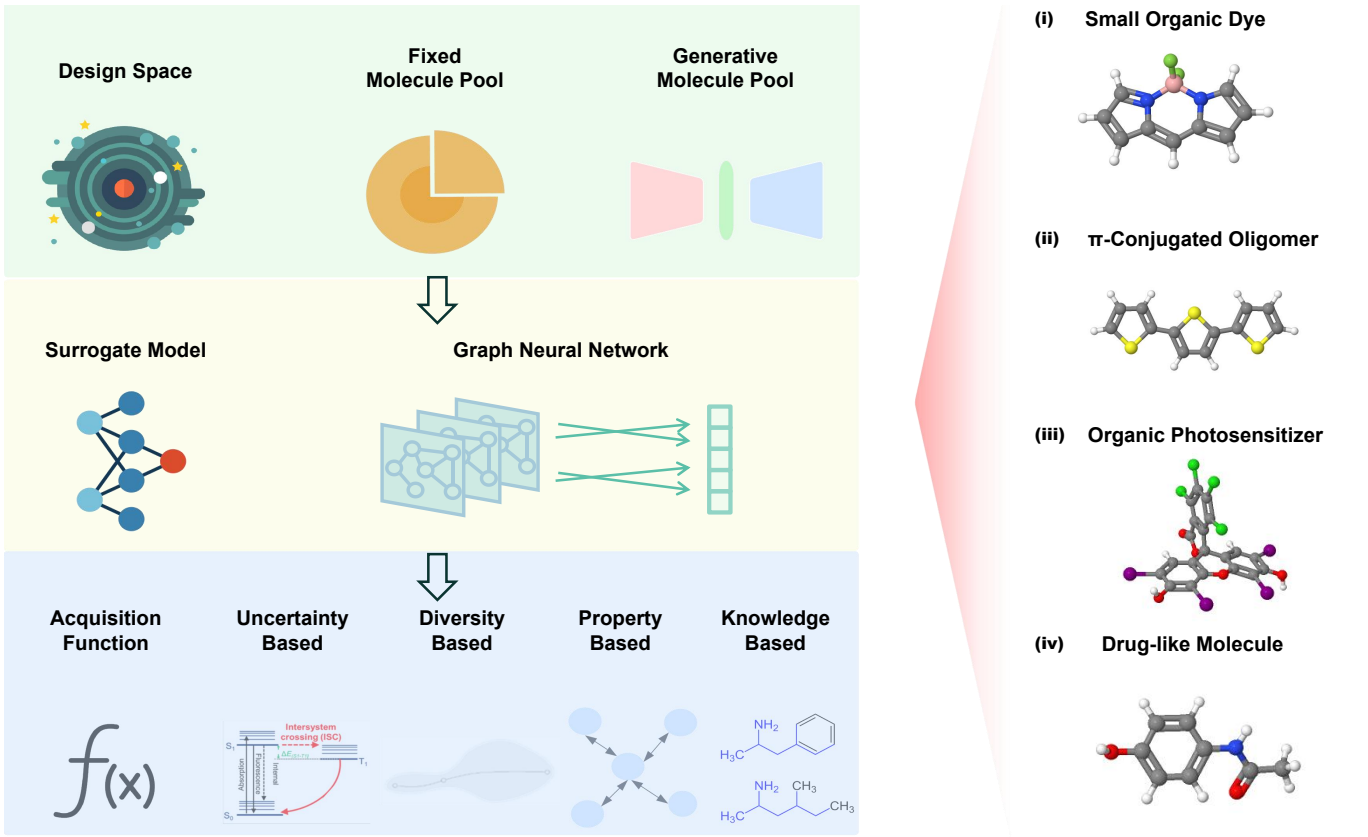

Fig. 1 Schematic of the active-learning platform for photosensitizer discovery. (Left) The design space can combine a fixed molecule pool and a generative molecule pool (shown here only as a conceptual example to illustrate future possibilities to further expand chemical diversity via generative models). A graph neural network (GNN) surrogate rapidly predicts key photophysical properties such as singlet–triplet energy gap (ΔE_{ST}) and absorption wavelength. Four complementary acquisition strategies—uncertainty-based, diversity-based, property-based, and knowledge-based selection—are employed to efficiently select new molecules for accurate calculations or experimental validation. (Right) Four representative molecular classes that can be directly processed by the GNN model: (i) small organic dyes (e.g., BODIPY), (ii) π-conjugated oligomers (e.g., α-terthiophene), (iii) organic photosensitizers (e.g., Rose Bengal), and (iv) drug-like molecules (e.g., paracetamol).

structure-property relationships (QSPRs), ML models can predict key PS characteristics like singlet-triplet energy gaps (ΔE_{ST}) with millisecond inference times.[10] Nevertheless, existing ML approaches face two critical limitations: (1) Public datasets contain less than 0.1% of the required photophysical data for PS design, creating severe data scarcity issues; (2) Conventional ML workflows prioritize passive learning from static datasets, inefficiently allocating computational resources to chemically redundant regions.[12 13 14]

Active learning (AL) is a machine learning approach where the model selects the most informative data points for labeling, aiming to improve performance with fewer labeled examples.[18]It addresses these limitations through iterative cycles of prediction and targeted data acquisition.[15 19 20 21] Unlike traditional methods that treat all molecules equally, AL algorithms dynamically identify the most informative candidates for quantum chemical calculations—those with high prediction uncertainty or high potential to improve model performance.[22 23 24]Recent demonstrations in catalyst discovery achieved 32× acceleration over random screening by prioritizing metal alloys with optimal d-band centers[15 25] For PS design specifically, AL's ability to navigate high-dimensional chemical spaces while respecting synthetic constraints could revolutionize molecular discovery pipelines.[26 27 28 29 30]

This work establishes a unified AL framework integrating two key innovations. First, we generate a calibrated dataset of 655,197 PS candidates using ML-xTB—a hybrid quantum mechanics/machine learning approach that reduces computational costs by 99% compared to TD-DFT while maintaining chemical accuracy (ΔE_{ST} MAE <0.08 eV).[31] Second, we implement novel acquisition strategies that combine uncertainty sampling with photophysical objective functions. Experimental benchmarks demonstrate that sequential AL strategies, which explore chemical diversity in early cycles before exploiting target regions, outperform static approaches by 15-20% in mean absolute error (MAE) metrics in test sets. This framework provides a generalizable platform for data-efficient molecular discovery, with immediate applications in solar fuels and optoelectronic device engineering.

2 The unified active learning framework

Our unified active learning framework for molecular material design comprises three primary components: Design Space Generation, Surrogate Model Construction, and Acquisition Function

Formulation.

2.1 Design Space Generation

The construction of a chemically relevant and computationally tractable design space forms the foundation of our active learning framework. Traditional approaches relying solely on expert intuition or brute-force enumeration fail to address the dual challenges of chemical diversity and computational feasibility.[32,33]

We combined Simplified Molecular-Input Line-Entry System (SMILES) data from numerous public molecular datasets to construct a unified library of 655,197 candidate photosensitizer molecules. Each source dataset was chosen because it contributes molecules with relevant excited-state or optical properties, thereby ensuring that our merged collection covers a broad range of photophysical characteristics. Starting from an initial seed set of 50,000 molecules, we expanded the library by integrating many diverse data sources (computational, experimental, and even patent-derived). We then predicted the lowest singlet and triplet energies (S_1 and T_1) for all candidates using our ML-xTB workflow, achieving DFT-level accuracy at 1% of the typical cost. We performed the analysis of the dataset in Fig. 3. (Full details of the datasets, including their names, references, and selection criteria, are provided in the *Supplementary Information (Text S5)*.)

ML-xTB Pipeline

The ML-xTB workflow comprises three stages:

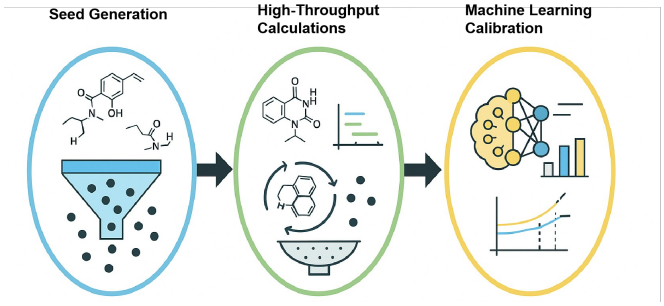

Fig. 2 ML-xTB pipeline for large-scale molecular property calculation. The process consists of initial seed generation from public databases and expert scaffolds, rapid pre-screening of excited-state energies using semi-empirical methods, and subsequent machine learning calibration to achieve DFT-level accuracy at substantially reduced computational cost.

1. **Initial Seed Generation**: A diverse set of 50,000 molecules was curated from public databases (PubChemQC[34], QM-spin[35]) and expert-designed scaffolds (porphyrins, phthalocyanines). SMILES strings were standardized using RDKit, with stereochemistry and tautomer states normalized via Morgan fingerprint clustering (radius=2, 1024 bits).

2. **xTB-sTDA High-Throughput Calculations**: Each molecule underwent geometry optimization and excited-state calculation using the geometry, frequency, noncovalent–tight binding (GFN2-xTB)[36] method combined with the simplified Tamm–Dancoff approximation (sTDA)[37] implemented in xtb (GFN2-xTB/xtb-sTDA):

$$S_1 = E_{\text{singlet}} - E_{\text{ground}} \tag{1}$$

$$T_1 = E_{\text{triplet}} - E_{\text{ground}} \tag{2}$$

$$\Delta E_{ST} = S_1 - T_1 \tag{3}$$

For the initial seed set (50,000 molecules), additional TD-DFT calculations (B3LYP/6-31+G(d), Gaussian 16) were performed on the xTB-optimized geometries to provide accurate reference values (details provided in Supporting Information).

3. **Machine Learning Calibration**: A 10-model ensemble of Chemprop Message Passing Neural Networks (Chemprop-MPNN) was trained to correct systematic errors between the 50,000 xTB-sTDA and TD-DFT calculations for the S_1 and T_1 excitations separately. Each network dynamically generated molecular representations from SMILES strings without static fingerprints, predicting state-specific errors:

$$\Delta E_S(i) = E_{S_1}^{\text{DFT}}(i) - E_{S_1}^{\text{xTB}}(i) \tag{4}$$

$$\Delta E_T(i) = E_{T_1}^{\text{DFT}}(i) - E_{T_1}^{\text{xTB}}(i) \tag{5}$$

The multitask loss function minimized during training was:

$$\mathscr{L} = \frac{1}{N}\sum_{i=1}^{N}\left[(\Delta E_S(i) - f_S(\mathbf{x}_i))^2 + (\Delta E_T(i) - f_T(\mathbf{x}_i))^2\right] \tag{6}$$

Where $f_S(\mathbf{x}_i)$ and $f_T(\mathbf{x}_i)$ are MPNN-predicted corrections for singlet and triplet excitations respectively.

The calibrated energies were then computed as:

$$E_{S_1}^{\text{corr}} = E_{S_1}^{\text{xTB}} + f_S(\mathbf{x}) \tag{7}$$

$$E_{T_1}^{\text{corr}} = E_{T_1}^{\text{xTB}} + f_T(\mathbf{x}) \tag{8}$$

for all molecules in the full dataset (655,197 molecules). We performed only xTB-sTDA calculations for the remaining molecules beyond the seed set and then applied the ML correction model. This calibration approach reduced the mean absolute error (MAE) from 0.23 eV (raw xTB) to 0.08 eV (ML-corrected) with respect to TD-DFT for the 50,000 molecules in the calibration set.

Dataset Splitting and Active Learning Protocol

We first reserved a fixed external test set of molecules from the complete dataset before any model training. The remaining data were randomly split into training, validation, and internal test sets in a 94:5:1 ratio. An initial training set of 5,000 molecules was randomly selected and kept consistent across all strategies. In each active learning round, 20,000 additional molecules were sampled from the remaining pool, with a total of 8 rounds conducted for each acquisition strategy. This protocol enables model development and tuning on one portion of the data, while reserv-

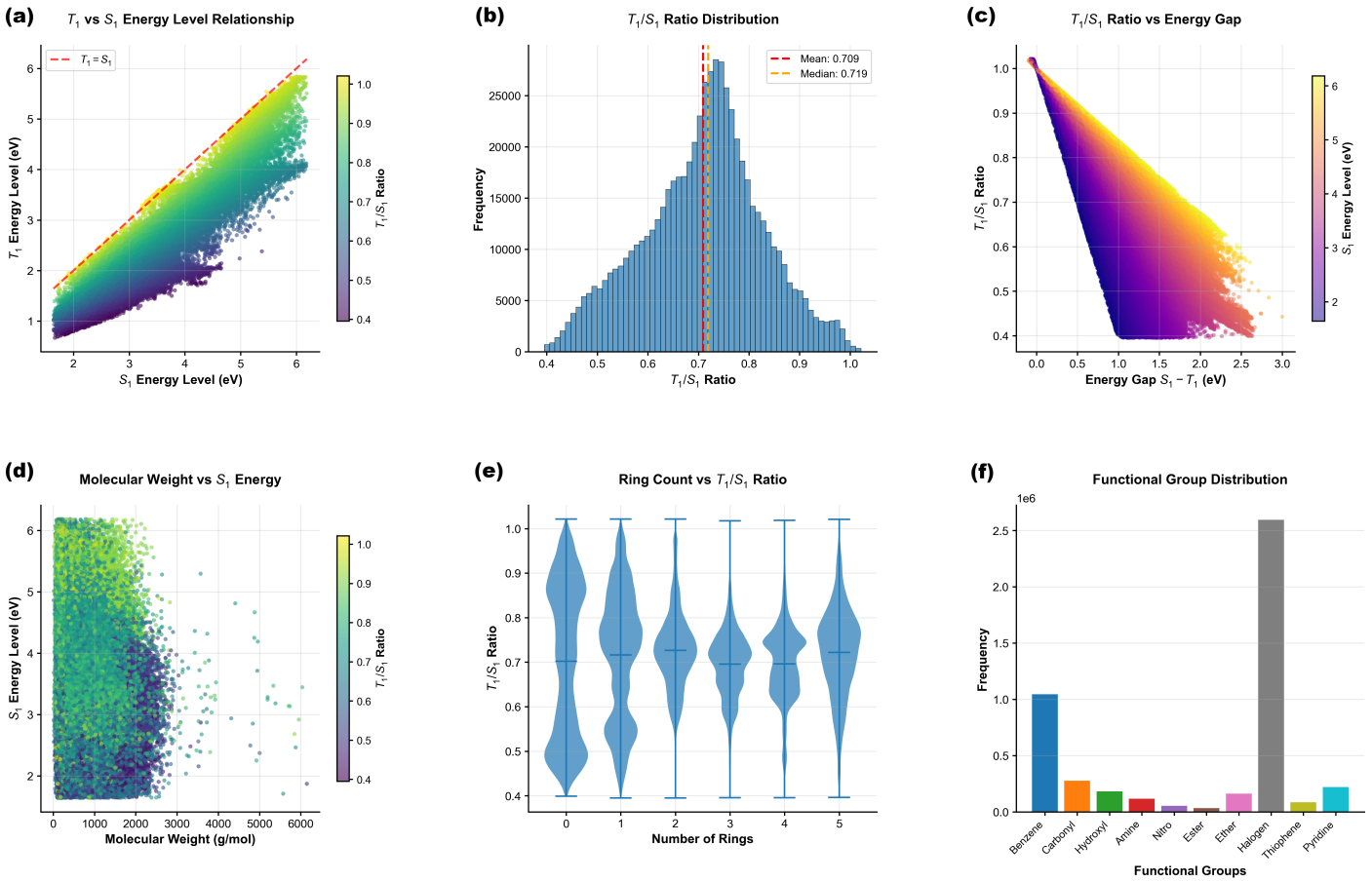

Fig. 3 Overview of key molecular and photophysical features in the unified active learning dataset. (a) T_1 vs S_1 energy levels for 655,197 candidates show a near-linear correlation (colored by T_1/S_1 ratio), with most molecules below the T_1=S_1 diagonal, highlighting typical singlet–triplet splitting and energetically favored structures. (b) Distribution of T_1/S_1 ratios, peaking near 0.7, provides a quantitative reference for selecting candidates with optimal photophysical properties. (c) T_1/S_1 ratio as a function of the S_1–T_1 energy gap (colored by S_1), showing a triangular distribution: larger gaps yield lower T_1/S_1 ratios, revealing a fundamental structure–property trade-off. (d) S_1 energy as a function of molecular weight (colored by T_1/S_1 ratio), indicating broad chemical diversity and little direct dependence between S_1 and molecular weight. (e) Violin plots of T_1/S_1 ratio for different ring counts, showing robust distribution across core structures with only minor variation in highly fused systems. (f) Functional group statistics of the dataset: halogenated, aromatic, and carbonyl-containing molecules dominate, ensuring chemical diversity for generalizable model development.

ing a representative set of promising candidates for unbiased final evaluation.

2.2 Surrogate Model: Chemprop-MPNN

Rationale for Model Selection

The directed message-passing neural network (D-MPNN) from the Chemprop framework was selected as the surrogate model for its strong performance in molecular property prediction.[38,39] **(The same Chemprop architecture is also used in Section 2.1 as a Δ-learning calibration model that predicts the TD-DFT–xTB error; here, by contrast, it directly outputs absolute S_1 and T_1 energies as an surrogate model.)**

This choice was driven by two key advantages: (1) Its explicit modeling of bond directionality (e.g., single, double, conjugated) reduces noise from undirected representations, which is critical for capturing electronic transitions in photoactive molecules; and (2) Its native ensemble support enables simultaneous quantification of model and data uncertainties, making it highly suitable for active learning strategies.

Uncertainty Quantification

An ensemble of five independently trained D-MPNNs provided epistemic uncertainty estimates:

$$\text{Total Variance} = \underbrace{\frac{1}{5}\sum_{k=1}^{5}(\hat{y}_k - \bar{y})^2}_{\text{Model Uncertainty}} \tag{9}$$

where $\bar{y}$ is the mean of the ensemble.

Bayesian Hyperparameter Optimization

Key hyperparameters were tuned via Gaussian process-based Bayesian optimization to minimize validation MAE:

$$\theta^* = \arg\min_{\theta} \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \left(|S_1^{(i)} - \hat{S}_1^{(i)}(\theta)| + |T_1^{(i)} - \hat{T}_1^{(i)}(\theta)| \right) \tag{10}$$

During the active learning process, as we incrementally increased the training set size from 5k to 165k samples, we conducted independent Bayesian hyperparameter optimizations at

each data scale to determine the optimal model architectures. The hyperparameter search employed Root Mean Square Error (RMSE), the default evaluation metric in Chemprop, as the criterion for selecting the best configuration.

Our analysis showed that the optimal architecture did not converge to a single universal configuration; instead, the best-performing hyperparameters varied distinctly with the dataset size, including model depths (ranging from 3 to 6), hidden layer widths (hidden_size ranging from 1600 to 1900), and dropout rates (0.05–0.25). In medium-to-large data regimes (65k–165k), shallow architectures (depth=3), wider hidden layers (hidden_size=1900), and lower dropout rates (0.05) consistently outperformed other configurations. Adopting these individually optimized hyperparameters at each training set size reduced the test RMSE by approximately 8%–15% compared to an untuned baseline model (e.g., depth=3, hidden_size=1200, dropout=0.10). This stage-specific optimization strategy ensured optimal architecture selection at each phase of the active learning cycle, thereby eliminating potential bias from a fixed architecture and enabling fair comparisons across different acquisition strategies.

2.3 Acquisition Strategies

Four strategies were systematically benchmarked to balance exploration and exploitation (Fig. 4)[40,41]:

2.3.1 Uncertainty Sampling

For each molecule x, we compute the ensemble variance of the two targets—S_1 and T_1—using Eq. (9). The acquisition score is the sum of these two variances:

$$A_{\text{score}}(x) \;=\; \sigma^2_{S_1}(x) \;+\; \sigma^2_{T_1}(x), \tag{11}$$

where $\sigma^2_{S_1}(x)$ and $\sigma^2_{T_1}(x)$ are the predictive variances of S_1 and T_1, respectively, obtained from the five-model ensemble. Candidates with the highest $A_{\text{score}}(x)$ probe regions where the surrogate is least confident, thereby accelerating error reduction.

2.3.2 Diversity-Enhanced Sampling

To avoid selecting many nearly identical molecules, we penalize intra-batch similarity:

$$A_{\text{score}}(x) = A_{\text{base}}(x) \prod_{x' \in B} I\big(\text{Tanimoto}(x,x') < 0.6\big), \tag{12}$$

where B is the set of molecules already chosen for the current batch. The indicator function $I(\cdot)$ discards a candidate if its radius-2, 2048-bit Morgan fingerprint yields a Tanimoto similarity greater than 0.6 to any member of B, thereby guaranteeing sufficient structural diversity while still prioritizing high-uncertainty points.

The Tanimoto similarity between two molecules x and x', based on their binary fingerprint vectors $f(x)$ and $f(x')$, is defined as:

$$\text{Tanimoto}(x,x') = \frac{c}{a+b-c} \tag{13}$$

where $a = \sum_i f_i(x)$ and $b = \sum_i f_i(x')$ are the numbers of 'on' bits in each fingerprint, and $c = \sum_i f_i(x) f_i(x')$ is the number of bits shared by both fingerprints. Here, $f(x)$ denotes the Morgan fingerprint of molecule x; the Tanimoto similarity measures the overlap between the binary features of two molecules.

2.3.3 Target-Property Optimisation

We guide sampling toward molecules whose photophysical ratio

$$m(x) = \frac{T_1(x)}{S_1(x)}$$

is close to either of two design targets: emitter and sensitizer.[31] The overall score is a weighted sum of the two targets, with tunable weights $w_e, w_s \geq 0$ satisfying $w_e + w_s = 1$ (default $w_e = w_s = 0.5$).

v1 — Probability-within-band kernel

$$\begin{aligned} A_{\text{score}}(x) = w_e \int_{\tau_e-\varepsilon}^{\tau_e+\varepsilon} \mathcal{N}\big(m;\mu(x),\sigma^2(x)\big)\,dm \\ + w_s \int_{\tau_s-\varepsilon}^{\tau_s+\varepsilon} \mathcal{N}\big(m;\mu(x),\sigma^2(x)\big)\,dm. \end{aligned} \tag{14}$$

v2 — Exponential alignment kernel

$$A_{\text{score}}(x) = w_e \exp\big[-\eta\,|m(x)-\tau_e|\big] + w_s \exp\big[-\eta\,|m(x)-\tau_s|\big]. \tag{15}$$

v3 — Expected-improvement kernel. An EI-style variant that accounts for correlation between T_1 and S_1 is derived in *Text S3*; its numerical results are reported in *Fig S3* and therefore omitted here for brevity.

Detailed derivations, hyperparameter sweeps, and a head-to-head comparison of v1–v3 are provided in *Supplementary Information (Text S1-S3)*.

Symbol Definitions (Default Values in Parentheses)

- $m(x)$ — predicted ratio T_1/S_1 for molecule x;
- $\mu(x), \sigma^2(x)$ — mean and variance of the ensemble prediction for $m(x)$;
- $\tau_e = 0.5$, $\tau_s = 1.0$ — emitter-like and sensitizer-like targets[42,43];
- w_e, w_s — weights for the two targets (0.5, 0.5);
- ε — half-width of the tolerance band (0.05);
- η — selectivity parameter in v2 (5).

2.3.4 Domain Knowledge Integration

To ensure that candidates proposed by our active learning (AL) loop are experimentally viable, we introduced a **synthetic-feasibility prior**. The approach is inspired by the *Retrosynthetic Accessibility Score (RAscore)*[44], which predicts synthesizability from AiZynthFinder output.[45] Instead of using the original, domain-agnostic RAscore model, we trained a more expressive Chemprop graph neural network on a *photoswitch-specific* dataset to obtain higher accuracy in the chemical space of interest.

2.7×10^5 emitter, sensitizer and analogue structures were subjected to AiZynthFinder. A molecule was labeled *synthesized* ($y = 1$) if the planner discovered at least one complete route to commercially available precursors within a user-defined search limit

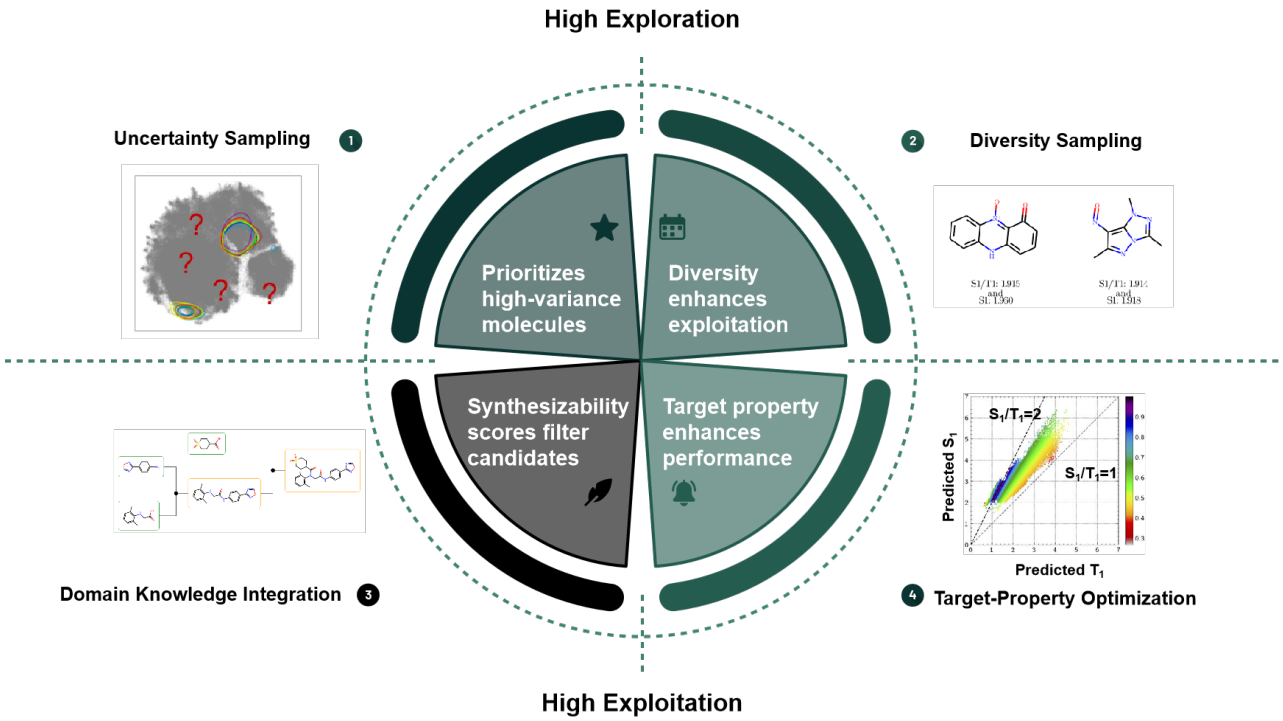

Fig. 4 Balancing exploration and exploitation in molecular selection. Four acquisition strategies are illustrated: uncertainty sampling prioritizes high-variance molecules; diversity sampling enhances exploration across chemical space; target-property optimization focuses on high-performance regions; and domain knowledge integration ensures synthetic feasibility. The combination of these strategies accelerates the identification of promising photosensitizer candidates.

(maximum n_{step} retrosynthetic steps or $t_{\max}$ seconds). Molecules for which no route was found under these restrictions were labeled *non-synthesizable* ($y = 0$). The resulting binary data set was used to train a Chemprop message-passing neural network that outputs a continuous probability $P_{\text{synth}}(x) \in [0,1]$. We refer to this domain-adapted score as the **PhotoSynthScore**.

During each AL iteration, we down-weight candidates that are unlikely to be synthesizable:

$$A_{\text{score}}(x) = A_{\text{base}}(x) \cdot I\big(P_{\text{synth}}(x) \geq 0.6\big), \tag{16}$$

where $A_{\text{base}}(x)$ is the score from Strategies 1–3 and $I(\cdot)$ is the indicator function. Only molecules with $P_{\text{synth}} \geq 0.6$ pass the filter, focusing computational effort on candidates that are both high-performing and practically attainable. Empirically, incorporating the PhotoSynthScore improves search efficiency and maintains chemical realism throughout the discovery campaign.

3 Results and Discussion

3.1 Sampling Strategy Benchmarking

Uncertainty Sampling: Reduced MAE from 0.091 eV to 0.077 eV within 4 rounds, with performance plateauing at Round 8. Deepening hidden layers further improved accuracy across all test sets, particularly for large molecules (>20 non-H atoms).

Diversity-Enhanced Sampling: Introducing Tanimoto similarity thresholds (<0.6) increased heterocyclic system coverage by 25%, but aggressive thresholds (<0.4) raised RMSE by 15% due to oversampling of chemically irrelevant regions.

Target-Property Optimization: Focused screening of T_1/S_1 ratios (1 for sensitizers, 0.5 for emitters) reduced emitter MAE from 0.065 eV $\rightarrow$ 0.061 eV versus uncertainty sampling.

Synthetic Feasibility Integration: Threshold filtering ($P_{\text{synth}} \leq 0.6$) eliminated over 20% of all candidates, as these were judged to be impractical for synthesis.

3.2 Model Performance and Dataset Size

As the size of the training dataset increases, the overall prediction accuracy of our model improves significantly, reflected by decreasing mean absolute error (MAE) and root mean square error (RMSE). However, this improvement demonstrates diminishing returns, indicating that once the dataset is sufficiently large, further additions yield limited performance gains. This trend is clearly depicted by the error curves across active learning rounds (Fig. 5), where initial rounds substantially reduce prediction errors, followed by a gradual plateau indicating saturation of the model's capacity to extract useful information.

Most predictions achieve high accuracy within acceptable chemical precision thresholds. However, a small fraction of compounds exhibit notably larger errors, likely arising from unique or rare structural motifs not adequately represented in the training set. This highlights areas for future model refinement, particularly addressing the "long-tail" of challenging cases.

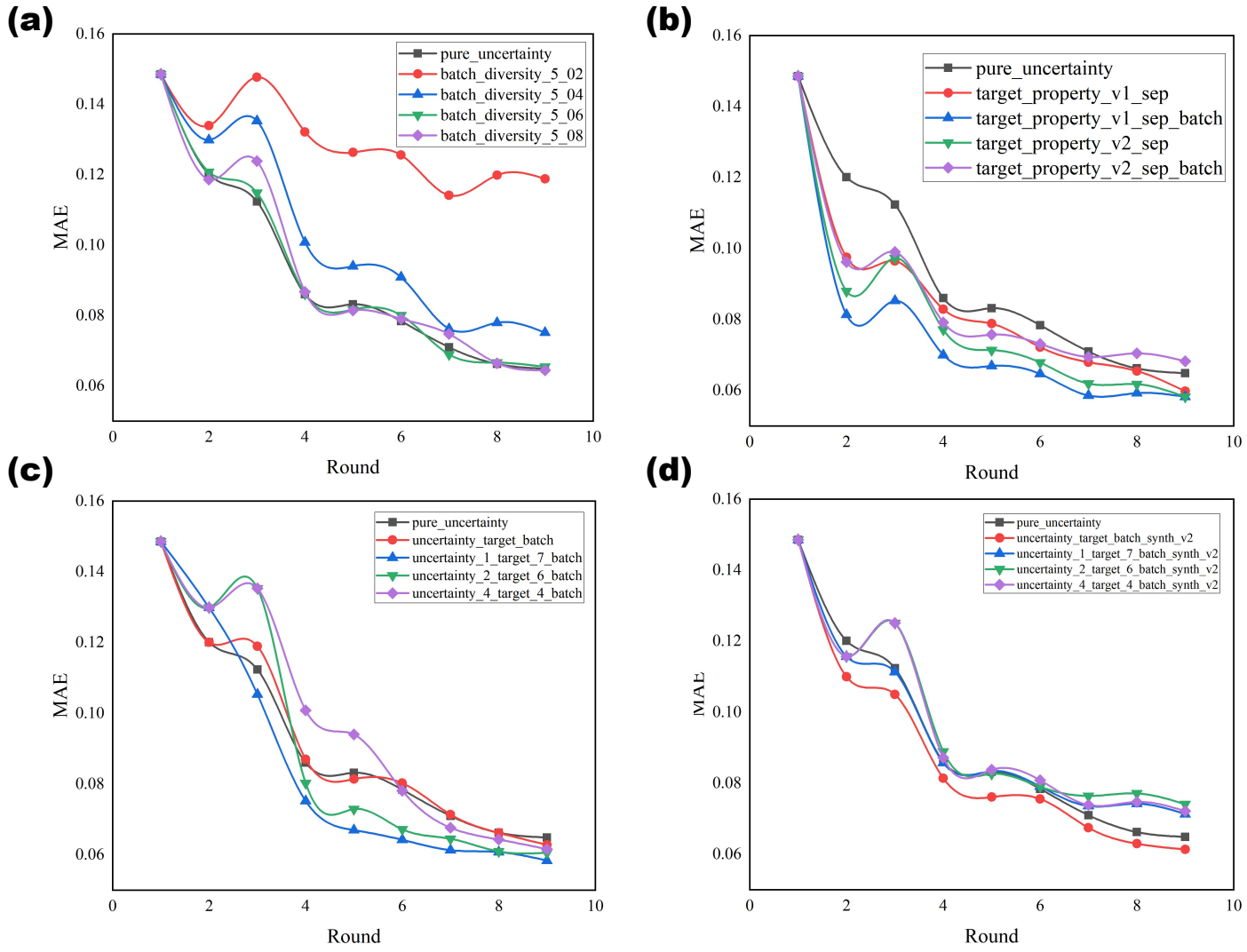

Fig. 5 Evolution of test-set MAE over active-learning round r for four representative acquisition strategies:
(a)
`uncertainty_batch_diversity_m_0`τ — Uncertainty-driven acquisition augmented by a diversity filter: a candidate is excluded if its Tanimoto similarity exceeds $\tau/10$ relative to at least m molecules already selected in the same batch;
(b)
`target_property_v`φ`_sep_batch` — Target-property-driven acquisition (version φ, see Section 2.3), combined with the diversity filter. Each batch split into half emitter-like and half sensitizer-like molecules;
`target_property_v`φ`_sep` — Target-property-driven acquisition (version φ, see Section 2.3), without applying the diversity filter;
(c)
`uncertainty_a_target_b_batch` — Sequential portfolio strategy: the first a rounds employ uncertainty sampling, followed by b rounds of target-property sampling, both applying the diversity filter;
`uncertainty_target_batch` — Weighted-sum portfolio strategy: uncertainty and target-property criteria are simultaneously considered in each round, applying the diversity filter;
(d)
`uncertainty_a_target_b_batch_synth_v2` — Same sequential portfolio strategy as (c), but incorporating an additional synthetic feasibility filter to exclude candidates predicted to be unsynthesizable;
`uncertainty_target_batch_synth_v2` — Weighted-sum portfolio strategy (as above), augmented with the synthetic feasibility filter .

3.3 Impact of Sampling Strategies on Data Distribution and Model Generalization

We investigated how various active learning sampling strategies, uncertainty sampling, diversity sampling, target property optimization, and synthesizability-guided sampling, impact data set composition and model generalization.

Uncertainty sampling prioritizes molecules that the model finds the most ambiguous, thus rapidly improving the model precision by identifying informative samples. **Diversity sampling** aims to maximize coverage of chemical space, ensuring structural heterogeneity and preventing redundancy. **Target-property optimization** selects molecules with extreme property values, facilitating efficient exploration of high-performance regions, but potentially neglecting broader chemical diversity.

Table 1 Performance comparison of different active learning sampling strategies.

Strategy	Initial MAE (eV)	Final MAE (eV)	Rounds to Convergence	Chemical Space Coverage
Pure uncertainty	0.149	0.077	8	Moderate
Uncertainty + Batch diversity	0.149	0.078	9	High
Target property + Batch diversity	0.149	0.062	9	Moderate (target-focused)
Portfolio strategy	0.149	0.059	9	High (balanced)
Portfolio + Synthesizability	0.149	0.074	9	High (practically feasible)

Lastly, **synthesizability-guided sampling** incorporates synthetic feasibility, prioritizing practical compounds that are experimentally accessible.

These differences directly affect model training dynamics and generalization performance. Uncertainty sampling yields rapid initial reductions in prediction errors, demonstrating superior efficiency. Diversity sampling ensures broader chemical generalization but shows moderate accuracy improvements. Target-property sampling initially contributes less to global error reductions but significantly enhances predictions for extreme property cases. Synthesizability-guided strategies moderately slow accuracy gains due to conservative selections but ensure higher practical relevance.

We further summarized efficiency and error metrics for each strategy in Table 1, providing comprehensive comparisons on test MAE, and chemical space coverage. This analysis facilitates the selection of strategies that align with research priorities, balancing prediction accuracy, scope of the study, and practical applicability.

3.4 Portfolio Strategies: Multi-Phase Active Learning and Synthesizability Constraints

Given the trade-offs above, an effective approach is to combine exploration and exploitation in a staged or simultaneous fashion—what we term a portfolio strategy. The idea is to first use uncertainty sampling to broadly train the model (exploration), and then gradually or abruptly shift toward selecting for the target property as the model becomes more reliable (exploitation). This latter approach dynamically balances exploration vs. exploitation within each iteration, rather than in separate phases.

3.4.1 Effect of Synthesizability Filtering on Portfolio Performance

Before comparing the exploration–exploitation schedules, we consider the role of a **synthesizability filter**. This filter excludes candidates predicted to be synthetically infeasible, adding a practical constraint to the selection process. Interestingly, we found that this constraint also has a pronounced effect on the learning dynamics. Both variants perform similarly in the early uncertainty-driven rounds (since those initial picks often tend to be relatively simple molecules or at least those the model is uncertain about, which may or may not be synthesizable).

Specifically, the **filtered portfolio strategy** shows a slight uptick or oscillation in MAE during the final rounds. After the switch to target-driven selection, the model begins choosing molecules that it predicts have extreme target values—some of these are likely unusual, highly complex structures that the model is less familiar with (and which might be chemically difficult to synthesize in reality). Incorporating such exotic compounds can momentarily *worsen* the model's overall performance: the newly added data might lie outside the domain where the model has predictive strength, leading to higher prediction errors for those points (and possibly disrupting the model's previously learned correlations). In other words, without the synthesizability filter, the model sometimes ventures into out-of-distribution regions in pursuit of high target property, and as a result the global MAE stops decreasing and even increases slightly in that phase. This phenomenon can be interpreted as a form of model extrapolation or overfitting issue—the model is essentially overextending into chemical space where its predictions are not reliable, analogous to how an overly aggressive exploitative strategy can mislead the model.

3.4.2 Balancing Exploration and Exploitation in Multi-Phase Strategies

Finally, we assess the general effectiveness of the portfolio strategies and how different balances between exploration (uncertainty sampling) and exploitation (target-based selection) influence outcomes. The multi-phase approaches allow us to tune the exploration–exploitation trade-off by deciding how many rounds to devote to each.

One notable advantage of the hybrid strategy is that it avoids an abrupt transition that might destabilize the model. In the sequential strategies, we sometimes see a kink or change in the MAE trend at the point of switching from uncertainty to target mode—if the switch happens too early, the model could struggle (as discussed, a slight MAE rise can occur if the model isn't ready to accurately handle the exploitation picks). The continuous strategy softens this by always maintaining a mix; effectively, it performs a dynamic rebalancing: as the model's confidence grows, more of the selected batch inherently contributes to exploitation (since fewer points will be at high uncertainty, the focus naturally shifts to high predicted property, without ever entirely ignoring uncertainty). This dynamic adjustability is a strong point of the hybrid method. The only slight drawback observed is that the hybrid strategy can be a bit slower to find the very top-performing molecules compared to a full-on exploitation in later rounds. Since it's never selecting only target-optimal candidates, it might miss a few opportunities to immediately test the absolute top predicted molecule in favor of an uncertain one. However, in practice this seems to be a minor penalty—the hybrid still discovers high-performing molecules throughout, just interspersed with exploratory picks. In exchange, it maintains the lowest MAE

curve among the methods, indicating it never compromises the model's learning too much in pursuit of the objective.

4 Conclusions

In this work, we present a unified computational framework to tackle the main challenges in photosensitizer discovery by introducing three key innovations:

- We developed a hybrid pipeline that combines fast semi-empirical quantum calculations with machine learning correction, enabling us to generate a large dataset of 655,197 diverse candidate molecules. This approach keeps the speed of computation high while ensuring the accuracy of quantum chemistry.

- We designed a sequential active learning strategy that separates the exploration and exploitation phases. First, the model explores chemical space broadly using uncertainty-driven sampling to find new promising areas, and then focuses on finding molecules with the desired T_1/S_1 energy ratios (1 for sensitizers, 0.5 for emitters).

Our open dataset and modular workflow offer a useful resource for the research community to speed up the development of new energy materials. Through detailed analysis, we show that an active learning approach which combines uncertainty sampling, target-driven optimization, diversity filters, and synthesizability constraints is most effective. This balanced strategy helps us find high-performing molecules faster, while also making sure the predictive model works well on different types of molecules. Adding diversity filters prevents the model from overfitting to very similar compounds, and considering synthesizability ensures the candidates found are likely to be made in real experiments.

While our model works well for small and moderate-sized molecules[46], it is less accurate for larger molecules with highly delocalized electronic structures or strong long-range interactions. This is mainly because the current graph neural networks are better at capturing local structure, and we only use 2D information rather than 3D shapes. For molecules whose properties depend on their 3D arrangement, this can limit our model's performance.

To overcome these issues, future work could explore hybrid architectures that insert multi-head Transformer self-attention directly into the message-passing layers of a GNN, going beyond the readout functions currently used in Chemprop, thus capturing both local chemical environments and long-range interatomic dependencies. Adding 3D geometric features, such as distances and angles between atoms, will also help the model better predict properties that depend on molecular shape.[47]

We also see value in combining multiple types of molecular information, such as electron density and orbital properties, and predicting a range of photophysical properties at once. This multi-task and multimodal approach can make the model stronger and easier to interpret.

Additionally, looking forward, computational methods for molecular discovery have evolved significantly, moving from simple virtual screenings to fully automated experimental platforms. Initially, high-throughput virtual screening (HTVS) allowed researchers to computationally evaluate thousands of candidate molecules from static libraries to quickly identify promising leads.[9] Next, active learning methods greatly improved efficiency by selectively and iteratively choosing the most informative candidates, thereby reducing computational cost and speeding up discovery. More recently, a new generation of closed-loop discovery systems has emerged, combining active learning with fully automated laboratories that autonomously synthesize, test, and validate molecular candidates in continuous cycles.[48] By incorporating these successive advancements into photosensitizer discovery, future workflows could efficiently and automatically discover effective molecules for practical applications in energy and medicine.

For active learning, specifically, adaptive strategies, those that dynamically adjust sampling rules throughout different learning stages, can further improve efficiency and generalization. By smoothly moving between uncertainty, diversity, and target-based sampling, the model can learn more with fewer experiments.

In summary, by applying smart active learning, we were able to further increase the accuracy and usefulness of our method. These steps helped bridge the gap between fast ML predictions and high-accuracy quantum methods, making our framework a reliable tool for discovering new photoactive molecules. Our work highlights the importance of balancing exploration and exploitation, including chemical diversity and synthesizability, and adjusting learning strategies as the model grows, laying a practical foundation for future AI-driven molecular discovery.

Author contributions

Yizhe Chen conceptualized the study, developed the overall framework, implemented the active learning algorithms, and drafted the manuscript. Shomik Verma provided the datasets, performed the implementation of the computational method. Kevin P. Greenman contributed to the algorithm design within Chemprop and provided critical analysis and suggestions regarding the active learning methodologies and reviewed and edited the manuscript. Haoyu Yin, Zhihao Wang, and Lanjing Wang assisted with data visualization, figure preparation, and content curation. Jiali Li, Rafael Gómez-Bombarelli, Aron Walsh, and Xiaonan Wang supervised the project, reviewed and edited the manuscript, and provided oversight throughout the study.

Conflicts of interest

The authors declare no competing interests.

Data availability

The curated dataset of 655,197 photosensitizer candidates and the complete active-learning code are available at (`https://github.com/jiali1025/A_General_Active_learning_framework_for_MoleDesign`). They are released under a CC-BY-4.0 licence (data) and MIT licence (code). No further restrictions apply.

Acknowledgements

K. P. G. was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1745302 and the DARPA Accelerated Molecular Discovery (AMD) program under contract HR00111920025. Other funding support details will be provided in the final published manuscript.

Notes and references

1 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.

2 Y. Cai, T. Chai, W. Nguyen, J. Liu, E. Xiao, X. Ran, Y. Ran, D. Du, W. Chen and X. Chen, *Signal Transduction and Targeted Therapy*, 2025, **10**, 115.

3 T. Froitzheim, S. Grimme and J.-M. Mewes, *Journal of Chemical Theory and Computation*, 2022, **18**, 7702–7713.

4 H. Kim, Y. R. Lee, H. Jeong, J. Lee, X. Wu, H. Li and J. Yoon, *Smart Molecules*, 2023, **1**, e20220010.

5 F. Hu, S. Xu and B. Liu, *Advanced Materials*, 2018, **30**, 1801350.

6 J. P. Janet, S. Ramesh, C. Duan and H. J. Kulik, *ACS Central Science*, 2020, **6**, 513–524.

7 S. Xu, J. Li, P. Cai, X. Liu, B. Liu and X. Wang, *Journal of the American Chemical Society*, 2021, **143**, 19769–19777.

8 K. Chen, X. Zhang, J. Wang, D. Li, T. Hou, W. Yang and Y. Kang, *Chemical Science*, 2025.

9 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu *et al.*, *Nature Materials*, 2016, **15**, 1120–1127.

10 P. Xu, X. Ji, M. Li and W. Lu, *npj Computational Materials*, 2023, **9**, 42.

11 X. Kang, Z. Du, S. Yang, M. Liang, Q. Liu and J. Qi, *Smart Molecules*, 2024, **2**, e20240033.

12 K. M. Jablonka, G. M. Jothiappan, S. Wang, B. Smit and B. Yoo, *Nature Communications*, 2021, **12**, 2312.

13 M. Sumita, X. Yang, S. Ishihara, R. Tamura and K. Tsuda, *ACS Central Science*, 2018, **4**, 1126–1133.

14 X. Li, P. M. Maffettone, Y. Che, T. Liu, L. Chen and A. I. Cooper, *Chemical Science*, 2021, **12**, 10742–10754.

15 J. Moon, W. Beker, M. Siek, J. Kim, H. S. Lee, T. Hyeon and B. A. Grzybowski, *Nature Materials*, 2024, **23**, 108–115.

16 Y. Zhao, Q. Liu, J. Du, Q. Meng and L. Zhang, *Smart Molecules*, 2023, **1**, e20230012.

17 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, *Nature*, 2023, **624**, 80–85.

18 B. Settles, 2009.

19 H. Chun, J. R. Lunger, J. K. Kang, R. Gómez-Bombarelli and B. Han, *npj Computational Materials*, 2024, **10**, 246.

20 C. Duan, A. Nandy, G. G. Terrones, D. W. Kastner and H. J. Kulik, *JACS Au*, 2022, **3**, 391–401.

21 R. Ding, J. Liu, K. Hua, X. Wang, X. Zhang, M. Shao, Y. Chen and J. Chen, *Science Advances*, 2025, **11**, eadr9038.

22 M. Kim, Y. Kim, M. Y. Ha, E. Shin, S. J. Kwak, M. Park, I.-D. Kim, W.-B. Jung, W. B. Lee, Y. Kim *et al.*, *Advanced Materials*, 2023, **35**, 2211497.

23 L. Wang, Z. Zhou, X. Yang, S. Shi, X. Zeng and D. Cao, *Drug Discovery Today*, 2024, 103985.

24 T. Yin, G. Panapitiya, E. D. Coda and E. G. Saldanha, *Journal of Cheminformatics*, 2023, **15**, 105.

25 H. H. Loeffler, S. Wan, M. Klahn, A. P. Bhati and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2024, **20**, 8308–8328.

26 B. Cree, M. K. Bieniek, S. Amin, A. Kawamura and D. J. Cole, *Digital Discovery*, 2025.

27 P. Shetty, A. Adeboye, S. Gupta, C. Zhang and R. Ramprasad, *Chemistry of Materials*, 2024, **36**, 7676–7689.

28 S. Thaler, F. Mayr, S. Thomas, A. Gagliardi and J. Zavadlav, *npj Computational Materials*, 2024, **10**, 86.

29 L. Kavalsky, V. I. Hegde, B. Meredig and V. Viswanathan, *Digital Discovery*, 2024, **3**, 999–1010.

30 D. Buterez, J. P. Janet, S. J. Kiddle, D. Oglic and P. Lió, *Nature Communications*, 2024, **15**, 1517.

31 S. Verma, M. Rivera, D. O. Scanlon and A. Walsh, *The Journal of Chemical Physics*, 2022, **156**, year.

32 G. Schneider and U. Fechner, *Nature Reviews Drug Discovery*, 2005, **4**, 649–663.

33 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J.-L. Reymond, *Journal of Chemical Information and Modeling*, 2012, **52**, 2864–2875.

34 M. Nakata, T. Shimazaki and H. Nakai, *Journal of Chemical Information and Modeling*, 2017, **57**, 1300–1308.

35 M. Schwilk, D. N. Tahchieva and O. A. von Lilienfeld, *The QMspin data set: Several thousand carbene singlet and triplet state structures and vertical spin gaps computed at MRCISD+Q-F12/cc-pVDZ-F12 level of theory*, 2020, `https://doi.org/10.24435/materialscloud:2020.0051/v1`.

36 C. Bannwarth, S. Ehlert and S. Grimme, *Journal of Chemical Theory and Computation*, 2019, **15**, 1652–1671.

37 S. Grimme and C. Bannwarth, *The Journal of Chemical Physics*, 2016, **145**, year.

38 E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green and C. J. McGill, *Journal of Chemical Information and Modeling*, 2024, **64**, 9–17.

39 J. Westermayr and P. Marquetand, *Chemical Reviews*, 2020, **121**, 9873–9926.

40 M. Martyka, L. Zhang, F. Ge, Y.-F. Hou, J. Jankowska, M. Barbatti and P. O. Dral, *npj Computational Materials*, 2025, **11**, 1–12.

41 A. Nigam, R. Pollice, G. Tom, K. Jorner, J. Willes, L. Thiede, A. Kundaje and A. Aspuru-Guzik, *Advances in Neural Information Processing Systems*, 2023, **36**, 3263–3306.

42 L. Naimovičius, P. Bharmoria and K. Moth-Poulsen, *Materials Chemistry Frontiers*, 2023, **7**, 2297–2315.

43 J. L. Weber, E. M. Churchill, S. Jockusch, E. J. Arthur, A. B. Pun, S. Zhang, R. A. Friesner, L. M. Campos, D. R. Reichman and J. Shee, *Chemical Science*, 2021, **12**, 1068–1079.

44 A. Thakkar, V. Chadimová, E. J. Bjerrum, O. Engkvist and J.-L. Reymond, *Chem. Sci.*, 2021, **12**, 3339–3349.

45 S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *Journal of Cheminformatics*, 2020, **12**, 70.

46 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Central Science*, 2018, **4**, 268–276.

47 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu and H. Wang, *Nature Machine Intelligence*, 2022, **4**, 127–134.

48 T. Wu, S. Kheiri, R. J. Hickman, H. Tao, T. C. Wu, Z.-B. Yang, X. Ge, W. Zhang, M. Abolhasani, K. Liu *et al.*, *Nature Communications*, 2025, **16**, 1473.