QUANTAMIND: A LARGE CHEMISTRY LANGUAGE MODEL FOR STRUCTURED MOLECULAR SCREENING AND DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models offer a flexible interface for scientific reasoning, yet general-purpose models remain poorly calibrated for quantitative molecular prediction, especially when targets depend on excited-state electronic structure. We introduce **QuantaMind**, a 32-billion-parameter chemistry language model that maps a SMILES string and a chemistry instruction to a structured, multi-property molecular profile for screening and discovery. QuantaMind is obtained by parameter-efficient instruction tuning of Qwen-2.5-32B on QuantumChem-200K, a supervision corpus of more than 214,000 organic molecules annotated with photophysical, quantum-chemical, safety, accessibility, and physicochemical properties. The model is evaluated on 3,000 unseen molecules from a hold-out testset using both property-prediction and decision-oriented screening metrics. Across the seven shared benchmark properties, QuantaMind achieves an overall weighted mean absolute error (wMAE) of 0.1287, compared with 2.8540 for its untuned Qwen backbone and 0.4581 for a Gemma-3-27B model fine-tuned under the same protocol. In multi-objective screening, it recovers 35 of the ground-truth top 100 candidates, reaches 10.50-fold enrichment, and obtains 3.26% normalized hypervolume regret. Data-scaling, training-duration, backbone-transfer, and failure-mode analyses further characterize where domain adaptation helps and where SMILES-only prediction could be improved. These results position QuantaMind as a chemistry language model for quantitatively grounded molecular screening, with photoinitiator discovery as a demanding case study. The project repository is available at `https://github.com/AnonymousUser-3/QuantaMind`, and the project website is `https://ravenllm.com/`.

1 INTRODUCTION

Large language models (LLMs) have become capable general-purpose interfaces for reasoning, code, and scientific text Yang et al. (2024); Kumar (2024); Patil & Gudivada (2024); Zhao et al. (2023). Chemistry, however, requires more than fluent discussion: useful models must connect molecular representations to numerical properties, respect units and output schemas, and rank candidates under several competing objectives. Off-the-shelf LLMs are not explicitly trained for this form of quantitatively grounded prediction, while task-specific graph models typically expose a narrow regression interface rather than a reusable language-model interface. A central question is therefore whether instruction tuning can turn a general LLM into a reliable chemistry model that predicts multiple properties directly from SMILES and supports downstream molecular selection.

We study this question in photoinitiator discovery, a demanding setting in which the relevant targets span excited-state photophysics and practical synthesis constraints. Photoinitiators absorb light and generate reactive radicals for processes such as direct laser writing (DLW) Ocier et al. (2020); Deubel et al. (2004). Norrish Type I photoinitiators Jagtap & More (2022); Rutsch (1996), often activated near 780 nm, undergo excited-state bond cleavage and enable two-photon polymerization (TPP) for high-resolution fabrication Lee et al. (2006); Nguyen & Narayan (2017); Zhou et al. (2015). Screening such molecules requires simultaneously estimating two-photon absorption (TPA), singlet–triplet intersystem-crossing (ISC) gaps, toxicity, synthetic accessibility, and physicochemical behavior. These labels are expensive to calculate or measure at scale Magnotta & Johnston (1980); Ye et al. (2002), making the task an informative test of whether a chemistry language model

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.

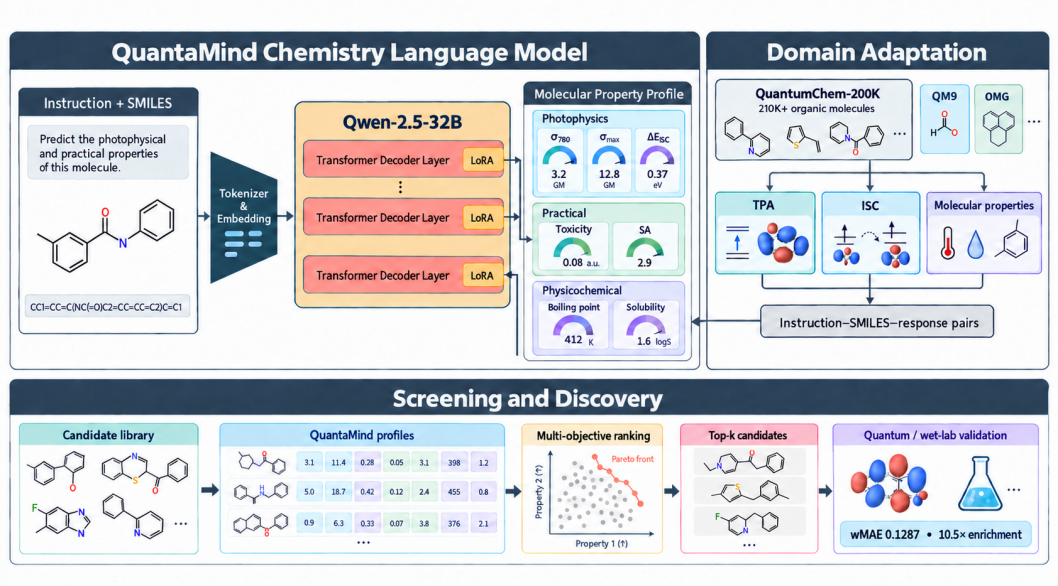


Figure 1: Overview of QuantaMind. QuantumChem-200K supplies instruction–SMILES–response supervision for parameter-efficient adaptation of Qwen-2.5-32B. The resulting chemistry language model produces structured seven-property profiles that support multi-objective ranking, candidate selection, and subsequent quantum-chemical or wet-lab validation.

can learn structure–property relationships beyond common descriptors. Detailed photodissociation mechanisms are provided in Appendix A.3 and Appendix Figure 8.

We introduce **QuantaMind**, a large chemistry language model built by instruction tuning Qwen-2.5-32B on structured molecular supervision. As illustrated in Figure 1, the model receives a chemistry instruction and a SMILES string, then emits a fixed-schema property profile that can be parsed for prediction or used to rank molecular candidates. The primary benchmark evaluates seven nontrivial targets: TPA cross sections at 780 nm and at the spectral maximum, the ISC energy gap, toxicity, synthetic accessibility, boiling point, and solubility. Molecular weight, aromaticity, and logP are retained in the training schema and auxiliary analyses but excluded from the primary benchmark because they are comparatively direct to infer from SMILES.

QuantaMind is trained using QuantumChem-200K, a corpus of more than 214,000 organic molecules whose target properties were produced with a hybrid workflow combining density functional theory, semi-empirical excited-state calculations, atomistic solvers, and cheminformatics predictors. The corpus is an enabling component of the model rather than the endpoint of the study: it supplies the rare TPA and ISC supervision needed to adapt a general text model to quantitative photochemistry. Figure 2 and Table 1 summarize how the supervision is constructed.

We evaluate QuantaMind at two levels: absolute numerical error and multi-objective candidate retrieval. The external 3,000-molecule testbank is disjoint from the tuning corpus and includes broader scaffolds from VQM24 and ZINC20. Comparisons include the untuned Qwen and Gemma backbones, a same-protocol tuned Gemma-3-27B control, six API-based reference LLMs, and five graph neural networks trained on the same molecular targets. This design tests both the effect of domain adaptation and the utility of the resulting model for discovery decisions rather than only point prediction.

Overall, our **main contributions** are as follows:

- We introduce QuantaMind, a 32B chemistry language model that converts SMILES-based instructions into structured, multi-property predictions for organic molecular screening.
- We develop a parameter-efficient adaptation recipe using QuantumChem-200K as domain supervision, including rare excited-state targets that are largely absent from general language-model pretraining.

- We evaluate the model across numerical prediction, out-of-distribution behavior, and multi-objective top-k discovery, with controlled backbone, data-scale, training-duration, LLM, and GNN comparisons.

2 Related Work

Language models for chemistry. Recent work has adapted language models to molecules by treating SMILES as a chemical language or by aligning molecular representations with natural-language instructions Jacobs et al. (2024); Liu et al. (2024); Xiao et al. (2024). These models provide a flexible interface across tasks, but accurate numerical regression remains challenging because token likelihood is not naturally aligned with physical error, uncommon quantum targets are sparse in pretraining corpora, and valid-looking text can still contain poorly calibrated numbers. QuantaMind focuses on this gap: it uses a fixed output schema and targeted instruction tuning to turn a general decoder-only LLM into a quantitative, multi-property chemistry model.

Graph-based molecular prediction. Graph neural networks (GNNs) and message-passing neural networks remain strong molecular predictors because they operate directly on atomic connectivity and can incorporate three-dimensional structure Hirschfeld et al. (2020); Tang et al. (2023); Miao et al. (2022). Hybrid LLM–GNN systems have also been proposed to combine linguistic flexibility with graph inductive biases Li et al. (2025). We therefore compare QuantaMind with five GNN architectures trained on the same targets. This comparison distinguishes the benefit of language-model adaptation from the benefit of denser structural input.

Supervision for excited-state chemistry. General molecular resources such as QM7, QM9, GDB-17, ZINC20, and the Open Macromolecular Genome (OMG) provide structures and common descriptors Hoja et al. (2021); Letovsky (1998); Irwin et al. (2020); Ramakrishnan et al. (2014). ANI-1ccx, VQM24, and PubChemQC add quantum-optimized geometries, energies, or conformer information Smith et al. (2020); Khan et al. (2024); Nakata & Shimazaki (2017). However, paired TPA and ISC targets suitable for instruction tuning are not generally available at the scale required here. QuantumChem-200K fills this supervision gap specifically to support QuantaMind; Appendix Table 6 contrasts it with the source resources.

3 QuantaMind: Chemistry Language Model

3.1 Task Formulation and Model Interface

Given a molecular SMILES string x and a chemistry instruction q, QuantaMind generates a structured response y containing a molecular-property profile. The model represents this task autoregressively as

$$p_\theta(y \mid q, x) = \prod_{t=1}^{|y|} p_\theta(y_t \mid q, x, y_{<t}),$$

where every field appears in a fixed order with an explicit unit. This formulation preserves the natural-language interface of a large language model while making its output deterministically parseable for regression and screening. Unlike independent single-property predictors, one QuantaMind response jointly describes the photophysical, safety, accessibility, and physicochemical attributes needed to compare a candidate against a user-defined discovery objective.

The model is initialized from Qwen-2.5-32B. We use SMILES as the molecular input because it is compact, ubiquitous in chemistry workflows, and compatible with the backbone tokenizer. The primary evaluation covers seven nontrivial targets: σ_{780}, $\sigma_{\max}$, ΔE_{ISC}, toxicity, synthetic accessibility, boiling point, and solubility. Figure 1 shows the corresponding instruction–response interface and screening workflow.

3.2 Domain Adaptation

We instruction-tune Qwen-2.5-32B with parameter-efficient LoRA adapters Hu et al. (2022) and quantization Dettmers et al. (2022). Each training example serializes an instruction, an input

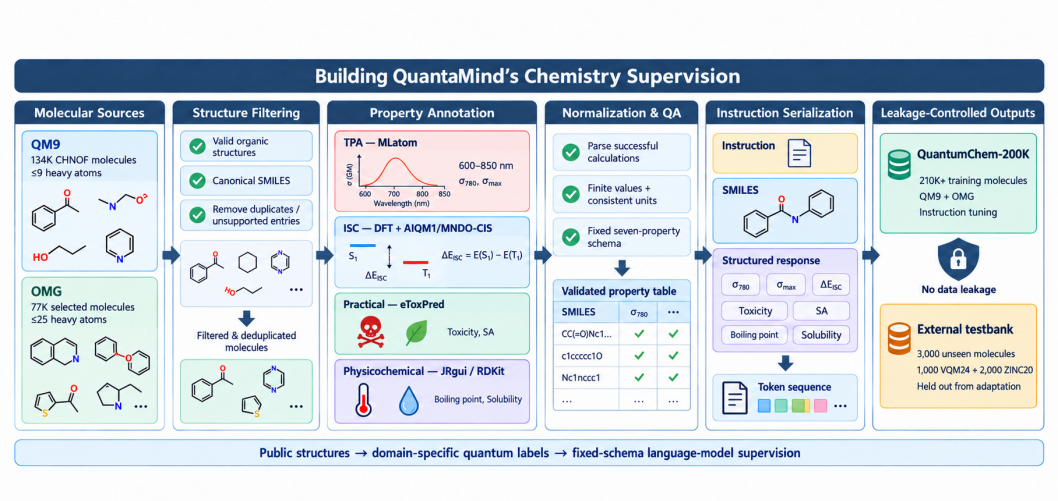


Figure 2: Construction of the chemistry supervision used to adapt QuantaMind. Structures selected from QM9 and OMG undergo validity and duplicate filtering before photophysical, excited-state, practical, and physicochemical annotation. Validated labels are normalized into a fixed seven-property schema and serialized as instruction–SMILES–response examples. The external VQM24/ZINC20 testbank remains fully held out from model adaptation.

SMILES string, and the target property profile in the same fixed response schema used at evaluation time. Training minimizes the causal language-model loss over the response tokens,

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{t=1}^{|y|} \log p_\theta(y_t \mid q, x, y_{<t}).$$

Experiments ran for approximately four to five days on compute nodes with dual AMD 7763 CPUs, 512 GB RAM, and four A100 80 GB GPUs. The final model is stored as LoRA adapters. Appendix Figures 5 and 11 report the optimization behavior and hyperparameter analyses.

3.3 Chemistry Supervision

QuantumChem-200K provides QuantaMind's chemistry supervision through the multi-stage pipeline in Figure 2. DFT, semi-empirical quantum chemistry, atomistic solvers, and cheminformatics predictors produce the targets in Table 1, which we treat as screening surrogates rather than uniformly experimental measurements. More than 210K molecules are used for training; an external 3K testbank is fully held out.

Table 1: Target properties (labels) in the QuantumChem-200k corpus used to supervise QuantaMind. Arrows show preferred directions (↑ larger; ↓ smaller).

Property	Description	Annotation Tool
TPA cross section σ (maximum and at 780 nm) ↑	Probability of simultaneously absorbing two photons, measured in GM units	MLatom atomistic platform Dral et al. (2024)
Wavelength range of absorption ↑	Wavelength range with a cross section above 20 GM	MLatom atomistic platform
Intersystem-crossing energy (ISC) ↓	Singlet–triplet energy gap, $\Delta E_{\mathrm{ISC}} = E(S_1) - E(T_1)$	AIQM1/MNDO-CIS and DFT Thiel (2014)
Toxicity ↓ and synthetic accessibility (SA) ↑	Estimated biological harm and ease of synthesis	eToxPred Pu et al. (2019)
Solubility and boiling point (BP) ↑	Solubility and boiling temperature in °C	JRgui Shi & Borchardt (2017)

Detailed source-dataset composition, target calculations, corpus statistics, and independent label validation are provided in Appendix A.1. Appendix Figure 7 visualizes the molecule-wise ISC comparison, while Appendix Table 8 summarizes the independent ISC and TPA validation statistics.

4 EXPERIMENTAL EVALUATION

We evaluate QuantaMind as both a quantitative chemistry model and a molecular screening system. Seven complementary metrics measure absolute prediction error and top-k multi-objective retrieval on a 3,000-molecule external testbank. Comparisons include three API-based reference LLMs, the untuned Qwen-2.5-32B backbone, a same-protocol fine-tuned Gemma-3-27B control, three pretrained molecular encoders, and five GNNs.

4.1 BASELINES AND CONTROLLED COMPARISONS

To isolate the effect of chemistry instruction tuning from backbone choice, we fine-tune Gemma-3-27B using the same training split, response schema, and optimization settings as QuantaMind. Fine-tuned Gemma and QuantaMind are evaluated on identical holdout inputs with the same parser and scoring pipeline; their results are reported in Tables 2 and 3. The untuned Qwen-2.5-32B result provides a direct before–after comparison for the QuantaMind backbone.

Full-InChIKey and Bemis–Murcko-scaffold matching finds 41 exact training-set matches among the 3,000 test molecules (1.37%); source-stratified overlap statistics are reported in Appendix Table 9.

Graph-neural-network baselines. We additionally compare against GCN, GIN, GAT, PNA, and EdgeCNN models trained on the same targets with explicit molecular graphs Kipf & Welling (2017); Xu et al. (2019); Veličković et al. (2018); Corso et al. (2020); Wang et al. (2019). Architecture, loss, and optimization details are provided in Appendix A.7.

Pretrained molecular-encoder baselines. We also include MoLFormer, ChemBERTa, and ChemFormer Ross et al. (2022); Chithrananda et al. (2020); Irwin et al. (2022). These SMILES encoders use multi-task prediction heads over the same seven targets and are evaluated on the same 3,000-molecule testbank. They provide a focused comparison with pretrained chemical sequence representations rather than instruction-following generative LLMs.

Table 2: wMAE benchmark (lower is better) for 10 LLMs, 3 pretrained molecular encoders, and 5 GNNs on 7 shared properties. Claude Haiku 4.5 and DeepSeek-V3.2 are few-shot prompted; all other untuned LLMs are zero-shot prompted. Fine-tuned Gemma and QuantaMind were trained on QuantumChem-200K. The rightmost column lists overall wMAE.

Model	**Excited-State quantum properties**			**Scores**		**Physicochemical properties**		**Overall wMAE**
	σ_{780}	$\sigma_{\max}$	ISC(S_1-T_1)	Tox	SA	BP	Sol	
Claude-Haiku-4.5	0.024	0.031	1.367	0.209	0.555	0.121	0.025	2.3320
Gemma-3-27B	0.026	0.026	1.794	0.329	0.582	0.096	0.026	2.8790
GPT-5.2	0.027	0.042	0.102	0.208	0.361	0.097	0.041	0.8780
DeepSeek-V3.2	0.029	0.030	0.228	0.303	1.020	0.079	0.025	1.7140
Qwen-2.5-32B	0.021	0.032	1.621	0.310	0.621	0.199	0.050	2.8540
Llama-3.1-70B	0.027	0.028	0.429	0.415	0.216	0.212	0.025	1.3520
Mistral-3.2-24B	0.197	0.582	0.308	0.133	0.231	0.172	0.040	1.6630
Phi-4	0.026	0.027	0.564	0.475	0.187	0.210	0.031	1.5200
Fine-tuned Gemma-3-27B	0.015	0.014	0.307	0.070	0.034	0.009	0.009	0.4581
MoLFormer Ross et al. (2022)	0.013	0.013	0.120	0.062	0.073	0.029	0.024	0.3340
ChemBERTa Chithrananda et al. (2020)	0.017	0.016	0.122	0.059	0.065	0.026	0.024	0.3290
ChemFormer Irwin et al. (2022)	0.014	0.013	0.092	0.059	0.086	0.023	0.027	0.3140
EdgeCNN Wang et al. (2019)	0.012	0.012	0.052	0.088	0.066	0.093	0.026	0.3490
GCN Kipf & Welling (2017)	0.015	0.015	0.055	0.098	0.093	0.099	0.031	0.4060
GIN Xu et al. (2019)	0.017	0.017	0.074	0.092	0.071	0.093	0.025	0.3890
PNA Corso et al. (2020)	0.017	0.016	0.055	0.097	0.062	0.116	0.027	0.3900
GAT Veličković et al. (2018)	0.015	0.015	0.057	0.114	0.103	0.134	0.031	0.4690
QuantaMind (Ours)	0.011	0.010	0.027	0.045	0.023	0.007	0.006	0.1287

4.2 MOLECULAR PROPERTY PREDICTION

We evaluate the seven primary properties on the 3,000-molecule testbank with wMAE, RMSE, and top-k screening metrics. Haiku and DeepSeek are few-shot prompted; Gemma, GPT, Qwen, Llama, Mistral, and Phi are zero-shot prompted. All models share the same response schema and parser; protocol and metric details appear in Appendix A.2 and Appendix A.4.

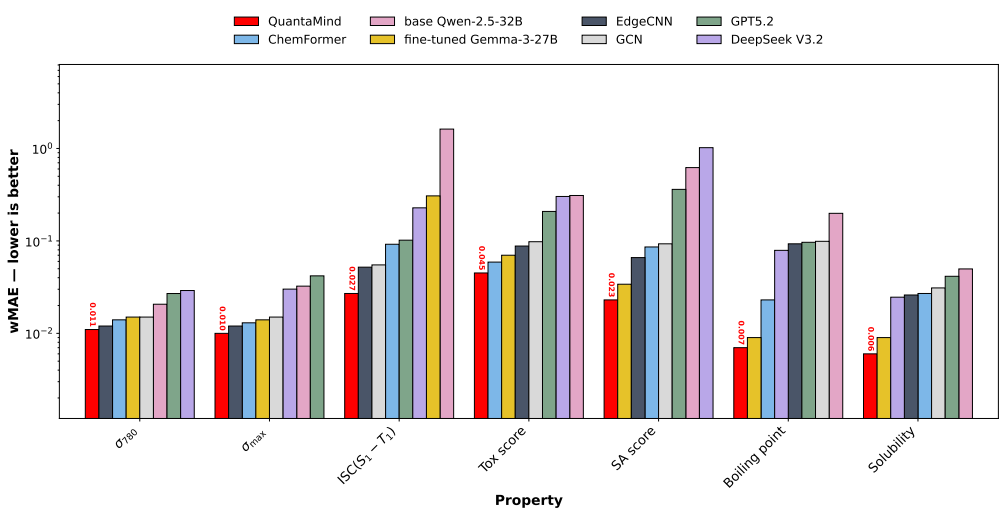


Figure 3: Property-level prediction error for representative language-model and graph-neural-network baselines (lower is better; logarithmic scale). QuantaMind attains the lowest wMAE on all seven screening properties shown. Table 2 reports aggregate wMAE for the complete benchmark.

Aggregate prediction accuracy. Table 2 and Figure 3 show that QuantaMind achieves the lowest aggregate and property-level wMAE across the evaluated systems. Across the seven shared properties in Table 2, instruction tuning reduces the Qwen-2.5-32B backbone's summed wMAE from 2.8540 to 0.1287; the same protocol lowers Gemma-3-27B from 2.8790 to 0.4581. The figure makes the source of this gain explicit: QuantaMind leads on every displayed property, with especially large improvements over base Qwen for ISC, toxicity, synthetic accessibility, boiling point, and solubility.

QuantaMind also remains accurate on the photophysics-critical targets, with wMAE values of 0.0106 for σ_{780}, 0.0104 for $\sigma_{\max}$, and 0.0273 for ISC. MoLFormer, ChemBERTa, and ChemFormer are competitive on TPA and scalar properties, but their ISC wMAE remains 0.120, 0.122, and 0.092, respectively. EdgeCNN is the closest TPA comparator, supporting the value of explicit graph structure, whereas QuantaMind provides the most balanced error profile across photophysical and practical screening targets.

RMSE, which penalizes large deviations more strongly than wMAE, provides a complementary view (Appendix Table 12). QuantaMind achieves the lowest RMSE among all evaluated systems for toxicity (0.051), synthetic accessibility (0.041), boiling point (20.559), and solubility (6.814×10^4), indicating that its aggregate gains are not driven only by many small errors. For the photophysical targets, molecular encoders and GNNs retain modest advantages: QuantaMind obtains RMSE values of 120.285 for σ_{780}, 143.846 for $\sigma_{\max}$, and 0.052 for ISC, close to EdgeCNN's 107.631, 128.728, and 0.047 and lower than the untuned language-model baselines. Together, wMAE and RMSE show that QuantaMind combines a balanced normalized error profile with reduced sensitivity to large errors on the practical screening properties.

Repeated-split/seed robustness. To provide a larger test set and quantify uncertainty beyond a single partition, we combined the 211,137-molecule training corpus with the 3,000-molecule testbank, yielding 214,137 molecules. We constructed three independent 90%/10% train/test splits (thus 21,414 molecules per test set) and trained QuantaMind with three seeds for each split, producing nine repeated runs in total. The resulting per-property means, standard deviations, and 95% confidence intervals are reported in Appendix Table 14.

4.3 Multi-Objective Molecular Screening

Prediction error is only a proxy for discovery utility. We therefore evaluate recovery of the 100 highest-ranked reference candidates using Precision@100, Recall@100, and Enrichment@100, and assess multi-objective quality using ϵ-Pareto precision and hypervolume regret. Definitions are provided in Appendix A.4; Figure 4 gives a traceable held-out example.

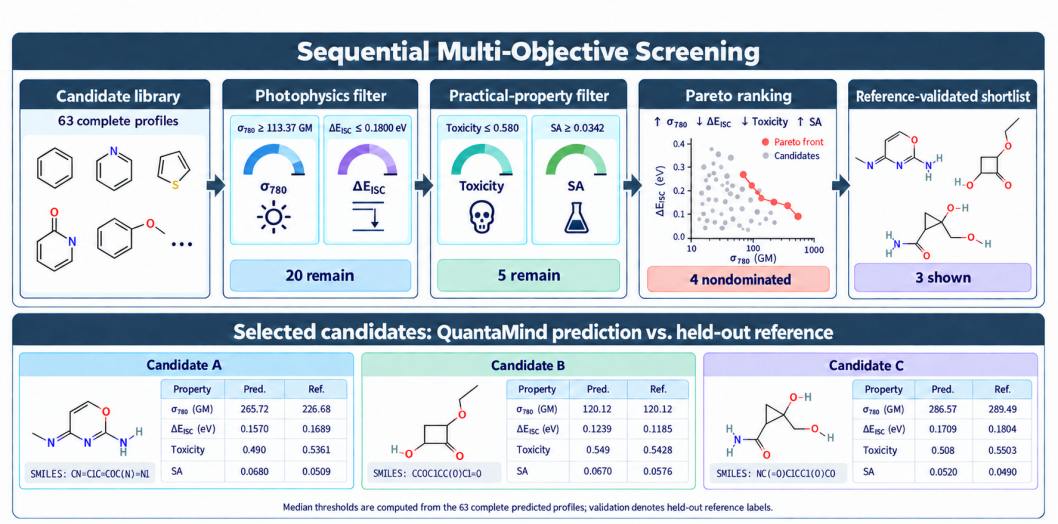


Figure 4: Sequential multi-objective screening with QuantaMind. Median thresholds computed from 63 complete predicted profiles first select candidates with high two-photon absorption at $780\,\mathrm{nm}$ and small singlet–triplet gaps, then impose toxicity and synthetic-accessibility constraints. Pareto ranking maximizes σ_{780} and synthetic accessibility while minimizing ΔE_{ISC} and toxicity. Three top-ranked nondominated candidates are shown with QuantaMind predictions and held-out reference values.

Using the same multi-objective rule for every model, QuantaMind recovers 35 of the reference top 100 and reaches 10.50 Enrichment@100 (Table 3), versus 31 and 9.30 for fine-tuned Gemma, 30 and 9.00 for EdgeCNN, and 29 and 8.70 for MoLFormer. Fine-tuned Gemma has slightly higher ϵ-Pareto Precision@20 (0.85 versus 0.80), whereas QuantaMind has substantially lower normalized HV regret (3.26% versus 42%). The five-property result is reported in Appendix Table 13.

Figure 4 follows 63 held-out examples with complete predictions for σ_{780}, ΔE_{ISC}, toxicity, and synthetic accessibility. Photophysical filtering retains 20 molecules, practical-property constraints retain five, and four are Pareto-nondominated. The three displayed molecules have the highest equal-weight percentile scores. This is ranking of a supplied library, not molecular generation; "reference-validated" means agreement with benchmark labels or DFT calculations.

Known-photoinitiator sanity check of screening. We identified known photoinitiators, including DMPA/BDK (2,2-dimethoxy-2-phenylacetophenone), 2-isopropylthioxanthone, and Irgacure 184, 1173, 369, and 907 that are absent from the testbank, where every photoinitiator exists on the markert and are commercially available. We then tested QuantaMind's multi-objective screening capabilities using those real known molecules, included as part of the whole test records. Irgacure 907 ranked best at the top 15.0%, while the other molecules all occupied the top 24.4–36.6% (lower is better), so QuantaMind does consistently recover established photoinitiators among other candidates. The benchmark therefore establishes fidelity to its surrogate ranking, with real-photoinitiator validation.

4.4 Model Behavior and Failure Analysis

Aggregate error alone can obscure which molecular decisions are reliable. Table 4 therefore reports representative high-TPA and low-ISC successes and failures from the external testbank. QuantaMind accurately recovers examples governed by relatively local functional-group patterns, but underpredicts TPA for extended conjugated scaffolds and shows larger ISC errors for nitrogen-rich structures. These cases reinforce the limitation of inferring delocalized electronic structure and heteroatom interactions from SMILES alone. A compact decomposition of the evaluated outputs by scientific role is provided in Appendix Table 11.

Table 3: Seven-property multi-objective screening. Higher is better except HV regret. Haiku and DeepSeek are few-shot; other untuned LLMs are zero-shot.

Model	Precision@100	Recall@100	Enrichment@100	ϵ-Pareto Precision@20	Normalized HV Regret@20
Claude-Haiku-4.5	0.06	0.06	1.80	0.40	91%
Gemma-3-27B	0.10	0.10	3.00	0.55	14%
GPT-5.2	0.01	0.01	0.30	0.50	68%
DeepSeek-V3.2	0.01	0.01	0.30	0.55	69%
Qwen-2.5-32B	0.06	0.06	1.80	0.45	89%
Llama-3.1-70B	0.03	0.03	0.90	0.45	65%
Mistral-3.2-24B	0.02	0.02	0.60	0.75	59%
Phi-4	0.00	0.00	0.00	0.65	76%
Fine-tuned Gemma-3-27B	0.31	0.31	9.30	0.85	42%
MoLFormer	0.29	0.29	8.70	0.60	69%
ChemBERTa	0.26	0.26	7.80	0.45	24%
ChemFormer	0.27	0.27	8.10	0.60	22%
EdgeCNN	0.30	0.30	9.00	0.60	8.80%
GCN	0.11	0.11	3.30	0.65	13.91%
GIN	0.23	0.23	6.90	0.60	15.18%
PNA	0.22	0.22	6.60	0.65	12.79%
GAT	0.14	0.14	4.20	0.65	14.33%
QuantaMind (Ours)	0.35	0.35	10.50	0.80	3.26%

Table 4: Representative success and failure cases selected from the testbank. Errors are rounded.

Case	Representative SMILES	Ground Truth	Prediction	Absolute error	Failure-mode insight
High TPA Success	`CC(NC(=O)OC(C)(C)C)C1CN(C(=O)CCO)CCO1`	$\sigma_{780} = 271.93$ $\sigma_{\max} = 283.70$	$\hat{\sigma}_{780} = 271.92$ $\hat{\sigma}_{\max} = 283.70$	$\lvert\hat{\sigma}_{780} - \sigma_{780}\rvert \approx 0.01$ $\lvert\hat{\sigma}_{\max} - \sigma_{\max}\rvert \approx 0.00$	The model succeeds on simpler high-TPA molecules with fewer aromatic carbons, suggesting that local functional-group patterns are easier to learn.
High TPA Failure	`CC(=Cc1cccc(Cl)c1)C(=O)NC(C)c1ccc(N2CCOCC2)cc1`	$\sigma_{780} = 293.25$ $\sigma_{\max} = 430.79$	$\hat{\sigma}_{780} = 48.20$ $\hat{\sigma}_{\max} = 119.41$	$\lvert\hat{\sigma}_{780} - \sigma_{780}\rvert = 245.05$ $\lvert\hat{\sigma}_{\max} - \sigma_{\max}\rvert = 311.38$	The molecule contains many aromatic carbons (c), indicating an extended conjugated scaffold; the model underpredicts TPA when the optical response depends on delocalized π-electron structure.
High TPA Failure	`O=C(C=Cc1ccc(OC(F)F)cc1)NCCC1CCC[NH2+]C1`	$\sigma_{780} = 346.61$ $\sigma_{\max} = 465.59$	$\hat{\sigma}_{780} = 29.10$ $\hat{\sigma}_{\max} = 64.92$	$\lvert\hat{\sigma}_{780} - \sigma_{780}\rvert = 317.51$ $\lvert\hat{\sigma}_{\max} - \sigma_{\max}\rvert = 400.67$	
Low ISC Success	`O=C1C=2OC12`	$\mathrm{ISC} = 0.0852$	$\widehat{\mathrm{ISC}} = 0.1333$	$\lvert\widehat{\mathrm{ISC}} - \mathrm{ISC}\rvert = 0.0481$	The model is more accurate for smaller low-ISC molecules with fewer nitrogen atoms, where the singlet–triplet gap is less affected by complex heteroatom interactions.
Low ISC Success	`O=N1=C2C=C12`	$\mathrm{ISC} = 0.0852$	$\widehat{\mathrm{ISC}} = 0.1376$	$\lvert\widehat{\mathrm{ISC}} - \mathrm{ISC}\rvert = 0.0524$	
Low ISC Failure	`CN=NN=O`	$\mathrm{ISC} = 0.0644$	$\widehat{\mathrm{ISC}} = 0.1833$	$\lvert\widehat{\mathrm{ISC}} - \mathrm{ISC}\rvert = 0.1189$	The molecule is nitrogen-rich, suggesting that N lone-pair and heteroaromatic effects strongly perturb the singlet–triplet gap and make ISC prediction less precise.
Low ISC Failure	`CC1=NN=N1`	$\mathrm{ISC} = 0.0612$	$\widehat{\mathrm{ISC}} = 0.2117$	$\lvert\widehat{\mathrm{ISC}} - \mathrm{ISC}\rvert = 0.1505$	

4.5 Model Ablations

Training-data scale. We train QuantaMind with one-third (70K), two-thirds (140K), and all (210K) of the supervision corpus. Figure 5 reports the corresponding optimization curves and evaluation results. Overall wMAE decreases as supervision grows and reaches the Table 2 value of 0.1287 at full scale, showing that the model continues to benefit from additional chemistry supervision.

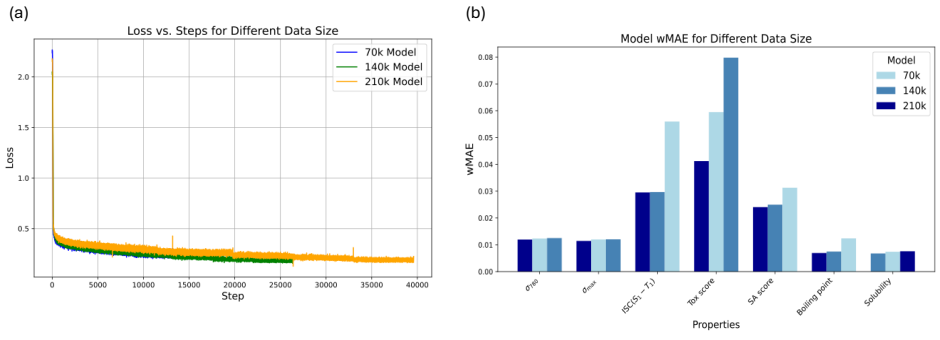


Figure 5: Training-data-scale ablation. Training loss and wMAE for QuantaMind models adapted with 70K, 140K, and 210K supervision examples. More domain supervision consistently improves held-out prediction accuracy.

Backbone and optimization controls. Domain adaptation is not specific to Qwen: applying the same protocol to Gemma-3-27B reduces its overall wMAE from 2.8790 to 0.4581. QuantaMind reaches 0.1287 under the same evaluation, indicating both a transferable benefit from chemistry instruction tuning and a remaining dependence on the pretrained backbone. Extending QuantaMind training from three to six epochs also reduces σ_{780} wMAE to the Table 2 value of 0.0106. We select a learning rate of 2×10^{-4} after comparing convergence profiles. Additional ablations, including the full backbone-transfer, learning-rate, and training-duration results, are presented in Appendix A.8.

Table 5: Implementation summary for the final QuantaMind model. Values not established by the experiments, such as a calibrated uncertainty threshold, are intentionally omitted.

Backbone	Qwen-2.5-32B decoder-only language model
Adaptation	LoRA adapters with a quantized backbone
Supervision	More than 210K instruction–SMILES–response examples
Objective	Causal language-model loss over ordered response tokens
Selected schedule	Six epochs; learning rate 2×10^{-4}
Compute	Four A100 80 GB GPUs; approximately four to five days
Model output	Fixed-schema profile for seven primary screening properties

At inference time, each candidate is evaluated independently: QuantaMind emits one ordered profile, the shared field-wise parser converts named values to numerical columns, and downstream objectives are applied only after parsing. This separation is important because it keeps molecular prediction distinct from the user's discovery policy. The same predicted profiles can support a photoinitiator-focused Pareto ranking, a toxicity-constrained shortlist, or another weighting without retraining the model. In other words, the system ranks a supplied library and exposes the property estimates used for that decision. The exact serialization template and parser policy are documented in Appendix A.2.

4.6 Scope and Intended Use

QuantaMind is designed as a high-throughput prioritization layer for structured molecular screening. It complements electronic-structure calculations and experiments by rapidly ranking supplied candidate libraries under user-defined objectives. Its predictions are most appropriate within the selected organic chemical space and seven-property schema used for adaptation; shortlisted candidates can then be advanced to independent quantum-chemical or wet-lab validation. Future work will extend the model with uncertainty calibration, three-dimensional and tool-augmented inputs, broader chemical supervision, and closed-loop experimental feedback.

5 Conclusion and Future Work

We introduced QuantaMind, a 32B chemistry language model for structured molecular-property prediction and multi-objective organic molecular screening. Parameter-efficient instruction tuning of Qwen-2.5-32B on QuantumChem-200K reduces overall wMAE from 2.8540 to 0.1287 on 3,000 unseen molecules. Same-protocol Gemma and GNN controls and training-scale ablations attribute the gains to domain adaptation while exposing the limits of SMILES-only prediction for excited-state and biological targets. QuantaMind recovers 35 of the reference top 100 candidates with 10.50-fold enrichment and low hypervolume regret, showing that its outputs support surrogate-guided multi-objective screening. The known-photoinitiator audit further identifies the potential of Quantamind to perform real-world photoinitiator-discovery tasks. Future work will add three-dimensional inputs, reinforcement-learning-enabled reasoning, experimental feedback, and retrieval over validated chemistry resources. Project website: `https://ravenllm.com/`.

Data and Code Availability

The QuantumChem-200K training corpus is available at https://huggingface.co/datasets/QuantumChem/QuantumChem-200k-new; the 3,000-molecule hold-out testbank is available at https://huggingface.co/datasets/QuantumChem/QuantumChem_Testbank_3000; the QuantaMind model is available at https://huggingface.co/QuantumChem/QuantaMind; and the anonymous code repository is available at https://github.com/AnonymousUser-3/QuantaMind.

Statement on GenAI Usage

The authors have used GenAI solely to proofread and polish the paper, and have not used GenAI to fabricate / falsify data, experiments, results, figures, or tables. The authors take full responsibility in the content of the paper.

Ethics Statement

The authors acknowledge the ICLR Code of Ethics and the responsibility to conduct and report research honestly and transparently. This work uses public molecular-structure resources and computed or model-derived property annotations; it does not report human-subject experiments or personal-data collection. QuantaMind is intended to prioritize organic molecules for subsequent expert evaluation, not to establish biological safety or replace real experiments. Molecular-screening models can also be misused to prioritize harmful substances; deployment should therefore include domain-expert oversight, appropriate chemical-safety controls, and compliance with applicable regulations. The paper documents data provenance, and reuse or redistribution of the resources should respect the licenses and attribution requirements of their original sources.

Reproducibility Statement

To support reproducibility, Section 3 describes the model formulation, response interface, adapter-based training objective, and chemistry supervision, while Table 5 summarizes the final model configuration. Figure 2 and Appendices A.1 and A.6 document the source structures, annotation workflows, normalization, and quality-control procedures. Appendix A.5 describes the external testbank; Appendices A.2 and A.4 specify the shared prompt, parser policy, and evaluation metrics. GNN baseline details and additional controlled experiments are provided in Appendices A.7 and A.8. Links to the training corpus and anonymous code repository are provided in the Data and Code Availability section.

REFERENCES

Stephen J. Blanksby and G. Barney Ellison. Bond dissociation energies of organic molecules. *Accounts of Chemical Research*, 36(4):255–263, 2003. doi: 10.1021/ar020230d.

S. E. Braslavsky. Glossary of terms used in photochemistry, 3rd edition (iupac recommendations 2006). *Pure and Applied Chemistry*, 79(3):293–465, 2007. doi: 10.1351/pac200779030293.

M. F. Budyka. Photodissociation of aromatic azides. *Russian Chemical Reviews*, 77(8):709–723, 2008. doi: 10.1070/RC2008v077n08ABEH003793.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets, 2020. URL https://arxiv.org/abs/2004.05718.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

Markus Deubel, Georg Von Freymann, Martin Wegener, Suresh Pereira, Kurt Busch, and Costas M Soukoulis. Direct laser writing of three-dimensional photonic-crystal templates for telecommunications. *Nature materials*, 3(7):444–447, 2004.

P. O. Dral, F. Ge, Y.-F. Hou, P. Zheng, Y. Chen, M. Barbatti, O. Isayev, C. Wang, B.-X. Xue, Jr. Pinheiro, M., Y. Su, Y. Dai, Y. Chen, L. Zhang, S. Zhang, A. Ullah, Q. Zhang, and Y. Ou. Mlatom 3: A platform for machine-learning-enhanced computational chemistry simulations and workflows. *Journal of Chemical Theory and Computation*, 20(3):1193–1213, 2024. doi: 10.1021/acs.jctc.3c01203.

Steven Farmer and Dietmar Kennepohl. Describing a reaction—bond dissociation energies. *Chemistry LibreTexts*, 2024. URL https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Organic_Chemistry_(Morsch_et_al.)/06%3A_An_Overview_of_Organic_Reactions/6.08%3A_Describing_a_Reaction_-_Bond__Dissociation_Energies. Last updated April 4 2024.

Lior Hirschfeld, Kevin Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020. doi: 10.1021/acs.jcim.0c00502.

Johannes Hoja, Leonardo Medrano Sandonas, Brian G. Ernst, Alvaro Vazquez-Mayagoitia, Robert A. DiStasio Jr., and Alexandre Tkatchenko. Qm7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific Data*, 8(1):43, 2021. doi: 10.1038/s41597-021-00812-2.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073, 2020. doi: 10.1021/acs.jcim.0c00675.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: A pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, 2022.

Ryan Jacobs, Maciej P. Polak, Lane E. Schultz, Hamed Mahdavi, Vasant Honavar, and Dane Morgan. Regression with large language models for materials and molecular property prediction. *arXiv preprint arXiv:2409.06080*, 2024. URL https://arxiv.org/abs/2409.06080.

Ameya Rajendra Jagtap and Aarti P. More. A review on self-initiated and photoinitiator-free system for photopolymerization. *Polymer Bulletin*, 79(10):8057–8091, 2022. doi: 10.1007/s00289-021-03887-4.

Nina Jeliazkova, Charalampos Chomenidis, Philip Doganis, Bengt Fadeel, Roland Grafström, Barry Hardy, Janna Hastings, Markus Hegi, Vedrin Jeliazkov, Nikolay Kochev, Pekka Kohonen, Cristian R. Munteanu, Haralambos Sarimveis, Bart Smeets, Pantelis Sopasakis, Georgia Tsiliki, David Vorgrimmler, and Egon Willighagen. The enanomapper database for nanomaterial safety information. *Beilstein Journal of Nanotechnology*, 6:1609–1634, 2015. doi: 10.3762/bjnano.6.165.

Joonyoung F. Joung, Minhi Han, Minseok Jeong, and Sungnam Park. Experimental database of optical properties of organic compounds. *Scientific Data*, 7:295, 2020. doi: 10.1038/s41597-020-00634-8.

Danish Khan, Anouar Benali, Scott Y. H. Kim, Guido F. von Rudorff, and O. Anatole von Lilienfeld. Quantum mechanical dataset of 836k neutral closed shell molecules with up to 5 heavy atoms from cnofsipsclbr. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2405.05961. arXiv:2405.05961.

S. Kim, C. M. Schroeder, and N. E. Jackson. Open macromolecular genome (version 1.0b). Data set, Zenodo, 2023. URL https://zenodo.org/records/7556992.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017. URL https://arxiv.org/abs/1609.02907.

Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260, 2024. doi: 10.1007/s10462-024-10888-y.

Kwang-Sup Lee, Dong-Yol Yang, Sang Hu Park, and Ran Hee Kim. Recent developments in the use of two-photon polymerization in precise 2d and 3d microfabrications. *Polymers for advanced technologies*, 17(2):72–82, 2006.

S. I. Letovsky. Gdb: The human genome database. *Nucleic Acids Research*, 26(1):94–99, 1998. doi: 10.1093/nar/26.1.94.

Youjia Li, Vishu Gupta, Muhammed Nur Talha Kilic, Kamal Choudhary, Daniel Wines, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Hybrid-llm-gnn: integrating large language models and graph neural networks for enhanced materials property prediction. *Digital Discovery*, 4:376–383, 2025. doi: 10.1039/D4DD00199K.

Yuyan Liu, Sirui Ding, Sheng Zhou, Wenqi Fan, and Qiaoyu Tan. Moleculargpt: Open large language model (llm) for few-shot molecular property prediction. *arXiv preprint arXiv:2406.12950*, 2024. URL https://arxiv.org/abs/2406.12950.

Frank Magnotta and Harold S. Johnston. Photodissociation quantum yields for the no_3 free radical. *Geophysical Research Letters*, 7(10):769–772, 1980. doi: 10.1029/GL007i010p00769.

Anna Mauri, Pascal Kiefer, Wolfgang Wenzel, and Mariana Kozlowska. Photoreactivity of norrish type photoinitiators for 3d laser printing via first principles calculations. *Macromolecular Rapid Communications*, -:e2500231, 2025. doi: 10.1002/marc.202500231. URL https://doi.org/10.1002/marc.202500231.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International conference on machine learning*, pp. 15524–15543. PMLR, 2022.

Michael D. Morse. Predissociation measurements of bond dissociation energies. *Accounts of Chemical Research*, 52(1):119–126, 2019. doi: 10.1021/acs.accounts.8b00526.

Maho Nakata and Tomomi Shimazaki. Pubchemqc project: A large-scale first-principles electronic structure database for data-driven chemistry. *Journal of Chemical Information and Modeling*, 57 (6):1300–1308, 2017. doi: 10.1021/acs.jcim.7b00083.

Alexander K. Nguyen and Roger J. Narayan. Two-photon polymerization for biological applications. *Materials Today*, 20(6):314–322, 2017. ISSN 1369-7021. doi: 10.1016/j.mattod.2017.06.004.

Christian R Ocier, Corey A Richards, Daniel A Bacon-Brown, Qing Ding, Raman Kumar, Tanner J Garcia, Jorik Van De Groep, Jung-Hwan Song, Austin J Cyphersmith, Andrew Rhode, et al. Direct laser writing of volumetric gradient index lenses and waveguides. *Light: Science & Applications*, 9(1):196, 2020.

Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074, 2024. doi: 10.3390/app14052074.

L. Pu, M. Naderi, T. Liu, H.-C. Wu, S. Mukhopadhyay, and M. Bryliński. etoxpred: A machine-learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacology & Toxicology*, 20:2, 2019. doi: 10.1186/s40360-018-0282-6.

Raghunathan Ramakrishnan, Michael Hartmann, Emanuela Tapavicza, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 k molecules. *Scientific Data*, 1:140022, 2014. doi: 10.1038/sdata.201422.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Wolfgang Rutsch. Recent developments in photoinitiators. *Progress in Organic Coatings*, 28(1–4): 147–153, 1996. doi: 10.1016/0300-9440(95)00539-0.

C. Shi and T. B. Borchardt. Jrgui: A python program of joback and reid method. *ACS Omega*, 2 (12):8682–8688, 2017. doi: 10.1021/acsomega.7b01464.

Justin S. Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E. Roitberg, Olexandr Isayev, and Sergei Tretiak. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data*, 7:134, 2020. doi: 10.1038/s41597-020-0473-z.

Miru Tang, Baiqing Li, and Hongming Chen. Application of message passing neural networks for molecular property prediction. *Current Opinion in Structural Biology*, 81:102616, 2023. doi: 10.1016/j.sbi.2023.102616.

Walter Thiel. Semiempirical quantum–chemical methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(2):145–157, 2014.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. URL https://arxiv.org/abs/1710.10903.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds, 2019. URL https://arxiv.org/abs/1801.07829.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2408.11363. arXiv:2408.11363.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks?, 2019. URL https://arxiv.org/abs/1810.00826.

Xiliang Yan, Alexander Sedykh, Wenyi Wang, Bing Yan, and Hao Zhu. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nature Communications*, 11:2519, 2020. doi: 10.1038/s41467-020-16413-3.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):160, 2024. doi: 10.1145/3649506.

Xiong Ye, Andrey Demidov, and Paul M. Champion. Measurements of the photodissociation quantum yields of mbno and mbo_2 and the vibrational relaxation of the six-coordinate heme species. *Journal of the American Chemical Society*, 124(20):5914–5924, 2002. doi: 10.1021/ja017359n.

Andriy Zakutayev, Nick Wunder, Marcus Schwarting, John D. Perkins, Robert White, Kristin Munch, William Tumas, and Caleb Phillips. An open experimental database for exploring inorganic materials. *Scientific Data*, 5:180053, 2018. doi: 10.1038/sdata.2018.53.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. URL https://arxiv.org/abs/2303.18223.

Xiaoqin Zhou, Yihong Hou, and Jieqiong Lin. A review on the processing accuracy of two-photon polymerization. *AIP Advances*, 5(3):030701, 2015. doi: 10.1063/1.4916886. URL https://pubs.aip.org/aip/adv/article/5/3/030701/279889.

Appendix

A Appendix

A.1 Supervision Corpus and Target Construction

QuantumChem-200K uses two public resources as structure pools. The first component contains 77K selected molecules from the Open Macromolecular Genome (OMG), covering synthetically relevant monomer and polymer motifs with up to 25 heavy atoms Kim et al. (2023). The second contains 134K CHNOF molecules from QM9 with up to nine heavy atoms and optimized ground-state geometries Ramakrishnan et al. (2014). Invalid, duplicated, unsupported, or failed entries are removed before serialization. These sources do not supply the complete photoinitiator targets used here; the value of the resulting supervision corpus is the additional standardized photophysical and practical annotation. To better align QuantumChem-200K with practical photoinitiator screening, we added 50 known photoinitiators, including structures with up to 95 heavy atoms, to expose the model to relevant molecular scaffolds and sizes that may be underrepresented in the QM9 and OMG-derived subsets.

TPA cross sections are obtained with MLatom over 600–850 nm in an Et(30)=55.4 solvent setting. We retain the maximum response $\sigma_{\max}$ and the response at 780 nm, σ_{780}. For ISC supervision, geometries are optimized at B3LYP/6-31G(2df,p), followed by AIQM1/MNDO-CIS singlet and triplet calculations Thiel (2014); the label is $\Delta E_{\mathrm{ISC}} = E(S_1) - E(T_1)$. Toxicity and synthetic accessibility are predicted with eToxPred Pu et al. (2019), while boiling point and solubility are obtained with JRgui/RDKit Shi & Borchardt (2017). The complete workflow, including calculation validation, unit normalization, fixed-schema serialization, and external-testbank separation, is summarized in Figure 2 and detailed further in Appendix A.6.

Table 6: Comparison of QuantaMind's supervision corpus with existing open molecular resources. QuantumChem-200K contributes the photophysical targets and screening labels used for model adaptation.

Dataset	Scale	Optimized geometry	TPA σ	ISC	Photoinitiator screening ready	Top-k ranking
QM9 Hoja et al. (2021)	134K	✓	×	×	×	×
OMG Kim et al. (2023)	12M	×	×	×	×	×
ZINC20 Irwin et al. (2020)	1.4B	×	×	×	×	×
VQM24 Khan et al. (2024)	836K	✓	×	×	×	×
QuantumChem-200K	210K+	✓	✓	✓	✓	✓

Table 7: Statistical summary of QuantumChem-200K and the external testbank.

Property	QuantumChem-200K				External testbank			
	Min	Max	Mean	Std	Min	Max	Mean	Std
σ_{780}	1.651	6217.846	121.550	99.540	2.365	3642.280	124.074	129.308
$\sigma_{\max}$	2.887	6778.190	151.222	114.872	3.501	4431.900	156.478	154.717
ISC($S_1 - T_1$)	0.036	0.307	0.173	0.031	0.021	0.305	0.153	0.048
Tox	0.195	0.914	0.595	0.095	0.164	0.895	0.458	0.191
SA	0.001	1.000	0.122	0.121	0.001	0.726	0.113	0.099
BP	26.280	2286.210	270.182	125.976	−75.010	1095.650	428.657	308.403
Sol	4.740×10^{-19}	9.994×10^{5}	1.096×10^{5}	1.908×10^{5}	8.110×10^{-7}	9.941×10^{5}	4.443×10^{4}	1.539×10^{5}
Heavy atoms	1.000	95.000	10.837	3.916	1.000	33.000	17.623	9.757

Table 7 shows that the external testbank is compositionally more demanding than the training corpus: its molecules contain more heavy atoms on average (17.623 versus 10.837) and exhibit broader variation in several photophysical and physicochemical targets. Figure 6 complements these summary statistics. The training corpus is concentrated around a median of nine heavy atoms but retains a long molecular-size tail, while the continuous PCA projections indicate broad, overlapping variation rather than a few isolated property clusters. Together, these views motivate treating the external testbank as a partially out-of-distribution evaluation rather than an in-distribution random split.

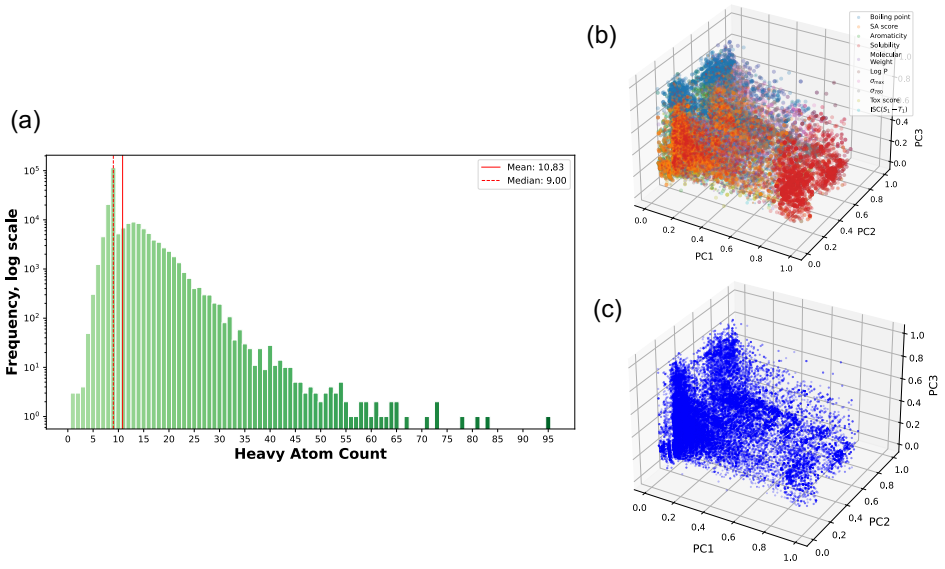


Figure 6: QuantumChem-200K statistics. (a) Heavy-atom-count distribution. (b–c) Principal-component projections of the property space.

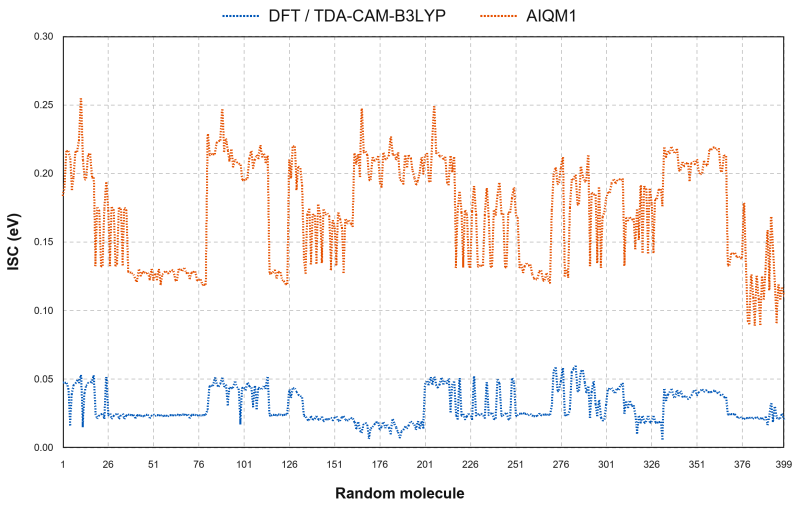


Figure 7: Validation of ISC energies against PySCF calculations. Both traces show comparable patterns across randomly sampled molecules.

Independent supervision validation. We compare the semi-empirical ISC labels against PySCF TDA–CAM-B3LYP calculations for 400 randomly selected CHNO molecules and the ML-derived TPA labels against compatible experimental measurements for 690 molecules from 49 publications. Figure 7 shows the molecule-wise ISC traces: the two methods share broad changes across the sampled molecules but retain substantial local deviations. Table 8 quantifies both checks. ISC gives MSE 0.021 and MAE 0.142; TPA gives MSE 0.218 and MAE 0.332. These error statistics support using the labels as noisy screening supervision, but neither comparison establishes experimental accuracy for individual predictions.

Table 8: Independent validation of ISC and TPA supervision.

Property	Validation reference	Validation set	MSE	MAE
ISC	PySCF TDA–CAM-B3LYP	400 molecules	0.021	0.142
TPA	Experimental measurements	690 molecules	0.218	0.332

A.2 Evaluation Prompt and Parser Policy

Every language model received the same instruction, property order, units, and response schema:

> "Based on the given SMILES string, predict the monomer's relevant properties. Given toxicity and SA score ranging from 0 to 1, aromaticity can only take 0 or 1. Answer the task in this format: The monomer compound has sigma of ? GM at 780 nm, maximum sigma of ? GM, ISC of ? eV, toxicity score of ?, SA score of ?, boiling point of ? °C, logP of ?, aromaticity of ?, solubility of ? ug/mol, molecular weight of ? g/mol."

Claude Haiku 4.5 and DeepSeek-V3.2 were evaluated with few-shot prompting. Gemma-3-27B, GPT-5.2, Qwen-2.5-32B, Llama-3.1-70B, Mistral-3.2-24B, and Phi-4 were evaluated zero-shot. None of these reference systems received model-specific prompt optimization or access to training labels. The same field-wise numerical parser was applied to every response. Syntactically numeric but degenerate outputs such as "0000" were recorded literally as zero rather than discarded or retrospectively corrected. QuantaMind and the fine-tuned Gemma control are reported as adapted systems; Qwen-2.5-32B provides the direct untuned-backbone comparison.

A.3 Photo-dissociation quantum yield

Photo-dissociation quantum yield is the most direct and informative indicator of photoinitiator performance, as it measures the fraction of excited molecules that successfully undergo bond cleavage after absorbing photons of a given frequency. Formally, the quantum yield is defined as the rate of the desired dissociation pathway divided by the total rate of all competing photophysical and photochemical processes Braslavsky (2007); Budyka (2008). Traditionally, determining quantum yield relies mostly on experimental measurements: researchers must design a controlled irradiation setup, choose an appropriate resin system, photodissociate the sample, and then infer reaction rates from the resulting polymerization or residual monomer content Mauri et al. (2025).

$$\Phi_{\text{diss}} = \frac{k_{\text{diss}}}{k_{\text{diss}} + k_f + k_{\text{IC}} + k_{\text{ISC}} + \sum k_{\text{other}}} \tag{1}$$

Figure 8 shows how two-photon polymerization works in advanced subsurface lithography to manufacture nano-scale photonic and quantum devices.

Another critical quantity governing photodissociation efficiency is the bond dissociation energy (BDE) of the cleavable bond in the excited state. After a molecule absorbs one or two photons and reaches an excited electronic state, it can either relax radiatively through fluorescence or undergo bond cleavage. For dissociation to occur, the absorbed energy must exceed the BDE of the initiating bond, which is why many photoinitiators contain relatively weak, photolabile linkages Blanksby & Ellison (2003); Morse (2019). For example, common photoinitiator: DMPA (2,2-dimethoxy-2-phenylacetophenone), undergoes Norrish Type I α-cleavage at the carbonyl–α-carbon bond to produce a benzoyl radical and an α-alkoxy radical. The BDE for a given cleavage pathway is computed as

$$\text{BDE} = H^\circ(\text{rad}_1) + H^\circ(\text{rad}_2) - H^\circ(\text{parent}) \tag{2}$$

where H denotes the electronic enthalpy of each species Farmer & Kennepohl (2024). Lower BDE values generally correlate with more efficient photodissociation and serve as a mechanistic indicator of promising photoinitiator candidates.

A.4 Evaluation Metrics

Definitions of the 7 evaluation metrics adopted in this work are described here:

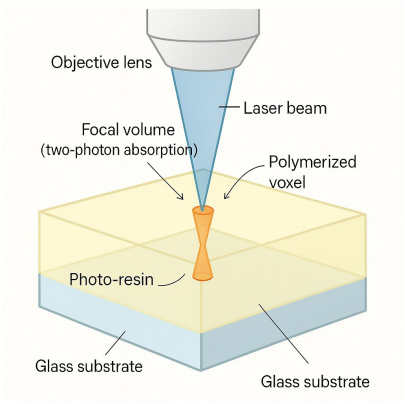


Figure 8: Schematic of TPP lithography: A focused photon beam is confined to a single voxel within the photo-resin. At this focal point, the photoinitiator absorbs light, undergoes dissociation, and generates reactive radicals. These radicals initiate polymerization with nearby monomers in the resin, enabling the formation of solid structures with nanoscale precision.

wMAE:

$$\mathrm{wMAE} = \frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} \sum_{i \in \mathcal{I}(M)} w_i \, |\hat{y}_i(M) - y_i(M)| \tag{3}$$

$$w_i = \left(\frac{1}{r_i}\right) \left(\frac{K \sqrt{\frac{1}{n_i}}}{\sum_{j=1}^{K} \sqrt{\frac{1}{n_j}}} \right) \tag{4}$$

Symbols:

$\mathcal{M}$: set of evaluated molecules;

$\mathcal{I}(M)$: index set of properties available for molecule M (missing properties are skipped).

$\hat{y}_i(M)$ and $y_i(M)$: predicted and ground-truth value of property i for M.

w_i: weighting factor for property i.

$r_i = \max(y_i) - \min(y_i)$: observed range of property i (computed on the evaluation split).

n_i: number of evaluation samples with a valid y_i.

K: total number of property prediction tasks included in the metric.

The factor $1/r_i$ rescales properties to reduce the effect of differing units/ranges, while $\frac{K\sqrt{1/n_i}}{\sum_{j=1}^{K}\sqrt{1/n_j}}$ allocates relatively larger weight to rarer properties (smaller n_i) and is normalized across tasks before the range scaling is applied.

Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{5}$$

where y_i denotes the ground-truth value, $\hat{y}_i$ denotes the predicted value, and N is the total number of samples.

Ranking Metrics: The Precision@100, Recall@100, Enrichment@100, ϵ-Pareto Precision@20, and normalized hypervolume (HV) regret are defined and calcuated as such:

By assigning different weights to different properties, we calculate a composite score for each of the 3000 molecules. We then rank the molecules based on their scores and select the top 100 candidates from the ground-truth testbank. This list is then compared with the top 100 candidates

ranked by QuantaMind and other LLMs. By comparing these scores and candidate lists, we further demonstrate our model's ability to correctly screen and identify strong candidates for potential photoinitiators. For a larger dataset, QuantaMind could potentially be used to efficiently narrow down a list of promising photoinitiator candidates for further wet-lab verification.

The value of each property is first percentile-normalized to a range of 1 to 100. Then, the score S_i is calculated using the following weights:

$$\begin{aligned}\text{Score } S_i = {} & 0.30\,\sigma_{780\,\text{nm}} + 0.15\,\sigma_{\max} - 0.20\,\text{ISC} \\ & + 0.10\,\text{Solubility} + 0.05\,\text{Boiling Point} \\ & + 0.10\,\text{SA} - 0.10\,\text{Toxicity}.\end{aligned} \tag{6}$$

A weight of 0.30 is assigned to $\sigma_{780\,\text{nm}}$ because it is the most important property for this task. Maximum σ reflects the overall absorption strength and is assigned a weight of 0.15. A lower intersystem crossing energy and a lower toxicity score are also preferred, so they are assigned negative weights. Other less critical properties are assigned smaller positive weights based on their relevance.

$$\text{Precision@k} = \frac{|\text{Top-}k_{\text{pred}} \cap \text{Top-}\alpha\%_{\text{true}}|}{k}$$

$$\text{Recall@k} = \frac{|\text{Top-}k_{\text{pred}} \cap \text{Top-}\alpha\%_{\text{true}}|}{|\text{Top-}\alpha\%_{\text{true}}|}$$

$$\text{Enrichment@k} = \frac{\dfrac{|\text{Top-}k_{\text{pred}} \cap \text{Top-}\alpha\%_{\text{true}}|}{k}}{\dfrac{|\text{Top-}\alpha\%_{\text{true}}|}{N}}$$

where N is the total number of molecules.

$$\epsilon\text{-Pareto Precision@k} = \frac{|\text{Top-}k_{\text{pred}} \cap \mathcal{P}_{\epsilon}^{\text{true}}|}{k}$$

where $\mathcal{P}_{\epsilon}^{\text{true}}$ is the set of ϵ-Pareto-optimal (near-Pareto) molecules.

$$\text{Normalized HV Regret@k} = \frac{HV(\mathcal{P}^{\text{true}}) - HV(\mathcal{S}_{\text{pred},k})}{HV(\mathcal{P}^{\text{true}})}$$

where:

$$\begin{aligned}\mathcal{P}^{\text{true}} &: \text{ true Pareto front} \\ \mathcal{S}_{\text{pred},k} &: \text{top-}k \text{ molecules selected by the model} \\ &\quad \text{evaluated with ground truth}\end{aligned}$$

A.5 Holdout Test Set Setup

To rigorously evaluate the generalization ability of our model beyond the training dataset, we constructed a hold-out testbank of 3,000 externally sourced molecules from two independent resources. We first selected 1,000 molecules from the VQM24 dataset Khan et al. (2024). Only CHNOF molecules were retained to ensure compatibility with our ISC workflow. The second subset consists of 2,000 molecules extracted from ZINC20 Irwin et al. (2020), a large and chemically diverse database containing drug-like structures with up to 30 heavy atoms (CHNOFBrSiSCl). This subset introduces significantly broader chemical diversity than the training set, including functional groups not present in QuantumChem-200K, making it an ideal source for testing robustness and out-of-distribution generalization. Together, these sources provide a property-complete benchmark for evaluating QuantaMind and the baseline models.

InChIKey and scaffold overlap analysis. We assessed separation between the 211,137-record QuantumChem-200K training corpus and the external testbank using full InChIKeys and Bemis–Murcko scaffolds derived from SMILES. Only 41 test molecules (1.37%) shared a full InChIKey with the training corpus. Among the 2,531 ring-bearing test molecules, 395 (15.61%) shared a scaffold with at least one training molecule, so most ring-bearing test molecules retained previously unseen scaffolds. At the unique-scaffold level, 121 of the testbank's 1,963 unique scaffolds (6.16%) occurred in training (Table 9). The VQM24 and ZINC20 test subsets shared no full InChIKeys with one another and only two scaffolds. Together with the testbank's higher mean heavy-atom count (17.62 versus 10.84 in training), these results indicate a meaningful shift in molecular size and structure for assessing generalization.

Table 9: Overlap of the test subsets with the training corpus. Unique-scaffold percentages use the number of unique scaffolds in each test subset as the denominator.

Test subset	Molecules	Shared full InChIKeys	Shared unique scaffolds
Entire testbank	3,000	41 (1.37%)	121/1,963 (6.16%)
VQM24	1,000	40 (4.00%)	21/260 (8.08%)
ZINC20	2,000	1 (0.05%)	102/1,705 (5.98%)

A.6 Data Quality Control

We applied several quality-control steps during the construction of QuantumChem-200K to improve the reliability of both molecular structures and computed labels. First, source molecules from OMG and QM9 were treated as structure pools rather than directly merged datasets. Molecules were filtered to retain valid organic structures compatible with the downstream quantum-chemical workflows, and invalid, duplicated, or unsupported molecular entries were removed before property annotation. Canonical molecular identifiers were used where applicable to reduce duplicate structures across source datasets.

Second, we performed calculation-level checks for the generated photophysical labels. For TPA annotation, each molecule was processed through the same wavelength window and solvent setting, and the final labels were extracted only from successfully generated wavelength-resolved spectra. Molecules with failed calculations, missing spectra, non-parsable output files, or invalid numerical values were excluded from the final labeled corpus. For ISC annotation, only molecules with valid optimized geometries and successfully computed singlet and triplet excited-state energies were retained. The ISC label was computed consistently as $\Delta E_{\mathrm{ISC}} = E(S_1) - E(T_1)$, and entries with missing excited-state energies or failed electronic-structure calculations were removed.

Repeated TPA values can occur because ML-TPA accepts SMILES at the workflow interface but predicts from a compressed fingerprint-derived descriptor representation Dral et al. (2024). Different structures can therefore map to the same effective feature pattern and receive the same prediction; repeated TPA labels do not by themselves indicate duplicated molecular records.

Third, we applied label-level sanity checks to ensure that all properties were numerically valid and physically interpretable. Values with missing entries, non-finite numbers, parsing errors, or inconsistent units were excluded or recomputed. For downstream statistical analysis and PCA visualization, extreme outlier effects were reduced by clipping each property to the 2–98% percentile range, while the original numerical labels were retained for benchmark evaluation. We also inspected the distributions of key properties, including TPA cross sections, ISC energies, physicochemical descriptors, and heavy-atom counts, to verify that the resulting dataset covers a broad but chemically plausible molecular space.

Finally, we controlled evaluation quality by constructing the 3K holdout testbank from external molecular sources. We verify its separation from the QuantumChem-200K fine-tuning set through full-InChIKey and Bemis–Murcko-scaffold matching (Appendix A.5, Table 9). The low exact-match rate limits direct train–test duplication while retaining a structurally shifted evaluation of generalization. Together, these structure-level, calculation-level, label-level, and split-level checks make the corpus suitable for adapting and evaluating QuantaMind.

A.7 GNN Architecture and Training Details

1. GNN Training Configuration

For fair comparison, all GNN baselines are trained under the same general configuration. Each model uses a shared graph encoder with three graph convolution layers, followed by task-specific MLP prediction heads. Graph-level molecular embeddings are obtained using global mean pooling over node embeddings. Dropout with a rate of 0.2 is applied to reduce overfitting.

The models are trained with a batch size of 128 using the AdamW optimizer. The learning rate is set to 1e-3, and the weight decay is set to 5e-4. These settings are kept consistent across all GNN architectures so that performance differences mainly reflect architectural differences rather than differences in training setup.

2. GNN Loss Function and Missing Label Handling

Because QuantumChem-200K contains missing values for some properties, we use a masked-MSE loss for regression targets. During training, missing labels are excluded from the loss calculation, so the model is only penalized for properties with valid ground-truth values. This prevents missing cells from affecting gradient updates while still allowing the model to train on molecules with partially available labels.

For binary aromaticity prediction, the model uses BCEWithLogitsLoss, which directly operates on raw logits and improves numerical stability. The final training loss is computed from the sum of the task-specific losses across the regression heads and the binary aromaticity head.

A.8 Additional Experiments and Ablation Studies

Table 10 provides a compact structural comparison of the two supervision pools. QM9 contributes smaller CHNOF molecules with optimized geometries, whereas the OMG-derived subset extends the model's exposure to longer SMILES and additional organic elements; this table complements the broader resource comparison in Appendix Table 6.

Table 10: Composition of the QuantumChem-200k training set.

Traits	**134k Subset**	**77k Subset**
Length of SMILES	Up to 9 heavy atoms	Up to 25 heavy atoms
Chemical elements	CHNOF	CHNOFSBrClSi

Radar-summary analysis. Figure 9 provides an auxiliary benchmarking view of the various language models in a radar plot. Tables 2 and 12 remain the complete error comparisons for the seven primary screening targets.

Panel (a) shows that QuantaMind preserves the most balanced structure–property ordering among the plotted LLMs. Its correlations on the difficult excited-state targets are moderate but positive ($r = 0.433$ for σ_{780}, 0.416 for $\sigma_{\max}$, and 0.551 for ISC), whereas several untuned LLMs are near zero or negative. The much stronger correlations for toxicity (0.966), synthetic accessibility (0.911), and the physicochemical descriptors indicate that domain adaptation is particularly effective when SMILES contains recognizable structural motifs, while excited-state behavior remains harder because it depends on electronic structure that is only indirectly encoded in the molecular string.

Panel (b) compresses the ten-property errors into an inverse overall-wMAE view, so a larger radius corresponds to lower aggregate error. The separation of QuantaMind from the untuned models is consistent with the property-level comparison in Figure 3: chemistry instruction tuning improves the aggregate profile rather than only one convenient descriptor. Panel (c) emphasizes large deviations through RMSE. Each property-wise RMSE is normalized by the corresponding cross-model average and inversely transformed so that larger radii again indicate better performance. QuantaMind shows the clearest advantages on the two TPA targets, ISC, toxicity, synthetic accessibility, and boiling point. The smaller gaps for logP and molecular weight are consistent with those descriptors being common in public chemistry resources and comparatively direct to infer from composition. Together, the three panels show that the largest benefit of domain adaptation occurs on the photophysical and practical screening targets that are least represented in generic language-model pretraining.

Table 11: Error decomposition by property group.

Group	Properties	Insight
Excited-state / photophysics	σ_{780}, $\sigma_{\max}$, ISC	Most important and difficult to predict since these labels depend on excited-state electronic structure and quantum-chemical simulations.
Practical screening scores	Toxicity, SA score	Determine usable candidates, since low toxicity and high synthetic accessibility are crucial for downstream validation.
Physicochemical descriptors	Solubility, boiling point	Easier to compute and infer from molecular structure, size, and functional groups.

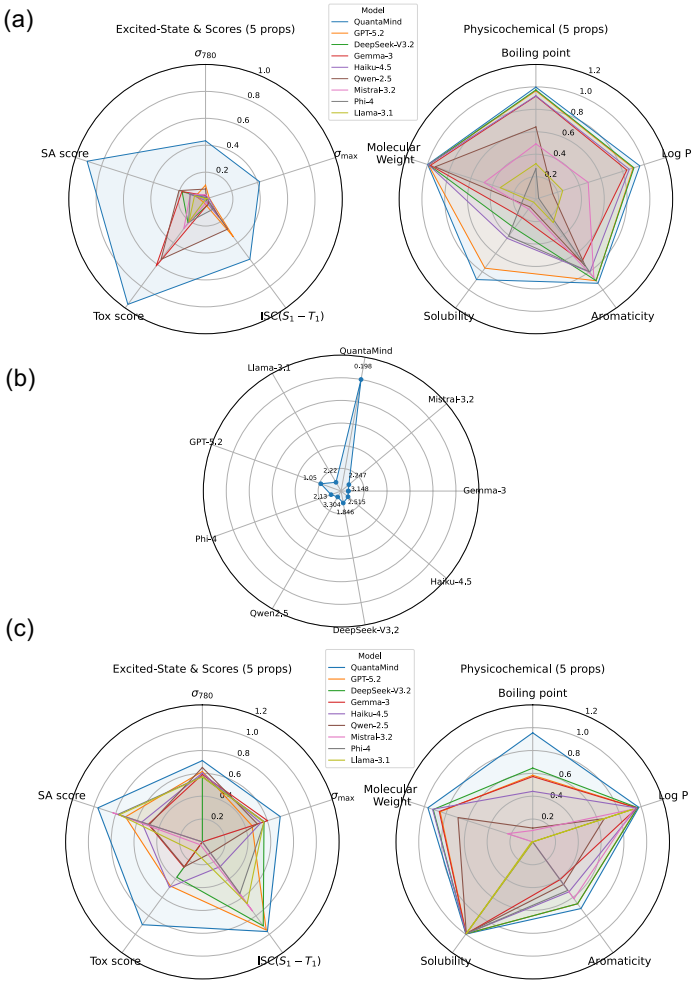


Figure 9: Auxiliary ten-property comparison of nine plotted language models using (a) Pearson r, (b) inverted overall wMAE, and (c) inverted RMSE. Larger radial distance indicates better performance. Fine-tuned Gemma, molecular encoders, and GNN baselines are not included in this visualization; the complete seven-property error comparisons are reported in Tables 2 and 12.

Error-group summary. Table 11 groups the evaluated properties by their role in molecular screening. The excited-state targets depend most strongly on electronic structure, while the practical scores and physicochemical descriptors impose complementary constraints on candidate selection.

Table 12: RMSE benchmark (lower is better) across 10 language-model systems, 3 pretrained molecular encoders, and 5 GNNs for 7 properties. Claude Haiku 4.5 and DeepSeek-V3.2 are few-shot prompted; all other untuned LLMs are zero-shot prompted.

Model	Excited-State quantum properties			Scores		Physicochemical properties	
	σ_{780}	$\sigma_{\max}$	ISC(S_1-T_1)	Tox	SA	BP	Sol
Claude-Haiku-4.5	165.683	235.271	1.288	0.250	0.509	255.466	1.598×10^5
Gemma-3-27B	171.077	203.713	1.757	0.356	0.580	196.078	1.591×10^5
GPT-5.2	160.110	276.194	0.082	0.257	0.336	190.876	2.028×10^5
DeepSeek-V3.2	179.455	220.179	0.160	0.304	1.157	162.027	1.550×10^5
Qwen-2.5-32B	145.509	251.175	1.544	0.359	0.591	402.947	2.280×10^5
Llama-3.1-70B	179.713	221.053	0.582	0.440	0.269	458.417	1.605×10^5
Mistral-3.2-24B	417.515	510.420	0.442	0.477	0.200	412.561	2.700×10^7
Phi-4	177.951	218.265	0.772	0.492	0.255	457.420	1.963×10^5
Fine-tuned Gemma-3-27B	186.817	226.582	5.565	0.087	0.060	24.091	9.225×10^4
MoLFormer	118.392	138.814	0.082	0.064	0.083	54.360	1.350×10^5
ChemBERTa	135.775	162.269	0.097	0.064	0.085	75.217	1.358×10^5
ChemFormer	111.884	132.773	0.078	0.065	0.090	47.115	1.413×10^5
EdgeCNN	107.631	128.728	0.047	0.088	0.075	205.481	1.269×10^5
GCN	120.491	145.347	0.048	0.099	0.104	226.675	1.486×10^5
GIN	455.482	483.792	0.105	0.095	0.079	248.876	1.289×10^5
PNA	199.707	222.925	0.053	0.090	0.078	275.214	1.355×10^5
GAT	125.500	150.418	0.050	0.114	0.114	268.933	1.466×10^5
QuantaMind (Ours)	120.285	143.846	0.052	0.051	0.041	20.559	6.814×10^4

Table 13: Top-k ranking and multi-objective discovery across the LLMs using five properties. Higher is better except for hypervolume regret. Claude Haiku 4.5 and DeepSeek-V3.2 are few-shot prompted; all other untuned LLMs are zero-shot prompted.

Model	Precision@100	Recall@100	Enrichment@100	ϵ-Pareto Precision@20	Normalized HV Regret@20
Claude-Haiku-4.5	0.08	0.08	2.40	0.20	99%
Gemma-3-27B	0.10	0.10	3.00	0.10	51%
GPT-5.2	0.04	0.04	1.20	0.50	99%
DeepSeek-V3.2	0.07	0.07	2.10	0.20	99%
Qwen-2.5-32B	0.07	0.07	2.10	0.15	99%
Llama-3.1-70B	0.08	0.08	2.40	0.20	98%
Mistral-3.2-24B	0.03	0.03	0.90	0.20	99%
Phi-4	0.05	0.05	1.50	0.05	99%
QuantaMind (Ours)	0.38	0.38	11.40	0.75	71%

Nested split-and-seed robustness. We additionally evaluate QuantaMind using three independently constructed data splits and three optimization seeds per split, for nine training runs in total. Because the three seeds within a split share the same test set, we first average the seeds within each split and treat the three split-level means as the independent observations. Table 14 reports the mean, the standard deviation across split-level means, and the resulting 95% confidence interval, computed as $\bar{x} \pm t_{0.975,2} s/\sqrt{3}$ with $t_{0.975,2} = 4.303$.

Table 14: Nested split-and-seed uncertainty for QuantaMind (wMAE across 9 total runs: 3 independent data splits $\times$ 3 optimization seeds).

Property	Mean wMAE (9 runs)	SD across splits	95% CI (nested, $df = 2$)
σ_{780}	0.0112	0.0008	[0.0092, 0.0132]
$\sigma_{\max}$	0.0102	0.0007	[0.0085, 0.0119]
ISC	0.0278	0.0032	[0.0199, 0.0357]
Toxicity	0.0465	0.0054	[0.0331, 0.0599]
SA	0.0234	0.0015	[0.0197, 0.0271]
Boiling point	0.0071	0.0003	[0.0064, 0.0078]
Solubility	0.0061	0.0002	[0.0056, 0.0066]
Overall	**0.1323**	**0.0072**	**[0.1144, 0.1502]**

Additional ablations

Backbone-transfer ablation. To verify that the improvement is not specific to one backbone, we fine-tuned Gemma-3-27B under the same training conditions and evaluated it with the same holdout protocol. Fine-tuning reduces Gemma's overall wMAE from 2.8790 to 0.4581, while QuantaMind achieves 0.1287. Figure 10 visualizes the property-level wMAE and RMSE comparisons together with the training-loss trajectories. The results confirm that chemistry instruction tuning transfers across backbones, while the Qwen-based QuantaMind model retains the lower aggregate error.

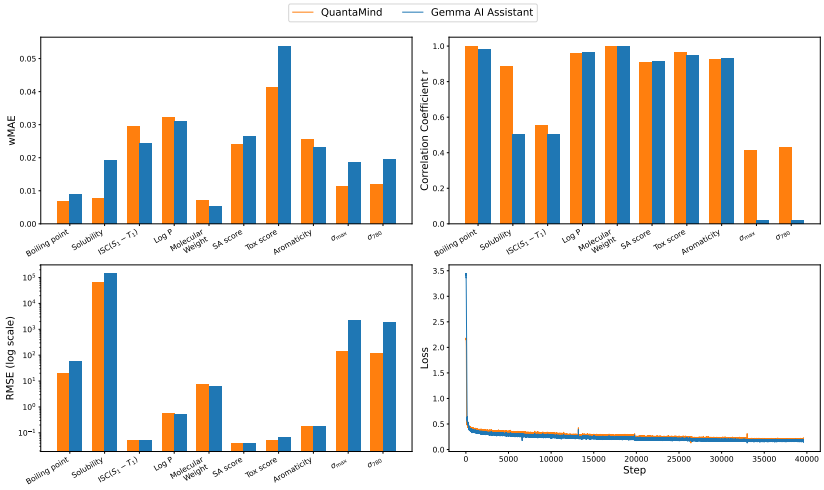


Figure 10: Backbone-transfer comparison between a Gemma-3-27B model fine-tuned under the same protocol (blue) and the Qwen-based QuantaMind model (orange), showing property-level wMAE, RMSE, and training-loss trajectories.

Learning-rate ablation. Figure 11 shows the training loss for different learning rates during fine-tuning. We use a learning rate of 2×10^{-4} for subsequent experiments after balancing convergence speed and training cost.

Training-duration ablation. A separate 100-molecule checkpoint analysis compares QuantaMind after three and six fine-tuning epochs. The six-epoch checkpoint has lower wMAE across the evaluated properties, with particularly clear improvements for the TPA cross section at $780\,\mathrm{nm}$ and ISC. Toxicity remains less stable, consistent with the difficulty of inferring biological behavior from molecular strings alone. These results show that additional optimization helps the model associate SMILES-level structural cues with photochemical targets.

Additional PCA Plots Figure 12 is the 2D PCA plots of all 10 properties in the QuantumChem-200K dataset. Principal component analysis was performed on the standardized 10-dimensional property space. Each panel shows the same PC1–PC2 projection of all molecules, with points colored by the value of each property (color bar). Across panels, only the color mapping changes while the PCA coordinates remain fixed, enabling direct comparison of how different properties vary over the shared low-dimensional embedding.

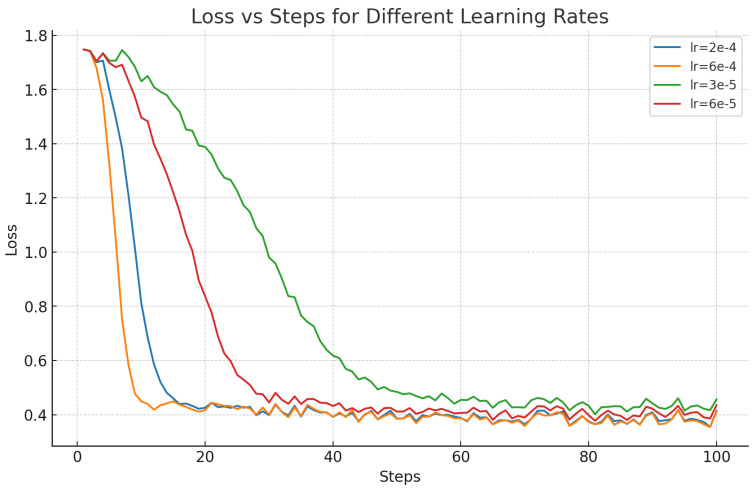


Figure 11: Ablation study 3: Training loss of using different learning rates.

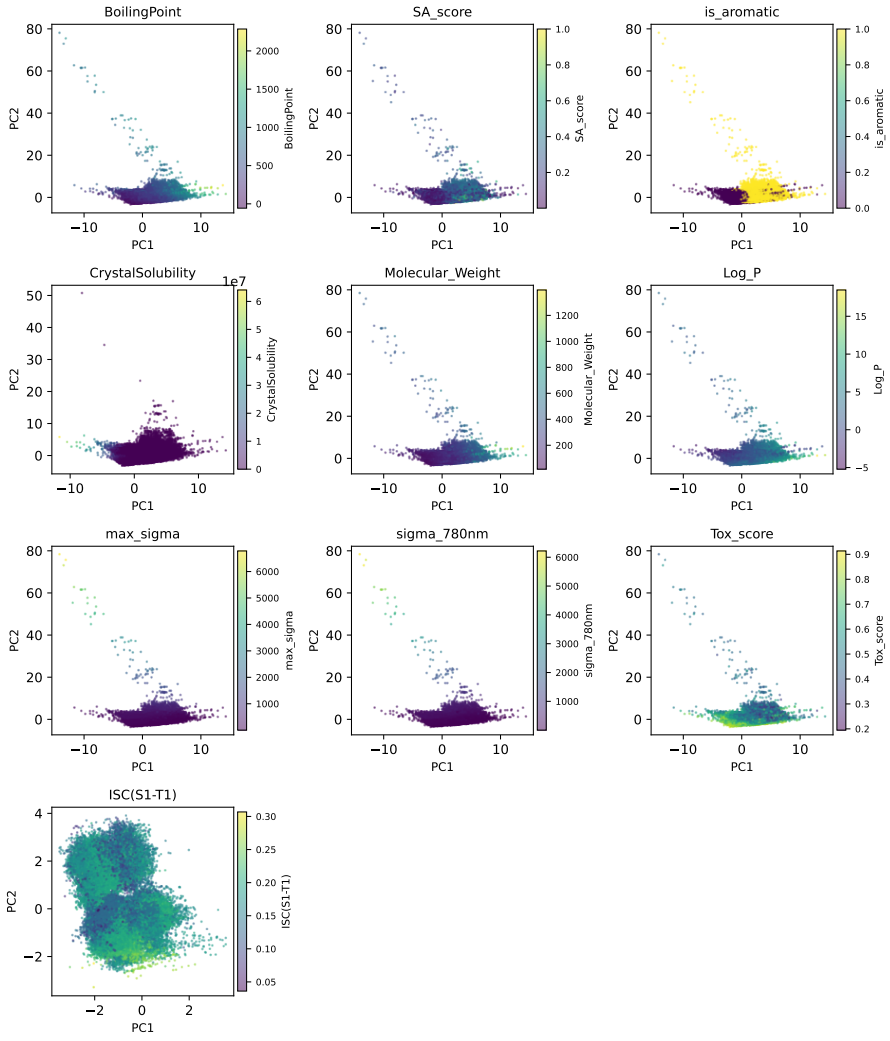


Figure 12: 2D PCA plots of the QuantumChem-200K dataset. Color bar denotes the magnitude of each property in their respective subplots.

A.9 Other Related Work

In parallel, a variety of domain-specific datasets have been developed. For example, datasets have been curated focusing on inorganic and solid-state materials, organic optoelectronic or nanomaterial systems Zakutayev et al. (2018); Joung et al. (2020); Yan et al. (2020); Jeliazkova et al. (2015). Such datasets extend general molecular collections into specific material classes and support applications that require targeted chemical or structural representations. For example, properties including the TPA spectra, ISC energy, and excited-state energetics, which are crucial for photoinitiator and photochemical-material design, need both optimized geometry and TD-DFT to compute.