BOLEK: A MULTIMODAL LANGUAGE MODEL FOR MOLECULAR REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

AI models and platforms for molecular science underpin high-stakes applications in drug discovery, yet the systems delivering them are largely opaque: they expose either a score and a binary answer with no rationale, or fluent prose rarely anchored in molecular structure. We introduce BOLEK, a compact multimodal language model that injects a molecular embedding into an instruction-tuned decoder to predict molecular properties and explain them in auditable terms. BOLEK improves over its Qwen3 base on every task group of a 30-endpoint TDC panel and outperforms the chemistry-specialist TxGemma-9B-Chat on 11 of 17 ADMET endpoints and on four of the five endpoints held out from its training. Together, these results suggest that targeted modality injection paired with reasoning supervision tied to verifiable features can match domain-specialist systems at a fraction of the parameter count while producing auditable explanations for downstream decisions.

1 INTRODUCTION

Machine Learning models have become a central component of modern molecular discovery (Wu et al., 2018; Huang et al., 2021; Yang et al., 2019; Méndez-Lucio et al., 2024), supporting the whole pipeline from hit identification through the late preclinical stage with tasks such as virtual screening, activity & ADME prediction, toxicity screening, retro-synthesis planning, and lab-in-the-loop optimization cycles.

Yet in high-stakes scientific workflows, a label or a calibrated score is rarely sufficient on its own. A medicinal chemist deciding whether to synthesise a compound, a pharmacologist triaging a toxicity flag, or a translational scientist building a candidate's profile must understand *why* a molecule is predicted to cross the blood–brain barrier, engage a target, or fail a drug-likeness criterion: downstream decisions hinge not on the prediction alone, but on whether the evidence behind it is mechanistically plausible and chemically meaningful (Jiménez-Luna et al., 2020). When a model returns only a score, the chemist's role collapses into accepting or overriding it; when the model exposes the reasoning behind the score, that reasoning becomes something the chemist can interrogate, correct, and build on, turning prediction into a substrate for design. The central challenge is therefore not merely to build molecular models that answer accurately, but to build models whose answers are accompanied by explanations that can be inspected, challenged, and connected to chemically meaningful features.

The need for grounded explanation is sharpened as molecular prediction moves from standalone benchmark systems into interactive tools used by both human experts and autonomous agents. In pharmaceutical research, auditable reasoning is necessary for prioritising compounds, diagnosing model failures, and communicating evidence across experimental, computational, and regulatory teams (Jiménez-Luna et al., 2020). At the same time, agentic systems increasingly rely on language models to plan experiments, call tools, and justify decisions (M. Bran et al., 2024; Boiko et al., 2023; Wang et al., 2025); without grounded reasoning, these systems can produce fluent but chemically unsupported chains-of-thought, errors that then propagate silently through every downstream step. Moreover, even frontier models are known to occasionally produce contradictory reasoning (Arcuschin et al., 2026).

Existing approaches tend to address only part of this requirement. Dedicated molecular predictors built on fingerprints, graph neural networks, and molecular foundation models achieve strong

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.

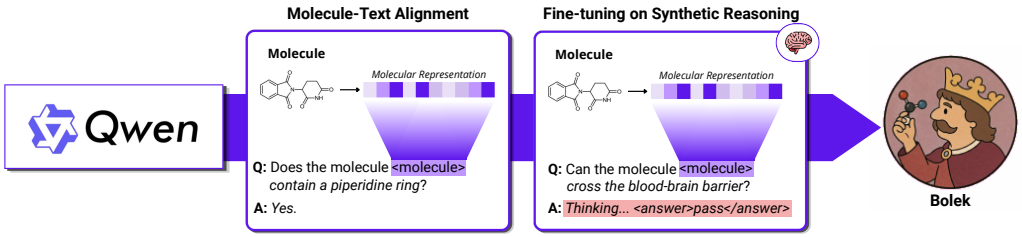

Figure 1: BOLEK is built from Qwen3-4B-Instruct in two supervised stages. A projector maps a fixed-size molecular vector into the token-embedding space at the `<molecule>` position. Alignment teaches the model to answer checkable questions about that molecule; downstream fine-tuning then trains it to predict endpoints, reasoning before it answers.

benchmark performance (Rogers & Hahn, 2010; Yang et al., 2019; Xiong et al., 2020; Xu et al., 2019; Zhou et al., 2023; Méndez-Lucio et al., 2024), but typically expose only a score, leaving the chemist no way to inspect the evidence behind it. General-purpose and chemistry-oriented language models can articulate fluent reasoning, but they struggle to read molecular structure from SMILES strings alone (Bartmann et al., 2026), and their chains-of-thought routinely cite functional groups, properties, or mechanisms not actually present in the input (Zhang et al., 2024; Li et al., 2026a; Guo et al., 2026; Li et al., 2026b; Yu et al., 2024). Tool-calling agents offer a partial remedy by querying external calculators or databases, yet their reasoning remains dependent on orchestration and prompt context rather than on an internal representation that connects molecular evidence to the final decision; recent evaluations show that tool augmentation does not consistently improve over the base LLM on general chemistry tasks, and that the bottleneck lies in the model's own chemical reasoning, not the availability of tools (Yu et al., 2025a). What high-stakes applications need is competitive prediction paired with explanations that are natively in language and verifiable against the molecule. Current approaches largely leave this combination unaddressed.

LLM-native multimodal molecular models (Liu et al., 2024b; Cao et al., 2025; Park et al., 2024; Le et al., 2026) appear well placed to close this gap: by attaching a structural encoder directly to a language model, they combine the structural fidelity of dedicated predictors with the explanatory capacity of LLMs and read the molecule natively, without external tools. Their alignment recipe, however, is largely inherited from image–language pretraining (Radford et al., 2021; Liu et al., 2023a): the molecular token is trained to map to descriptive prose (PubChem captions, IUPAC names), an objective well suited to naming and describing molecules but not to chemical reasoning, which runs on concrete quantitative evidence (logP, polar surface area, descriptor and substructure counts) rather than qualitative description. A multimodal LLM that is to reason about molecules, rather than narrate them, needs an alignment objective built around verifiable chemical features from the start.

Contributions. Our main contributions are the following.

- **BOLEK**, a multimodal language model with a deliberately minimal architecture, a molecular embedding (here a Morgan fingerprint (Rogers & Hahn, 2010)) projected onto a single token in an otherwise unmodified Qwen3-4B-Instruct (Yang et al., 2025a), which solves downstream endpoints by combining supervision from training labels with literature-derived priors, and generalizes to endpoints excluded from training (subsection 3.1);
- A **suite of alignment tasks**, checkable questions of graded difficulty over more than 850,000 molecules, ranging from atoms and bonds, through substructures and computed properties, to descriptions of the whole molecule and of each of its building blocks (subsection 3.2);
- **Synthetic reasoning generation**, a procedure that builds chains-of-thought from literature-derived knowledge, structural decomposition, and the values of task-relevant descriptors, on which the model is then trained to solve downstream tasks (subsection 3.3).

We evaluate BOLEK experimentally and show that (i) trained by supervised fine-tuning alone, it is competitive with a chemistry specialist more than twice its size (subsection 4.2), (ii) the recipe

transfers across molecular embeddings and backbone sizes, so it is not tied to this representation or this model, and (iii) generating a chain-of-thought before the final prediction helps in some settings, its effect depending on both the endpoint and the molecular representation (subsection 4.3).

2 Related Work

We position our contribution at the intersection of three research lines: **text-aligned molecular models** that pair molecular structures with natural-language supervision, **chemistry- and therapeutic-specialized LLMs** that operate purely on strings, and **supervised molecular property and toxicity prediction**, which defines our evaluation surface.

Text-aligned molecular models. Text-aligned molecular work couples structures with natural-language supervision either through shared structure–text spaces or through LLM-native projector bridges. The first group includes KV-PLM (Zeng et al., 2022), MoleculeSTM (Liu et al., 2023b), MoMu (Su et al., 2022), and FineMolTex (Li et al., 2025); the second includes MolCA (Liu et al., 2024b), InstructMol (Cao et al., 2025), LLaMo (Park et al., 2024), 3D-MoLM (Li et al., 2024), BioMedGPT (Luo et al., 2023), GIT-Mol (Liu et al., 2024a), MolX (Le et al., 2026), MR-MoL (Park et al., 2026), and EDT-Former (Jing et al., 2026). Across both groups, alignment is still mostly *molecule-token to caption* (PubChem descriptions, ChEBI-20 captions (Liu et al., 2025), IUPAC names), with only sporadic substructure or property objectives (Le et al., 2026; Park et al., 2024). BOLEK instead pairs a deliberately minimal, fixed Morgan-fingerprint token with feature alignment that asks thousands of checkable questions about molecular structure, descriptors, and substructures.

Chemistry- and therapeutic-specialized LLMs. A second line of work adapts general LLMs to chemistry purely through instruction-tuned text. Galactica (Taylor et al., 2022) trains a decoder transformer on a scientific corpus including SMILES; Mol-Instructions (Fang et al., 2024) provides a biomolecular instruction corpus that subsequent models build on; LlaSMol (Yu et al., 2024) fine-tunes Galactica/Llama-2/Code Llama/Mistral on the SMolInstruct corpus. Tx-LLM (Chaves et al., 2024), fine-tuned from PaLM-2 over 709 datasets covering 66 TDC tasks, achieves near- or exceeding-SOTA performance on $43/66$; its open successor TxGemma (Wang et al., 2025) (Gemma-2; 2B/9B/27B) matches or beats best-in-class on $50/66$ tasks and exceeds specialist models on 26. ChemDFM (Zhao et al., 2025) and BioT5 (Pei et al., 2024) are further data-centric exemplars. None of these models, however, exposes the LLM to a *separate* molecular-embedding modality, which is our central design choice.

Molecular property and toxicity prediction. Our evaluation lives in the MoleculeNet ecosystem (Wu et al., 2018), including Tox21, ToxCast, ClinTox, and SIDER, and the Therapeutics Data Commons (TDC) (Huang et al., 2021), which standardize splits and metrics. Two baseline families dominate. Supervised graph and descriptor models include D-MPNN/Chemprop (Yang et al., 2019), AttentiveFP (Xiong et al., 2020), GIN (Xu et al., 2019), and Morgan fingerprints (Rogers & Hahn, 2010) paired with random forests or XGBoost. Self-supervised pretraining methods include ChemBERTa-2 (Ahmad et al., 2022), MolCLR (Wang et al., 2022), GROVER (Rong et al., 2020), Uni-Mol (Zhou et al., 2023), and MolE (Méndez-Lucio et al., 2024).

Positioning. BOLEK occupies the gap left by these lines: it replaces molecule-as-caption alignment with atomic questions about structure, descriptors, and substructures, posed against a parameter-free Morgan-fingerprint interface (Rogers & Hahn, 2010). Prediction quality competitive with chemistry-specialized LLMs such as TxGemma (Wang et al., 2025) then emerges as a consequence of alignment rather than a separate design target.

3 Method

BOLEK combines three components, shown in Figure 1. The **molecular embedding** is projected onto a single token in the token-embedding space, so the model reads the structure natively rather than from a SMILES string. The **suite of alignment tasks** (Figure 2) teaches it to answer checkable questions about that token, ascending from atoms and bonds through substructures and computed properties to whole molecules and their parts. **Synthetic reasoning generation** (Figure 3) then ties

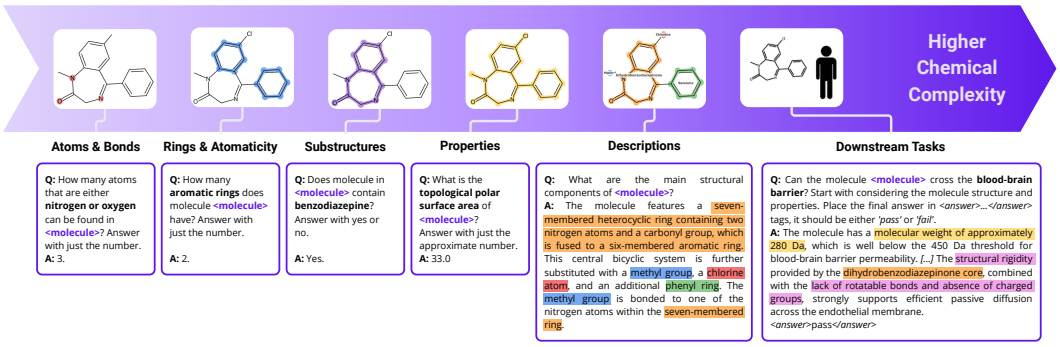

Figure 2: The alignment mixture asks checkable questions about one molecule, ordered by the chemical complexity of the evidence each answer needs. Structured questions have a single verifiable answer; the description and downstream tasks are free text.

those facts to endpoint labels, grounding each rationale in a literature-derived mechanism and in concrete RDKit descriptors relevant for the endpoint. The architecture is minimal and universal, admitting any fixed-size molecular embedding and any base language model, so the two protocols carry over unchanged to other representations and backbones.

3.1 Molecular Embedding

BOLEK starts from Qwen3-4B-Instruct (Yang et al., 2025a) and adds a two-layer projector that turns a fixed-size molecular vector into one additional token at the `<molecule>` position, described in Appendix A.2. Everything else is embedded and decoded as usual, so the language model is otherwise unchanged. We instantiate the vector with 2048-bit, radius=2 Morgan fingerprints (Rogers & Hahn, 2010), which are well established in cheminformatics and carry strong predictive signal for the QSAR-style tasks considered here.

Nothing in the design depends on that choice, so the representation is a swappable input. Table 2 compares Morgan fingerprints with Uni-Mol (Zhou et al., 2023) and Molformer-XL (Ross et al., 2022) pooled embeddings of 512 and 768 dimensions, and with the canonical SMILES string (Weininger, 1988) given in the prompt. SMILES is the control, removing the projector entirely. All four follow the same two-stage training procedure, so differences between them primarily reflect the input representation rather than the instruction-tuning recipe.

3.2 Alignment Tasks

The alignment tasks teach the model to read the molecule token, climbing a ladder of chemical abstraction (see Figure 2), from atoms and bonds, through substructures and computed properties, to descriptions of the whole molecule and of its named building blocks in terms of local physicochemical properties such as lipophilicity, polarity, and protonation. The mixture draws on more than 850,000 molecules from public corpora described in Appendix B.1, asking the model to classify 1,403 substructure patterns and regress 88 RDKit or Mordred descriptors (Landrum et al., 2024; Moriwaki et al., 2018) alongside the whole-molecule and per-part descriptions. The free-text molecule descriptions are therefore grounded in verifiable chemical facts, unlike the image captions used in vision–language alignment.

3.3 Synthetic Reasoning Generation

Downstream supervision then teaches the model to predict what the molecule does in an assay or an organism. The suite has 30 TDC binary classification tasks (Huang et al., 2021), of which 25 are used for training and 5 are held out for evaluation only. The 25 comprise 17 ADMET and bioactivity endpoints, spanning mutagenicity, blood–brain-barrier penetration, bioavailability, intestinal absorption, hERG liability, HIV activity, CYP inhibition and substrate status, clinical toxicity, skin reaction, muscarinic M1 agonism, and SARS-CoV-2 activity, together with 8 Tox21 nuclear-receptor and stress-response assays. Each endpoint is trained in two answer formats, a direct yes/no format

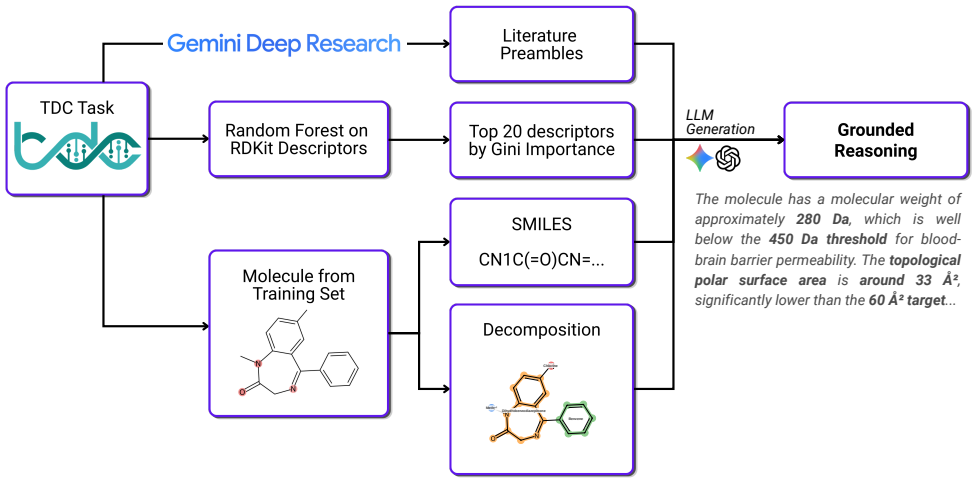

Figure 3: Building a chain-of-thought target for one endpoint. The teacher writes the rationale offline, from inputs BOLEK itself never sees at training time.

that asks only for the binary label, and a chain-of-thought (CoT) format that reasons over structure and properties before emitting the label in `<answer>...</answer>` tags (see representative prompts in Appendix B.5). Training labels are balanced by upsampling the minority class, while validation and test labels follow the TDC scaffold splits unchanged.

The held-out endpoints are chosen one per task family, each for a specified relationship to the trained panel, so that a transfer result can be interpreted and not only measured. They form a ramp of transfer distance. M1 antagonist inverts the polarity of the trained M1 agonist task, CYP2D6 inhibition shares the Veith assay with the trained CYP inhibition tasks and its enzyme with a trained CYP substrate task but belongs to a different gene subfamily with its own pharmacophore, PAMPA NCATS permeability moves from the trained human absorption endpoints to an in vitro assay format, Tox21 PPAR-γ is a lipid-sensing receptor unlike the steroid-binding receptors among the trained Tox21 assays, and Pgp Broccatelli has no trained relative and requires transfer across task families.

Descriptor-grounded reasonings. The chain-of-thought targets are synthetic but chemically constrained. A teacher model receives the molecule's SMILES and its bond-cutting decomposition, the values of task-relevant descriptors, and assay context (see Figure 3). The assay context is a *literature preamble* for the ADMET endpoints and an *assay card* for the Tox21 assays, both mined from the literature as Appendix A describes. The descriptors are the top 20 ranked by a random forest fit on that task's own training split, so the rationale is steered towards the features that carry signal for that endpoint. Tox21 rationales are additionally structured as an adverse-outcome pathway. A rationale is kept only when its tagged final answer matches the ground-truth label, and molecules that never yield one are excluded from chain-of-thought training but still appear in yes/no training. The teacher alone sees these inputs, and BOLEK is trained by teacher forcing on the generated rationale, receiving only the downstream question and the molecule token. The high-throughput screens (Tox21, HIV, M1, SARS-CoV-2) are large and dominated by inactives, so generating a rationale for every molecule would let them swamp the downstream mixture. We therefore cap how many inactives accompany each active, selecting them for structural relevance on the Tox21 assays, with the full selection rules, teacher models, and attempt budgets given in Appendix B.3.

Literature-mined reasonings. A second source of chain-of-thought targets is mined from published literature rather than derived from descriptors, one molecule and one endpoint at a time. An agent searches PubMed and returns a report whose every step cites a source and carries an evidence level, which a second call condenses into a short rationale running from structure to conclusion (Appendix B.4). Mining is restricted to molecules matched to ChEMBL compounds, since anonymous screening compounds carry no literature, and yields 2,214 rationales over 13 ADMET endpoints, about 9% of the chain-of-thought draws on those endpoints.

Table 1: Mean ROC AUC per task group. An underlined cell differs from BOLEK CoT by more than a 95% paired bootstrap interval over the shared test molecules; a plain cell does not. YN = yes/no answer, CoT = chain-of-thought.

Model	Trained on	Format	ADMET (17)	Tox21 (8)	Held-out (5)
Qwen3-4B-Instruct	none	YN	0.559	0.520	0.588
Qwen3-4B-Instruct	none	CoT	0.653	0.557	0.690
GPT-6-Sol[†§]	none	YN	0.687	0.576	0.690
GPT-6-Sol[†]	none	CoT	0.730	0.607	0.714
BOLEK	ADMET + Tox21	YN	0.743	0.747	0.770
BOLEK	ADMET + Tox21	CoT	0.760	0.762	0.799
TxGemma-9B-Chat	all	–	0.714	0.792	0.738
TxGemma-9B-Predict	all	–	0.800	0.899	0.855
Morgan RF	all	–	0.763	0.737	0.812
Descriptor RF	all	–	0.780	0.787	0.822
Uni-Mol	all	–	0.746	0.732	0.757
Molformer-XL	all	–	0.754	0.722	0.791
Chemprop D-MPNN	all	–	0.746	0.748	0.824

[†]Scored from 5 rollouts per molecule rather than 50. [§]Averaged over 16 of the 17 ADMET endpoints: the API refused SARS-CoV-2 Touret in this format.

3.4 TRAINING

Training runs in two supervised stages, 8,250 steps on the alignment tasks alone, then 1,750 steps on downstream supervision alone, in both cases training the full backbone and the projector with the same next-token objective over the answer (Appendix A.3). Every downstream training prompt also carries the assay context, injected with 50% dropout, so the model learns to answer both with and without assay context, and remains usable when none is available. Table 2 compares this recipe against training alignment and downstream supervision together in one stage, and against continuing it with reinforcement learning on answer correctness (Appendix A.4). Appendix D.3 compares backbones from 0.6B to 14B parameters under the same recipe, where downstream gains plateau from 4B while training cost grows roughly linearly, motivating 4B as the main configuration.

4 EXPERIMENTS

4.1 EVALUATION

We evaluate downstream molecular endpoints with ROC AUC for classification and Spearman correlation plus mean absolute error for regression (Huang et al., 2021). All downstream metrics use the corresponding TDC scaffold-split test set unless otherwise noted. For BOLEK and Qwen3, AUC scores are estimated from repeated stochastic generations by parsing final labels, while TxGemma-9B-Chat scores use the answer-choice token logit ratio with the TDC prompts provided in the model release. Unless stated otherwise, evaluation supplies the same assay context the model was trained with, and BOLEK and Qwen3 receive it alike, with Table 2 ablating it on both the training and the inference side. Appendices A and C.5 give the full scoring, rollout, and baseline details, and the TxGemma-9B-Chat protocol.

4.2 PREDICTION QUALITY

Table 1 places BOLEK against other models, grouped by their respective training panel.

BOLEK gains 0.11 on ADMET, 0.21 on Tox21 and 0.11 on held-out over Qwen3-4B-Instruct answering in the same format with the same assay knowledge, significant in every group. GPT-6-Sol, a frontier model given the same molecular context, falls a significant 0.030 behind BOLEK on ADMET, 0.16 on Tox21 and 0.09 on held-out.

Table 2: Ablations of BOLEK, reported as the mean ROC AUC per task group as in Table 1. Each block changes one design choice and repeats BOLEK as its reference row. Variants trained without chain-of-thought are not evaluated on it (–). An underlined cell differs from that reference by more than a 95% paired bootstrap interval over the shared test molecules. All variants are evaluated under the default protocol, except the last block, which varies that protocol on the main model.

	ADMET (17)		Tox21 (8)		Held-out (5)	
Variant	CoT	YN	CoT	YN	CoT	YN
Molecular representation						
Morgan fingerprint (BOLEK)	0.760	0.743	0.762	0.747	0.799	0.770
Uni-Mol	0.743	0.733	0.747	0.751	0.769	0.742
Molformer-XL	0.753	0.747	0.770	0.761	0.775	0.776
SMILES	0.746	0.749	0.742	0.760	0.777	0.736
Alignment stage and supervision						
Alignment, then CoT and YN (BOLEK)	0.760	0.743	0.762	0.747	0.799	0.770
Alignment, then YN only	–	0.716	–	0.646	–	0.640
No alignment, CoT and YN	0.748	0.726	0.766	0.729	0.788	0.720
No alignment, YN only	–	0.683	–	0.634	–	0.636
No alignment, projector only	0.654	0.650	0.668	0.684	0.660	0.639
Training pipeline						
Alignment then downstream (BOLEK)	0.760	0.743	0.762	0.747	0.799	0.770
Alignment then downstream, then GRPO RL	0.759	0.742	0.769	0.724	0.792	0.765
Single joint stage	0.755	0.755	0.770	0.773	0.793	0.784
In-context assay knowledge during training						
Injected with 50% dropout (BOLEK)	0.760	0.743	0.762	0.747	0.799	0.770
Never injected	0.736	0.718	0.719	0.741	0.795	0.794
In-context assay knowledge at inference						
In the prompt (BOLEK)	0.760	0.743	0.762	0.747	0.799	0.770
Absent from the prompt	0.756	0.750	0.763	0.744	0.784	0.744

Against TxGemma-9B-Chat, BOLEK is ahead by a significant 0.046 on ADMET and by 0.061 on the five endpoints that are held out from BOLEK but not from TxGemma, and behind by 0.030 on Tox21.

A similar split appears against the supervised specialists. On the endpoints BOLEK trains on it holds its own: on ADMET it is level with all of them, and on Tox21 it is ahead of Morgan RF, Uni-Mol and Molformer-XL, and behind only Descriptor RF. On the held-out five none of them beats it and Uni-Mol falls behind, though each was refitted on those endpoints' own train splits. Finally, TxGemma-9B-Predict is ahead everywhere. It is the same backbone without a conversational objective, so that gap estimates what BOLEK gives up producing an explanation rather than a score.

Three TDC regression endpoints, none of them trained on, test whether the same interface transfers to a different answer type (Appendix C.2). BOLEK is the only model that ranks plasma protein binding at all, 0.340 Spearman against 0.001 for Qwen3 and -0.001 for TxGemma-9B-Chat, while TxGemma leads on lipophilicity and solubility.

The two answer formats differ by endpoint rather than uniformly (cf. Sprague et al., 2025), as Figure 6 in Appendix C.4 resolves per endpoint. Chain-of-thought is at least as strong as yes/no in every task group, and it gains most exactly where the model has never seen the endpoint: 0.029 on the held-out five, against 0.017 on ADMET and 0.015 on Tox21. Its significant losses concentrate on the Veith CYP-inhibition tasks, where abundant labels let direct training learn finer substructure evidence.

4.3 ABLATIONS

The representation variants each repeat the alignment stage against their own encoder, so every downstream stage begins from a checkpoint aligned to its representation.

Molecular representation. No representation separates from the Morgan fingerprint in its own favour on any task group. Uni-Mol falls behind it in four cells and SMILES in three, both held-out cells among them in each case, and Molformer-XL trails only on held-out chain-of-thought. Where the learned encoders lead, on Tox21, they lead by 0.014 or less and no interval excludes zero. The fingerprint is also the cheapest of the four, computed without a learned encoder, and it keeps the substructure vocabulary of the alignment stage interpretable.

Alignment stage and supervision. The alignment stage and chain-of-thought supervision both raise downstream ROC AUC, and their gains add up: BOLEK, trained with both, scores highest in every yes/no column of the block, and no variant scores significantly above it in any column. Chain-of-thought supervision is the larger of the two; removing the chain-of-thought traces from downstream training lowers ROC AUC most on Tox21 and on the held-out five, with or without the alignment stage. The alignment stage shows its effect most clearly when downstream training uses yes/no questions only: the aligned model then scores higher than a model trained from Qwen3-4B without alignment on the ADMET and Tox21 endpoints seen during training, significantly so on ADMET. When the chain-of-thought traces are present, removing the alignment stage lowers ROC AUC less, because the traces restate alignment-stage properties such as polar surface area. Training only the projector scores significantly lower than BOLEK in every column, so the language model itself has to adapt to the molecular input. Appendix D.1 resolves this block per task.

Training pipeline. Continuing the supervised model with GRPO reinforcement learning (Shao et al., 2024) trades one Tox21 format for the other, 0.022 off yes/no and 0.007 onto chain-of-thought, and changes nothing on ADMET or the held-out five. Training both mixtures together in a single stage is ahead on Tox21 in both formats, by 0.008 on chain-of-thought and 0.027 on yes/no, and is level everywhere else. We report the two-stage model because it leads on chain-of-thought for ADMET and for the held-out five, by 0.005 to 0.006 without separating, and treat the single-stage recipe as an open question rather than a settled one.

Assay knowledge during training. Injecting cards and preambles during training, with 50% dropout, is what makes the model able to use them at all. It is worth 0.025 on ADMET chain-of-thought, 0.025 on ADMET yes/no and 0.042 on Tox21 chain-of-thought. The gain does not reach every group: the never-injected variant is ahead by 0.024 on held-out yes/no, and the remaining cells are level.

Assay knowledge at inference. Supplying the same knowledge at inference matters most where the model has no trained-in knowledge of the endpoint. It is worth 0.026 on held-out yes/no, the only cell that separates, against 0.014 on held-out chain-of-thought and nothing on any trained group. This is the expected shape if the preamble supplies assay context that training already provides for the trained endpoints.

5 DISCUSSION AND LIMITATIONS

Molecular alignment exposes reusable signal. The main result of this study is that molecular alignment can turn a general instruction-tuned language model into a stronger molecular predictor while preserving its ability to produce natural-language rationales. BOLEK improves substantially over Qwen3 on downstream binary classification, transfers to held-out endpoints, and remains competitive with stronger chemistry-oriented baselines on several tasks. Some endpoints improved more than others. This may reflect how much can be predicted from molecular structure and descriptors alone, and how much depends on assay conditions or biological factors that the model does not observe.

Reasoning supervision is still imitation rather than verification. The chain-of-thought stage teaches BOLEK to organize molecular evidence into task-specific reasoning, but it does not directly optimize the truth of each intermediate statement. That truth matters, because the answer follows the rationale. When its rationale is replaced by one written for another molecule of the same endpoint, with the verdict removed, BOLEK gives the label that rationale argues for over 98% of test molecules (Appendix D.4). This limitation is visible in the mixed effect of chain-of-thought supervision: it helps on some small-data pharmacophore-like tasks, but can hurt when the supervised

rationale prior is coarser than the signal available in the training labels. Because CoT targets are kept only when the teacher reaches the ground-truth label, the CoT subset can also be biased toward examples that are easier for the teacher, making more complete and less selective CoT generation an important direction for improving reasoning supervision. Consistent with this, continuing the supervised model with GRPO on answer correctness alone trades one answer format for the other rather than improving either overall (Table 2), which suggests the useful reward signal is not the final label. Rewards that check the rationale itself are a natural next step, and in chemistry much of a rationale can be checked exactly (Guo et al., 2026): rewarding a correct answer only when its rationale contains no fabricated molecular claims reduces fabrication without costing accuracy (Li et al., 2026b). Imitation nonetheless carries further than the endpoints it was written for: the gain from reasoning is largest on the five held-out endpoints (Figure 6), where the model has no endpoint-specific training signal to fall back on.

The molecular projector is intentionally minimal. BOLEK uses a single projected Morgan fingerprint token, which makes the architecture simple and the contribution easy to isolate. However, Morgan fingerprints (Rogers & Hahn, 2010) are also a bottleneck: they compress molecular structure into hashed local environments and discard information that may matter for stereochemistry, conformation, long-range geometry, target binding, quantum effects, and assay-specific context. The same interface accepts richer fixed-size embeddings, and Table 2 already swaps in two learned ones: on the task groups measured so far a Molformer-XL embedding is ahead of the fingerprint, so the representation rather than the interface looks like the binding constraint. Multi-token representations (Jing et al., 2026), protein-conditioned representations, and ensembles that combine several molecular views remain open (Li et al., 2024; Liu et al., 2024b; Park et al., 2024; Méndez-Lucio et al., 2024).

Native multimodality and tool use are complementary. Tool-calling systems can compute exact descriptors, retrieve external facts, and verify specific claims (M. Bran et al., 2024; Boiko et al., 2023), while BOLEK places compact molecular evidence directly inside the language model. The broader point is that competitive prediction and auditable reasoning need not trade off against each other. BOLEK is one step toward that combination: it improves prediction, exposes reusable molecular signal on held-out endpoints, and answers in a form a chemist can interrogate rather than simply accept.

6 CONCLUSION

BOLEK pairs a molecular embedding projected onto a single token with a suite of alignment tasks that ground that token in checkable chemical facts, and with synthetic reasoning generation that supervises its chains-of-thought. Alignment allows the model to apply its literature priors to a molecule it reads directly, and the results show BOLEK combining both sources of knowledge, the training labels and those priors, and generalizing to endpoints excluded from training. With a 4B backbone it achieves performance comparable to the evaluated TxGemma-9B-Chat, with higher mean scores on the ADMET endpoints, similar results on the five held-out endpoints, and lower scores on Tox21.

The benefit of answering in chain-of-thought rather than yes/no depends on the endpoint. It is largest on the five endpoints the model never trained on. One possible explanation is that a rationale lets the model fall back on general chemical reasoning where it has no endpoint-specific training signal to draw on. The ablation results are consistent with this interpretation.

We also observed smaller chain-of-thought gains on larger datasets. The clearest losses occurred on four of the five Veith CYP-inhibition tasks, while HIV improved despite being one of the largest datasets. This pattern suggests that the amount of available supervision may influence the benefit of a rationale. In our panel, size varies together with assay type and with the rationale-generation procedure, so its contribution cannot be cleanly separated.

These results show that a simple molecular projector and supervised training can support competitive prediction in a compact language model. They also show that generating a rationale can help some tasks and hurt others. Understanding these differences, and evaluating the chemical correctness of the rationales themselves, are the next steps.

REPRODUCIBILITY STATEMENT

To support reproducibility, the supplementary material includes two release components. First, it includes metric-reproduction code that downloads the stored per-molecule predictions of every model in Table 1 and Table 2 and recomputes both tables, including their significance marks. The stored predictions include the TxGemma-9B scores, so the TxGemma weights are not redistributed. Second, it includes BOLEK inference code for local vLLM execution of the released checkpoint. A script in it re-runs the evaluation of any single task with the prompt, sampling settings, and answer parsing used for the reported results.

During anonymous review, the model weights and evaluation artifacts are provided through anonymized private repositories with a read-only reviewer access token. After the anonymity period, the BOLEK model weights, inference code, evaluation artifacts, and metric-reproduction scripts will be made available under the artifact-specific terms stated in each file header or accompanying documentation. Users are required to comply with all applicable terms, including those of upstream dependencies. In particular, Qwen3-4B-Instruct-2507 is subject to the Apache License 2.0, while the authors' proprietary BOLEK model weights and model-inference code are subject to separate terms restricting use to research purposes only. Obtaining access to any artifact does not grant rights beyond those explicitly stated in its license, file header, or accompanying documentation.

ETHICS STATEMENT

Models that improve molecular prediction and explanation can support beneficial work in drug discovery, toxicology, and chemical safety, but the same capabilities can also be misused. Prior work has shown that AI systems developed to avoid toxicity in therapeutic design can be redirected toward the search for harmful molecules (Urbina et al., 2022). BOLEK is not a de novo molecular generator and does not provide synthesis routes, but stronger chemical reasoning could still lower barriers for prioritizing, interpreting, or optimizing molecules with hazardous biological activity when combined with external generation, retrosynthesis, or laboratory services. This risk is especially relevant for models that explain predictions in chemically meaningful terms, because explanations can make harmful optimization more legible as well as make beneficial screening more auditable. Responsible release should therefore consider access controls for high-capability endpoints, monitoring for misuse-oriented queries, documentation of intended and prohibited uses, and evaluation protocols that include dual-use chemical safety alongside predictive accuracy and groundedness.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets, to clean and reformat data, to implement methods, and to design and give feedback on experiments. The first three are components of the method rather than aids to reporting it: the synthetic chain-of-thought targets of subsection 3.3, the per-part decomposition supervision of the alignment mixture, and the literature preambles and assay cards are all model-generated, by the systems named in Appendix A and listed with their terms of use in Table 22. Generative AI additionally assisted in writing the training, evaluation and analysis code, in choosing and reviewing analyses such as the paired bootstrap procedure of subsection 4.1, and in summarising differences across large sets of per-endpoint results. We have not used generative AI tools to interpret those results, which the authors did themselves, nor to propose or refine hypotheses, to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide ingredients for or write proofs, to assist with translation, or to support qualitative and thematic data analysis; the remaining tasks requiring disclosure are not applicable to this work.

Additionally, we used generative AI tools to create and modify scientific figures, to create and edit software code, to draft and edit parts of this paper for readability, to search for and summarise existing literature, and to format references.

We have reviewed all AI-assisted work. Every AI-assisted change was directed by an author and reviewed by at least one other author before it entered the repository, under a commit policy that records the level of AI involvement on each commit. All numbers reported in this paper are regenerated from stored model outputs by the scripts released with it rather than written into the text by a

model, and the script that produces each table is named in its caption. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jakub Adamczyk, Jakub Poziemski, Franciszek Job, Mateusz Król, and Maciej Makowski. Molpile – large-scale, diverse dataset for molecular representation learning, 2025. URL https://arxiv.org/abs/2509.18353.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models, 2022. URL https://arxiv.org/abs/2209.01712.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. In *ICML*, 2026.

Alex M. Aronov. Predictive in silico modeling for herg channel blockers. *Drug Discovery Today*, 10(2):149–155, January 2005. ISSN 1359-6446. doi: 10.1016/s1359-6446(04)03278-7. URL http://dx.doi.org/10.1016/S1359-6446(04)03278-7.

Christoph Bartmann, Johannes Schimunek, Mykyta Ielanskyi, Philipp Seidl, Günter Klambauer, and Sohvi Luukkonen. MolecularIQ: Characterizing chemical reasoning capabilities through symbolic verification on molecular graphs, 2026. URL https://arxiv.org/abs/2601.15279.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0. URL http://dx.doi.org/10.1038/s41586-023-06792-0.

Fabio Broccatelli, Emanuele Carosati, Annalisa Neri, Maria Frosini, Laura Goracci, Tudor I. Oprea, and Gabriele Cruciani. A novel approach for predicting p-glycoprotein (abcb1) inhibition using molecular interaction fields. *Journal of Medicinal Chemistry*, 54(6):1740–1751, February 2011. ISSN 1520-4804. doi: 10.1021/jm101421d. URL http://dx.doi.org/10.1021/jm101421d.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. InstructMol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 354–379, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.25/.

Miriam Carbon-Mangels and Michael C. Hutter. Selecting relevant descriptors for classification by bayesian estimates: A comparison with decision trees and support vector machines approaches for disparate data sets. *Molecular Informatics*, 30(10):885–895, September 2011. ISSN 1868-1751. doi: 10.1002/minf.201100069. URL http://dx.doi.org/10.1002/minf.201100069.

Juan Manuel Zambrano Chaves, Eric Wang, Tao Tu, Eeshit Dhaval Vaishnav, Byron Lee, S. Sara Mahdavi, Christopher Semturs, David Fleet, Vivek Natarajan, and Shekoofeh Azizi. Tx-llm: A large language model for therapeutics, 2024. URL https://arxiv.org/abs/2406.06316.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994. ISBN 9780429246593. doi: 10.1201/9780429246593.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, November 2018. ISSN 0893-6080. doi: 10.1016/j.neunet.2017.12.012. URL http://dx.doi.org/10.1016/j.neunet.2017.12.012.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models, 2024. URL https://arxiv.org/abs/2306.08018.

Hongyu Guo, Hao Li, He Cao, Gongbo Zhang, and Li Yuan. From answers to states: Verifiable process-level evaluation of chemical reasoning in large language models, 2026. URL https://arxiv.org/abs/2606.03660.

Tingjun Hou, Junmei Wang, Wei Zhang, and Xiaojie Xu. Adme evaluation in drug discovery. 7. prediction of oral absorption by correlation and classification. *Journal of Chemical Information and Modeling*, 47(1):208–218, January 2007. ISSN 1549-960X. doi: 10.1021/ci600343x. URL http://dx.doi.org/10.1021/ci600343x.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development, 2021. URL https://arxiv.org/abs/2102.09548.

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, 12 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00675. URL https://doi.org/10.1021/acs.jcim.0c00675.

José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, October 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00236-4. URL http://dx.doi.org/10.1038/s42256-020-00236-4.

Zihao Jing, Qiuhao Zeng, Ruiyi Fang, Yan Sun, Boyu Wang, and Pingzhao Hu. Entropy-guided dynamic tokens for graph-LLM alignment in molecular understanding, 2026. URL https://arxiv.org/abs/2602.02742.

Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1):312–320, January 2005. ISSN 1520-4804. doi: 10.1021/jm040835a. URL http://dx.doi.org/10.1021/jm040835a.

Thierry Kogej, Christos Kannas, Samuel Genheden, Eike Caldeweyher, and Mikhail Kabeshov. Smarts-rx: a smarts-based representation of chemical functions for reactivity analysis. *Journal of Cheminformatics*, 17(1):177, 2025. doi: 10.1186/s13321-025-01136-8. URL https://doi.org/10.1186/s13321-025-01136-8.

Gregory Landrum et al. RDKit: Open-source cheminformatics, 2024. URL https://www.rdkit.org. Release 2024.03.1.

Khiem Le, Zhichun Guo, Kaiwen Dong, Xiaobao Huang, Bozhao Nan, Roshni Iyer, Xiangliang Zhang, Olaf Wiest, Wei Wang, Ting Hua, and Nitesh V. Chawla. Molx: Enhancing large language models for molecular understanding with a multi-modal extension, 2026. URL https://arxiv.org/abs/2406.06777.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations, 2026a. URL https://arxiv.org/abs/2505.21318.

Jiatong Li, Yuxuan Ren, Weida Wang, Xiaoyong Wei, and Yatao Bian. Chemical chain-of-thought functions as a hallucination-prone molecular scratchpad, 2026b. URL https://arxiv.org/abs/2607.20935.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3d molecule-text interpretation in language models, 2024. URL https://arxiv.org/abs/2401.13923.

Yibo Li, Yuan Fang, Mengmei Zhang, and Chuan Shi. Advancing molecular graph-text pre-training via fine-grained alignment. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 1589–1599. ACM, August 2025. doi: 10.1145/3711896.3736834. URL http://dx.doi.org/10.1145/3711896.3736834.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a. URL https://arxiv.org/abs/2304.08485.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171:108073, 2024a. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2024.108073. URL https://www.sciencedirect.com/science/article/pii/S0010482524001574.

Pengfei Liu, Jun Tao, and Zhixiang Ren. A quantitative analysis of knowledge-learning preferences in large language models in molecular science, 2025. URL https://arxiv.org/abs/2402.04119.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023b. doi: 10.1038/s42256-023-00759-6. URL https://doi.org/10.1038/s42256-023-00759-6.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter, 2024b. URL https://arxiv.org/abs/2310.12798.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine, 2023. URL https://arxiv.org/abs/2308.09442.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8. URL http://dx.doi.org/10.1038/s42256-024-00832-8.

Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics*, 10(1):4, 2018. doi: 10.1186/s13321-018-0258-y. URL https://doi.org/10.1186/s13321-018-0258-y.

Oscar Méndez-Lucio, Christos A. Nicolaou, and Berton Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1), November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53751-y. URL http://dx.doi.org/10.1038/s41467-024-53751-y.

Hassan Pajouhesh and George R. Lenz. Medicinal chemical properties of successful central nervous system drugs. *NeuroRX*, 2(4):541–553, October 2005. ISSN 1545-5343. doi: 10.1602/neurorx.2.4.541. URL http://dx.doi.org/10.1602/neurorx.2.4.541.

George Papadatos, Mark Davies, Nathan Dedman, Jon Chambers, Anna Gaulton, James Siddle, Richard Koks, Sean A. Irvine, Joe Pettersson, Nicko Goncharoff, Anne Hersey, and John P. Overington. Surechembl: a large-scale, chemically annotated patent document database. *Nucleic Acids Research*, 44(D1):D1220–D1228, 01 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1253. URL https://doi.org/10.1093/nar/gkv1253.

Jinyoung Park, Minseong Bae, Dohwan Ko, and Hyunwoo J Kim. Llamo: Large language model-based molecular graph assistant. In *NeurIPS*, 2024.

Junwoo Park, Minyoung Shin, Cheol Soon Lee, and Sujee Lee. Multi-granular rationale-guided molecular LLM for property prediction, 2026. URL https://arxiv.org/abs/2608.10480.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations, 2024. URL https://arxiv.org/abs/2310.07276.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://arxiv.org/abs/2103.00020.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, April 2010. ISSN 1549-960X. doi: 10.1021/ci100050t. URL http://dx.doi.org/10.1021/ci100050t.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data, 2020. URL https://arxiv.org/abs/2007.02835.

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, December 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00580-7. URL http://dx.doi.org/10.1038/s42256-022-00580-7.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *ICLR*, 2025.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. A molecular multimodal foundation model associating molecule graphs with natural language, 2022. URL https://arxiv.org/abs/2209.05481.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022. URL https://arxiv.org/abs/2211.09085.

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, March 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00465-9. URL http://dx.doi.org/10.1038/s42256-022-00465-9.

Daniel F. Veber, Stephen R. Johnson, Hung-Yuan Cheng, Brian R. Smith, Keith W. Ward, and Kenneth D. Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12):2615–2623, May 2002. ISSN 1520-4804. doi: 10.1021/jm020017n. URL http://dx.doi.org/10.1021/jm020017n.

Henrike Veith, Noel Southall, Ruili Huang, Tim James, Darren Fayne, Natalia Artemenko, Min Shen, James Inglese, Christopher P Austin, David G Lloyd, and Douglas S Auld. Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries. *Nature Biotechnology*, 27(11):1050–1055, October 2009. ISSN 1546-1696. doi: 10.1038/nbt.1581. URL http://dx.doi.org/10.1038/nbt.1581.

Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. Txgemma: Efficient and agentic llms for therapeutics, 2025. URL https://arxiv.org/abs/2504.06196.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287, March 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00447-x. URL http://dx.doi.org/10.1038/s42256-022-00447-x.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, February 1988. ISSN 1520-5142. doi: 10.1021/ci00057a005. URL http://dx.doi.org/10.1021/ci00057a005.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 01 2018. ISSN 2041-6520. doi: 10.1039/c7sc02664a. URL https://doi.org/10.1039/c7sc02664a.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, August 2020. ISSN 1520-4804. doi: 10.1021/acs.jmedchem.9b00959. URL http://dx.doi.org/10.1021/acs.jmedchem.9b00959.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 08 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00237. URL https://doi.org/10.1021/acs.jcim.9b00237.

Zaifei Yang, Hong Chang, Ruibing Hou, Shiguang Shan, and Xilin Chen. Knowmol: Advancing molecular large language models with multi-level chemical knowledge, 2025b. URL https://arxiv.org/abs/2510.19484.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset, 2024. URL https://arxiv.org/abs/2402.09391.

Botao Yu, Frazier N. Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning, and Huan Sun. Chemtoolagent: The impact of tools on language agents for chemistry problem solving, 2025a. URL https://arxiv.org/abs/2411.07228.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025b. URL https://arxiv.org/abs/2503.14476.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13(1):862, 2022. doi: 10.1038/s41467-022-28494-3. URL https://doi.org/10.1038/s41467-022-28494-3.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. Chemllm: A chemical large language model, 2024. URL https://arxiv.org/abs/2402.06852.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Kai Yu, and Xin Chen. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4):102523, April 2025. ISSN 2666-3864. doi: 10.1016/j.xcrp.2025.102523. URL http://dx.doi.org/10.1016/j.xcrp.2025.102523.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. *ChemRxiv*, 2023(0307), 2023. doi: 10.26434/chemrxiv-2022-jjm0j-v4. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2022-jjm0j-v4.

S.-F. Zhou, Z.-W. Zhou, L.-P. Yang, and J.-P. Cai. Substrates, inducers, inhibitors and structure-activity relationships of human cytochrome p450 2c9 and implications in drug development. *Current Medicinal Chemistry*, 16(27):3480–3675, September 2009. ISSN 0929-8673. doi: 10.2174/092986709789057635. URL http://dx.doi.org/10.2174/092986709789057635.

A METHOD

A.1 EVALUATION PROTOCOL

We quantify test-set sampling uncertainty with a 95% percentile bootstrap over test molecules, using 2,000 resamples (Efron & Tibshirani, 1994). Classification resamples are stratified by label, drawing positives and negatives to their original counts, while regression resamples are unstratified. Point estimates are always computed on the full test set, and per-molecule scores are held fixed across resamples, so the intervals exclude generation noise, which Appendix C.3 characterizes separately. Model comparisons are paired, scoring both models on the same resamples and restricting to the molecules they share, and a difference is called significant when its 95% interval excludes zero. Means over tasks average the per-task bootstrap distributions, so their intervals reflect molecule sampling within a fixed set of endpoints.

Alignment binary tasks are evaluated by accuracy, and alignment regression tasks by Pearson correlation. Downstream classification tasks use ROC AUC, and downstream regression tasks use Spearman correlation and mean absolute error. Unless stated otherwise, all downstream metrics are computed on the corresponding TDC scaffold-split test set.

For TxGemma-9B-Chat, we use the TDC prompts provided in the model release and compute classification scores from the answer-choice token logit ratio. For BOLEK and Qwen3, CoT AUC scores are estimated from repeated stochastic generations because the final-answer probability is conditioned on the generated rationale and can saturate after the model has committed to an answer. This rollout fraction is a Monte Carlo estimate of the model's positive-answer probability, analogous in the limit to direct probability or logit scoring when those scores are accessible. We run 50 generations per molecule at temperature 0.6, parse the final label, and use the fraction of parseable positive labels as the score. For regression endpoints, we parse the numeric prediction from the generated answer and compare it with the ground-truth value in the original TDC units. We use a 300-tree random forest trained on RDKit descriptors (Landrum et al., 2024) to rank per-task descriptors for CoT construction.

A.2 PROJECTOR

A two-layer projector g_ϕ, with SiLU activation (Elfwing et al., 2018) and hidden width twice the Qwen3 embedding dimension, maps the molecular vector $\mathbf{m}$ into the token-embedding space:

$$g_\phi(\mathbf{m}) = W_2\,\mathrm{SiLU}(W_1\mathbf{m} + b_1) + b_2,$$
$$\mathbf{e}_i = \begin{cases} g_\phi(\mathbf{m}) & \text{if } x_i = \texttt{<molecule>}, \\ E(x_i) & \text{otherwise.} \end{cases}$$

The mixed sequence is then processed by the standard causal decoder, leaving the language-model architecture otherwise unchanged.

A.3 TRAINING CONFIGURATION

All variants are optimized with the same autoregressive instruction-tuning objective in both stages. Prompt tokens are masked from the loss and supervision is applied only to assistant responses, so alignment tasks, direct yes/no downstream tasks, and chain-of-thought downstream tasks are all trained as next-token prediction over the answer. We train the full Qwen3 backbone in every variant except the projector-only ablation of Table 2, and the projector alongside it wherever the molecule enters as a vector. The downstream stage does not replay alignment data.

Table 3: Training configuration for the two supervised stages of BOLEK. The ablation variants of Table 2 keep these settings and change only the aspect under test.

	Alignment	Downstream
Initialized from	Qwen3-4B-Instruct	alignment checkpoint
Steps	8,250	1,750
Maximum sequence length	256	1,280
Effective batch size	256	256
Backbone learning rate	5e−6	5e−6
Projector learning rate	5e−5	5e−5
Schedule	linear decay, no warmup, zero weight decay	
Precision	bf16 fully sharded data parallelism	
Hardware	4×H100	8×H100
Wall time	≈5 h	≈2.5 h

The longer downstream sequence length accommodates chain-of-thought targets, which the yes/no format does not need.

A.4 REINFORCEMENT LEARNING

The reinforcement-learning variant of Table 2 continues the supervised model rather than replacing it. It starts from the downstream checkpoint and runs 2,000 steps of group-relative policy optimization (Shao et al., 2024) with the DAPO loss (Yu et al., 2025b), as implemented in TRL, generating with vLLM. Only the chain-of-thought targets are optimized; the yes/no format is never trained, which is worth keeping in mind when reading that variant's yes/no column. Prompts are balanced by label within each endpoint and equalized across the 25 of them, capped by the smallest available label pool, so neither a class nor an endpoint dominates a batch.

The reward is dominated by answer correctness. A completion scores 1 when the label inside its <answer> tags matches the ground truth and 0 otherwise, at weight 1.0. A second term scores 0.5 for a parseable <answer> tag, at weight 0.1, which keeps the output format from drifting. A third term, which would penalize a completion that opens with <answer> and so skips the rationale entirely, carries weight 0 in this run. No term inspects the content of the rationale, so this variant tests answer correctness alone.

Table 4: Reinforcement-learning configuration for the ablation variant of Table 2.

Setting	Value
Initialized from	downstream checkpoint, step 1,750
Steps	2,000
Generations per prompt	8
Generation batch size	256
Maximum completion length	512
Sampling temperature	0.7
Learning rate	`5e-7`
KL coefficient	0.01
Effective batch size	64
Gradient-norm clipping	1.0
Reward weights (accuracy / format / rationale)	1.0 / 0.1 / 0.0

B Data

B.1 Alignment Data

Alignment exposes the molecule to the model through natural-language question–answer exchanges. The molecule pool combines four sources: 700,000 sampled molecules from MolPILE (Adamczyk et al., 2025), 90,000 KnowMol molecules with multi-level structural and property annotations (Yang et al., 2025b), 26,402 ChEBI-20-MM molecules with natural-language descriptions (Liu et al., 2025), and a name-prediction set with 32,936 molecules from an RCSB PDB-derived curated docking library (Berman et al., 2000) plus 796 ZINC20 molecules (Irwin et al., 2020). Each alignment task uses 5–20 paraphrased templates; representative examples are shown in Appendix B.5.

Free-text tasks include whole-molecule descriptions, aspect-specific descriptions of decomposed molecular parts, molecule naming, and SMILES recovery from the fingerprint. For the eight decomposition tasks, each molecule is decomposed into named structural parts, each part is annotated with local properties such as logP, hydrogen-bond donors and acceptors, stereochemistry, polarity, protonation, and partial charge, and Gemini 2.5 Flash generates the supervised answer from this annotated decomposition.

Substructure classification asks whether a molecule contains a named pattern. The vocabulary contains 534 simple atom-and-bond patterns, 153 MACCS keys, 73 RDKit fragment descriptors (Landrum et al., 2024), and 643 textbook or alert patterns assembled from RDKit functional-group definitions, RDKit hierarchical filters, SureChEMBL structural alerts (Papadatos et al., 2016), and SMARTS-RX (Kogej et al., 2025). Items are partitioned into seen and unseen subsets, and labels are balanced through targeted molecule sampling. Descriptor prediction covers 88 RDKit and Mordred descriptors (Moriwaki et al., 2018), including atom and bond counts, ring counts, molecular weight, and logP.

B.2 Molecule Decomposition

The decomposition alignment tasks use a deterministic RDKit-based preprocessing step (Landrum et al., 2024) to break each molecule into chemically named building blocks: scaffolds, linkers, functional groups, and substituents. These structured notes are then fed to a large language model to generate natural-language molecule descriptions. The supervision for tasks such as `describe this molecule` is therefore based on a deterministic molecular breakdown and local molecular features, rather than on an unconstrained autoregressive description of the input molecule. Given a parsed molecule, the pipeline first collects candidate bonds, i.e., bonds that are plausible split points. These candidates come from five sources: the Murcko scaffold boundary, ring boundaries, the boundary of the union of Ertl functional-group atoms, BRICS retrosynthetic bonds, and boundaries of heteroatom-containing functional-group SMARTS matches. To avoid cutting through chemically meaningful entities, the pipeline then excludes protected bonds from the cut set: aromatic bonds, bonds inside rings with at most 12 atoms, bonds that would separate exocyclic atoms retained with the ring scaffold, and bonds internal to a functional-group match unless that bond is also a scaffold

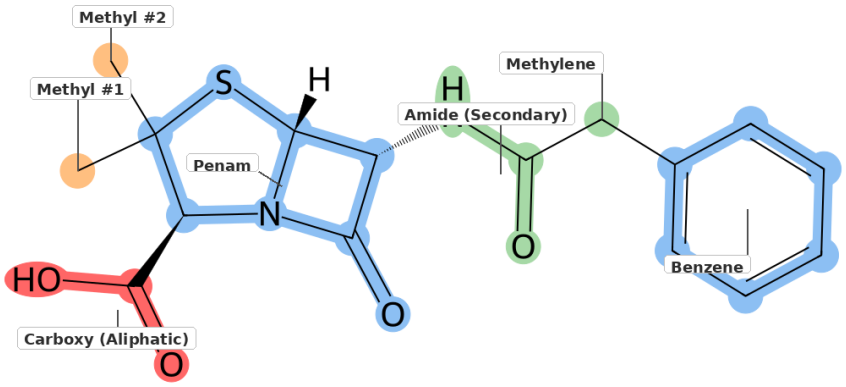

Figure 4: Example bond-cutting decomposition for Penicillin G. Blue highlights ring or ring-system parts, green highlights linker parts, red highlights a functional group, and orange highlights substituents.

or ring boundary. The remaining candidate bonds are cut with `FragmentOnBonds`; each part keeps its canonical fragment SMILES, original atom indices, and cross-fragment connection atoms.

The specification passed to the language model includes a type annotation for every molecular part. These types make the decomposition readable as chemistry rather than as an unlabelled list of fragments. To make the cuts sensible, the typing rules use boundary conditions derived from the Murcko scaffold, small rings, macrocycles, Ertl functional groups, and connection count. Atoms outside the scaffold become functional groups when they are fully covered by an Ertl functional group and otherwise become substituents, or chains when no scaffold exists. Atoms inside the scaffold become rings, macrocycles, linkers, exocyclic groups, or ring systems depending on whether they are contained in small rings, large rings, or non-ring scaffold paths. Mixed fragments that cross these boundaries are kept as generic fragments, which exposes failures of the cutting rules instead of silently assigning them to a chemically precise class.

Each typed part is named by a fixed priority cascade: lookup in a local canonical-SMILES structure library, exact match against a local functional-group SMARTS library, single-atom element name with charge label, ring summary by size, aromaticity, heteroatom content, and charge, and finally a carbon-scaffold heuristic such as `Methyl`, `Methylene`, or a branched alkyl label. The structure library is a curated table of common names for frequent ring systems and scaffolds, using many PubChem-derived names but with manual replacements for names that are too verbose or chemically unhelpful. The SMARTS library is separate from the Ertl functional-group detector: it is a local, ordered set of patterns adapted from RDKit's functional-group hierarchy and extended with custom patterns for decomposition-specific protection and naming. The formatted notes are ordered by a depth-first traversal that starts from the largest ring-like part and keeps decorations near their parent. When the molecule contains multiple instances of the same named part, such as two methyl substituents on the same scaffold, the notes distinguish them with `#1`, `#2`, and so on. For every part, the notes list the fragment name, type, SMILES, and either ring substituents or non-ring connections; task-specific annotations add only the local feature needed for that alignment task, namely hydrogen-bond acceptors, hydrogen-bond donors, logP contribution, polar surface area, protonation at pH 7.4, stereochemical sites, or Gasteiger partial charges. Gemini 2.5 Flash receives these structured notes for eight alignment datasets—structure, hydrogen-bond acceptors, hydrogen-bond donors, lipophilicity, polarity, protonation, stereochemistry, and partial charges—and generates the supervised assistant response used in the training mixture.

B.3 SYNTHETIC COT CONSTRUCTION

The downstream panel contains 25 trained TDC binary classification tasks (Huang et al., 2021): AMES, BBB Martins, Bioavailability Ma, HIA Hou, hERG, HIV, ClinTox, skin reaction, M1 agonist, SARS-CoV-2 Touret, four Veith CYP-inhibition tasks, three Carbon-Mangels CYP-substrate tasks, and eight Tox21 assays. The five held-out endpoints are CYP2D6 Veith, Pgp Broccatelli, PAMPA NCATS, M1 antagonist, and Tox21 PPAR-γ. Each task is trained in direct yes/no and CoT formats. Training labels are balanced by upsampling the minority class, while validation and test sets keep the TDC scaffold splits unchanged.

CoT targets are synthesized rather than human-written. The generation prompt contains the canonical SMILES, the same bond-cutting decomposition used in alignment, and the top 20 RDKit descriptors ranked by Gini importance from a 300-tree random forest fit on the task training split. The ADMET endpoints add the task-specific literature preamble of subsection 3.1, whose label criterion follows the TDC-linked source paper (Kazius et al., 2005; Pajouhesh & Lenz, 2005; Veber et al., 2002; Veith et al., 2009; Zhou et al., 2009; Carbon-Mangels & Hutter, 2011; Hou et al., 2007; Wu et al., 2018; Broccatelli et al., 2011; Aronov, 2005); the Tox21 assays add the assay card instead and require the rationale to follow an adverse-outcome-pathway chain from molecular initiating event to assay outcome.

Both are mined by a Gemini Deep Research agent that surveys the literature for mechanistic priors translating the assay into reasoning over molecular structure, after which Claude Opus 4.6 condenses the report into approximately 600 tokens, to keep training and inference affordable. The Tox21 assays share one qHTS reporter design, so their reports condense further into a fixed schema of roughly 400 tokens: target family, assay and label definition, biology, ligand chemotypes, and reporter-artifact caveats. The teacher generates a rationale ending with the task label inside `<answer>...</answer>` tags: Gemini 3.8 Flash for the ADMET endpoints, Gemini 3.5 Flash for Tox21. These inputs are used only to construct the supervised CoT targets and are not included in the BOLEK training prompt.

Molecule selection differs between the two groups, and is a property of the generation recipe rather than of the endpoint type: the other high-throughput screens in the panel, including HIV, the M1 muscarinic assays, and SARS-CoV-2, follow the ADMET recipe. For the ADMET endpoints, molecules are deduplicated by InChIKey and keys carrying contradictory labels are dropped, after which the majority class is strided down to at most 40 molecules per minority-class molecule. Among the endpoints this model trains on, the cap bound only on the M1 agonist screen, whose training split is 0.29% active; the others were generated in full. For the Tox21 assays, `select_curated` keeps every active plus a matched number of inactives, half of them the Morgan-fingerprint nearest neighbours of an active and half selected by MaxMin for diversity, under a fixed seed. Figure 5 shows the resulting selection on each assay.

Attempts are generated in one pass rather than conditionally retried, and every attempt whose tagged answer matches the ground-truth label is kept as a separate training row. The ADMET endpoints use one attempt per molecule, with four further attempts for minority-class molecules on three sparse endpoints. The Tox21 assays use two label-blind attempts, followed by up to five label-conditioned attempts for molecules that no blind attempt covered, in which the teacher is told the label and asked to derive it. Those fallback rationales are therefore not label-blind, and their answer match is largely true by construction. Molecules that yield no matching rationale contribute no CoT training row.

B.4 LITERATURE-MINED REASONINGS

Each TDC task is given an endpoint specification, comprising the permissible answers, the meaning of each label as defined in the source assay publication, a catalogue of mechanisms supporting or opposing each class, and an evidence hierarchy ordered by experimental system, from human measurements down to in silico computations. For blood–brain-barrier penetration, for instance, the mechanism catalogue contrasts passive transcellular diffusion and LAT1-mediated influx with P-glycoprotein efflux. An agent based on Gemini 3.8 Flash searches PubMed and produces a schema-validated report in which every hypothesis is a chain of typed steps, for that endpoint the ionisation state, the permeation event, transporter recognition, and the resulting brain partition. Each step cites its source and carries its own evidence level, and the confidence of a chain is capped by that of its weakest step. The benchmark label guides the search but is never treated as evidence, and reports



Figure 5: Molecules selected for Tox21 CoT generation, in a t-SNE embedding of each assay's training split. The embedding uses Tanimoto distance between the 2048-bit radius-2 Morgan fingerprints that the selection is computed on. Hard inactives are the nearest inactive neighbours of an active and sit next to the actives they were matched to. Diverse inactives are picked by MaxMin from the remaining inactives and also cover regions with few actives. Grey points are training molecules that receive no CoT. Tox21 PPAR-γ is held out from training and is shown for completeness.

contradicting it are retained in a separate subset. A second model call condenses each report into a four- to eight-sentence rationale tracing the reasoning from chemical structure alone to the report's conclusion.

B.5 Training Task Examples

Tables 5, 6, 7, and 8 show representative supervised examples from the training mixture. Each row gives the task type, the natural-language prompt, the molecule represented by its SMILES string, and the target assistant answer. The SMILES column identifies the molecule for readers; in Bolek training, the molecule is provided through the `<molecule>` token and projected molecular vector rather than as prompt text. The downstream rows use BBB Martins as a representative endpoint,

Table 5: Representative alignment examples for binary, numeric, and list-style tasks.

Task	Question	SMILES	Answer
FACCS substructure yes/no	In the structure of <molecule>, are there more than three ring bonds? Answer only with yes or no.	CC(C)OCCOC(C)C(C)C	No.
MACCS key yes/no	Does N in molecule <molecule> reach O through a chain of three atoms? Answer with yes or no.	CC(=O)NC(C)c1ccccn1	No.
RDKit fragment yes/no	Does the structure of <molecule> include a furan? Answer with yes or no.	COCC(O)CCNCCCF	No.
RDKit count descriptor	In <molecule>, how many rings are there? Answer with just the number.	O=PCCCCCC1CCSS1	1.
RDKit float descriptor	Tell me the LogP of <molecule>. Answer with just the approximate number.	COCCNCC(C)CCCN	0.6
Mordred count descriptor	Count the I atoms in <molecule>. Answer with just the number.	IC#CC#CC#CC#CC#CI	2.
Mordred float descriptor	For molecule <molecule>, what is the estimated LogS? Answer with just the approximate number.	CCOCC(N)c1cscc1C	-2.4
Mordred ring descriptor	How many heterocyclic rings with twelve members does <molecule> have? Answer with just the number.	C1CNCCNCCNCCN1	1.
SMILES recovery	Provide the SMILES notation of <molecule>.	CCCSCC(S)CSCCC	CCCSCC(S)CSCCC
Textbook substructure yes/no	Does molecule <molecule> contain X Fluoride Phenyl? Answer with yes or no.	CCC(CCN)Sc1nncs1	No.
Substructure list	List all chemical substructures contained by molecule <molecule>. Do not output anything else except a semicolon-separated list of substructures.	CCN(CC(=O)NC)C(C)=O	=O; Amide

Table 6: Representative free-text generation examples for molecular descriptions and naming tasks.

Task	Question	SMILES	Answer
KnowMol structure description	Provide a clear structural overview of <molecule> and explain how its parts fit together.	C1C(C(C(C(C1[NH3+])O)O)O)[NH3+]	The molecule consists of a cyclohexane ring as the main chain. Attached to the ring are three hydroxyl groups (-OH) at positions 2, 3, and 4. There are two ammonium ions (-NH3+) attached at positions 1 and 5. The hydroxyl groups and ammonium ions are connected directly to the carbon atoms of the cyclohexane ring. The molecule has no side chains, and all functional groups are directly bonded to the ring. The structure includes alkyl groups as part of the cyclohexane ring, hydroxyl groups, and ammonium ions.
KnowMol property description	Analyze how the functional groups and overall structure of <molecule> influence its physical and chemical properties, including polarity, acidity or basicity, solubility, reactivity, stereochemistry, and electrophilicity.	CCC1=C(N=C(C(=N1)CC)C)C	The molecule is nonpolar due to its symmetrical structure, which causes the dipole moments to cancel out. It exhibits basicity because of the nitrogen atoms in the pyrazine ring, which can donate electron pairs. The molecule is more soluble in nonpolar solvents due to the presence of nonpolar alkyl groups (ethyl and methyl). Its reactivity is relatively low as it lacks highly reactive functional groups, but the nitrogen atoms can participate in nucleophilic reactions. The molecule does not have chiral centers or double bonds that would lead to stereoisomerism, so it lacks stereochemistry concerns. The electrophilicity is low because the electron-donating alkyl groups reduce the electron deficiency at the nitrogen atoms.
ChEBI molecule description	Please describe the molecule <molecule>.	C(=O)(N)NS(=O)(=O)NC(=O)N	The molecule is classified as a sulfuric amide and belongs to the urea family.
Molecule naming (docking)	What is the IUPAC name of the compound <molecule>? Write only the name.	CCCCCCCCCCC=O	undecanal
Molecule naming (ZINC20)	What systematic name is assigned to the compound <molecule>? Do not output anything else except the name.	CCCCCCCCCCC#CCCO	tetradec-3-yn-1-ol

and show the prompt without the assay knowledge of subsection 3.1, which is injected ahead of the question with 50% dropout.

Table 7: Representative decomposition examples for local structural and physicochemical descriptions.

Task	Question	SMILES	Answer
Decomposition: structure	What are the main structural components of <molecule>?	F[C](F)C(F)CC(F)(F)C(F)F	A five-carbon chain is substituted with seven fluorine atoms.
Decomposition: hydrogen-bond acceptors	Where can <molecule> accept hydrogen bonds?	CC1CC2CC3(C)C1CCC23	No hydrogen bond acceptors are present.
Decomposition: hydrogen-bond donors	What hydrogen bond donor groups does <molecule> have?	COC(C)c1cc(F)ccc1C	No hydrogen bond donors are present.
Decomposition: lipophilicity	Describe the lipophilicity of <molecule>.	O=c1ccc2ccnc(=O)c-2co1	The molecule exhibits a slight hydrophilic character. The two external oxygen atoms present contribute to this property.
Decomposition: polarity	Is <molecule> polar or nonpolar?	CCC(S)CSC1SCSCS1	The molecule is nonpolar.
Decomposition: protonation	Describe the protonation state of <molecule> at physiological pH.	CC1C=NNC1 =C1[C+]=CC=[C-]C1	The molecule is uncharged.
Decomposition: stereochemistry	Identify the stereogenic sites in <molecule>.	O=S(=O)(CCl)NCc1cc[nH]c1	There are no stereogenic sites.
Decomposition: partial charges	Are there any regions of high or low electron density in <molecule>?	CC(C)(C)C=Cc1ccccc1	Electron density is distributed relatively evenly across this molecule.

Table 8: Representative downstream examples for direct and chain-of-thought binary classification.

Task	Question	SMILES	Answer
Downstream yes/no classification	Can the molecule <molecule> penetrate the blood-brain barrier? Answer with yes or no.	NCCc1cn2ccccc2n1	No.
Downstream chain-of-thought classification	Can the molecule <molecule> penetrate the blood-brain barrier? Start with considering the molecule structure and properties. Place the final answer in <answer>...</answer> tags, it should be either pass or fail.	CNCC(O) c1ccc(O)c(O)c1	Molecular weight is low (∼180), but lipophilicity is low (logP near 0) and polarity is fairly high (TPSA ∼70). It has many hydrogen-bond donors (4) and multiple acceptors, increasing desolvation cost. The secondary amine is protonated at physiological pH, adding a positive charge that strongly disfavors passive diffusion across the BBB despite the aromatic ring.<answer>fail</answer>

C RESULTS

C.1 PER-TASK CLASSIFICATION RESULTS

Table 9 and Table 10 break Table 1 down into its 30 endpoints. Every model is scored on the molecules it shares with BOLEK, so a row compares models on one molecule set; `papers/bolek/scripts/per_task_table.py` regenerates both tables.

Table 9: Per-task scores for the language models of Table 1, scored by ROC AUC on every task. GPT-6-Sol is scored from 5 rollouts per molecule rather than 50, and its yes/no run has no SARS-CoV-2 Touret, which the API refused.

Task	Metric	Qwen YN	Qwen CoT	GPT-6-Sol YN	GPT-6-Sol CoT	TxGemma-Chat	TxGemma-Predict	Bolek YN	Bolek CoT
ADMET (17)									
AMES	ROC	0.583	0.700	0.721	0.790	0.680	0.776	0.793	0.807
BBB Martins	ROC	0.628	0.857	0.805	0.871	0.713	0.822	0.882	0.884
Bioavailability Ma	ROC	0.478	0.651	0.590	0.815	0.685	0.562	0.725	0.751
CYP1A2 Veith	ROC	0.716	0.808	0.752	0.783	0.884	0.968	0.903	0.854
CYP2C19 Veith	ROC	0.614	0.637	0.719	0.760	0.829	0.963	0.851	0.784
CYP2C9 Substrate	ROC	0.591	0.550	0.639	0.705	0.632	0.600	0.604	0.630
CYP2C9 Veith	ROC	0.597	0.617	0.720	0.728	0.830	0.886	0.865	0.813
CYP2D6 Substrate	ROC	0.500	0.765	0.700	0.768	0.743	0.737	0.790	0.800
CYP3A4 Substrate	ROC	0.487	0.561	0.635	0.715	0.655	0.630	0.668	0.666
CYP3A4 Veith	ROC	0.570	0.686	0.711	0.744	0.788	0.870	0.845	0.805
hERG	ROC	0.497	0.674	0.752	0.776	0.808	0.897	0.698	0.774
HIA Hou	ROC	0.663	0.891	0.930	0.983	0.906	0.892	0.962	0.984
HIV	ROC	0.511	0.462	0.641	0.699	0.687	0.894	0.671	0.735
ClinTox	ROC	0.516	0.592	0.511	0.596	0.502	0.856	0.626	0.713
M1 agonist	ROC	0.500	0.516	0.499	0.504	0.660	0.799	0.701	0.685
SARS-CoV-2 Touret	ROC	0.508	0.569	–	0.482	0.581	0.739	0.418	0.573
Skin reaction	ROC	0.536	0.559	0.670	0.692	0.503	0.682	0.627	0.662
Tox21 (8)									
Tox21 AhR	ROC	0.537	0.639	0.636	0.615	0.821	0.932	0.809	0.832
Tox21 AR	ROC	0.495	0.487	0.507	0.536	0.721	0.832	0.762	0.702
Tox21 ARE	ROC	0.481	0.548	0.565	0.615	0.795	0.903	0.758	0.747
Tox21 ATAD5	ROC	0.502	0.507	0.582	0.587	0.695	0.932	0.704	0.763
Tox21 ER	ROC	0.525	0.551	0.574	0.594	0.735	0.805	0.705	0.687
Tox21 HSE	ROC	0.558	0.540	0.540	0.632	0.839	0.899	0.705	0.731
Tox21 p53	ROC	0.528	0.562	0.591	0.631	0.848	0.954	0.712	0.819
Tox21 MMP	ROC	0.536	0.622	0.611	0.644	0.870	0.932	0.818	0.813
Held-out (5)									
Pgp Broccatelli	ROC	0.649	0.795	0.766	0.849	0.816	0.929	0.911	0.900
PAMPA NCATS	ROC	0.661	0.648	0.726	0.673	0.583	0.689	0.708	0.726
M1 antagonist	ROC	0.534	0.709	0.634	0.641	0.766	0.904	0.778	0.812
CYP2D6 Veith	ROC	0.590	0.714	0.686	0.720	0.794	0.849	0.781	0.784
Tox21 PPAR-γ	ROC	0.504	0.584	0.641	0.686	0.728	0.904	0.669	0.769

Table 10: Per-task scores for the supervised baselines of Table 1, each trained on every endpoint including the five held out from BOLEK. Metrics follow Table 9.

Task	Metric	Morgan RF	Descriptor RF	Chemprop	Uni-Mol	MolFormer
ADMET (17)						
AMES	ROC	0.841	0.825	0.815	0.828	0.502
BBB Martins	ROC	0.880	0.908	0.883	0.845	0.883
Bioavailability Ma	ROC	0.699	0.712	0.572	0.685	0.664
CYP1A2 Veith	ROC	0.917	0.925	0.919	0.916	0.933
CYP2C19 Veith	ROC	0.860	0.872	0.870	0.860	0.873
CYP2C9 Substrate	ROC	0.566	0.615	0.558	0.592	0.611
CYP2C9 Veith	ROC	0.863	0.870	0.868	0.863	0.860
CYP2D6 Substrate	ROC	0.795	0.807	0.825	0.724	0.763
CYP3A4 Substrate	ROC	0.617	0.666	0.603	0.656	0.591
CYP3A4 Veith	ROC	0.856	0.856	0.885	0.852	0.876
hERG	ROC	0.815	0.829	0.737	0.762	0.813
HIA Hou	ROC	0.945	0.962	0.962	0.941	0.971
HIV	ROC	0.813	0.816	0.796	0.732	0.771
ClinTox	ROC	0.733	0.702	0.588	0.751	0.857
M1 agonist	ROC	0.645	0.679	0.676	0.579	0.709
SARS-CoV-2 Touret	ROC	0.474	0.578	0.575	0.507	0.533
Skin reaction	ROC	0.648	0.652	0.591	0.595	0.604
Tox21 (8)						
Tox21 AhR	ROC	0.822	0.833	0.805	0.833	0.808
Tox21 AR	ROC	0.702	0.762	0.765	0.672	0.700
Tox21 ARE	ROC	0.721	0.761	0.724	0.732	0.745
Tox21 ATAD5	ROC	0.701	0.675	0.665	0.632	0.703
Tox21 ER	ROC	0.713	0.786	0.704	0.717	0.723
Tox21 HSE	ROC	0.651	0.784	0.688	0.714	0.587
Tox21 p53	ROC	0.776	0.806	0.789	0.718	0.668
Tox21 MMP	ROC	0.800	0.879	0.833	0.837	0.843
Held-out (5)						
Pgp Broccatelli	ROC	0.920	0.901	0.909	0.871	0.913
PAMPA NCATS	ROC	0.760	0.734	0.717	0.700	0.734
M1 antagonist	ROC	0.824	0.831	0.828	0.742	0.828
CYP2D6 Veith	ROC	0.856	0.855	0.865	0.834	0.837
Tox21 PPAR-γ	ROC	0.697	0.788	0.798	0.640	0.642

C.2 Held-Out Regression

Table 11 reports the three held-out TDC regression endpoints, which the main text does not discuss.

Table 11: Held-out TDC regression transfer. Spearman correlation: higher is better; MAE: lower is better. For BOLEK, Lipophilicity and Solubility transfer from descriptor-alignment regression, including LogP- and LogS-like targets, rather than testing fully zero-shot regression. The *no preamble* column withholds the endpoint's literature preamble at inference. TxGemma-9B-Chat is included as a specialist reference. The best score per metric is bolded.

	Qwen3-4B-Instruct		BOLEK, no preamble		BOLEK		TxGemma-9B-Chat	
Task	Spearman	MAE	Spearman	MAE	Spearman	MAE	Spearman	MAE
Lipophilicity	0.222	2.09	0.313	1.49	0.368	1.23	**0.418**	**1.04**
PPBR AZ	0.001	15.56	0.183	25.84	**0.340**	13.73	-0.001	**10.82**
Solubility	0.440	2.21	0.443	2.21	0.595	1.90	**0.768**	**1.14**

C.3 Rollout-Count Sensitivity

This analysis was run early in BOLEK's development, to settle how many rollouts an evaluation needs, and is not repeated for the published model. Its setup and its numbers therefore differ from the rest of the paper: it uses that earlier recipe and its 15 trained endpoints rather than the 30-endpoint panel of Table 1. What it settled is the relative question, how much the rollout count moves a score, and the answer carried forward as the 50-rollout protocol used throughout.

Rollout scoring is a Monte Carlo estimate of the model's positive-answer probability, so a score computed from few generations is noisier than one computed from many. Table 12 quantifies that sensitivity by restricting BOLEK to the first five attempts from the same 50-rollout prediction files. The mean falls from 0.753 to 0.735, and four of the 15 endpoints score higher at five rollouts than at 50, so the rollout count contributes estimation noise rather than a systematic bias.

Table 12: Sensitivity of trained-endpoint AUCs to the number of BOLEK rollouts. BOLEK 5-rollout scores use attempts 0–4 from the same prediction files as the 50-rollout column. Scores keep the TDC task metric, ROC AUC or PR AUC, rather than the ROC AUC used elsewhere in the paper, because this table reuses the earlier runs' prediction files; the best score per row is bolded.

Task	Metric	BOLEK (YN), 50 rollouts	BOLEK (YN), 5 rollouts
AMES	ROC AUC	**0.761**	0.739
BBB Martins	ROC AUC	**0.857**	0.824
Bioavailability Ma	ROC AUC	**0.774**	0.712
CYP1A2 Veith	PR AUC	**0.904**	0.885
CYP2C19 Veith	ROC AUC	**0.847**	0.809
CYP2C9 Substrate	PR AUC	0.449	**0.473**
CYP2C9 Veith	PR AUC	0.716	**0.729**
CYP2D6 Substrate	PR AUC	0.607	**0.628**
CYP2D6 Veith	PR AUC	**0.612**	0.584
CYP3A4 Substrate	ROC AUC	**0.618**	0.573
CYP3A4 Veith	PR AUC	**0.792**	0.775
hERG	ROC AUC	**0.827**	0.800
HIA Hou	ROC AUC	0.934	**0.939**
HIV	ROC AUC	**0.679**	0.665
Pgp Broccatelli	ROC AUC	**0.923**	0.894
Mean	–	**0.753**	0.735

C.4 ANSWER FORMAT PER ENDPOINT

Figure 6 resolves the task-group means of Table 1 to the endpoint level, ranked by how much the chain-of-thought format gains over yes/no.

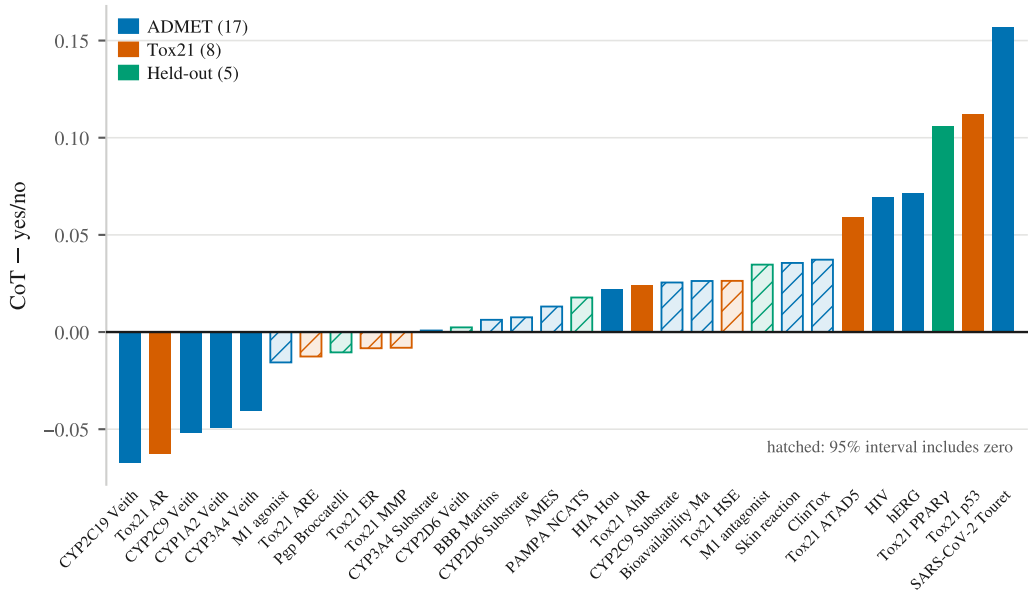

Figure 6: Difference between the chain-of-thought and yes/no answer formats on each of the 30 endpoints, ranked. Each difference carries a 95% paired bootstrap interval over test molecules. A hatched bar is one whose interval includes zero, so the difference is not significant; 17 of the 30 are. Scores are ROC AUC on every task.

C.5 TxGemma-9B-Chat Reproduction Comparison

The TxGemma-9B-Chat scores in the main tables come from our local evaluation of `google/txgemma-9b-chat`. For these runs, we use the TDC prompts provided in the model release, extract classification scores from the answer-choice token logits, and parse regression values from the model output. The TxGemma paper (Wang et al., 2025) reports results from 10-shot prompting, where the prompt includes the most similar examples drawn from training and validation data. Table 13 therefore compares our direct-question evaluation with the paper's 10-shot setting where a task-level paper value is available, but the two columns should be read as related rather than identical evaluation protocols. The comparison was run earlier in this work, against the 15 tasks the TxGemma paper reports at task level, so its task set is that of the earlier panel rather than the 30 endpoints of Table 1.

Table 13: TxGemma-9B-Chat direct-question comparison against the TxGemma paper's 10-shot results on the 15 standard downstream tasks. Metrics follow the TDC task-specific metrics, not the ROC AUC used elsewhere in the paper, since the comparison is against the TxGemma paper's published values.

Task	Metric	Direct question	10-shot (Wang et al., 2025)	Difference
AMES	ROC AUC	0.680	0.723	-0.043
BBB Martins	ROC AUC	0.717	0.832	-0.115
Bioavailability Ma	ROC AUC	0.685	0.666	$+0.024$
CYP1A2 Veith	PR AUC	0.886	0.862	$+0.019$
CYP2C19 Veith	ROC AUC	0.829	0.844	-0.015
CYP2C9 Substrate	PR AUC	0.369	0.414	-0.045
CYP2C9 Veith	PR AUC	0.684	0.700	-0.016
CYP2D6 Substrate	PR AUC	0.560	0.653	-0.093
CYP2D6 Veith	PR AUC	0.461	0.517	-0.056
CYP3A4 Substrate	ROC AUC	0.655	0.644	$+0.011$
CYP3A4 Veith	PR AUC	0.718	0.760	-0.042
hERG	ROC AUC	0.808	0.855	-0.047
HIA Hou	ROC AUC	0.906	0.906	0.000
HIV	ROC AUC	0.688	0.641	$+0.047$
Pgp Broccatelli	ROC AUC	0.816	0.873	-0.057
Mean	–	0.698	0.726	-0.028

On the standard tasks, our direct-question evaluation is lower on average by 0.028, with four task-level scores higher than the reported 10-shot TxGemma values and one exactly matched score. The largest gaps are on BBB Martins, CYP2D6 Substrate, CYP2D6 Veith, and Pgp Broccatelli. The standard-task conclusion is therefore not driven by a single outlier, but exact margins against the paper values should be read with the prompt-format difference in mind.

C.6 Regression Prediction Scatterplots

Figure 7 shows that the aggregate regression metrics in Table 11 hide qualitatively different prediction behaviors. On solubility, all models produce a visibly continuous range of values, and TxGemma-9B-Chat is both well ranked and reasonably calibrated relative to the other models. On PPBR AZ and lipophilicity, however, TxGemma-9B-Chat collapses much of the output space into a small number of repeated values. For PPBR AZ, it is close to a single-class predictor, with most predictions concentrated near 100% binding despite broad variation in the ground truth. For lipophilicity, it behaves more like a few-class classifier than a continuous regressor, assigning predictions to several horizontal bands.

This discretization matters for interpreting the held-out regression table. TxGemma-9B-Chat has low MAE on PPBR AZ because the endpoint is heavily concentrated near high binding values, but its Spearman correlation is approximately zero and the scatterplot shows little molecule-level ranking signal. On lipophilicity, its rank correlation remains the best in the table, yet the banded predictions indicate that much of the performance comes from coarse binning rather than calibrated numeric regression. This pattern is consistent with a task-specific response prior or overfitting to common endpoint ranges, although the scatterplot alone cannot distinguish memorization from a deliberately discretized generation strategy. BOLEK and its SMILES variant are less well calibrated in absolute terms, but their predictions remain more continuous on PPBR AZ and solubility, which is useful for rank-ordering molecules under held-out transfer.

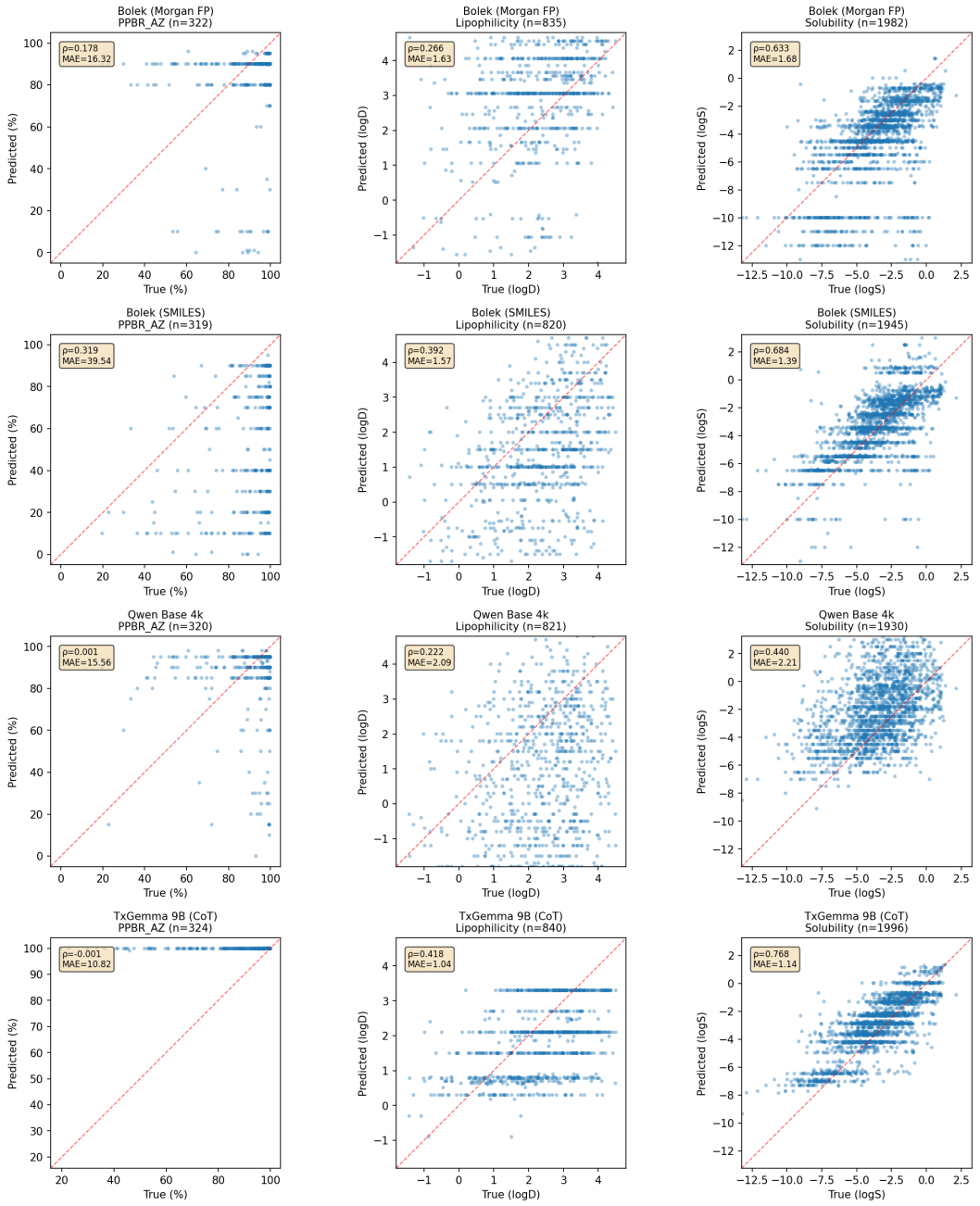

Figure 7: Per-molecule held-out regression predictions on PPBR AZ, lipophilicity, and solubility. Each row is one model, each column is one task, and the dashed line marks perfect calibration. The inset reports Spearman ρ and MAE for the corresponding panel.

D Analyses

D.1 Alignment Stage and Supervision

The variants of this block of Table 2 run the downstream stage of Table 3 and differ only in its initialization, its trainable parameters, and its data. Variants without alignment start from Qwen3-4B-Instruct with a newly initialized projector and skip the alignment stage rather than spending its steps on downstream training. The projector-only variant freezes the lan-

Table 14: Differences from BOLEK in mean ROC AUC for the alignment block of Table 2, with 95% paired bootstrap intervals. The last column compares the aligned and unaligned yes/no-only variants directly.

		Difference from BOLEK				Aligned −
Group	Format	Align., YN	No align.	No align., YN	Projector only	unaligned, YN
ADMET (17)	CoT	–	-0.012 $[-0.027, +0.002]$	–	-0.106 $[-0.131, -0.080]$	–
	YN	-0.027 $[-0.048, -0.007]$	-0.016 $[-0.036, +0.003]$	-0.060 $[-0.083, -0.035]$	-0.093 $[-0.120, -0.066]$	$+0.033$ $[+0.013, +0.052]$
Tox21 (8)	CoT	–	$+0.004$ $[-0.007, +0.015]$	–	-0.094 $[-0.112, -0.075]$	–
	YN	-0.100 $[-0.117, -0.083]$	-0.018 $[-0.033, -0.001]$	-0.113 $[-0.130, -0.096]$	-0.063 $[-0.082, -0.044]$	$+0.013$ $[-0.002, +0.026]$
Held-out (5)	CoT	–	-0.011 $[-0.027, +0.006]$	–	-0.138 $[-0.168, -0.107]$	–
	YN	-0.130 $[-0.156, -0.104]$	-0.049 $[-0.076, -0.022]$	-0.133 $[-0.164, -0.102]$	-0.131 $[-0.164, -0.099]$	$+0.003$ $[-0.024, +0.029]$

Table 15: Per-task difference from BOLEK in ROC AUC for the alignment block of Table 2. Underlined where the 95% paired bootstrap interval excludes zero.

	YN				CoT	
Task	Align., YN	No align.	No align., YN	Projector only	No align.	Projector only
ADMET (17)						
AMES	-0.018	-0.035	-0.042	-0.158	-0.029	-0.174
BBB Martins	-0.063	-0.021	-0.151	-0.239	$+0.016$	-0.185
Bioavailability Ma	-0.045	$+0.024$	-0.149	-0.080	-0.018	-0.140
CYP1A2 Veith	-0.007	-0.006	-0.028	-0.125	$+0.014$	-0.065
CYP2C9 Substrate	-0.060	$+0.018$	$+0.037$	-0.093	-0.011	-0.090
CYP2C9 Veith	-0.020	-0.014	-0.002	-0.098	$+0.005$	-0.045
CYP2D6 Substrate	-0.071	-0.001	-0.113	-0.195	-0.004	-0.203
CYP2C19 Veith	$+0.028$	-0.003	$+0.009$	-0.082	$+0.008$	-0.026
CYP3A4 Substrate	$+0.034$	$+0.031$	$+0.008$	-0.100	$+0.028$	-0.011
CYP3A4 Veith	$+0.011$	-0.013	-0.015	-0.067	$+0.003$	-0.024
hERG	$+0.025$	$+0.069$	-0.031	-0.122	$+0.034$	-0.199
HIA Hou	-0.099	-0.036	-0.180	-0.067	-0.038	-0.105
HIV	-0.039	$+0.023$	-0.041	$+0.047$	$+0.005$	-0.071
ClinTox	-0.035	0.000	-0.094	-0.009	-0.072	-0.047
M1 agonist	-0.144	-0.146	-0.106	-0.077	-0.047	-0.096
SARS-CoV-2 Touret	$+0.037$	-0.098	$+0.041$	$+0.069$	-0.047	-0.033
Skin reaction	$+0.002$	-0.070	-0.163	-0.183	-0.053	-0.287
Tox21 (8)						
Tox21 AhR	-0.115	-0.001	-0.099	-0.097	-0.006	-0.137
Tox21 AR	-0.130	-0.054	-0.178	-0.139	$+0.002$	-0.120
Tox21 ARE	-0.094	-0.015	-0.072	-0.060	$+0.017$	-0.059
Tox21 ATAD5	-0.151	-0.113	-0.157	-0.021	-0.012	-0.140
Tox21 ER	-0.027	$+0.029$	-0.057	-0.005	$+0.022$	-0.010
Tox21 HSE	-0.065	-0.021	-0.100	-0.052	$+0.011$	-0.088
Tox21 p53	-0.131	$+0.043$	-0.119	-0.043	$+0.002$	-0.132
Tox21 MMP	-0.088	-0.009	-0.124	-0.088	-0.006	-0.064
Held-out (5)						
CYP2D6 Veith	-0.068	-0.055	-0.142	-0.194	-0.001	-0.148
Pgp Broccatelli	-0.100	-0.022	-0.116	-0.151	-0.011	-0.073
Tox21 PPAR-γ	-0.105	-0.033	-0.019	$+0.026$	-0.011	-0.134
PAMPA NCATS	-0.161	$+0.008$	-0.114	-0.119	$+0.003$	-0.151
M1 antagonist	-0.215	-0.145	-0.274	-0.216	-0.036	-0.185

guage model, embeddings included, and trains 43M of its 4.1B parameters. Yes/no-only variants drop the chain-of-thought traces and keep the yes/no data at its original relative weights, so they see about 1.5 times as many yes/no examples as BOLEK. Table 14 gives the group-level differences with their intervals, and Table 15 breaks them down per task; `papers/bolek/scripts/alignment_ablation_appendix.py` regenerates both.

Table 16: Alignment metrics for the four molecular representations, at the alignment checkpoint each downstream run continues from. Substructure-classification rows report accuracy on the unseen pattern split, list and naming rows report token-set IoU, and descriptor rows report Pearson r. The best score per row is bolded.

Metric	Morgan fingerprint	SMILES	Uni-Mol	Molformer-XL
Substructure classification, unseen-pattern accuracy				
FACCS substructure	0.890	**0.968**	0.894	0.906
MACCS key	0.850	**0.900**	0.826	0.871
RDKit fragment	0.888	**0.939**	0.844	0.876
Textbook substructure	0.869	**0.941**	0.829	0.872
Group mean	0.874	**0.937**	0.848	0.881
IoU of substructures / name parts				
Substructure list	0.717	**0.908**	0.479	0.674
Molecule naming (docking)	0.362	**0.704**	0.191	0.309
Molecule naming (ZINC20)	0.184	**0.378**	0.116	0.155
Group mean	0.421	**0.663**	0.262	0.379
Descriptor prediction, Pearson r				
RDKit count descriptor	0.813	**0.960**	0.759	0.871
RDKit float descriptor	0.836	**0.930**	0.860	0.866
Mordred count descriptor	0.724	**0.922**	0.841	0.877
Mordred float descriptor	0.900	**0.979**	0.922	0.907
Mordred ring descriptor	0.619	**0.834**	0.600	0.757
Group mean	0.778	**0.925**	0.796	0.856

D.2 Alignment Metrics by Molecular Representation

Table 16 compares the four molecular representations of Table 2 on alignment itself, at the stage-one checkpoint each downstream run continues from. All four start from Qwen3-4B-Instruct (Yang et al., 2025a) and follow the same training procedure as each other; only the molecular representation differs. The fingerprint variant consumes a 2048-bit, radius-2 Morgan fingerprint through the projector, while the SMILES variant consumes the canonical SMILES string in the prompt.

Tasks and metrics. Substructure-classification rows (FACCS substructure, MACCS key, RDKit fragment, textbook substructure) ask binary yes/no questions of the form "does this molecule contain pattern X?" and report accuracy on patterns held out from training (see *Seen vs unseen splits* below). Substructure list reports the IoU between the semicolon-separated substructure tokens generated by the model and those in the reference list, treating both as multisets. Molecule naming (docking) and molecule naming (ZINC20) report the IoU between whitespace-separated parts of the predicted IUPAC or systematic name and the reference name. Descriptor rows report the per-descriptor Pearson correlation r between predicted and ground-truth values, averaged across all descriptors in the family. The SMILES recovery task is not included: the SMILES variant sees the canonical SMILES directly in the prompt, so emitting it is trivial and the task is dropped from the SMILES training mixture, which makes any FP-versus-SMILES comparison on this row uninformative.

Seen vs unseen splits. For each substructure-classification dataset (FACCS, MACCS, RDKit fragment, textbook substructure), we hold out a fixed subset of patterns from training and use them only at evaluation; the complementary patterns are the *seen* subset that appears in training questions. The unseen rows therefore measure transfer to new patterns rather than memorization of training-time questions, hence we report unseen accuracy. The remaining alignment tasks (substructure list, molecule naming, descriptor regression) cover the full task space at training time, so they only require a held-out molecule split, taken from the MolPILE validation set.

Discussion. SMILES leads every alignment row, with the largest gaps on tasks that require reading specific characters, atoms, or counts off the molecule (descriptor regression, substructure list, molecule naming). This is consistent with SMILES being a more natural representation for an LLM than a 2048-bit hashed fingerprint: the SMILES string exposes atoms, bonds, and ring closures directly to the decoder, while the fingerprint compresses local environments into a fixed-width vector

that the projector has to unpack. Substructure-classification accuracy on unseen patterns also favours SMILES, but by a smaller margin, because the yes/no answer can be derived from coarse fingerprint features. Among the three vector encoders the ranking depends on the task: Molformer-XL, itself trained on SMILES, is second on substructure classification and on descriptors, while the Morgan fingerprint is second on the IoU tasks and Uni-Mol is last on both. These gaps do not carry downstream, where the four sit within roughly 0.02 of each other per task group and the fingerprint leads the trained ADMET endpoints (Table 2).

D.3 BACKBONE SIZE ABLATION

This ablation was run early in BOLEK's development, to settle the choice of backbone, and is not repeated for the published model. Its setup and its numbers therefore differ from the rest of the paper: the six runs share a joint alignment-plus-downstream mixture trained for 10,000 steps rather than the two supervised stages of subsection 3.1, and they predate both the 25:5 panel and the assay cards. The columns are comparable with each other but not with Table 1, and what they support is the relative question the decision turned on: how performance scales with backbone size, holding the rest of that earlier recipe fixed.

To isolate the effect of backbone size, we re-train BOLEK on five base Qwen3 checkpoints from 0.6B to 14B parameters, plus the Qwen3-4B-Instruct-2507 backbone used by the main BOLEK model (Yang et al., 2025a). All six runs share the fingerprint adapter, mixture, optimizer, and schedule, and are evaluated with the same 50-rollout protocol used in the main paper. The 4B-Instruct column is that earlier recipe's BOLEK model and serves as the reference point for the rest of this appendix. A 32B run did not fit on a single 8×80 GB H100 node and was not pursued.

Training cost grows roughly linearly with backbone parameters (Table 17).

Table 17: Approximate H100 training cost per backbone for the size-sweep ablation.

Backbone	0.6B	1.7B	4B	8B	14B
GPU-hours	14	18	26	52	102

Alignment quality scales monotonically with backbone size. Table 18 reports alignment metrics across the sweep, grouped by metric type with a per-group mean. Substructure-classification accuracy on unseen patterns, IoU of listed substructures and predicted name parts, descriptor Pearson correlation, and SMILES recovery from the fingerprint all rise nearly monotonically from 0.6B to 14B. The single largest step is between 1.7B and 4B; subsequent doublings still help but at a slower rate. The 4B-Instruct backbone edges the 4B base backbone on the accuracy, Pearson, and SMILES means but loses to it on the IoU mean, so the instruct boost on alignment is small and not uniform.

Downstream gains plateau from 4B onward. Table 19 and Table 20 report ROC or PR AUC across the same six backbones for direct yes/no and chain-of-thought formats. The yes/no mean improves from 0.708 at 0.6B to 0.749 at 4B, then plateaus: 4B-Instruct reaches 0.753, 8B reaches 0.759, and 14B reaches 0.757. Chain-of-thought scaling is weaker still, with all backbones from 4B upward within 0.008 of each other on the mean (4B 0.709, 4B-Instruct 0.704, 8B 0.706, 14B 0.712).

A few task-level patterns survive across formats. HIA Hou is the cleanest positive scaler in yes/no mode, climbing from 0.889 at 0.6B to 0.988 at 14B. HIV inverse-scales in yes/no mode, dropping from 0.763 at 0.6B to 0.680 at 14B. The 4B-Instruct backbone wins outright on Bioavailability Ma, CYP2C9 Substrate, and Pgp Broccatelli in yes/no mode and on Bioavailability Ma, BBB Martins, CYP2D6 Substrate, and CYP3A4 Veith in chain-of-thought mode. However, most of those tasks are small, so the observed wins might be attributed to noisy test sets.

The sweep supports three observations:

- **Alignment quality scales monotonically with backbone size**, so the largest backbone we tested is the strongest on alignment.
- **Downstream gains plateau from 4B onward**: doubling parameters from 4B to 8B improves the yes/no mean by 1.0 pp (0.749 to 0.759), and a further doubling to 14B does not

Table 18: Alignment metrics across backbone sizes. Substructure-classification rows report unseen-class accuracy, substructure-list and naming rows report IoU between predicted and reference token sets, descriptor rows report Pearson r, and SMILES rows report Tanimoto similarity to the canonical SMILES and the fraction of generated SMILES that parse. The 4B-Instruct column is the backbone used by the main BOLEK model. The best score per row is bolded.

Metric	0.6B	1.7B	4B	4B-Instr (BOLEK)	8B	14B
Substructure classification, unseen-pattern accuracy						
FACCS substructure	0.792	0.830	0.890	0.895	0.893	**0.908**
MACCS key	0.764	0.801	0.825	0.843	0.846	**0.855**
RDKit fragment	0.813	0.842	0.870	0.879	0.885	**0.892**
Textbook substructure	0.811	0.847	0.859	**0.878**	0.851	0.877
Group mean	0.792	0.824	0.855	0.868	0.863	**0.875**
IoU of substructures / name parts						
Substructure list	0.638	0.690	0.726	0.709	**0.750**	0.746
Molecule naming (docking)	0.236	0.273	0.340	0.349	0.381	**0.401**
Molecule naming (ZINC20)	0.169	0.170	0.199	0.168	**0.201**	0.191
Group mean	0.348	0.378	0.422	0.409	0.444	**0.446**
Descriptor prediction, Pearson r						
RDKit count descriptor	0.672	0.757	0.793	0.813	0.823	**0.829**
RDKit float descriptor	0.645	0.713	0.841	0.846	0.861	**0.863**
Mordred count descriptor	0.614	0.637	0.719	0.704	0.722	**0.745**
Mordred float descriptor	0.825	0.850	0.833	0.890	0.891	**0.897**
Mordred ring descriptor	0.539	0.549	0.587	0.586	0.616	**0.643**
Group mean	0.659	0.701	0.755	0.768	0.783	**0.795**
SMILES recovery from fingerprint						
SMILES Tanimoto	0.306	0.337	0.387	0.387	**0.425**	0.415
SMILES parsed	0.816	0.846	0.891	**0.894**	0.885	0.887
Group mean	0.561	0.592	0.639	0.641	**0.655**	0.651

improve it (0.757); chain-of-thought scaling is weaker still, with 14B only 0.3 pp above 4B on the mean.

- **Training cost grows roughly linearly with backbone size**, so a 14B run uses about four times the H100-hours of a 4B run for an essentially flat downstream return.

The 4B-Instruct checkpoint adds a small additional boost over the 4B base on top of these trends, plausibly because the joint alignment and downstream mixture already includes a large body of chemical instruction-tuning supervision and partly overrides whatever instruction priors Qwen3-4B-Instruct contributes. **Together, these observations motivate Qwen3-4B-Instruct as the main backbone for BOLEK.**

D.4 COUNTERFACTUAL RATIONALE TEST

A chain-of-thought answer is worth auditing only if the rationale actually guides it. BOLEK could instead answer from the molecular token and treat the rationale as a distraction written alongside a decision it has already made. Correcting such a rationale would then leave the prediction unchanged. We test which is the case by replacing the model's rationale with one written for a different molecule.

For each test molecule we keep the molecular token and the question fixed. We then replace the model's own rationale with a *donor rationale* that BOLEK generated for another test molecule of the same endpoint. Donors come only from rollouts that reached the correct label for their own molecule, so each donor argues a known label. We remove the donor's final verdict sentence and

Table 19: Downstream yes/no AUC across backbone sizes. Each entry is estimated from 50 rollouts per molecule. Scores keep the TDC task metric, ROC AUC or PR AUC, rather than the ROC AUC used elsewhere in the paper, because this sweep predates it. The best score per row is bolded.

Task	Metric	0.6B	1.7B	4B	4B-Instr (BOLEK)	8B	14B
AMES	ROC AUC	0.640	0.712	0.740	0.761	**0.786**	**0.786**
BBB Martins	ROC AUC	0.785	0.802	0.841	0.857	**0.869**	0.864
Bioavailability Ma	ROC AUC	0.689	0.658	0.737	**0.774**	0.731	0.679
CYP1A2 Veith	PR AUC	0.860	0.891	0.904	0.904	0.907	**0.909**
CYP2C19 Veith	ROC AUC	0.824	0.847	0.860	0.847	0.867	**0.872**
CYP2C9 Substrate	PR AUC	0.330	0.347	0.427	**0.449**	0.374	0.402
CYP2C9 Veith	PR AUC	0.674	0.715	0.745	0.716	0.748	**0.763**
CYP2D6 Substrate	PR AUC	0.599	0.594	**0.678**	0.607	0.605	0.621
CYP2D6 Veith	PR AUC	0.563	0.591	0.593	0.612	**0.632**	0.615
CYP3A4 Substrate	ROC AUC	0.620	**0.676**	0.573	0.618	0.675	0.627
CYP3A4 Veith	PR AUC	0.749	0.778	0.815	0.792	0.819	**0.827**
hERG	ROC AUC	0.804	0.780	0.780	0.827	**0.844**	0.802
HIA Hou	ROC AUC	0.889	0.871	0.928	0.934	0.955	**0.988**
HIV	ROC AUC	**0.763**	0.704	0.705	0.679	0.664	0.680
Pgp Broccatelli	ROC AUC	0.836	0.906	0.907	**0.923**	0.911	0.914
Mean	–	0.708	0.725	0.749	0.753	**0.759**	0.757

Table 20: Downstream chain-of-thought AUC across backbone sizes. Conventions match Table 19.

Task	Metric	0.6B	1.7B	4B	4B-Instr (BOLEK)	8B	14B
AMES	ROC AUC	0.652	0.667	0.714	0.729	0.726	**0.741**
BBB Martins	ROC AUC	0.784	0.809	0.825	**0.852**	0.808	0.819
Bioavailability Ma	ROC AUC	0.728	0.692	0.687	**0.730**	0.729	0.668
CYP1A2 Veith	PR AUC	0.859	0.868	0.874	**0.876**	**0.876**	0.871
CYP2C19 Veith	ROC AUC	**0.774**	0.755	0.728	0.713	0.717	0.731
CYP2C9 Substrate	PR AUC	0.387	0.364	**0.489**	0.416	0.394	0.434
CYP2C9 Veith	PR AUC	0.614	**0.616**	0.607	0.579	0.580	0.601
CYP2D6 Substrate	PR AUC	0.610	0.618	0.651	**0.668**	0.663	0.649
CYP2D6 Veith	PR AUC	0.477	0.503	0.555	0.551	0.556	**0.560**
CYP3A4 Substrate	ROC AUC	0.599	0.606	0.626	0.623	0.618	**0.665**
CYP3A4 Veith	PR AUC	0.722	0.718	0.728	**0.737**	0.724	0.733
hERG	ROC AUC	0.808	0.808	0.827	0.779	**0.837**	**0.837**
HIA Hou	ROC AUC	0.888	0.860	0.888	0.869	**0.929**	0.928
HIV	ROC AUC	0.533	**0.535**	0.515	0.509	0.531	0.528
Pgp Broccatelli	ROC AUC	0.889	0.896	0.919	**0.923**	0.896	0.916
Mean	–	0.688	0.688	0.709	0.704	0.706	**0.712**

discard any donor that still names a label. The donor argues either the target's true label or the opposite one. Each molecule receives five donors per condition. We score both answer labels from the model's log-probabilities after the donor rationale, average the positive-label probability over the five donors, and predict positive at 0.5. The analysis covers the full test splits of all 30 endpoints in Table 1 under the default protocol, so both the scored prompt and the donor rollouts carry the literature preamble or assay card.

The answer follows the donor rationale almost without exception (Table 21). Accuracy is 99.8% when the donor argues the true label and 1.8% when it argues the opposite one, and ROC AUC is 1.000 and 0.000 in every task group. With a true-label donor no endpoint falls below 97.2%.

Table 21: Counterfactual rationale test on BOLEK, macro-averaged over the endpoints of each task group. The model's rationale is replaced by a verdict-stripped rationale that BOLEK wrote for another molecule of the same endpoint, arguing either the target's true label or the opposite one. Accuracy uses the mean positive-label probability over five donors per molecule. Donor consistency is the share of single donors that the model agrees with.

	Accuracy		Donor consistency	
Task group	True label	Opposite	True label	Opposite
ADMET (17)	99.8%	2.1%	97.4%	94.0%
Tox21 (8)	99.6%	1.8%	98.4%	93.0%
Held-out (5)	100.0%	0.6%	98.5%	94.4%
All (30)	99.8%	1.8%	97.9%	93.8%

With an opposite donor only M1 agonist rises above 12.6%, to 30.7%, because the model follows rationales arguing agonism for only about 60% of molecules. The molecular token and the question are unchanged and the verdict is removed, so the rationale, rather than the molecule alone, decides the answer.

E LICENSES AND TERMS OF USE FOR EXTERNAL ASSETS

Asset	Type	License / Terms	Use in this work
Qwen3-4B-Instruct	Model	Apache 2.0	SFT base model
TxGemma-9B Chat and Predict	Model	Health AI Dev. Foundations ToU (Gemma 2 lineage)	Frozen eval baselines
GPT-6-Sol	API	OpenAI Terms of Use	Frozen eval baseline
Uni-Mol	Model	MIT	Representation ablation, eval baseline
Molformer-XL	Model	Apache 2.0	Representation ablation, eval baseline
Chemprop	Code	MIT	Eval baseline
Gemini 2.5 Flash	API	Google Generative AI ToS	Per-part decomposition supervision
Gemini 3.5 / 3.8 Flash	API	Google Generative AI ToS	Synthetic CoT generation
Gemini 3.8 Flash	API	Google Generative AI ToS	Literature mining and rationale distillation
Claude Opus 4.6	API	Anthropic Usage Policies	Literature preamble generation
TDC (codebase)	Code	MIT	Task definitions, scaffold splits
TDC (datasets)	Data	Per-dataset (see TDC repo)	Train/eval molecules and labels
MoleculeNet	Data	MIT via DeepChem	Lipophilicity regression
AqSolDB	Data	CC0	Solubility regression
ZINC20	Data	Free to use; terms unspecified	Alignment molecules
PubChem (names)	Data	U.S. public domain	Name-prediction supervision
NCI AIDS Antiviral	Data	U.S. public domain	HIV classification
MolPILE	Data	MIT	Alignment molecules
KnowMol	Data	MIT	Multi-level annotations
ChEBI-20-MM	Data	MIT	Molecule descriptions
SureChEMBL alerts	Data	CC BY 4.0	Substructure vocabulary
SMARTS-RX	Data	CC BY 4.0	Substructure vocabulary
RCSB PDB (Berman et al., 2000)	Data	CC0 1.0 Public Domain Dedication	Name-prediction supervision
RDKit	Code	BSD 3-Clause	Descriptors, fingerprints
Mordred	Code	BSD 3-Clause	Descriptors

Table 22: External assets used in this work, with licenses and terms of use.