LLM-GUIDED EVOLUTION UNDER SEARCH-INDUCED DISTRIBUTION SHIFT FOR ORGANIC SOLAR CELL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Organic solar cell (OSC) discovery requires searching an enormous chemical space using only sparse experimental supervision. Efficient discovery requires exploring beyond the limited set of experimentally observed molecules, while optimization progressively concentrates candidates in the sparsely sampled high-performance tail, where conventional predictors become less reliable. To address these challenges, we introduce a closed-loop molecular discovery framework that couples feedback-conditioned LLM evolutionary search with a discovery-oriented high-tail evaluator. The LLM guides semantic mutations over experimentally grounded molecules, enabling adaptive structural exploration beyond fixed recombination heuristics while preserving chemically meaningful scaffolds. For candidate evaluation, we propose Aggressive Tail Fusion, which combines a data-driven model with an experimentally calibrated physics model to improve reliability in the high-PCE regime. It reduces MAE on the experimentally observed top-5% subset from 2.472 to 0.686.

1 INTRODUCTION

Organic solar cells (OSCs) have emerged as a promising photovoltaic technology owing to their lightweight, mechanical flexibility, and compatibility with solution-based fabrication. With the discovery of new non-fullerene acceptors (NFAs), recent advances have pushed their power conversion efficiencies (PCEs) beyond 20%, with optimized D18-Cl/BTP-C3F devices reaching 20.80% (Dong et al., 2025). Yet current efficiencies remain well below the detailed-balance limit of an ideal single-junction photovoltaic device (Shockley & Queisser, 1961), leaving substantial room for materials innovation. Closing this performance gap requires efficient discovery of new donor–acceptor materials from an enormous and highly combinatorial chemical space (Sanchez-Lengeling & Aspuru-Guzik, 2018). However, experimental coverage of this space remains extremely sparse: DeepAcceptor curated only 1,027 experimentally characterized NFAs from 508 publications, while OPV2D contains only 2,000 experimental donor–acceptor pairs (Sun et al., 2024; Lam et al., 2026). In contrast, even chemically constrained design spaces rapidly grow to millions or more candidates. DeepAcceptor generated 4.8 million molecules through BRICS recombination and a variational autoencoder, with approximately 3.6 million remaining after basic screening (Sun et al., 2024). Y6-type NFAs featuring an A–D–A′–D–A architecture have become a major focus in current experimental research. In our fragment space derived from experimentally reported A–D–A′–D–A molecules, combinatorial assembly yields over 300 million possible structures. Exhaustive synthesis and characterization are therefore infeasible, making computational discovery dependent on two tightly coupled capabilities: an effective search mechanism for exploring promising structures and a reliable evaluator for prioritizing candidates before expensive experimental validation.

Machine learning has advanced both sides of this pipeline. Early studies primarily focused on property prediction and virtual screening of donor–acceptor systems (Wu et al., 2020; Meftahi et al., 2020; Seifrid et al., 2024), while recent work has incorporated experimental fine-tuning, multi-fidelity learning, and large-scale screening (Sun et al., 2024; Zhang et al., 2025; Yang et al., 2026). Molecular design has further been explored through genetic algorithms and fragment-based search (Greenstein & Hutchison, 2023; Sun et al., 2024), generative Transformers with reinforcement learning (Lam et al., 2026), and more recently, LLM-based autonomous discovery agents (Hu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

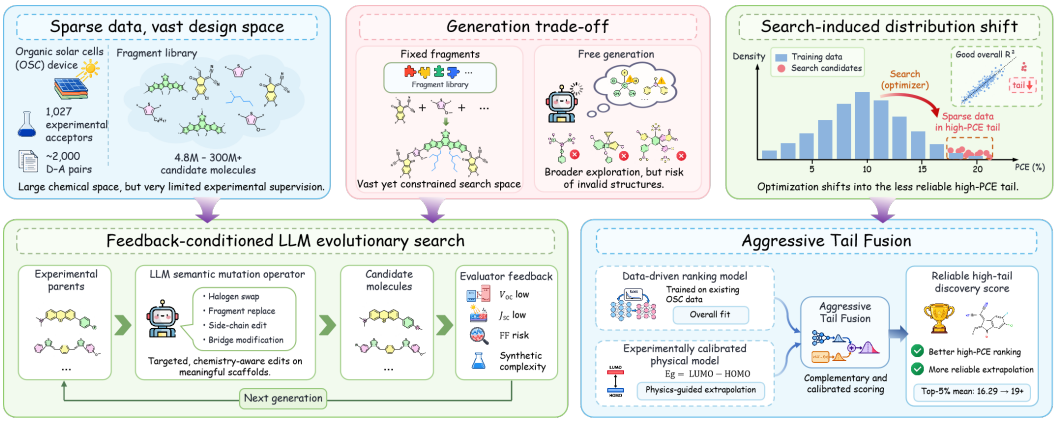

Figure 1: **Key challenges motivating reliable OSC discovery.** Sparse experimental supervision, search-induced distribution shift, and the trade-off between constrained and less controlled molecular generation jointly limit reliable optimization toward high-performance candidates.

2026). More broadly, recent work on recursive self-improvement (RSI) explores how LLM agents can use execution experience and feedback to progressively improve their own strategies, algorithms, or reusable skills (Yin et al., 2025; Ishibashi et al., 2025; Wang et al., 2026). However, scientific discovery introduces a failure mode that is less visible in conventional supervised prediction: optimization deliberately pushes candidates toward the sparsely observed high-performance frontier, where training support is limited. Closely related failures have been described as reward hacking in data-driven molecular design, where optimization exploits extrapolation errors of predictive models (Yoshizawa et al., 2025); broader molecular-optimization benchmarks have likewise emphasized the importance of realistic and reliable optimization oracles (Gao et al., 2022; Nigam et al., 2023). We refer to this phenomenon as *search-induced distribution shift*. Using P^3 (Lam et al., 2026) as a strong baseline, we observe an overall $R^2 = 0.718$, yet its mean prediction for the experimentally observed top 5% is only 16.458%, compared with the true mean of 18.930%. Thus, strong average predictive performance does not necessarily imply a reliable discovery oracle. At the same time, molecular generation faces a complementary tension: predefined fragment recombination preserves chemical priors but restricts exploration, whereas less constrained generation can drift toward unrealistic or poorly controlled structures (Sun et al., 2024; Hu et al., 2026).

Our key insight is that reliable molecular discovery requires an evaluator that remains trustworthy as search moves beyond the experimental training distribution. We therefore develop a closed-loop framework with two complementary components: feedback-conditioned LLM evolutionary search and a discovery-oriented high-tail evaluator for the extrapolative regime induced by search. The LLM acts as a semantic mutation operator over experimentally grounded parent molecules, enabling adaptive exploration over a chemically grounded fragment vocabulary while preserving meaningful scaffolds. For candidate evaluation, we introduce Aggressive Tail Fusion, which combines a data-driven model with an experimentally calibrated physics model. While maintaining strong overall predictive performance, the fused evaluator reduces MAE on the experimental Top-5% subset from 2.472 to 0.686, providing a more reliable selection signal for iterative discovery.

Our main contributions are threefold:

- **Search-induced distribution shift in molecular discovery.** We show that optimization drives candidates toward sparsely observed high-performance regions, where strong average predictive accuracy can fail to translate into reliable discovery performance.
- **A discovery-oriented evaluator for the high-PCE regime.** We introduce Aggressive Tail Fusion, which combines data-driven ranking with an experimentally calibrated physics model to improve evaluation in the high-PCE regime.
- **Feedback-conditioned LLM evolutionary search.** We use the LLM to guide semantic molecular edits over experimentally grounded molecules, enabling adaptive exploration within a chemically grounded fragment space rather than relying on fixed fragment-selection heuristics.

2 Related Work

Data-Driven and Physics-Based OSC Performance Prediction. Machine learning has become a standard approach for predicting photovoltaic properties and screening organic solar cell (OSC) materials. Early studies used molecular descriptors and fingerprints with conventional machine-learning models to predict PCE and related properties (Sahu et al., 2018; Zhao et al., 2020; Eibeck et al., 2021). Conventional models such as random forests, XGBoost, and LightGBM provide strong tabular baselines when combined with molecular descriptors or learned representations (Breiman, 2001; Chen & Guestrin, 2016; Ke et al., 2017). Deep molecular predictors instead learn directly from molecular structure, ranging from graph convolutional and attention networks (Kipf & Welling, 2016; Veličković et al., 2017) to message-passing models such as Chemprop (Heid et al., 2024) and self-supervised molecular representations such as MolCLR (Wang et al., 2022). OSC-specific models further incorporate domain data: DeepAcceptor fine-tunes pretrained molecular representations on experimentally characterized acceptors (Sun et al., 2024), while P^3 models donor–acceptor PCE through a hierarchical learning framework (Lam et al., 2026). In parallel, physics-based approaches such as the Scharber model estimate device efficiency from frontier orbital energies and optical gaps (Scharber et al., 2006), providing physically interpretable inductive biases but relying on simplifying assumptions. In contrast to methods optimized primarily for average predictive accuracy, we focus on reliable evaluation in the sparsely sampled high-PCE regime reached during molecular search.

Molecular Search and Generative Design for OSCs. Computational OSC discovery has evolved from virtual screening toward increasingly adaptive molecular optimization. Genetic algorithms and fragment-based methods explore chemical space through predefined mutation, crossover, or recombination operators (Greenstein & Hutchison, 2023; Sun et al., 2024; Cao et al., 2025), but their exploration is inherently constrained by the available fragment vocabulary and hand-designed transformations. More flexible generative approaches employ molecular language models and reinforcement learning to optimize candidates toward desired photovoltaic properties; for example, CycleChemist combines MatGPT-based generation with multi-objective reinforcement learning (Lam et al., 2026). Recent LLM-based systems further introduce autonomous scientific agents: OSCAgent coordinates planning, molecular generation, and candidate evaluation for OSC discovery (Hu et al., 2026). A related line of work studies recursive self-improvement (RSI), where agents use accumulated experience to revise their own reasoning procedures, algorithms, or reusable skills over successive iterations (Yin et al., 2025; Ishibashi et al., 2025; Wang et al., 2026). Our approach instead uses the LLM as a feedback-conditioned semantic mutation operator, enabling adaptive molecular edits within a predefined chemically grounded fragment space while retaining meaningful scaffolds.

3 Method

We present a closed-loop framework for high-performance organic solar cell (OSC) acceptor discovery, combining a discovery-oriented PCE evaluator with feedback-conditioned LLM evolutionary search. The framework is designed to provide reliable guidance as optimization progressively moves toward the sparsely observed high-PCE regime.

3.1 Problem Formulation

Let $x \in \mathcal{X}$ denote a candidate non-fullerene acceptor (NFA), and let $c = (D, H, E, Y)$ denote the device context, including donor, hole-transport layer, electron-transport layer, and device year. In this study, we fix the device context as $c^\star = (\mathrm{D18}, \mathrm{2PACz}, \mathrm{PNDIT{-}F3N}, 2026)$ and optimize

$$x^\star = \arg\max_{x \in \mathcal{X}_{\mathrm{valid}}} F(x; c^\star),$$

where $\mathcal{X}_{\mathrm{valid}}$ contains candidates satisfying molecular validity and predefined structural constraints, and F is the discovery-oriented PCE evaluator.

3.2 Framework Overview

Figure 2 illustrates the closed-loop framework. Starting from experimentally observed NFAs under a fixed device context, the LLM proposes local structural modifications to parent molecules. Valid

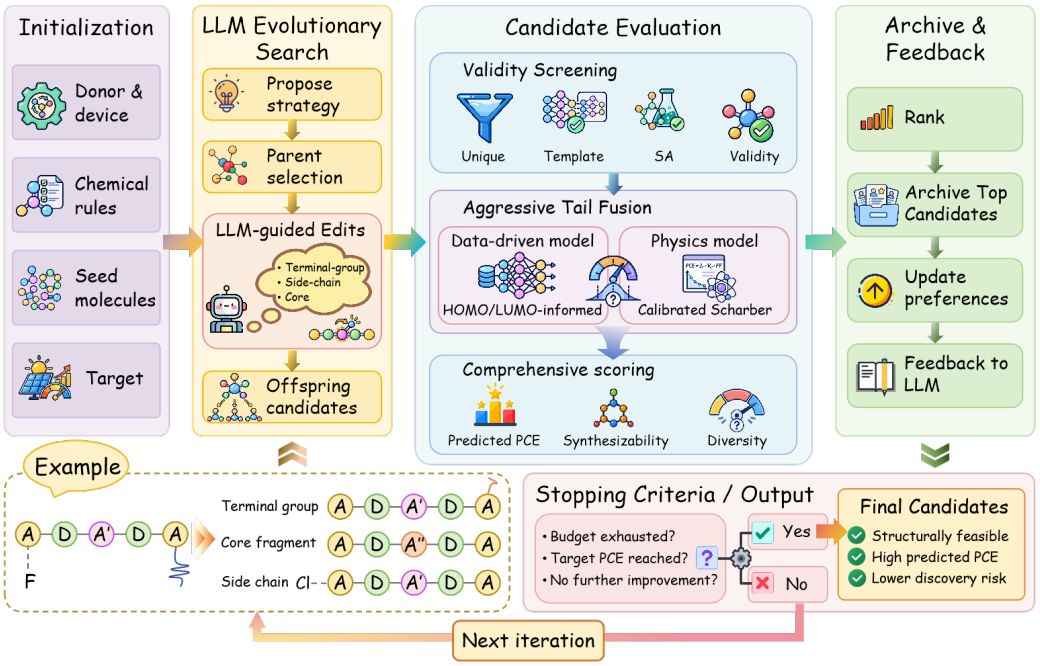

Figure 2: **Overview of the proposed LLM-guided evolutionary discovery framework.** LLM-based local mutation and Aggressive Tail Fusion form a closed-loop search that iteratively generates, evaluates, and refines high-performance molecular candidates.

candidates are evaluated by Aggressive Tail Fusion and selected according to predicted performance, feasibility, and structural diversity. Generation-wise outcomes are summarized as feedback for subsequent rounds, forming an iterative generation–evaluation–selection loop.

3.3 Discovery-Oriented PCE Evaluation

Molecular discovery requires reliable prediction in the high-performance tail, where iterative optimization progressively shifts candidates away from experimentally supported regions. As quantified in Appendix A, increasing predicted PCE is accompanied by decreasing similarity to experimental molecules and larger evaluator disagreement in later generations. We therefore combine a flexible data-driven predictor with a calibrated physics model through a tail-dependent fusion rule.

Data-Driven Ranking Model. Our data-driven model builds on the P^3 predictor in CycleChemist (Lam et al., 2026), but replaces direct PCE regression with structured multi-task prediction. Given a molecular representation, the model jointly predicts V_{oc}, J_{sc}, and FF, and constructs

$$\widehat{P}_i = \widehat{V}_{\mathrm{oc},i}\widehat{J}_{\mathrm{sc},i}\widehat{FF}_i + \alpha_P r_{P,i},$$

where $r_{P,i}$ captures residual device and fabrication effects.

High-PCE samples are relatively scarce in experimental datasets and can be under-emphasized by conventional regression objectives. We upweight them using

$$w_i \propto 1 + \lambda_{\mathrm{tail}}\sigma\left(\frac{z_i - \tau}{T_{\mathrm{tail}}}\right),$$

with weights normalized within each mini-batch. The model is trained with tail-weighted PCE regression together with auxiliary device-property supervision, pairwise ranking, and a consistency regularizer. Detailed objectives and hyperparameters are provided in Appendix B.1.

Physics-Based PCE Model. We construct an independently calibrated Scharber-style branch (Scharber et al., 2006), retaining explicit relationships between frontier energy levels, optical properties, and photovoltaic quantities while fitting empirical coefficients on experimental devices.

For example, the original Scharber model estimates the open-circuit voltage as

$$V_{\mathrm{oc}}^{\mathrm{Sch}} = \mathrm{clip}\left(|H_D| - |L_A| - 0.30,\ 0,\ 2.50\right),$$

where H_D and L_A denote the donor HOMO and acceptor LUMO, respectively. Thus, the donor–acceptor energy difference is converted to voltage using a fixed 0.30 eV empirical loss.

We instead fit

$$V_{\mathrm{oc}} = \mathrm{clip}(a_V E_{\mathrm{CT}} - \ell_V, 0, V_{\max}), \qquad \ell_V = b_0 - b_g(E_{\mathrm{opt}} - E_{\mathrm{ref}}) + c_L|L_A|^{p_L},$$

where E_{CT} denotes the donor–acceptor charge-transfer energy and ℓ_V is a molecular-property-dependent voltage loss. A fold-specific fitted example and additional calibration details are provided in Appendix B.3.

Aggressive Tail Fusion. Neither branch is uniformly optimal across the complete PCE range: the learned model is more accurate in well-supported regions, whereas the calibrated physics model becomes more informative in the high-performance tail. We therefore introduce *Aggressive Tail Fusion*, which activates the physics-based correction only when a candidate enters the high-score regime:

$$F(x) = r(x) + g_r(x)\left[\max\left(a(x) - r(x), 0\right) + \lambda g_a(x)\right],$$

where $r(x)$ and $a(x)$ denote the predictions of the learning and calibrated physics models, respectively. The smooth gates g_r and g_a are calibrated from the training-side validation prediction distributions of the two branches. For ordinary candidates, the gates remain small and the fused score closely follows the data-driven model. As the candidate enters the high-score tail, the physics-based correction is progressively activated. Unlike static averaging, this asymmetric fusion preserves interpolation accuracy while selectively introducing the physics prior in the regime most relevant to molecular search. Detailed fusion settings are provided in Appendix B.4.

3.4 LLM-Guided Evolutionary Search

We formulate molecular discovery as an iterative evolutionary search over NFAs under a fixed donor and device configuration. The search is initialized from experimentally observed molecules and maintains a population of high-performing and structurally diverse candidates throughout evolution.

At each generation, the LLM guides molecular editing by selecting promising fragments for three structural regions: the *core*, *end group*, and *side chain*. Its decisions are conditioned on representative parent molecules, their predicted photovoltaic properties, chemically compatible fragments, and feedback from previous generations. Rather than directly generating complete molecules, the LLM operates within a predefined fragment space, while molecular construction is performed through graph-level genetic operators, including fragment replacement, crossover, and local structural mutations. This design combines the chemical reasoning capability of LLMs with explicit structural constraints, reducing the likelihood of invalid molecular proposals.

Generated candidates are filtered for chemical validity, duplication, and synthetic accessibility before being scored by the discovery-oriented PCE evaluator described above. The resulting population is updated by jointly considering predicted performance and structural diversity, while a small fraction of initial molecules is retained to prevent premature convergence.

Importantly, the search forms a generation-wise feedback loop. For each molecular edit, we record its performance change relative to the parent and whether the resulting molecule survives subsequent selection. These observations are summarized and provided to the LLM in later generations, allowing it to progressively favor productive structural modifications while retaining sufficient exploration. Detailed evolutionary operators, population settings, and LLM selection hyperparameters are provided in Appendix C.1.

4 Experiments

Our experiments evaluate both components of the proposed framework: the reliability of Aggressive Tail Fusion in the high-PCE regime and the effectiveness of 40-generation LLM-guided evolutionary search. We further analyze search dynamics and the chemical characteristics of the generated high-scoring candidates.

Table 1: **PCE prediction performance on the overall distribution and the high-PCE Top-20% subset.** Overall results are reported as mean ± standard deviation across five folds, while Top-20% metrics are computed from aggregated out-of-fold predictions.

Method	Overall				High-PCE Top-20%			
	R^2 ↑	MAE↓	RMSE↓	Spearman ρ↑	R^2 ↑	MAE↓	RMSE↓	Spearman ρ↑
Traditional ML Baselines								
RF	0.728 ± 0.042	1.988 ± 0.124	2.706 ± 0.185	0.844 ± 0.021	-8.808	2.091	2.671	0.163
XGBoost	0.719 ± 0.042	2.010 ± 0.097	2.748 ± 0.164	0.837 ± 0.022	-8.199	1.987	2.587	0.167
LightGBM	0.722 ± 0.033	1.981 ± 0.064	2.739 ± 0.112	0.839 ± 0.017	-7.755	1.926	2.524	0.171
Deep-Learning Baselines								
GCN	0.678 ± 0.056	2.212 ± 0.145	2.945 ± 0.259	0.815 ± 0.038	-7.392	2.195	2.471	0.227
GAT	0.694 ± 0.076	2.118 ± 0.262	2.859 ± 0.369	0.830 ± 0.041	-5.945	1.894	2.248	0.170
Chemprop	0.714 ± 0.050	2.060 ± 0.169	2.774 ± 0.239	0.839 ± 0.029	-6.580	1.943	2.348	0.203
MolCLR	0.667 ± 0.048	2.229 ± 0.100	2.998 ± 0.213	0.805 ± 0.017	-6.294	1.966	2.304	0.066
DeepAcceptor	0.487 ± 0.038	2.995 ± 0.171	3.724 ± 0.093	0.474 ± 0.055	-19.514	3.738	3.863	-0.003
P^3	0.718 ± 0.047	1.987 ± 0.129	2.758 ± 0.234	0.836 ± 0.028	-4.775	1.682	2.050	0.285
Physics-Based Baseline								
Original Scharber	-0.607 ± 0.073	5.164 ± 0.157	6.598 ± 0.178	-0.309 ± 0.047	-9.892	2.667	2.815	-0.129
Ours								
Data-Driven Model	$\mathbf{0.732 \pm 0.045}$	$\mathbf{1.928 \pm 0.088}$	$\mathbf{2.686 \pm 0.214}$	$\mathbf{0.850 \pm 0.025}$	-4.399	1.617	1.982	0.295
Physics Model	0.364 ± 0.074	3.008 ± 0.211	4.147 ± 0.240	0.740 ± 0.029	-1.521	0.880	1.354	0.448
Aggressive Tail Fusion	0.683 ± 0.064	2.025 ± 0.164	2.917 ± 0.289	0.847 ± 0.026	$\mathbf{-1.051}$	**0.847**	**1.222**	**0.455**

4.1 Experimental Setup

Datasets. We curate 746 structurally distinct, experimentally synthesized Y6-type NFAs from 342 publications between 2019 and 2025, including J_{SC}, V_{OC}, FF, PCE, and literature provenance. We combine these data with the experimental dataset released with CycleChemist (Lam et al., 2026) and remove duplicates, resulting in 1,508 donor–acceptor devices. For molecular representation pretraining, we additionally use 51,256 DFT-evaluated molecules from the Harvard Clean Energy Project (Lopez et al., 2017).

Baselines. For PCE evaluation, we compare against three groups of prediction baselines. Traditional machine-learning methods include Random Forest (RF), XGBoost, and LightGBM, using the same learned molecular embeddings for a controlled comparison. Deep-learning baselines include GCN, GAT, Chemprop, MolCLR, and two OSC-specific predictors, DeepAcceptor (Sun et al., 2024) and P^3 (Lam et al., 2026). We additionally include the original Scharber model (Scharber et al., 2006) as a physics-based baseline. Our evaluator is reported through its three components: the Data-Driven Model, the experimentally calibrated Physics Model, and Aggressive Tail Fusion. For molecular search, we compare against BRICS (Degen et al., 2008), Graph-GA (Jensen, 2019), SMILES-GA (Yoshikawa et al., 2018), and an OSC-trained VAE (Gómez-Bombarelli et al., 2018), covering fragment recombination, graph-based genetic optimization, SMILES-based genetic optimization, and latent-space generation, respectively. We further include Direct LLM and Few-shot LLM as non-iterative generation baselines. Rule-based EA uses the same evolutionary operators, candidate screening, and evaluator as our method but removes LLM-guided fragment selection. Our evolutionary search is evaluated with three LLM backbones: DeepSeek-V4.1-Flash (Xu et al., 2026), GPT-5.6 Sol (OpenAI, 2026), and Claude Opus 5 (Anthropic, 2026).

Evaluation Protocol. For PCE evaluation, we use strict molecule-level five-fold cross-validation on the experimental PCE dataset. Within each fold, the held-out partition is used only for testing, and 10% of the remaining data are reserved for validation. We report fold-averaged performance on the full distribution and additionally evaluate aggregated out-of-fold predictions on the experimental Top-20% PCE subset. For molecular evolution, the three LLM backbones are evaluated with D18 as the donor. Previously unseen generated molecules are evaluated by averaging the predictions of the five fold-specific Aggressive Tail Fusion evaluators. All search methods therefore use the same five-fold ensemble throughout optimization. Unless otherwise specified, all evolutionary methods are run for 40 generations with 256 offspring candidates per generation, corresponding to 10,240 submitted candidates, and use the same candidate-screening criteria. Detailed evolutionary settings are provided in Appendix C.1. To assess robustness to search stochasticity, we additionally repeat the DeepSeek-V4.1-Flash evolutionary search with three random seeds and observe consistent high-PCE portfolios across independent runs; detailed statistics are reported in Appendix C.2.

Table 2: **Molecular optimization performance.** All candidates are evaluated by Aggressive Tail Fusion. AUC Top-10 is computed over the matched 40-generation search budget, and Q95 Excess uses the empirical 95th-percentile PCE threshold (18.2921%). † denotes short-budget baselines evaluated with 1,024 submitted candidates; AUC is not applicable.

Method	Search Eff.	Optimization		High-Tail Quality
	AUC Top-10 ↑	Best PCE ↑	Top-10 PCE ↑	Q95 Excess ↑
Traditional Molecular Optimization				
BRICS	18.1243	18.8871	18.8127	0.1885
SMILES-GA	18.2910	19.0490	19.0271	0.5925
Graph-GA	18.4174	19.4691	19.3830	0.8879
VAE (OSC)	12.1918	15.4301	13.2898	0.0000
Direct LLM-Based Generation				
Direct LLM†	/	12.7655	12.4872	0.0000
Few-shot LLM†	/	18.8959	18.6363	0.0128
Rule-Based Evolutionary Search				
Rule-based EA	18.5143	19.4425	19.4029	0.9370
Ours: LLM-Guided Evolutionary Search				
DeepSeek-V4.1-Flash	18.4543	19.5428	19.5259	0.7678
GPT-5.6 Sol	**18.6056**	<u>19.5569</u>	<u>19.5556</u>	**1.1452**
Claude Opus 5	<u>18.5810</u>	**19.5813**	**19.5797**	<u>1.1191</u>

Metrics. For PCE evaluation, we report R^2, MAE, RMSE, and Spearman correlation ρ on both the overall experimental distribution and the high-PCE Top-20% subset. For molecular optimization, we report AUC Top-10 for search efficiency, Best PCE and Top-10 PCE for optimization quality, Q95 Excess for performance beyond the experimental 95th-percentile boundary. Detailed definitions are provided in Appendix D.1.

4.2 Discovery-Oriented PCE Evaluation

Following the staged representation-learning procedure of CycleChemist (Lam et al., 2026), we first train the molecular encoder on frontier-orbital prediction before optimizing it for PCE prediction. The computed-domain stage uses more than 51,000 molecules from the dataset of Lopez et al. (Lopez et al., 2017), followed by fine-tuning on experimentally measured HOMO/LUMO values from our curated dataset described above. For both supervised stages, we adopt strict molecule-level 8:1:1 train/validation/test splits to prevent molecular overlap across partitions. On the held-out test sets, the model achieves $R^2 = 0.825$ and 0.876 for computed HOMO and LUMO, respectively, and $R^2 = 0.505$ and 0.349 after fine-tuning on experimental HOMO and LUMO measurements. Detailed prediction metrics are provided in Appendix B.2.

As shown in Table 1, the Data-Driven Model provides the strongest performance over the full experimental distribution, achieving $R^2 = 0.732$, MAE $= 1.928$, RMSE $= 2.686$, and Spearman $\rho = 0.850$. Aggressive Tail Fusion incurs only a modest reduction in bulk predictive accuracy while substantially improving high-tail prediction, retaining a strong overall Spearman correlation of $\rho = 0.847$. Because R^2 is sensitive to the reduced target variance within truncated high-PCE subsets, we interpret tail performance jointly using MAE, RMSE, and Spearman ρ. On the experimental Top-20% subset, the Data-Driven Model has an MAE of 1.617, whereas the Physics Model reduces it to 0.880. Aggressive Tail Fusion further improves the MAE to 0.847, together with an RMSE of 1.222 and Spearman $\rho = 0.455$, corresponding to a 47.6% MAE reduction relative to the Data-Driven Model.

This advantage becomes more pronounced in the extreme high-PCE tail. On the experimental Top-5% subset, Aggressive Tail Fusion achieves an MAE of 0.686, reducing the error by 69.3% relative to the Data-Driven Model (2.231) and by 72.3% relative to P^3 (2.472). These results further demonstrate the benefit of incorporating the calibrated physics prior when evaluation moves toward the

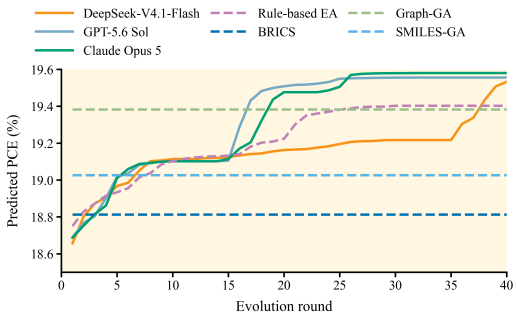

Figure 3: **Evolution trajectory under different search strategies.** Generation-wise cumulative Top-10 predicted PCE under the matched 40-generation budget.

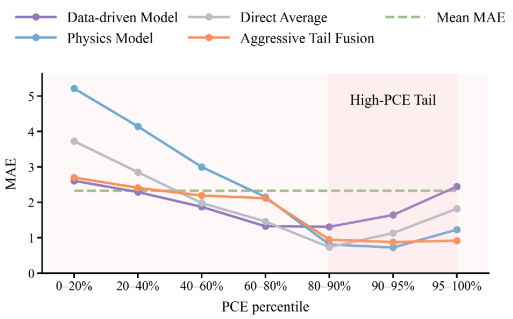

Figure 4: **Prediction error across the PCE distribution.** MAE of the Data-Driven, Physics, Fusion, and direct-average models across PCE percentiles.

sparsely observed high-performance frontier. Complete Top-5% results for all baselines are provided in Appendix D.2.

4.3 LLM-Guided Molecular Evolution

We next evaluate whether evolutionary search can consistently identify high predicted-PCE acceptor candidates rather than relying on isolated high-scoring samples. We use three LLM backbones, DeepSeek-V4.1-Flash, GPT-5.6 Sol, and Claude Opus 5, with D18 as the donor. All generated candidates are evaluated by the same Aggressive Tail Fusion evaluator.

Table 2 compares molecular search methods in terms of search efficiency, final optimization quality, and high-tail portfolio quality. Among conventional optimization methods, Graph-GA achieves a Best PCE of 19.4691% and a Top-10 PCE of 19.3830%, while the matched Rule-based EA reaches 19.4425% and 19.4029%, respectively. LLM-guided evolution further improves the elite population with the stronger backbones: GPT-5.6 Sol reaches a Best PCE of 19.5569% and a Top-10 PCE of 19.5556%, while Claude Opus 5 achieves the highest values of 19.5813% and 19.5797%. The close agreement between Best PCE and Top-10 PCE indicates that these gains are supported by a consistently high-performing elite set rather than a single isolated candidate.

The benefit of LLM guidance is also reflected over the search trajectory and in the depth of the discovered high-performance region. Under the matched 40-generation budget, GPT-5.6 Sol and Claude Opus 5 achieve AUC Top-10 values of 18.6056 and 18.5810, respectively, compared with 18.5143 for Rule-based EA and 18.4174 for Graph-GA. Their retained Top-200 portfolios also extend further beyond the experimental 95th-percentile boundary: GPT-5.6 Sol and Claude Opus 5 obtain Q95 Excess values of 1.1452 and 1.1191, exceeding Rule-based EA (0.9370) and Graph-GA (0.8879). Together, these results show that stronger LLM backbones improve not only the final elite candidates but also the accumulated optimization trajectory and the depth of the high-PCE portfolio. DeepSeek-V4.1-Flash shows a weaker early-search trajectory despite competitive final PCE, indicating that the benefit of LLM guidance depends on the backbone and its convergence behavior.

4.4 Analysis

Evolution dynamics. Figure 3 tracks the generation-wise cumulative Top-10 PCE, defined as the predicted PCE of the 10th-ranked candidate among all candidates observed up to each generation, while horizontal dashed lines denote the final Top-10 PCE reference levels attained by conventional optimization baselines. The LLM-guided trajectories exhibit stepwise improvements over successive generations rather than obtaining their final scores immediately. GPT-5.6 Sol and Claude Opus 5 progressively move beyond the Graph-GA reference level, while DeepSeek-V4.1-Flash exhibits a later improvement and reaches a similar high-PCE regime toward the end of the displayed trajectory. In contrast, Rule-based Evolution improves rapidly at early generations but saturates at a lower level. This behavior is consistent with gains accumulating through repeated feedback-conditioned selection of productive molecular edits, rather than arising from a single early-generation improve-

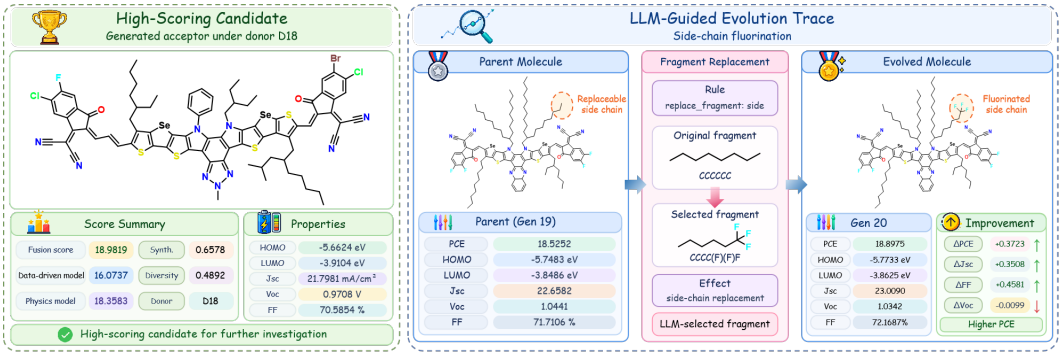

Figure 5: **Molecular evolution case study. Left:** a representative high-scoring candidate under D18. **Right:** a GPT-5.6 Sol alkyl-to-trifluoroalkyl side-chain substitution from Generation 19 to 20.

ment. These trends are consistent with the AUC Top-10 results in Table 2, indicating that the LLM backbone affects not only final candidate quality but also search convergence.

Ablation on tail-aware fusion. Figure 4 compares the Data-Driven Model, Physics Model, their direct average, and Aggressive Tail Fusion across experimental PCE percentiles. The Data-Driven Model achieves lower error over most of the lower and intermediate PCE range, while the Physics Model becomes increasingly competitive toward the high-PCE tail. Direct averaging introduces the less accurate physics prediction even in regions where the Data-Driven Model is already reliable. In contrast, Aggressive Tail Fusion preserves data-driven predictions over the better-supported region while increasingly incorporating the physics branch toward the high-PCE regime, yielding the lowest error in the extreme tail. This ablation shows that the improvement comes from the tail-dependent fusion mechanism rather than from simply averaging two predictors.

Case study. Figure 5 illustrates a representative local modification proposed during GPT-5.6 Sol evolution. At Generation 19, the LLM selects an alkyl-to-trifluoroalkyl side-chain substitution, replacing `CCCCCC` with `CCCC(F)(F)F` while preserving the main molecular scaffold. The evaluator predicts an increase in fused PCE from 18.525% to 18.898%, associated with higher predicted J_{sc} and FF despite a slight decrease in predicted V_{oc}. This example illustrates how feedback-conditioned evolution can identify local structural edits that improve the overall photovoltaic objective while allowing trade-offs among intermediate properties.

Chemical characteristics. As detailed in Appendix E, the high-scoring candidates largely retain the characteristic Y6-type scaffold, while most structural variation arises from localized side-chain, end-group, and halogen modifications. These localized edits preserve the overall molecular architecture while allowing the search to explore chemically plausible variations around experimentally grounded structures.

5 Conclusion

We study organic solar cell discovery under search-induced distribution shift, where iterative optimization moves candidate molecules toward a sparsely observed high-performance regime. We introduce Aggressive Tail Fusion to combine data-driven prediction with an experimentally calibrated physics prior, and integrate it with feedback-conditioned LLM evolutionary search. Experiments demonstrate improved high-tail evaluation and consistent discovery of high-scoring NFA candidates across multiple LLM backbones. Since no established high-fidelity computational oracle currently provides experimental-level PCE validation for unseen OSC molecules, the generated structures should be viewed as high-priority candidates for subsequent experimental validation. Overall, our results highlight the importance of reliability along the optimization frontier, beyond average predictive accuracy.

AI USE STATEMENT

Large language models, including DeepSeek-V4.1-Flash, GPT-5.6 Sol, and Claude Opus 5, are used as experimental backbones in our molecular evolutionary framework, with their roles and evaluation protocols described in the main text and appendix. Generative AI tools were additionally used only for language polishing and drafting visual elements during manuscript preparation. The authors retain full responsibility for the technical content, experimental results, citations, conclusions, and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

We describe the data sources, evaluation protocol, evaluator construction, molecular-search procedure, and key experimental settings in the main text and appendix. All LLM-based experiments use the official APIs of DeepSeek-V4.1-Flash, GPT-5.6 Sol, and Claude Opus 5, with fixed prompts and search settings within each experimental setting. GPT-5.6 Sol is used with `high` reasoning effort. Anonymous code and additional implementation details are provided with the submission.

REFERENCES

Anthropic. Introducing claude opus 5. https://www.anthropic.com/news/claude-opus-5, July 2026. Accessed: 2026-09-25.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Rui Cao, Cai-Rong Zhang, Xiao-Meng Liu, Ji-Jun Gong, Mei-Ling Zhang, Zi-Jiang Liu, You-Zhi Wu, and Hong-Shan Chen. Molecular design of organic photovoltaic donors and non-fullerene acceptors: a combined machine learning and genetic algorithm approach. *Journal of Materials Chemistry C*, 13(24):12150–12168, 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503, 2008.

Junwei Dong, Yinfeng Li, Chentong Liao, Xiaopeng Xu, Liyang Yu, Ruipeng Li, and Qiang Peng. Dielectric constant engineering of nonfullerene acceptors enables a record fill factor of 83.58% and a high efficiency of 20.80% in organic solar cells. *Energy & Environmental Science*, 18(10): 4982–4995, 2025.

Andreas Eibeck, Daniel Nurkowski, Angiras Menon, Jiaru Bai, Jinkui Wu, Li Zhou, Sebastian Mosbach, Jethro Akroyd, and Markus Kraft. Predicting power conversion efficiency of organic photovoltaics: models and data analysis. *ACS omega*, 6(37):23764–23775, 2021.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):8, 2009.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in neural information processing systems*, 35: 21342–21357, 2022.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268, 2018.

Brianna L Greenstein and Geoffrey R Hutchison. Screening efficient tandem organic solar cells with machine learning and genetic algorithms. *The Journal of Physical Chemistry C*, 127(13): 6179–6191, 2023.

Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: a machine learning package for chemical property prediction. *Journal of chemical information and modeling*, 64(1): 9–17, 2024.

Zhaolin Hu, Zhiliang Wu, Kun Li, Hehe Fan, and Yi Yang. Oscagent: Accelerating the discovery of organic solar cells with llm agents. *arXiv preprint arXiv:2602.04510*, 2026.

Yoichi Ishibashi, Taro Yano, and Masafumi Oyamada. Can large language models invent algorithms to improve themselves?: Algorithm discovery for recursive self-improvement through reinforcement learning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10332–10363, 2025.

Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Hou Hei Lam, Jiangjie Qiu, Xiuyuan Hu, Wentao Li, Fankun Zeng, Siwei Fu, Hao Zhang, and Xiaonan Wang. Cyclechemist: A dual-pronged machine learning framework for organic photovoltaic discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38799–38807, 2026.

Steven A Lopez, Benjamin Sanchez-Lengeling, Julio de Goes Soares, and Alán Aspuru-Guzik. Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule*, 1(4):857–870, 2017.

Nastaran Meftahi, Mykhailo Klymenko, Andrew J Christofferson, Udo Bach, David A Winkler, and Salvy P Russo. Machine learning property prediction for organic photovoltaic devices. *npj computational materials*, 6(1):166, 2020.

AkshatKumar Nigam, Robert Pollice, Gary Tom, Kjell Jorner, John Willes, Luca Thiede, Anshul Kundaje, and Alan Aspuru-Guzik. Tartarus: A benchmarking platform for realistic and practical inverse molecular design. *Advances in Neural Information Processing Systems*, 36:3263–3306, 2023.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, July 2026. Accessed: 2026-09-25.

Harikrishna Sahu, Weining Rao, Alessandro Troisi, and Haibo Ma. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Advanced Energy Materials*, 8(24):1801032, 2018.

Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

Markus C Scharber, David Mühlbacher, Markus Koppe, Patrick Denk, Christoph Waldauf, Alan J Heeger, and Christoph J Brabec. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Advanced materials*, 18(6):789–794, 2006.

Martin Seifrid, Stanley Lo, Dylan G Choi, Gary Tom, My Linh Le, Kunyu Li, Rahul Sankar, Hoai-Thanh Vuong, Hiba Wakidi, Ahra Yi, et al. Beyond molecular structure: critically assessing machine learning for designing organic photovoltaic materials and devices. *Journal of Materials Chemistry A*, 12(24):14540–14558, 2024.

William Shockley and Hans J Queisser. Detailed balance limit of efficiency of p-n junction solar cells. *Journal of applied physics*, 32(3):510–519, 1961.

Jinyu Sun, Dongxu Li, Jie Zou, Shaofeng Zhu, Cong Xu, Yingping Zou, Zhimin Zhang, and Hongmei Lu. Accelerating the discovery of acceptor materials for organic solar cells by deep learning. *npj Computational Materials*, 10(1):181, 2024.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Zefeng Wang, Minxi Yan, Jinhe Bi, Sikuan Yan, Volker Tresp, and Yunpu Ma. Metaskill-evolve: Recursive self-improvement of llm agents via two-timescale meta-skill evolution. *arXiv preprint arXiv:2607.05297*, 2026.

Yao Wu, Jie Guo, Rui Sun, and Jie Min. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Computational Materials*, 6(1):120, 2020.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. DeepSeek-V4.1-Flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Haizhou Yang, Adam Wold, Junlin Ou, Jeromy J Rech, Wei You, and Yi Wang. Osc-net: a multi-fidelity machine learning model for organic solar cells. *Journal of Materials Chemistry A*, 14(2): 1208–1220, 2026.

Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, and William Yang Wang. Gödel agent: A self-referential agent framework for recursively self-improvement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27890–27913, 2025.

Naruki Yoshikawa, Kei Terayama, Masato Sumita, Teruki Homma, Kenta Oono, and Koji Tsuda. Population-based de novo molecule generation, using grammatical evolution. *Chemistry Letters*, 47(11):1431–1434, 2018.

Tatsuya Yoshizawa, Shoichi Ishida, Tomohiro Sato, Masateru Ohta, Teruki Honma, and Kei Terayama. A data-driven generative strategy to avoid reward hacking in multi-objective molecular design. *Nature communications*, 16(1):2409, 2025.

Shizhao Zhang, Shuixing Li, Siqin Song, Yang Zhao, Liang Gao, Hongzheng Chen, Hanying Li, and Jiaping Lin. Deep learning-assisted design of novel donor–acceptor combinations for organic photovoltaic materials with enhanced efficiency. *Advanced Materials*, 37(4):2407613, 2025.

Zhi-Wen Zhao, Marcos del Cueto, Yun Geng, and Alessandro Troisi. Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells. *Chemistry of Materials*, 32(18):7777–7787, 2020.

A SEARCH-INDUCED DISTRIBUTION SHIFT

To examine whether evolutionary optimization progressively moves candidates beyond regions well supported by experimental observations, we track both molecular-space proximity and evaluator disagreement throughout the search.

For each generated candidate x, we measure its proximity to the experimental training distribution using the maximum Tanimoto similarity to any experimentally observed molecule,

$$s_{\mathrm{NN}}(x) = \max_{x_i \in \mathcal{D}_{\mathrm{exp}}} \mathrm{Tanimoto}(x, x_i),$$

where molecular similarity is computed using the same fingerprint representation for generated and experimental molecules. A lower s_{NN} therefore indicates that the candidate lies farther from its nearest experimentally supported analogue.

We additionally measure disagreement between the two independently constructed evaluator branches as

$$\Delta_{\mathrm{eval}}(x) = \left| F_{\mathrm{data}}(x) - F_{\mathrm{phys}}(x) \right|,$$

where F_{data} and F_{phys} denote the predictions of the data-driven and calibrated physics models, respectively. Large disagreement does not by itself imply prediction error, but provides a complementary indicator that the two models extrapolate differently for a candidate.

Figure 6 shows these quantities across generations for a representative Claude Opus 5 search with D18 as the donor and random seed 7. As optimization proceeds, the median fused PCE increases from approximately 17.8% in the first generation to above 19.1% in the final generations, while the best-so-far score reaches approximately 19.6%. At the same time, the median nearest-neighbor similarity decreases substantially, from about 0.93 at the beginning of the search to about 0.52 by Generation 30. Thus, the improvement in the optimization objective is accompanied by a systematic movement away from the experimentally observed molecular distribution.

Evaluator disagreement exhibits a complementary pattern. Although it fluctuates across intermediate generations, the disagreement between the data-driven and physics-based branches becomes larger toward the later stage of evolution, reaching its highest sustained levels when the population has moved furthest from the experimental set. Together, these observations indicate that evolutionary optimization increasingly concentrates candidates in a high-scoring but less experimentally supported region of chemical space. We refer to this optimization-induced change in the evaluated candidate distribution as *search-induced distribution shift*.

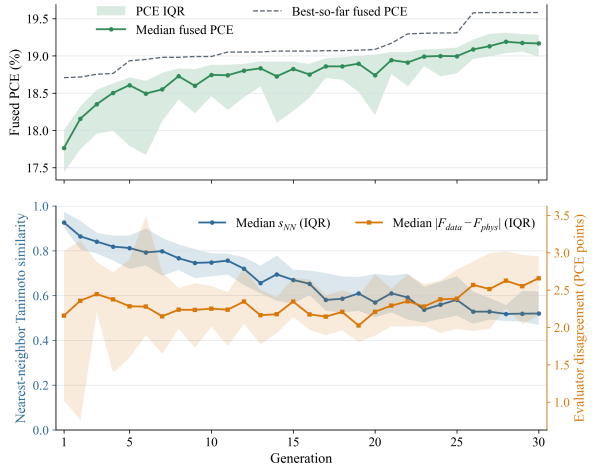

Figure 6: **Search-induced distribution shift during molecular evolution.** Generation-wise fused PCE, nearest-neighbor similarity to experimental molecules, and evaluator disagreement. Shaded regions denote the IQR.

B Discovery-Oriented PCE Evaluator Details

B.1 Data-Driven Ranking Model Details

Our data-driven evaluator builds upon the P^3 molecular predictor in CycleChemist (Lam et al., 2026), while modifying its training objective for high-performance molecular discovery. Rather than directly regressing PCE as a single scalar target, the model jointly predicts the photovoltaic quantities V_{oc}, J_{sc}, and FF from a shared molecular representation. The final PCE prediction is constructed as

$$\widehat{P}_i = \widehat{V}_{\mathrm{oc},i}\widehat{J}_{\mathrm{sc},i}\widehat{FF}_i + \alpha_P r_{P,i},$$

where $\widehat{P}_i$ denotes the predicted PCE for sample i, $\widehat{V}_{\mathrm{oc},i}$, $\widehat{J}_{\mathrm{sc},i}$, and $\widehat{FF}_i$ are the corresponding predicted device quantities, $r_{P,i}$ is a learnable residual term, and α_P controls the magnitude of the residual correction. The residual accounts for device and fabrication effects that are not explicitly captured by the product of the three intermediate photovoltaic quantities.

Tail-aware regression. High-PCE devices constitute only a small fraction of the experimental dataset. Uniform regression can therefore under-emphasize precisely the region that becomes most important during molecular optimization. We assign each training sample a smooth tail-dependent weight

$$\widetilde{w}_i = 1 + \lambda_{\mathrm{tail}}\sigma\left(\frac{z_i - \tau}{T_{\mathrm{tail}}}\right),$$

where z_i is the standardized experimental PCE, $\sigma(\cdot)$ denotes the sigmoid function, λ_{tail} determines the strength of tail emphasis, τ controls the onset of the high-PCE region, and T_{tail} controls the smoothness of the transition. The weights are normalized within each mini-batch:

$$w_i = \frac{\widetilde{w}_i}{\frac{1}{B}\sum_{j=1}^{B}\widetilde{w}_j},$$

where B is the mini-batch size. The resulting tail-aware PCE regression objective is

$$\mathcal{L}_{\mathrm{PCE}} = \frac{1}{B}\sum_{i=1}^{B} w_i\left(\widehat{z}_i - z_i\right)^2,$$

where $\widehat{z}_i$ is the standardized predicted PCE.

Auxiliary supervision and ranking. In addition to the PCE objective, the model receives auxiliary supervision from the experimentally measured photovoltaic quantities. We denote the corresponding objective by

$$\mathcal{L}_{\mathrm{aux}} = \mathcal{L}_{V_{\mathrm{oc}}} + \mathcal{L}_{J_{\mathrm{sc}}} + \mathcal{L}_{FF}.$$

We additionally introduce a pairwise ranking objective $\mathcal{L}_{\mathrm{rank}}$ that encourages the predicted ordering of candidate molecules to agree with their experimental PCE ordering. This is particularly relevant for evolutionary search, where candidate selection depends primarily on relative ranking rather than absolute prediction alone.

Finally, a consistency regularizer $\mathcal{L}_{\mathrm{cons}}$ discourages excessive deviation between the final PCE prediction and the physically interpretable prediction anchor formed by $\widehat{V}_{\mathrm{oc}}\widehat{J}_{\mathrm{sc}}\widehat{FF}$.

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{PCE}} + 0.35\mathcal{L}_{\mathrm{aux}} + 0.15\mathcal{L}_{\mathrm{rank}} + 0.10\mathcal{L}_{\mathrm{cons}}.$$

Table 3: HOMO/LUMO prediction after computed-domain pretraining.

Evaluation Split	Target	n	$R^2 \uparrow$	MAE↓	RMSE↓
Stage split / combined	HOMO	5,256	0.828	0.151	0.251
Stage split / combined	LUMO	5,256	0.857	0.160	0.210
Raw test / computed	HOMO	5,126	0.825	0.151	0.252
Raw test / computed	LUMO	5,126	0.876	0.152	0.197
Raw test / experimental	HOMO	130	-0.659	0.145	0.185
Raw test / experimental	LUMO	130	-16.301	0.476	0.505

Table 4: HOMO/LUMO prediction after experimental-domain fine-tuning.

Evaluation Split	Target	n	$R^2 \uparrow$	MAE↓	RMSE↓
Stage split / combined	HOMO	5,256	0.594	0.284	0.385
Stage split / combined	LUMO	5,256	0.423	0.339	0.422
Raw test / computed	HOMO	5,126	0.581	0.289	0.390
Raw test / computed	LUMO	5,126	0.415	0.346	0.427
Raw test / experimental	HOMO	130	0.505	0.078	0.101
Raw test / experimental	LUMO	130	0.349	0.075	0.098

This objective combines high-tail emphasis, device-level supervision, relative ranking, and structured consistency while retaining a flexible data-driven predictor.

B.2 Molecular Representation Pretraining

Tables 3 and 4 report the detailed HOMO/LUMO prediction results after computed-domain pretraining and experimental-domain fine-tuning, respectively. We report R^2, MAE, and RMSE on the stage-level combined split, the held-out computed test set, and the held-out experimental test set. MAE and RMSE are reported in eV.

B.3 Calibrated Physics-Based PCE Model

Our physics evaluator follows the analytical structure of the Scharber model (Scharber et al., 2006), but replaces fixed empirical constants with coefficients fitted from experimental donor–acceptor devices. The model therefore retains explicit physical relationships between molecular properties and photovoltaic quantities while adapting their empirical calibration to the experimental domain.

Open-circuit voltage calibration. As a representative example, the original Scharber-style relation used in our baseline estimates the open-circuit voltage as

$$V_{\mathrm{oc}}^{\mathrm{Sch}} = \mathrm{clip}\left(|H_D| - |L_A| - 0.30, 0, 2.50\right),$$

where H_D denotes the donor HOMO energy and L_A denotes the acceptor LUMO energy. The fixed 0.30 eV term represents an empirical voltage loss.

The clipping operator is defined as

$$\mathrm{clip}(x, l, u) = \min\left(\max(x, l), u\right),$$

and constrains predicted device quantities to physically plausible ranges.

Instead of assuming a fixed voltage loss, we parameterize the open-circuit voltage as

$$V_{\mathrm{oc}} = \mathrm{clip}\left(a_V E_{\mathrm{CT}} - \ell_V, 0, V_{\max}\right),$$

where the effective voltage-loss term is

$$\ell_V = b_0 - b_g\,(E_{\mathrm{opt}} - E_{\mathrm{ref}}) + c_L|L_A|^{p_L}.$$

Here, $E_{\mathrm{CT}} \approx L_A - H_D$ denotes the donor–acceptor charge-transfer energy, a_V controls the conversion from charge-transfer energy to V_{oc}, b_0 is the fitted baseline voltage loss, E_{opt} denotes the effective optical gap, E_{ref} is a reference optical gap, and the final term allows an additional acceptor-LUMO-dependent correction.

All empirical coefficients are fitted using only the corresponding training partition, while physically plausible bounds are imposed on intermediate device quantities.

Example fitted relation. For example, in Fold 1 the fitted parameters are $a_V = 0.70$, $b_0 = 0.2093$, $b_g = 0.0552$, $E_{\mathrm{ref}} = 1.40$ eV, and $V_{\max} = 1.40$ V. The fitted LUMO-dependent coefficient is approximately zero in this fold, giving

$$V_{\mathrm{oc},1} = \mathrm{clip}\left[0.70E_{\mathrm{CT}} - 0.2093 + 0.0552\,(E_{\mathrm{opt}} - 1.40)\,, 0, 1.40\right].$$

Compared with the original Scharber expression, this calibration makes three principal changes. First, the response to the donor–acceptor charge-transfer energy is fitted rather than fixed to unit slope. Second, the constant 0.30 eV voltage loss is replaced by an experimentally fitted baseline loss. Third, molecular-property-dependent terms provide additional corrections to the effective voltage loss. The resulting model preserves the analytical structure of the physical relation while adapting its empirical parameters to experimentally observed OSC devices.

B.4 Aggressive Tail Fusion Details

Aggressive Tail Fusion combines the complementary predictions of the data-driven and calibrated physics branches. Let $r(x)$ denote the prediction of the Data-Driven Model and $a(x)$ denote that of the Physics Model. The fused prediction is defined as

$$F(x) = r(x) + g_r(x)\left[\max(a(x) - r(x), 0) + \lambda g_a(x)\right].$$

Specifically, we set the fusion coefficient to $\lambda = 0.625$, and use smooth sigmoid gates

$$g_r(x) = \sigma\left(\frac{r(x) - \tau_r}{T_r}\right), \qquad g_a(x) = \sigma\left(\frac{a(x) - \tau_a}{T_a}\right),$$

where τ_r and τ_a determine the onset of the high-score regime, and T_r and T_a control the transition smoothness. All parameters are selected using training-side validation predictions and fixed before held-out evaluation. The gates remain small in well-supported regions and increase in the high-score regime, while $\max(a(x) - r(x), 0)$ activates the physics correction only when it exceeds the data-driven prediction, preserving interpolation performance.

Five-fold ensemble for molecular search. The five cross-validation folds are also used to construct the final evaluator for previously unseen generated molecules. For each fold $k \in \{1, \ldots, 5\}$, the corresponding Data-Driven Model and Physics Model produce a fold-specific fused prediction $F^{(k)}(x)$. The final score used throughout molecular evolution is the ensemble average

$$F_{\mathrm{ens}}(x) = \frac{1}{5}\sum_{k=1}^{5} F^{(k)}(x).$$

All molecular-search baselines and LLM-guided evolutionary runs are evaluated using the same five-fold ensemble, ensuring that differences in optimization performance arise from the search strategy rather than from the evaluator.

C LLM-GUIDED EVOLUTIONARY SEARCH DETAILS

C.1 EVOLUTIONARY SEARCH CONFIGURATION

Table 5 summarizes the main hyperparameters used for LLM-guided evolutionary search. Unless otherwise specified, the same configuration is used across all LLM backbones.

Table 5: **Configuration of LLM-guided evolutionary search.**

Parameter	Setting
Population size	96
Initial candidate pool	200
Offspring per generation	256
Evolution generations	40
Random seed	7
LLM context candidates	24
Fragment types	core / end / side
Genetic operators	12 fixed operators
Final fragment ranking size	6
LLM-guided fragment sampling	0.75
Random immigrant fraction	0.15
Generated-molecule archive size	2,000

The genetic operators include fragment replacement and crossover for core, end-group, and side-chain regions, together with local mutations such as halogen substitution, fluorination adjustment, and side-chain modification. LLM guidance is applied to fragment-replacement operations, whereas crossover and local mutations follow predefined chemical operators. Fragment compatibility and molecular validity are checked using RDKit before property evaluation.

C.2 ROBUSTNESS ACROSS RANDOM SEEDS

We further evaluate the robustness of DeepSeek-V4.1-Flash evolutionary search across three independent runs with different random seeds. As shown in Table 6, the search consistently reaches a similar high-PCE regime across runs.

Across the three runs, Best PCE remains at $19.5548 \pm 0.0359\%$, while the Top-10 PCE, defined as the score of the 10th-ranked candidate, is $19.5206 \pm 0.0743\%$. The broader retained portfolio is also highly stable, with a Top-200 mean PCE of $19.0614 \pm 0.0098\%$ and Q95 Excess of 0.7694 ± 0.0096. These results indicate that the high-performance region discovered by DeepSeek-V4.1-Flash is robust to search stochasticity rather than being driven by a single favorable trajectory.

D EVALUATION PROTOCOL AND ADDITIONAL RESULTS

D.1 EVALUATION METRICS

We evaluate the framework from two complementary perspectives: *PCE prediction reliability* and *molecular optimization quality*. The former measures whether the evaluator accurately ranks and predicts experimentally observed devices, particularly in the high-PCE regime, whereas the latter measures the efficiency and quality of the resulting molecular search.

PCE prediction metrics. For an evaluation set of N experimental devices with measured PCE y_i and prediction $\hat{y}_i$, we report the coefficient of determination,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean experimental PCE.

Table 6: **Robustness of DeepSeek-V4.1-Flash evolution across random seeds.** Top-10 PCE denotes the predicted PCE of the 10th-ranked candidate, and Top-200 PCE denotes the mean predicted PCE of the retained Top-200 candidates. Q95 Excess measures the mean positive excess over the experimental 95th-percentile PCE threshold.

Seed	**Best PCE ↑**	**Top-10 PCE ↑**	**Top-200 PCE ↑**	**Q95 Excess ↑**
7	19.5428	19.5259	19.0599	0.7678
8	19.5265	19.4438	19.0718	0.7797
9	19.5952	19.5922	19.0524	0.7606
Mean ± Std.	19.5548 ± 0.0359	19.5206 ± 0.0743	19.0614 ± 0.0098	0.7694 ± 0.0096

Mean absolute error (MAE) is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|,$$

and root mean squared error (RMSE) as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}.$$

We additionally report Spearman rank correlation,

$$\rho = \mathrm{corr}\left(\mathrm{rank}(y), \mathrm{rank}(\hat{y})\right),$$

which measures agreement between predicted and experimental molecular ordering and is particularly relevant to selection-based molecular search.

Overall metrics are computed separately on each held-out cross-validation fold and reported as mean ± standard deviation across five folds. To evaluate the discovery-relevant high-performance regime, we aggregate the out-of-fold predictions from all five test partitions and compute the same metrics on molecules whose experimental PCE lies within the Top-20% of the experimental distribution. We additionally report Top-5% results in Appendix D.2 to characterize the extreme high-PCE tail.

Best PCE and Top-10 PCE. Let $\mathcal{C}$ denote all valid candidates generated by a search method, and for molecular-search metrics we use $F(x) \equiv F_{\mathrm{ens}}(x)$ for brevity. We define

$$\text{Best PCE} = \max_{x\in\mathcal{C}} F(x).$$

Let $x_{(k)}$ denote the k-th highest-scoring candidate in $\mathcal{C}$ under F. We define Top-10 PCE as the score of the 10th-ranked candidate:

$$\text{Top-10 PCE} = F\left(x_{(10)}\right).$$

Best PCE measures the strongest individual candidate, whereas Top-10 PCE characterizes the lower boundary of the elite Top-10 set: all ten highest-ranked candidates achieve a predicted PCE no lower than this value. This makes the metric less dependent on a single isolated optimum.

AUC Top-10. To quantify search efficiency throughout evolution rather than only at the final generation, we track the cumulative Top-10 PCE after each generation. Let $\mathcal{C}_{1:t}$ denote all candidates submitted through generation t, and let $x^{(t)}_{(10)}$ denote the 10th highest-scoring candidate in $\mathcal{C}_{1:t}$. We define

$$S_t = F\left(x^{(t)}_{(10)}\right).$$

For a search budget of T generations, AUC Top-10 is computed as

$$\text{AUC Top-10} = \frac{1}{T-1}\left[\frac{S_1 + S_T}{2} + \sum_{t=2}^{T-1} S_t\right],$$

Table 7: **PCE prediction performance on the experimental high-PCE Top-5% subset.** Results are computed from aggregated out-of-fold predictions over 76 samples with experimental PCE $\geq$ 18.366%.

Method	R^2 ↑	MAE↓	RMSE↓	Spearman ρ ↑
Traditional ML Baselines				
RF	-53.394	2.938	3.417	-0.027
XGBoost	-45.979	2.713	3.176	0.057
LightGBM	-48.123	2.675	3.248	0.080
Deep-Learning Baselines				
GCN	-49.024	3.124	3.277	0.044
GAT	-43.263	2.841	3.083	-0.068
Chemprop	-44.900	2.836	3.139	-0.081
MolCLR	-44.677	2.900	3.132	-0.121
DeepAcceptor	-110.412	4.867	4.891	-0.129
P^3	-32.118	2.472	2.667	0.055
Physics-Based Baseline				
Original Scharber	-70.325	3.878	3.913	-0.158
Ours				
Data-Driven Model	-27.879	2.231	2.490	0.059
Physics Model	-8.195	1.059	1.405	**0.378**
Aggressive Tail Fusion	**-3.262**	**0.686**	**0.957**	0.329

corresponding to the normalized trapezoidal area under the cumulative rank-10 trajectory. All iterative methods are compared over the same 40-generation budget, so larger values indicate that a strong Top-10 candidate set is established earlier and maintained throughout optimization.

Q95 Excess. Best PCE and Top-10 PCE characterize the highest-ranked candidates. To measure the depth of the discovered high-performance portfolio, we introduce Q95 Excess. Let

$$q_{95} = Q_{0.95}(\mathcal{D}_{\text{exp}}) = 18.2921\%$$

denote the empirical 95th percentile of the experimental PCE distribution, where $Q_{0.95}$ is computed using linear percentile interpolation. Let $\mathcal{P}_{200}$ denote the retained Top-200 generated candidates. We define

$$\text{Q95 Excess} = \frac{1}{200} \sum_{x \in \mathcal{P}_{200}} \max(F(x) - q_{95},\, 0)\,.$$

Candidates below the experimental 95th-percentile reference contribute zero, whereas candidates above it contribute in proportion to their predicted excess PCE. Q95 Excess therefore measures both the prevalence and the magnitude of candidates extending beyond the experimentally observed high-performance boundary.

D.2 HIGH-PCE TOP-5% EVALUATION

The experimental Top-5% subset contains 76 samples, with an exact lower-bound PCE of 18.366%.

The performance gap becomes substantially larger when evaluation is restricted to this extreme tail. Data-driven predictors systematically underestimate the highest-PCE samples: P^3 and our Data-Driven Model obtain MAEs of 2.472 and 2.231, respectively. The calibrated Physics Model reduces MAE to 1.059, demonstrating the value of explicit physical structure in this sparsely supported regime. Aggressive Tail Fusion further lowers MAE to 0.686 and RMSE to 0.957, while achieving the strongest R^2 among all compared methods. This corresponds to a 69.3% reduction in MAE relative to the Data-Driven Model and a 72.3% reduction relative to P^3. The Physics Model obtains the highest Spearman correlation, while Fusion retains competitive ranking performance with substantially better absolute accuracy, highlighting the complementary roles of the two branches.

Because the Top-5% subset has limited target variance, R^2 can remain negative despite substantially reduced prediction errors. We therefore assess tail performance jointly using MAE, RMSE, and Spearman correlation.

E CHEMICAL INTERPRETABILITY OF GENERATED MOLECULES

Structural characteristics. As shown in Figure 7, these candidates remain strongly concentrated within the Y6 design space. The most frequent core and end-group motifs largely preserve its characteristic structural patterns. Fragment usage is also highly concentrated: 6 of 23 core types and 4 of 29 end-group types account for about half of all occurrences, while 9 of 91 side-chain types account for half of the side-chain occurrences. Overall, the predicted high-performance structures occupy a relatively concentrated region of chemical space and are mainly obtained through local structural optimization rather than large changes in molecular architecture.

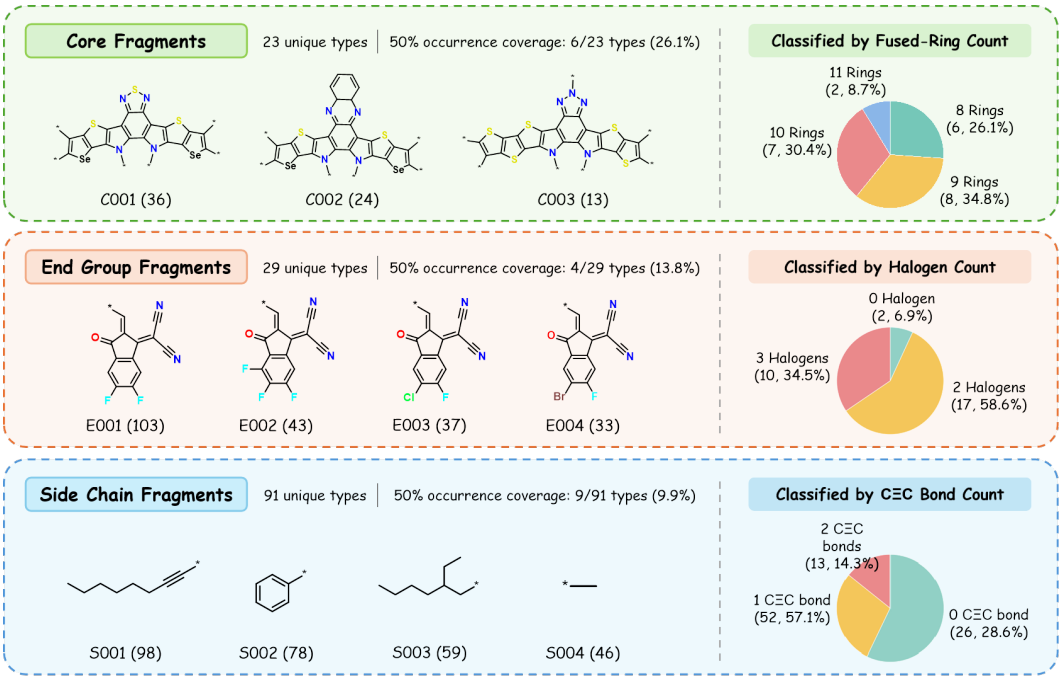

Figure 7: **Chemical characteristics of high-scoring molecules generated by LLM-guided evolution.** We summarize the most frequent core, end-group, and side-chain fragments among the retained candidates, together with structural statistics including fused-ring, halogen, and C=C bond counts. The generated molecules remain concentrated around characteristic Y6-type motifs, with most diversity arising from localized structural modifications.

Synthetic accessibility. As an additional characterization of the generated molecules, we compute the RDKit synthetic accessibility (SA) score (Ertl & Schuffenhauer, 2009) for the retained high-scoring candidates, where lower values indicate easier estimated synthesis. The average SA scores of our LLM-guided evolutionary search are 7.6156, 7.0702, and 7.2378 for DeepSeek-V4.1-Flash, GPT-5.6 Sol, and Claude Opus 5, respectively. GPT-5.6 Sol yields the lowest average SA score among the three backbones, while Claude Opus 5 and DeepSeek-V4.1-Flash obtain slightly higher values. These results provide a complementary view of the synthetic complexity of the generated high-PCE candidates.