Pareto-Guided Reinforcement Learning for Multi-Objective ADMET Optimization in Generative Drug Design

Hoang-My Nguyen
University College Cork
Cork, Ireland
mnguyen@ucc.ie

Nguyet-Hang Vu
ISY Labs
Hanoi, Vietnam
hang.vu@reliable-ai.org

Hoang Thanh Lam
IBM Research
Dublin, Ireland
t.l.hoang@ie.ibm.com

Hoang D. Nguyen*
University College Cork
Cork, Ireland
hn@cs.ucc.ie

Abstract

Multi-objective optimization is fundamental to early drug discovery, where improving one ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) property often degrades others. Existing generative approaches commonly rely on scalarized rewards or descriptor-based objectives, limiting their ability to capture complex pharmacokinetic trade-offs. We present RL-Pareto, a Pareto-guided reinforcement learning framework that directly optimizes predictor-driven ADMET objectives using a transformer-based SELFIES generator and a panel of LightGBM models. A compact reference Pareto set provides a dominance-based reward signal that preserves the structure of trade-offs while encouraging broad exploration. The framework scales flexibly to 1–22 simultaneous objectives without retraining and includes a natural-language interface that enables users to specify goals in plain text. In a benchmark involving simultaneous optimization of solubility and toxicity, RL-Pareto outperforms five strong baselines, PMMG, REINVENT, DrugEx-PCD, DrugEx-PTD, and GMD-MO-LSO, achieving 100% validity and novelty, strong diversity, and the highest hypervolume, reflecting the broadest Pareto-front expansion. RL-Pareto also reaches the best solubility and lowest toxicity extremes. These results highlight RL-Pareto with predictor-driven feedback as a principled, scalable, and practical approach for multi-objective molecular design.

1 Introduction

Designing effective drug molecules is a costly and complex process, often spanning several decades and requiring billions of dollars in investment (1). A key challenge arises because a candidate drug must not only demonstrate potency against its intended target but also simultaneously satisfy multiple pharmacokinetic and toxicity constraints, collectively known as ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity). These properties critically influence whether a compound can progress from a lead molecule to a viable therapeutic (1; 2).

Traditionally, adverse ADMET findings were frequently made at later stages of drug development, leading to costly project terminations or restarts, thereby imposing an unacceptable burden on

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Science.

pharmaceutical research and development budgets (2). This challenge has led to a widespread recognition of the critical need to consider ADMET and toxicity as early as possible in the drug discovery pipeline. Optimizing ADMET endpoints is inherently difficult due to their conflicting nature (1). For example, improving cell permeability may reduce solubility or increase toxicity. This interdependence transforms drug design into a multi-objective optimization (MOO) problem, where balancing trade-offs is essential (1; 2). The complexity of this task escalates rapidly with the number of objectives being considered (3). Recent advances in artificial intelligence (AI), including deep learning and large language models (LLMs), have unlocked new opportunities for de novo molecular design. However, existing generative frameworks still struggle to capture the true complexity of multi-objective optimization in drug discovery. Many of them rely on molecular descriptors that often fail to capture pharmacologically meaningful endpoints. And when it comes to multi-objective optimization, the prevalent use of reward linear scalarization can lead to the ignorance of trade-offs between target objectives.

1.1 Related Work

Despite notable progress, current approaches face three major limitations.

1.1.1 Reliance on molecular descriptors

Many frameworks (4; 5; 6) optimize molecules through molecular descriptors, such as logP, QED, TPSA, molecular weight, hydrogen bond donors and acceptors, or synthetic accessibility, typically aligned with Lipinski's rules (7). While computationally convenient and useful for benchmarking, these descriptors are proxies and fail to capture pharmacological relevance (1). Gao et. al. (8) implicitly categorized these as less complex or "trivial" compared to actual bioactivity predictions. In contrast, predictor-driven objectives (e.g., LightGBMs), which are built from machine learning models trained on ADMET datasets, provide direct pharmacological feedback (9). These models have demonstrated their reliability as surrogates for experimental endpoints (2; 9).

1.1.2 Reward linear scalarization

Many existing multi-objective generative frameworks face the limitation of using reward linear scalarization, where objective functions are aggregated into a single scalar objective using predefined scalarization functions such as weighted arithmetic means, geometric means, or Chebychev scalarization. Examples of such aggregations include the weighted sum (WS) and weighted product (WP) schemes (1; 3). This approach hides inherent trade-offs between properties, biases exploration toward predefined directions, and limits the diversity of generated solutions (1; 3). Although multiple scalarizations using different weight vectors can be applied to approximate different regions of the Pareto front, such approaches still require many independent optimization runs and remain sensitive to the coverage of chosen weight combinations (10). Pareto optimization offers a principled alternative by maintaining sets of non-dominated solutions that explicitly represent trade-offs across objectives, enabling broader exploration of chemical space (11). By using Pareto ranking, it is possible to explore chemical space while preserving diversity and trade-off structure, rather than enforcing arbitrary weightings. Recent work confirms that Pareto-based methods outperform scalarized approaches in generating balanced multi-objective molecules (12; 3; 13; 14).

1.1.3 Limited scalability of Pareto-RL frameworks

Some frameworks combine reinforcement learning (RL) with Pareto ranking, such as (15), which optimized three objectives (A1AR activity, A2AAR activity, and reduced hERG binding). While these methods demonstrate the potential of Pareto-RL, they are usually restricted to a small, fixed set of objectives and do not allow flexible expansion. More recent methods such as CPRL (12) also focus on a small group of properties and have not scaled to broad ADMET endpoints. This lack of scalability limits their adoption in real-world drug discovery, where different projects may require very different objective sets.

1.2 Our framework

To overcome these limitations, we propose a Pareto-guided reinforcement learning framework for multi-objective drug design. Our approach directly optimizes predictor-driven ADMET endpoints,

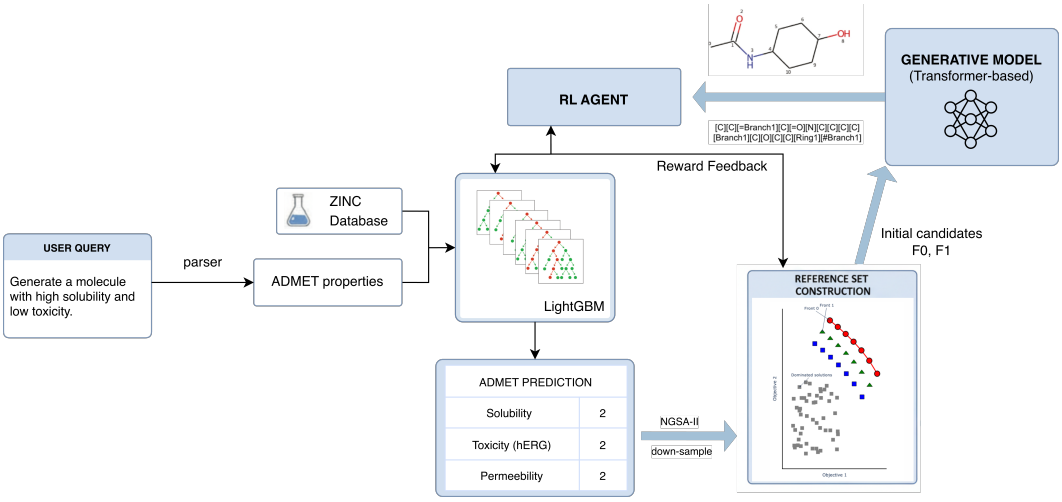

Figure 1: Workflow of the proposed framework. User queries specifying desired molecular properties are parsed into ADMET objectives. Predictor models trained with LightGBM provide real-time property feedback, which is integrated with Pareto-based reward signals derived from a reference set constructed via NSGA-II. A transformer-based generative model proposes candidate molecules, while the RL agent iteratively refines generation to balance trade-offs between ADMET properties.

scoring each molecule with a vector of trained property predictors. Rewards are derived from Pareto dominance with respect to a reference set, ensuring that trade-offs are preserved while maintaining diversity across solutions. In order to guide the search more effectively and more efficiently, our system builds indices of molecules from well-known databases such as ZinC or PubChem and the search starts from promissing candidates selected from the databases. The framework is scalable and flexible, with modular predictor models that can be easily extended to new ADMET endpoints without retraining. Our current implementation supports simultaneous optimization of 1–22 user-selected objectives. Furthermore, we integrate a natural language interface powered by large language models (LLMs). Users can specify objectives in plain language (e.g., “generate molecules that have high solubility and low toxicity and binding to a given target"). This makes the system more accessible, lowering the barrier for non-experts to interact with molecular design models.

2 Method

Our framework operates as a guided search model that integrates: (i) a panel of endpoint-specific ADMET predictors providing real-time feedback on candidate molecules, and (ii) a transformer-based autoregressive generative model employing SELFIES (16) as the molecular sequence representation. Within this feedback loop, guidance from the predictors continuously informs and refines the generative process, enabling the simultaneous pursuit of novelty, drug-likeness, and favorable ADMET profiles. In this section, we thoroughly examine the methodology components, spanning from the dataset to the generative model and the elements of the reinforcement learning paradigm.

2.1 Generative model

The ZINC database (17) is a collection of commercially available chemical compounds prepared for virtual screening, making it widely used for de novo molecule generation. From this resource, 35 million molecules were sampled and represented as SELFIES strings to train our generative model. A GPT-2 (18) architecture was adopted as the backbone, and training followed a masked autoregressive strategy: 15% of the tokens in each SELFIES string were randomly masked, and the model was trained to reconstruct the original sequence by continuing the decoding process. Training was done using 2 A100 GPUs, with minibatch size 128 per GPU and learning rate set as 5e-5. Training was done with 140000 steps and the checkpoints are available in HuggingFace as an open model[2].

[2] `https://huggingface.co/lamthuy/SelfiesGen`

2.2 ADMET predictors

ADMET endpoint predictors were trained on 22 datasets from the Therapeutics Data Commons (TDC) (19), covering a broad range of absorption, distribution, metabolism, excretion, and toxicity tasks. Molecules were encoded using a combination of Morgan, Avalon, and ErG fingerprints to capture complementary structural features. LightGBM (20) models were employed for supervised learning, with model performance estimated through 5-fold scaffold-based cross-validation resulting in the best performance reported in the corresponding TDC ADMET leaderboards (19). We use the raw predictor outputs without applying any normalization, rescaling, or additional transformations, and the Pareto-ranking reward operates directly on these unmodified property values.

2.3 Reinforcement Learning Setup

The `trl` (21) library (Transformers Reinforcement Learning) supports advanced fine-tuning strategies for large language models, including a technique called Guided Reward Policy Optimization (GRPO) (22). GRPO is designed to balance the stability of supervised fine-tuning with the flexibility of reinforcement learning by guiding the policy updates using a reward signal while still anchoring the model to its original supervised behavior. This is achieved by interpolating between the policy gradient objective and the supervised loss, allowing the model to improve with respect to a reward function (e.g., human preferences or task-specific scores) without drifting too far from its initial behavior. GRPO is particularly useful in scenarios where reward optimization alone may lead to instability or undesirable outputs, and it provides a more controllable framework for aligning language models with task goals or ethical constraints.

We use the `trl` library for our purpose. Given a query, GRPO trains the generative model to generate corresponding molecules that optimize the set of objectives specified in the queries.

Let x denote a candidate molecule. A panel of d learned ADMET predictors produces a score vector

$$\mathbf{s}(x) = \big(s_1(x), \ldots, s_d(x)\big) \in [0,1]^d,$$

where larger values are uniformly interpreted as better.

Reference set To provide stable rewards for evaluation, we construct a compact *reference Pareto set* $\mathcal{R} = \{\tilde{\mathbf{s}}^{(i)}\}_{i=1}^{N}$. This set is obtained in three steps: First, we begin with a large archive of molecules (2 millions samples from the ZINC database) and compute their predictor score vectors. The full pool of score vectors is partitioned into Pareto fronts using standard non-dominated sorting (23), producing a global Pareto frontier. Since using the full frontier is computationally prohibitive and may overweight densely clustered regions, we down-sample it to a fixed budget $N = 3000$. This is done by stratified sampling across fronts and within-front clustering: dense regions are thinned, while diverse representatives are retained. The result is a balanced set $\mathcal{R}$ that preserves both the *shape* of the global frontier and the *diversity* of trade-offs across objectives.

The design goal is then to discover candidate molecules whose score vectors are Pareto-competitive with respect to this reference set $\mathcal{R}$ while maintaining chemical diversity in the generated set.

We use standard Pareto dominance to compare vectors. For two vectors $u, v \in \mathbb{R}^d$ we write $u \succeq v$ iff $u_i \geq v_i$ for all i and $u \succ v$ if $u \succeq v$ and $u \neq v$. The reference set $\mathcal{R}$ is partitioned by non-dominated sorting into ordered fronts $F_0, F_1, \ldots$ where F_0 is the (reference) Pareto front. Given a candidate score vector $\mathbf{s}$ we define its *reference-front index*

$$m(\mathbf{s}) \;=\; \min\{\, f+1 \;:\; \exists\, u \in F_f \text{ with } u \preceq \mathbf{s} \,\}.$$

and adopt the convention that $m(\mathbf{s}) = 0$ if no reference vector dominates $\mathbf{s}$ (i.e., the candidate is non-dominated w.r.t. the reference). The index m is a discrete ordinal indicator of candidate quality relative to $\mathcal{R}$.

To accelerate convergence and anchor the initial exploration phase in high-quality chemical space, we employ a warm-start strategy for the RL fine-tuning. Instead of beginning with random or uninformative prompts, the initial generation steps are seeded with molecules sampled from the top tiers of our pre-computed reference set $\mathcal{R}$. Specifically, we use molecules from the first two Pareto fronts (F_0 and F_1) as starting points for the generator.

This initialization helps the learning process in two significant ways. First, it mitigates the cold-start problem, ensuring the model receives a strong, positive reward signal from the very first training

batches, which leads to more stable and meaningful policy updates early on. Second, it focuses the search on promising regions of the chemical space rather than exploring vast, low-reward regions from scratch.

Reward assignment To interface with policy-gradient methods we convert the discrete Pareto index m into a scalar base reward $r_{\text{base}}(m)$. The mapping used in experiments is intentionally coarse for top ranks and exponentially decaying for distant ranks:

$$r_{\text{base}}(m) = \begin{cases} 1.0, & m = 0, \\ 0.8, & m = 1, \\ 0.6, & m = 2, \\ 0.5, & m = 3, \\ \max\big(0.05,\ 0.5\,\exp\big[-\gamma(m-3)\big]\big), & m \geq 4, \end{cases} \tag{1}$$

where $\gamma > 0$ is a small decay constant $\gamma = 0.02$. This construction preserves ordinal Pareto information (priority to non-dominated and near-front candidates) while ensuring a non-vanishing yet diminishing signal for exploration.

Diversity filter The Pareto index provides the main quality signal, but without explicit control the generator may collapse onto a few scaffold families. To address this, we introduce a *scaffold-bin penalty* that limits over-exploitation of common chemotypes.

For each candidate molecule x, we extract its Bemis–Murcko scaffold (24) and a Morgan fingerprint. We then define occ as the number of molecules already in the archive that share the same scaffold and exceed a Tanimoto similarity threshold τ (we used $\tau = 0.5$). This gives a local occupancy measure of how overrepresented a scaffold class is.

If a molecule achieves a sufficiently high base reward $r_{\text{base}}(x)$ (e.g., Pareto-based score), we apply a linear penalty based on occ relative to a tolerance T and a hard buffer B. The final reward $R(x)$ is then:

$$R(x) = r_{\text{base}}(x) \cdot \begin{cases} 1, & \mathrm{occ} \leq T, \\ \dfrac{B - \mathrm{occ}}{B - T}, & T < \mathrm{occ} < B, \\ 0, & \mathrm{occ} \geq B, \end{cases} \tag{2}$$

where molecules below the minimum score threshold are not penalized.

3 Experiments

We present a case study demonstrating the performance of our framework in the ADMET and physicochemical property optimization scenario. The generative model is tasked with creating novel molecules that simultaneously exhibit **high aqueous solubility** and **low toxicity**. This task was chosen as it represents a common and challenging trade-off in drug development.

Evaluation Metrics We evaluated our framework RL-Pareto on a two-objective ADMET task designed to maximize predicted solubility while minimizing predicted toxicity. To provide a comprehensive view, we report both goal-directed optimization metrics, including success rate, hypervolume, and the extrema and means of each objective, as well as distribution-learning metrics such as validity, uniqueness, novelty, QED (25), and SAS (26). These distribution metrics follow the GuacaMol benchmark definitions (27). For comparison, we included four baselines:

- PMMG (3): Pareto-based Multi-objective Molecule Generation, a specialized framework designed explicitly for multi-objective optimization. As it also leverages Pareto principles, it serves as a strong baseline to determine the advantages of our specific reward and guidance mechanisms.

- SELFIES-REINVENT (28): The original REINVENT framework adopts a policy-based reinforcement learning (RL) approach to tune Recurrent Neural Networks (RNNs) for generating molecule strings, initially designed to operate on SMILES representations. For a fair comparison with our SELFIES-based generator, we utilize an adapted version that generates SELFIES strings. The implementation for this baseline, hereafter referred to as SELFIES-REINVENT, was sourced from the official code provided in the PMO benchmark paper. There is a newer version of REINVENT, which implements Pareto in its method (10). Although it interpolates across multiple weight combinations, it ultimately remains a weighted-sum strategy where scalarized objectives are optimized independently for each weight vector, and Pareto analysis is applied only as a post hoc filtering step. As the number of objectives increases, the number of weight combinations required to adequately explore the objective space grows. Even with subsampling strategies such as clustering, the method still requires running many independent RL trajectories, and omitting most combinations risks poor coverage of the chemical space. In contrast, our framework integrates Pareto dominance directly into the reward function, enabling a single RL run to explore trade-offs without relying on predefined weight vectors. For these reasons, we do not include REINVENT_MOO as a baseline in this study.
- DrugEx (15): a SMILES-based recurrent neural network (RNN) generative model trained under a Pareto-based multi-objective reinforcement learning framework. During training, scores from multiple predictive models are used to construct Pareto fronts via non-dominated sorting, and these ranks define the rewards that guide optimization. DrugEx incorporates mutation and crossover operators inspired by evolutionary algorithms to enhance exploration while steering the generator toward molecules that satisfy multiple pharmacological objectives simultaneously. Within the Pareto-based scoring scheme, DrugEx offers two complementary selection strategies—Pareto Crowding Distance (PCD) and Pareto Tanimoto Distance (PTD), which emphasize exploitation and exploration, respectively. In our benchmarking, we include both DrugEx_PCD and DrugEx_PTD configurations to provide a representative assessment of their ability to generate Pareto-efficient ADMET-optimized molecules.
- GMD-MO-LSO (29): a multi-objective latent-space optimization (MO-LSO) framework that improves deep generative models such as Junction-Tree Variational Autoencoder (JT-VAE) through iterative weighted retraining. At each iteration, molecules in the training set are ranked using non-dominated sorting, and higher-ranked molecules receive larger weights, biasing the latent space toward favorable multi-objective trade-offs. The retrained model then explores the latent space, via random sampling or Bayesian optimization, to propose improved candidates, which are incorporated back into the training set to progressively expand the Pareto front. This process enables MO-LSO to balance multiple objectives without relying on ad-hoc scalarization, making it an effective baseline for multi-objective molecular generation.

At each baseline, we generated 1000 molecules to ensure comparability across methods. For our method, the user query was explicitly phrased as "Generate 1000 molecules that have high solubility and low toxicity," which guided the predictor-driven reinforcement learning framework during molecule generation. Success rate was calculated under the condition of solubility > -2 and toxicity < 3, which is the condition of a molecule to be considered as soluble and low toxicity (30; 31). Solubility is reported as LogS, where S is solubility in mol/L (30) and LogS is dimensionless. Toxicity refers to the LD50 predictor trained on the Zhu et al. dataset (31), where raw LD50 values (mol/kg) are transformed into log(1/(LD50)), which is dimensionless.

4 Results and Discussion

Table 1 provides a comparison of our model (RL-Pareto) against five strong baselines: PMMG, REINVENT, DrugEx-PCD, DrugEx-PTD, and GMD-MO-LSO. Across all distribution-learning metrics, RL-Pareto achieves perfect validity (100%) and perfect novelty (100%), demonstrating that the optimization process preserves structural correctness and does not rely on memorizing training molecules. Our model achieves 67.13% uniqueness, which is lower than PMMG and the DrugEx variants but still substantially higher than REINVENT (40.85%). This aligns with the GuacaMol benchmark (27) definition of uniqueness as a measure of diversity across chemical space, confirming

	PMMG	REINVENT	DrugEx PCD	DrugEx PTD	GMD-MO LSO	Our model
No. molecules	1000	1000	1000	1000	1000	1000
QED ↑	0.43 ± 0.00	0.40 ± 0.04	0.44 ± 0.01	0.45 ± 0.03	$\mathbf{0.77} \pm 0.00$	0.40 ± 0.02
SAS ↓	5.58 ± 0.00	4.34 ± 0.16	2.50 ± 0.06	$\mathbf{2.44} \pm 0.08$	3.25 ± 0.01	5.18 ± 0.69
Validity (%) ↑	$\mathbf{100.00} \pm 0.00$	$\mathbf{100.00} \pm 0.00$	$\mathbf{100.00} \pm 0.00$	$\mathbf{100.00} \pm 0.00$	$\mathbf{100.00} \pm 0.00$	$\mathbf{100.00} \pm 0.00$
Uniqueness (%) ↑	90.40 ± 0.00	40.85 ± 7.28	$\mathbf{100.00} \pm 0.00$	$\mathbf{100.00} \pm 0.00$	$\mathbf{100.00} \pm 0.00$	67.13 ± 7.35
Novelty (%) ↑	$\mathbf{100.00} \pm 0.00$	$\mathbf{100.00} \pm 0.00$	99.93 ± 0.06	99.90 ± 0.00	99.97 ± 0.06	$\mathbf{100.00} \pm 0.00$
Diversity ↑	0.85 ± 0.00	0.74 ± 0.07	0.83 ± 0.02	0.82 ± 0.01	0.86 ± 0.00	$\mathbf{0.87} \pm 0.01$
Hypervolume ↑	30.17 ± 0.00	32.36 ± 0.09	33.25 ± 0.51	32.89 ± 0.22	23.85 ± 1.07	$\mathbf{35.25} \pm 0.09$
Success Rate (%) ↑	$\mathbf{100.00} \pm 0.00$	99.90 ± 0.14	78.33 ± 5.31	77.37 ± 14.03	9.90 ± 0.82	99.53 ± 0.47
Mean Sol. ↑	0.22 ± 0.00	0.65 ± 0.12	-0.75 ± 0.33	-0.93 ± 0.77	-3.11 ± 0.02	$\mathbf{1.17} \pm 0.14$
Max Sol. ↑	1.57 ± 0.00	1.53 ± 0.01	1.60 ± 0.10	1.49 ± 0.01	-0.06 ± 0.31	$\mathbf{2.12} \pm 0.12$
Mean Tox. ↓	2.19 ± 0.00	$\mathbf{1.75} \pm 0.07$	1.80 ± 0.07	1.82 ± 0.04	2.69 ± 0.01	1.77 ± 0.08
Min Tox. ↓	1.68 ± 0.00	1.47 ± 0.01	1.39 ± 0.02	1.40 ± 0.02	1.86 ± 0.09	$\mathbf{1.34} \pm 0.04$

Table 1: Benchmarking results for our model and baselines (PMMG, REINVENT). Metrics include both distribution-learning benchmarks (QED, SAS, validity, uniqueness, novelty) and goal-directed optimization benchmarks (success rate, hypervolume, solubility, toxicity). Values are reported as mean ± standard deviation, rounded according to IUPAC guideline. Solubility is reported as LogS, where S is solubility in mol/L (30) and LogS is dimensionless. Toxicity refers to the LD50 predictor trained on the Zhu et al. dataset (31), where raw LD50 values (mol/kg) are transformed into log(1/(LD50)), which is dimensionless.

that RL-Pareto avoids mode collapse and consistently generates diverse chemical scaffolds. Overall, these results highlight a balanced exploration strategy that prioritizes Pareto-front expansion while still producing diverse scaffolds.

On the solubility–toxicity optimization task, our model achieves a 99.53 percent success rate, which is the second highest, slightly below PMMG's perfect score (100 percent) but higher than all other baselines. More importantly, our method obtains a hypervolume score of 35.25, the highest among all models, outperforming PMMG (30.17), REINVENT (32.36), DrugEx-PCD (33.25), DrugEx-PTD (32.89), and GMD-MO-LSO (23.85). Hypervolume is a standard metric quantifying the dominated region of the objective space (11), so higher values indicate a Pareto set that spans a broader range of high-solubility and low-toxicity trade-offs. The superior hypervolume achieved by RL-Pareto reflects not only strong extreme values but also a well-distributed Pareto front rather than collapsing into narrow regions.

Figure 2 illustrates these trends. Our model (brown) generates a wide Pareto front extending toward high solubility and low toxicity. In contrast, PMMG (red) and REINVENT (purple) form clusters concentrated near solubility approximately 0 and toxicity above 2, suggesting limited movement toward optimal trade-off regions. DrugEx-PCD and DrugEx-PTD (blue and orange) successfully cover lower-toxicity regions but fail to reach higher solubility. GMD-MO-LSO (green) collapses almost entirely into low-solubility, high-toxicity space, consistent with its poor hypervolume reported in Table 1.

QED quantifies how closely a molecule matches the physicochemical profiles of approved oral drugs (25), while SAS estimates synthetic feasibility based on fragment frequency and structural complexity, with lower scores indicating easier synthesis (26). These metrics are widely used measures for assessing drug-likeness and synthetic accessibility. Although neither metric was included as an optimization target in our model, the generated molecules still achieve a QED of 0.40 (comparable to PMMG and REINVENT) and an SAS of 5.18, which falls within a reasonable synthesizability range relative to the baselines. These values show that even without explicit constraints, our model naturally proposes molecules that remain both drug-like and synthetically tractable, making them suitable candidates for downstream medicinal chemistry.

Solubility and toxicity statistics further highlight the strength of our approach. Our model attains the highest mean solubility (1.17) and highest maximum solubility (2.12) among all methods, showing its ability to reach high-solubility regions of chemical space. Molecules generated from our model also achieve the lowest minimum toxicity (1.34) across the baselines, lower than PMMG (1.68), REINVENT (1.47), DrugEx-PCD (1.39), DrugEx-PTD (1.40), and GMD-MO-LSO (1.86). These results confirm that the Pareto-guided reward successfully drives the model toward low-toxicity extremes while maintaining chemically meaningful structures.

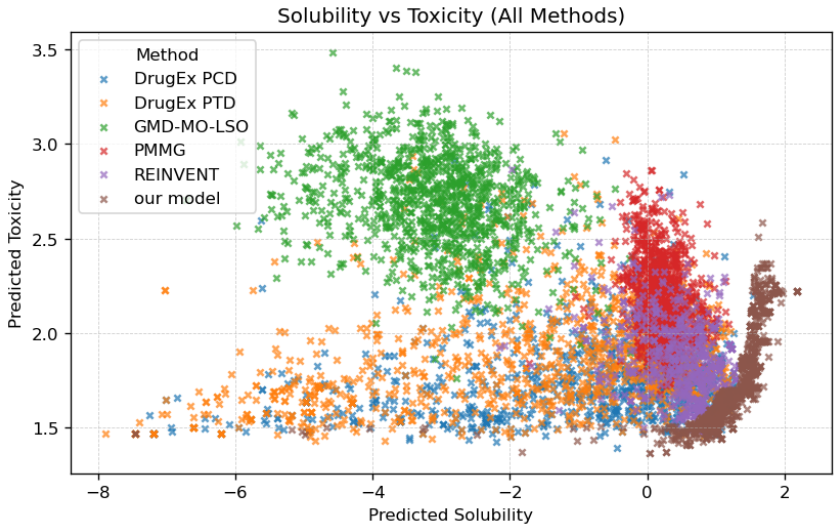

Figure 2: Scatter plot of predicted solubility versus predicted toxicity for all methods. Each point represents a generated molecule (1000 per method). Our model (brown) produces a broad Pareto front spanning higher solubility and lower toxicity regions, while PMMG (red) and REINVENT (purple) concentrate near a solubility of 0 and higher toxicity. DrugEx (blue and orange)'s molecules cover the acceptable toxicity region but have lower solubility, and GMD-MO-LSO (green) collapses into low-solubility, high-toxicity space.

The robustness of our model (RL-Pareto) stems from two design choices. First, the use of predictor-driven objectives ensures direct feedback on pharmacologically meaningful endpoints, unlike descriptor-based proxies. Second, the Pareto-guided reward mechanism, which integrates dominance ranking, preserves the natural structure of trade-offs and avoids the collapse observed in scalarized reward schemes. Together, these features allow the model to scale across multiple objectives while maintaining both effectiveness and diversity.

In summary, RL-Pareto achieved high performance in balancing solubility and toxicity and showed promise for multi-objective drug molecule optimization. It combined a high success rate, high diversity, and the best hypervolume with complete validity and novelty, drug-like QED and SAS scores, and a diverse Pareto front. PMMG produced useful trade-offs but lacked wide front coverage, REINVENT collapsed under the scalarized setup, while GMD-MO-LSO collapsed into a low-solubility and high-toxicity region, and the DrugEx variants achieved reasonable toxicity levels but failed to explore the high-solubility region, resulting in limited Pareto-front expansion. These findings support the central claim that our model, which uses Pareto-guided reinforcement learning with predictor-driven ADMET feedback, explores the chemical space more effectively than other baselines and shows strong potential as a robust framework for multi-objective optimization in drug design.

5 Conclusion

This work introduced RL-Pareto, a Pareto-guided reinforcement learning framework for multi-objective molecular generation with direct ADMET optimization. By combining predictor-driven property feedback with a dominance-based reward scheme, the framework preserves natural trade-offs between objectives and enables broad exploration of chemical space. In benchmarking against five strong baselines, RL-Pareto achieved complete validity and novelty, competitive uniqueness, and the highest hypervolume, indicating a well-distributed Pareto front. It also reached the highest solubility values and the lowest toxicity minima, demonstrating strong capability in navigating conflicting pharmacokinetic objectives.

Our results reinforce the growing consensus that Pareto-guided reinforcement learning offers a robust strategy for capturing trade-offs in molecular optimization. The modularity of our predictor panel and the ability to scale to up to, and potentially beyond, twenty-two objectives without retraining

further enhance the practical utility of the framework. The natural language interface additionally lowers the barrier for defining complex objective combinations, making the system more accessible for real-world drug discovery workflows.

Overall, RL-Pareto provides an effective and extensible solution for multi-objective de novo design and represents a promising direction for incorporating robust Pareto optimization into LLM-driven molecular generation pipelines. Future work may extend the framework toward larger objective spaces, incorporate structure-based or physics-informed predictors, and explore integration with closed-loop experimental validation.

6 Acknowledgment

This publication has emanated from research supported in part by a grant from Research Ireland under Grant number [12-RC-2289-P2] which is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] S. Luukkonen, H. W. van den Maagdenberg, M. T. Emmerich, and G. J. van Westen, "Artificial intelligence in multi-objective drug design," *Current Opinion in Structural Biology*, vol. 79, p. 102537, 2023.

[2] H. van de Waterbeemd and E. Gifford, "Admet in silico modelling: towards prediction paradise?," *Nature Reviews Drug Discovery*, vol. 2, no. 3, p. 192–204, 2003.

[3] Y. Liu, Y. Zhu, J. Wang, R. Hu, C. Shen, W. Qu, G. Wang, Q. Su, Y. Zhu, Y. Kang, P. Pan, C.-Y. Hsieh, and T. Hou, "A multi-objective molecular generation method based on pareto algorithm and monte carlo tree search," *Advanced Science*, vol. 12, no. 20, p. 2410640, 2025.

[4] T. Nguyen and A. Grover, "Lico: Large language models for in-context molecular optimization," 2025.

[5] H. Wang, M. Skreta, C.-T. Ser, W. Gao, L. Kong, F. Strieth-Kalthoff, C. Duan, Y. Zhuang, Y. Yu, Y. Zhu, Y. Du, A. Aspuru-Guzik, K. Neklyudov, and C. Zhang, "Efficient evolutionary search over chemical space with large language models," 2025.

[6] H. Kim, Y. Jang, and S. Ahn, "Mt-mol:multi agent system with tool-based reasoning for molecular optimization," 2025.

[7] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced Drug Delivery Reviews*, vol. 23, no. 1, pp. 3–25, 1997.

[8] W. Gao, T. Fu, J. Sun, and C. Coley, "Sample efficiency matters: A benchmark for practical molecular optimization," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 21342–21357, Curran Associates, Inc., 2022.

[9] X. Li, L. Tang, Z. Li, D. Qiu, Z. Yang, and B. Li, "Prediction of admet properties of anti-breast cancer compounds using three machine learning algorithms," *Molecules*, vol. 28, no. 5, 2023.

[10] L. Landolfi, B. Catalanotti, and J. P. Janet, "Finding balance: Multi-objective optimization in molecular generative modeling," 2025.

[11] J. Blank and K. Deb, "Pymoo: Multi-objective optimization in python," *IEEE Access*, vol. 8, pp. 89497–89509, 2020.

[12] J. Wang and F. Zhu, "Multi-objective molecular generation via clustered pareto-based reinforcement learning," *Neural Networks*, vol. 179, p. 106596, 2024.

[13] T. Suzuki, D. Ma, N. Yasuo, and M. Sekijima, "Mothra: Multiobjective de novo molecular generation using monte carlo tree search," *Journal of Chemical Information and Modeling*, vol. 64, no. 19, pp. 7291–7302, 2024. PMID: 39317969.

[14] Y. Yang, G. Chen, J. Li, J. Li, O. Zhang, X. Zhang, L. Li, J. Hao, E. Wang, and P.-A. Heng, "Enabling target-aware molecule generation to follow multi objectives with pareto mcts," *Communications Biology*, vol. 7, no. 1, 2024.

[15] X. Liu, K. Ye, H. W. T. van Vlijmen, M. T. M. Emmerich, A. P. IJzerman, and G. J. P. van Westen, "Drugex v2: de novo design of drug molecules by pareto-based multi-objective reinforcement learning in polypharmacology," *Journal of Cheminformatics*, vol. 13, no. 1, 2021.

[16] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (selfies): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, vol. 1, no. 4, p. 045024, 2020.

[17] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, and R. A. Sayle, "Zinc20—a free ultralarge-scale chemical database for ligand discovery," *Journal of Chemical Information and Modeling*, vol. 60, no. 12, pp. 6065–6073, 2020. PMID: 33118813.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[19] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik, "Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development," 2021.

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[21] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec, "Trl: Transformer reinforcement learning." `https://github.com/huggingface/trl`, 2020.

[22] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," 2024.

[23] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[24] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. molecular frameworks," *Journal of Medicinal Chemistry*, vol. 39, no. 15, pp. 2887–2893, 1996. PMID: 8709122.

[25] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, "Quantifying the chemical beauty of drugs," *Nature Chemistry*, vol. 4, no. 2, p. 90–98, 2012.

[26] P. Ertl and A. Schuffenhauer, "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions," *Journal of Cheminformatics*, vol. 1, no. 1, 2009.

[27] N. Brown, M. Fiscato, M. H. Segler, and A. C. Vaucher, "Guacamol: Benchmarking models for de novo molecular design," *Journal of Chemical Information and Modeling*, vol. 59, no. 3, pp. 1096–1108, 2019. PMID: 30887799.

[28] M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen, "Molecular de-novo design through deep reinforcement learning," *Journal of Cheminformatics*, vol. 9, no. 1, 2017.

[29] A. N. M. N. Abeer, N. M. Urban, M. R. Weil, F. J. Alexander, and B.-J. Yoon, "Multi-objective latent space optimization of generative molecular design models," *Patterns*, vol. 5, no. 10, p. 101042, 2024.

[30] M. C. Sorkun, A. Khetan, and S. Er, "Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds," *Scientific Data*, vol. 6, no. 1, 2019.

[31] H. Zhu, T. M. Martin, L. Ye, A. Sedykh, D. M. Young, and A. Tropsha, "Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure," *Chemical Research in Toxicology*, vol. 22, no. 12, pp. 1913–1921, 2009. PMID: 19845371.