PHARMACOLOGICAL MECHANISM-GUIDED COLLABORATIVE LLM AGENTS FOR MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular property prediction has advanced on the premise that properties follow from a sufficiently precise representation of structure. LLMs have extended this paradigm by aligning molecular and textual representations, retrieving analogous molecules, and learning step-wise reasoning over structure. In all of these, the property is still predicted directly from structure, without passing through the biological context in which it arises. A molecular property, however, is observed as the outcome of a molecule engaging targets in vivo and triggering downstream events, and a prediction that does not pass through that mechanism remains an extrapolation of structural similarity. We reformulate property prediction as mechanism-grounded evidence reasoning and propose PharmaLens, a multi-agent framework that realizes this formulation. For a given property, agents at the physicochemical, target-relation, and binding-structure levels establish the specified factors on the query molecule by tool-based measurement, and the resulting evidence is integrated over the relational logic. Across in vivo endpoints from TDC and organ-toxicity tasks from UniTox, PharmaLens achieves strong performance on both benchmarks, and it retains this margin on property cliffs, where structurally similar molecules carry opposite labels and label-trained predictors fall off most sharply. Its rationales are also consistently favored across all three evaluations. The source code is publicly available at https://github.com/anonymousqwerty350/PharmLens.

1 INTRODUCTION

Molecular property prediction has used fingerprints, molecular graphs, and three-dimensional geometry to capture structural information relevant to molecular properties (Li et al., 2026). However, learning structure–property relationships alone does not readily enable explicit assessment of the mechanistic factors underlying a property or the functional relationships between molecules and their targets. Such assessment is particularly important for absorption, distribution, metabolism, excretion, and toxicity (ADME/T), which depend jointly on molecular physicochemical properties and interactions with biological processes and targets (int, 2010). For example, absorption and distribution involve both molecular characteristics and transport processes, while toxicity can arise when specific molecule–target relationships lead to downstream biological events (Ankley et al., 2010). Because the relevant factors and relationships vary across properties, property-specific mechanistic knowledge can guide what to assess and which evidence to acquire.

Encouragingly, LLM-based approaches have broadened molecular property prediction through molecular–language alignment (Fang et al., 2024; Cao et al., 2025; Li et al., 2024) and chemical reasoning and retrieval (Zhuang et al., 2025; Lin et al., 2025; Lee et al., 2026; Xian et al., 2025). However, these approaches primarily improve molecular representations and reasoning from chemical features or retrieved information, rather than explicitly organize prediction around the biological processes underlying ADME/T outcomes. Approaches incorporating biological context have introduced molecule-associated pathway and Gene Ontology annotations (Park et al., 2025). Yet incorporating these associations does not by itself establish whether a query molecule exhibits the functional relations implicated in the mechanisms underlying a property. Reliance on pre-existing annotations further limits the available evidence for poorly annotated molecules. A procedure is therefore needed to translate property-specific mechanistic knowledge into molecule-specific assessments, acquiring

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.

evidence for relevant factors and functional relations and interpreting their implications for the endpoint.

We propose PharmLens, a framework that predicts ADME/T-related properties by explicitly assessing molecular factors and functional relations implicated in known biological mechanisms. To implement this approach, it constructs a literature-grounded specification of molecular factors, protein–relation hypotheses, and their endpoint-specific mechanistic implications. This specification guides molecule-specific evidence acquisition and interpretation according to each relation's biological role. PharmLens integrates physicochemical and relation evidence, adding auxiliary structural evidence when relation evidence is insufficient and preserving uncertainty in unresolved hypotheses. This automates the process from selecting what to assess to generating predictions and rationales, without requiring pre-existing biological annotations for query molecules.

We evaluate PharmLens on TDC and UniTox for predictive performance and property-cliff discrimination. Experiments across multiple backbones, evidence ablations, and specification swaps examine the consistency of performance gains, the contributions of individual evidence layers, and the effect of alignment between a property and its specification. We assess the generated rationales through downstream predictive utility, quality ratings from LLM judges, and comparative evaluations by experts.

Our contributions can be summarized as follows:

- We introduce **PharmLens**, which predicts ADME/T properties by assessing molecular factors and functional relations using property-specific mechanistic knowledge. Literature-grounded specifications guide evidence acquisition and interpretation without requiring pre-existing biological annotations for query molecules.
- PharmLens combines physicochemical descriptors and calculated relation estimates, selectively adding interaction evidence from reference-ligand comparisons when relation evidence is insufficient. An LLM integrates this evidence according to each relation's mechanistic role, preserving uncertainty in unresolved hypotheses.
- PharmLens outperforms evaluated LLM-based baselines on TDC and UniTox, improves UniTox property-cliff discrimination, and yields gains across four backbones. Knowledge ablations and specification swaps support the utility of acquired evidence and property-aligned specifications. Generated rationales improve downstream prediction and achieve the best mean expert rank on 40 Bioavailability molecules.

2 Related Works

2.1 Aligning Molecular Representations with Language

Molecular–language alignment connects molecular representations to language models, making structural information available for diverse molecular tasks. Mol-Instructions (Fang et al., 2024) uses instruction data to support adaptation to molecular and protein tasks, while InstructMol (Cao et al., 2025) aligns graph representations with the language space. 3D-MoLM (Li et al., 2024) incorporates three-dimensional geometry, Mol-LLaMA (Kim et al., 2026) combines 2D and 3D representations, and EDT-Former incorporates substructure information. These approaches contribute representations and alignment mechanisms that make molecular information accessible to language models.

2.2 Reasoning from Molecular Structure

Structure-based reasoning constructs evidence for property prediction by analyzing chemical features or retrieving similar examples. MPPReasoner (Zhuang et al., 2025) learns reasoning trajectories that apply chemical principles to SMILES and 2D images, while AttriLens-Mol (Lin et al., 2025) addresses the relevance and validity of reasoning through molecular attributes that can be verified with RDKit. MolRAG (Xian et al., 2025) retrieves similar molecules and their property labels to construct the prediction context. CLADD (Lee et al., 2026) determines whether to consult external information sources based on the query molecule and task, combining knowledge-graph information about similar anchor drugs with outputs from molecular annotation and captioning tools. These approaches integrate structural analysis, information retrieval, and tool use into molecular reasoning.

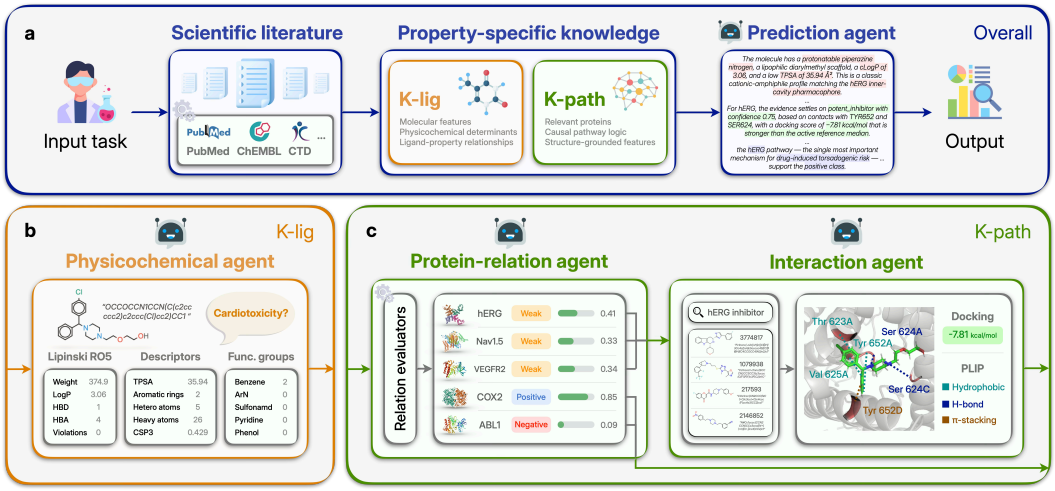


Figure 1: Overview of PharmaLens. (a) The Knowledge Agent constructs $\mathcal{K}_\tau^{\text{lig}}$ and $\mathcal{K}_\tau^{\text{path}}$ from the literature, guiding the evidence collected in (b) and (c), which the Prediction Agent then integrates. (b) Ligand-centric evidence, the descriptors diagnostic for τ. (c) Pathway-centric evidence, the relation judged for each protein with escalation to docking and contact comparison.

2.3 Reasoning with Biological Context

Biological-context approaches expand the evidence for property prediction using pathways and functions associated with molecules. CoTox (Park et al., 2025) selects compound-associated pathway and Gene Ontology annotations from CTD and combines them with chemical structures for organ-specific toxicity prediction. This illustrates the value of biological context, although the evidence available depends on the availability of molecule-specific annotations.

3 Method

3.1 Problem Formulation

PharmLens predicts molecular properties through knowledge-guided evidence reasoning: property-specific mechanistic knowledge specifies the factors and relations to assess, which are then evaluated for the query molecule: $\hat{y} = \mathcal{R}(\mathcal{E}(m, \mathcal{K}_\tau) \mid \mathcal{K}_\tau)$. Here, m denotes the query molecule, τ the target property, and $\mathcal{K}_\tau$ a specification defining the factors and relations to assess and their mechanistic relevance. $\mathcal{E}$ acquires molecule-specific evidence according to the specification, and $\mathcal{R}$ integrates this evidence in the context of property-specific knowledge. The procedure comprises property-level specification construction (Section 3.2), molecule-specific evidence acquisition and selective additional analysis (Sections 3.3–3.4), and prediction and rationale generation (Section 3.5). Tool outputs are generated by selecting assessment targets and executing tools according to the specification.

3.2 Mechanism Specification

Using only the target property τ as input, we construct a literature-grounded mechanistic specification. This explicit intermediate representation connects the molecular and biological factors to be assessed with the evidence required to assess them. It is constructed independently of the query molecule and consists of two components: $\mathcal{K}_\tau = \left(\mathcal{K}_\tau^{\text{lig}}, \mathcal{K}_\tau^{\text{path}}\right)$.

Ligand-centric knowledge. $\mathcal{K}_\tau^{\text{lig}}$ specifies physicochemical quantities, such as MW, LogP, TPSA, and hydrogen-bond donor count, their relevant ranges, and structural motifs relevant to the property. It covers features linked in the literature to the property through absorption and distribution, reactive metabolic intermediates, target binding, and substrate recognition by transporters (Stepan et al., 2011). It guides the Physicochemical Agent in selecting descriptors and structural motifs to examine.

Pathway-centric knowledge. $\mathcal{K}_\tau^{\text{path}}$ comprises a protein set $\mathcal{P}_\tau$ and target-specific pathway logic ℓ_p. Each ℓ_p describes the molecule–protein functional relation to assess and the direction of its contribution to the endpoint. Relation types include substrates in metabolism and transport, inhibitors of enzyme or transporter function, and antagonists and agonists that block or activate target function. These define protein–relation hypotheses (p, ℓ_p) and guide the selection of corresponding evaluators. For example, when considering impaired bile acid efflux in DILI, the specification identifies both whether the relevant transporter is inhibited and how this relation contributes to liver injury (Dawson et al., 2012). Thus, ℓ_p connects the relation being evaluated to the property-specific interpretation of its assessment.

Each knowledge item is linked to the identifiers and relevant passages of its supporting literature, making the basis for its selection traceable. The specification is fixed before benchmark evaluation and reused across molecules for the same endpoint. Molecule-specific evidence is acquired when the specified items are computed or evaluated.

3.3 Evidence Acquisition

The evidence agents compute or evaluate the specified items for each molecule to generate ligand-centric evidence ($\mathbf{E}_{\text{LCE}}$), pathway-centric evidence ($\mathbf{E}_{\text{PCE}}$), and interaction evidence ($\mathbf{E}_{\text{IE}}$). These provide intermediate evidence for the final endpoint prediction; additional analysis and relation decisions follow the rules in Section 3.4.

3.3.1 Ligand-Centric Evidence

The Physicochemical Agent selects a subset of the available descriptor space $\mathcal{F}$ according to $\mathcal{K}_\tau^{\text{lig}}$: $\mathcal{F}_\tau = \text{Select}(\mathcal{F} \mid \mathcal{K}_\tau^{\text{lig}})$. The specification guides the selection of descriptors needed to assess the property. Each $\varphi \in \mathcal{F}_\tau$ is computed deterministically with RDKit to construct $\mathbf{E}_{\text{LCE}}$. The available descriptor space is provided in Appendix A.2.

3.3.2 Pathway-Centric Evidence

The Protein-Relation Agent uses an evaluator for the functional relation specified by ℓ_p to predict whether that relation holds for the query molecule, yielding r_p. Evaluators trained on relation labels are ensembles of different algorithm–descriptor combinations; the mean predicted probability $P_p(m)$ serves as the relation signal. For CYP substrate relations, we use CypReact, which specializes in substrate prediction (Dawson et al., 2012). These outputs constitute $\mathbf{E}_{\text{PCE}}$, and the sufficiency of the evidence for each relation determines whether additional analysis is required (Section 3.4). Relation-data construction and evaluator design are described in Appendix A.3.

3.3.3 Protein–Ligand Interaction Evidence

For relation hypotheses requiring additional analysis, the Interaction Agent consults the binding patterns of reference ligands whose status with respect to the same relation is known. It generates auxiliary evidence E_{IE} to assess whether the query molecule's predicted binding pattern is structurally compatible with the hypothesis.

Reference profiling. We search PubMed abstracts using the protein and the functional relation specified by ℓ_p to identify reference ligands, and resolve their names to SMILES through PubChem (Kim et al., 2025). After excluding entries with unresolved structures, duplicates, and overlaps with evaluation molecules, we select k ligands per target. Each reference ligand is docked with UniDock (Yu et al., 2023), and its top-ranked pose is analyzed with PLIP (Schake et al., 2025) to represent contacts by residue, interaction type, and distance. The reference profile Π_p summarizes contacts from relation-positive reference ligands while excluding contacts also commonly observed in relation-negative reference ligands.

Query assessment. Applying the same analysis to the query yields a binding score $s(m, p)$ and a contact pattern C_m. The score is an approximate signal of the predicted pose's binding compatibility, while agreement between C_m and Π_p indicates the extent to which the query shares contact features with relation-positive reference ligands (Warren et al., 2006). The two signals are passed forward as supporting, contradicting, or uninformative structural evidence for the relation hypothesis.

Profile applicability. Profile applicability depends on the availability of reference ligands and the distinguishability of contact patterns between relation-positive and relation-negative reference ligands. For binding patterns that are difficult to characterize with a single profile, we first examine essential structural requirements supported by the literature. When the profile is not applicable, only the score is passed forward as limited evidence of binding compatibility. Each output records the structural signals used and their interpretive scope, distinguishing structural support for a relation hypothesis from evidence of binding compatibility. Structural preprocessing and analysis settings are provided in Appendix A.4.

3.4 Evidence Routing and Abstention

For each protein–relation hypothesis (p, ℓ_p), the sufficiency of $\mathbf{E}_{\text{PCE}}$ determines whether to acquire additional $\mathbf{E}_{\text{IE}}$. The interpretive scope of $\mathbf{E}_{\text{IE}}$ depends on the applicability of the reference profile Π_p.

1. $\mathbf{E}_{\text{PCE}}$ assessment. $\mathbf{E}_{\text{PCE}}$ is treated as insufficient if $P_p(m)$ falls within an ambiguity interval, no corresponding evaluator is available, or the prediction conflicts with previously acquired evidence. Otherwise, the relation prediction r_p is adopted.
2. Selective $\mathbf{E}_{\text{IE}}$ acquisition. Hypotheses with insufficient $\mathbf{E}_{\text{PCE}}$ are passed to the Interaction Agent to acquire $\mathbf{E}_{\text{IE}}$ (Section.3.3.3). When Π_p is applicable, the binding score and contact pattern are used jointly; otherwise, only the score is passed forward as limited evidence of binding compatibility.
3. Hypothesis-level decision. The available $\mathbf{E}_{\text{PCE}}$ and selectively acquired $\mathbf{E}_{\text{IE}}$ are assessed within their respective interpretive scopes to assign a positive, negative, or unresolved decision to each hypothesis. A hypothesis remains unresolved when the evidence is insufficient to distinguish whether the functional relation holds, even if $\mathbf{E}_{\text{IE}}$ supports the hypothesis.

Abstention defers the decision on an individual relation hypothesis. Unresolved relations are passed to the final endpoint assessment as items with insufficient evidence.

3.5 Mechanistic Integration

The Prediction Agent integrates $\mathbf{E}_{\text{LCE}}$, $\mathbf{E}_{\text{PCE}}$, and selectively acquired $\mathbf{E}_{\text{IE}}$ according to $\mathcal{K}_\tau$. $\mathbf{E}_{\text{LCE}}$ provides physicochemical context, while $\mathbf{E}_{\text{PCE}}$ is interpreted through ℓ_p according to each target's biological role; for example, influx and efflux substrate relations can imply opposite endpoint contributions. $\mathbf{E}_{\text{IE}}$ complements this assessment within the interpretive scope of the acquired signals, and unresolved relations retain their uncertainty. The final prediction synthesizes these evidence sources and the property-specific implications of each target relation. Integration is performed by an LLM conditioned on the specification and tool outputs.

4 Experiments

Our evaluation examines the predictive utility of evidence acquisition and integration guided by property-specific specifications, the contributions of the framework's design components, and the value of its generated rationales. Specifically, we analyze predictive performance across backbones (Section 4.1), property-cliff discrimination and the contributions of evidence and specifications (Section 4.2), and the downstream utility and assessed quality of rationales (Section 4.3). We use endpoints from TDC (Huang et al., 2021) and UniTox (Silberg et al., 2024) that concern the in vivo behavior and safety of drug candidates, with the labels defined by each benchmark as the prediction targets.

4.1 Property Prediction across LLM Backbones

Datasets and evaluation setting. We evaluate blood–brain barrier penetration (BBBP), drug-induced liver injury (DILI), human intestinal absorption (HIA), and oral bioavailability (Bioavailability) from the TDC ADMET collection. Comparisons cover four categories: general-purpose LLMs, domain-specific LLMs, molecular LLMs, and a multi-agent system. Model-level results are presented in Table 1. We report the Matthews correlation coefficient (MCC) (Chicco & Jurman, 2020) and Macro-F1 to assess performance under class imbalance. Given PharmLens's use of property-specific

Table 1: Molecular property prediction on four in vivo endpoints from TDC. MCC (left) and Macro-F1 (right) are reported for each endpoint. Overall is the average across the four endpoints. Best results in **bold**.

	MCC					Macro-F1				
Model	BBBP	DILI	HIA	Bioav	Overall	BBBP	DILI	HIA	Bioav	Overall
General LLMs										
Llama3.1	0.054	0.049	0.204	0.136	0.111	0.424	0.427	0.534	0.537	0.481
GPT-4o	0.339	0.131	0.631	0.255	0.339	0.613	0.553	0.637	0.626	0.607
DeepSeek-V4-Pro	0.341	0.334	0.541	0.359	0.394	0.668	0.659	0.718	0.659	0.676
Qwen3.7-max	0.471	0.230	0.687	0.293	0.421	0.701	0.615	0.805	0.616	0.684
Domain LLMs										
Galactica 6.7B	-0.120	-0.256	-0.188	-0.198	-0.191	0.199	0.333	0.388	0.401	0.330
Galactica 30B	0.304	0.00	0.194	-0.009	0.122	0.565	0.324	0.332	0.479	0.425
TxGemma-Chat-9B	0.449	0.011	0.333	0.211	0.251	0.712	0.495	0.594	0.582	0.596
TxGemma-Chat-27B	0.259	0.103	0.296	0.273	0.233	0.616	0.552	0.541	0.585	0.574
Molecular LLMs										
Mol-Instructions	-0.174	0.000	0.104	-0.127	-0.049	0.407	0.324	0.546	0.382	0.415
3D-MoLM	0.028	0.011	0.034	0.071	0.036	0.487	0.381	0.509	0.531	0.477
LLaMo	0.024	-0.130	0.222	-0.327	-0.053	0.489	0.354	0.596	0.261	0.425
Mol-LLaMA	0.161	0.151	0.034	-0.042	0.076	0.508	0.426	0.509	0.449	0.472
EDT-Former	0.154	-0.101	0.179	-0.134	0.024	0.549	0.350	0.466	0.433	0.449
MolecularGPT	0.231	0.131	0.167	0.000	0.132	0.417	0.401	0.302	0.195	0.329
Multi-Agents										
CLADD	0.244	0.187	0.098	0.090	0.155	0.603	0.411	0.538	0.538	0.522
PharmaLens (ours)	**0.481**	**0.357**	**0.756**	**0.410**	**0.501**	**0.726**	**0.669**	**0.875**	**0.694**	**0.741**

mechanistic knowledge to guide evidence acquisition and prediction, we restrict evaluation to TDC endpoints reflecting in vivo outcomes and follow the corresponding test set split protocols.

Predictive performance. PharmLens achieves the highest MCC and Macro-F1 on all four endpoints among the evaluated LLM-based systems (Table 1). Its mean MCC and Macro-F1 are 0.501 and 0.741, respectively, exceeding 0.421 and 0.684 for Qwen3.7-max, the strongest baseline on average. These results support the predictive utility of the complete procedure: specifying property-relevant factors and relations, acquiring molecule-specific evidence, and integrating that evidence.

Backbone ablation. We apply the same framework to Llama-3.1-8B, GPT-4o, DeepSeek-V4-Pro, and Qwen3.7-max, comparing each with its standalone counterpart. Mean Macro-F1 improves for all four backbones, with relative gains of 7.32–22.53% (backbone ablation table). Although the magnitude of improvement and the resulting performance vary across backbones, the gains across the evaluated LLMs indicate that specification-guided evidence acquisition and integration can improve prediction across multiple backbones.

4.2 Property-Cliff Discrimination and Specification Analysis

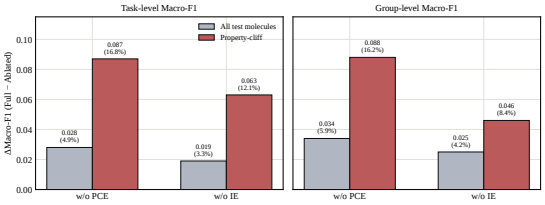
Dataset and evaluation setting. UniTox comprises eight organ- and system-specific toxicity endpoints, constructed by extracting toxicity information from the labeling documents of 2,418 FDA-approved drugs, with expert validation of a subset. We use a small-molecule evaluation set obtained after processing non-small-molecule entries, salts, and duplicates, retaining the baselines from Section 4.1 and an evaluation setting without additional fitting. Tasks are grouped into Circulatory (cardiac, pulmonary, and hematological toxicity), Metabolic (hepatic and renal toxicity), Sensory (dermatological toxicity and ototoxicity), and Reproductive (infertility). Group scores are molecule-count-weighted averages of the constituent tasks' Macro-F1 scores. Table 2 reports the group scores and Overall performance.

In addition to the full evaluation set, we evaluate a property-cliff subset comprising structurally similar molecules with opposite toxicity labels. A cliff pair (Van Tilborg et al., 2022) consists of two molecules with opposite labels and a similarity exceeding 0.9 under at least one of three measures: Morgan-fingerprint Tanimoto similarity, Bemis–Murcko scaffold-based similarity, or SMILES edit-based similarity. The subset contains molecules belonging to these pairs. This evaluation examines

Table 2: Molecular toxicity prediction on eight toxicity endpoints from UniTox, grouped by physiological system (**Circulatory**: Cardio, Pulmonary, Hemato; **Metabolic**: Liver, Renal; **Sensory**: Dermato, Oto; **Reproductive**: Infertility). Macro-F1 is reported for each group as the molecule-count weighted average of its constituent tasks. Overall is the average across the four groups. Best results in **bold**.

	All Test Molecules					Property-cliff Molecules				
Model	Circ.	Metab.	Sens.	Repro.	Overall	Circ.	Metab.	Sens.	Repro.	Overall
General LLMs										
Llama3.1	0.387	0.410	0.344	0.384	0.381	0.373	0.447	0.404	0.364	0.397
GPT-4o	0.463	0.526	0.437	0.532	0.489	0.458	0.503	0.513	0.410	0.471
DeepSeek-V4-Pro	0.491	0.523	0.486	0.587	0.522	0.428	0.470	0.479	0.487	0.466
Qwen3.7-max	0.473	0.506	0.539	0.607	0.531	0.426	0.477	0.515	0.515	0.483
Domain LLMs										
Galactica 6.7B	0.338	0.343	0.293	0.368	0.335	0.336	0.343	0.351	0.326	0.339
Galactica 30B	0.377	0.346	0.546	0.368	0.409	0.420	0.343	0.539	0.326	0.407
TxGemma-Chat-9B	0.379	0.439	0.360	0.488	0.417	0.368	0.435	0.413	0.351	0.392
TxGemma-Chat-27B	0.419	0.477	0.506	0.558	0.490	0.407	0.413	0.570	0.382	0.443
Molecular LLMs										
Mol-Instructions	0.322	0.337	0.287	0.368	0.328	0.373	0.343	0.351	0.326	0.348
3D-MoLM	0.380	0.391	0.300	0.465	0.384	0.380	0.436	0.389	0.439	0.401
LLaMo	0.376	0.351	0.279	0.499	0.376	0.453	0.389	0.324	0.497	0.416
Mol-LLaMA	0.426	0.420	0.415	0.485	0.437	0.418	0.443	0.441	0.428	0.433
EDT-Former	0.465	0.470	0.333	0.448	0.429	0.492	0.465	0.417	0.481	0.464
MolecularGPT	0.385	0.383	0.282	0.417	0.367	0.435	0.394	0.398	0.343	0.392
Multi-Agents										
CLADD	0.493	0.534	0.388	0.403	0.455	0.421	0.511	0.397	0.428	0.439
TheKnow (ours)	**0.507**	**0.558**	**0.619**	**0.670**	**0.588**	**0.516**	**0.554**	**0.583**	**0.531**	**0.546**

Table 3: **Left:** Evidence-layer ablation on UniTox, reported at task- and group- level for all test molecules and the property-cliff subset. PCE and IE denote Pathway-Centric and Interaction Evidence. **Right:** Backbone study, Macro-F1 averaged over four TDC tasks.



LLMs	LLM Only	+PharmLens	Avg. Gain
Llama3.1-8B	48.13	58.89	+22.36%
gpt-4o	60.73	69.48	+14.41%
deepseek-v4-pro	67.64	74.10	+9.55%
qwen3.7-max	68.41	73.42	+7.32%

the ability to distinguish opposite binary toxicity outcomes among similar molecules. We compare performance and the effects of evidence removal separately on the full set and the cliff subset.

Overall and cliff performance. Among the evaluated LLM-based systems,PharmLens achieves the highest performance in every group on the full evaluation set, with an Overall Macro-F1 of 0.588, exceeding the strongest baseline. It also outperforms the strongest baseline's score of 0.483 on the cliff subset (Table 2). These results support the ability to distinguish similar molecules with different labels using evidence newly acquired for each molecule, even when they share a property-level specification.

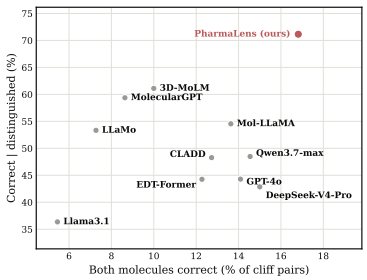


Figure 2: Distinguished rate and accuracy evaluation on property-cliff pairs.

Pair-level discrimination. Beyond per-molecule classification, we evaluate the proportion of all cliff pairs for which both molecules are classified correctly and the conditional accuracy among pairs assigned different predicted labels (Fig. 7). The former measures success across all pairs; the latter measures the correctness of a distinction when the model predicts different labels. PharmLens outperforms the evaluated LLM-based baselines on both metrics, correctly predicting opposite labels for more pairs while also improving the accuracy of the distinctions it makes.

Evidence-layer ablation. We remove $\mathbf{E}_{\text{PCE}}$ and $\mathbf{E}_{\text{IE}}$ separately and compare the resulting performance changes on the full evaluation set and the property-cliff subset. Removing either evidence layer

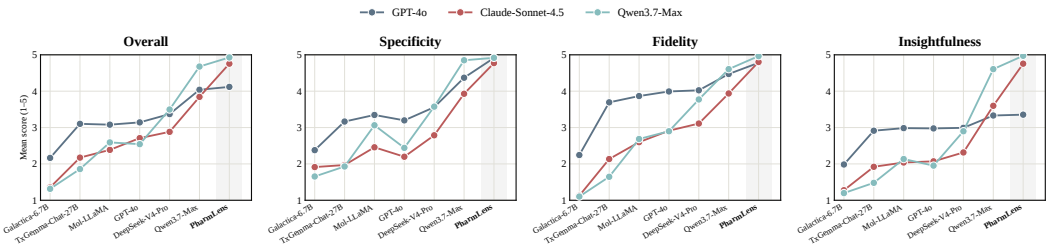


Figure 4: Evaluation of generated rationales on Bioavailability. (a) Mean scores assigned by three LLM judges to each model's rationales, across four criteria. (b) Distribution of expert rankings over 40 molecules. (c) Rank correlation between each LLM judge and the experts.

reduces performance, with larger decreases on the cliff subset. This pattern is consistent at both task and group aggregation levels (Table 3). These results support the particular contribution of property-relevant protein–relation evidence and auxiliary structural evidence to distinguishing similar molecules with different labels.

Mechanism-specification swap. To evaluate the effect of alignment between a target property and its specification, we replace each endpoint's specification with those of other UniTox endpoints, one at a time. The DeepSeek backbone, evaluation molecules, prediction target, and available tools are held fixed. For each replacement, evidence is reacquired according to the factors and relations specified by the substituted specification and used to generate the final prediction. We average Macro-F1 across conditions using other endpoints' specifications and report the relative change from the task-matched condition (Fig. 3).

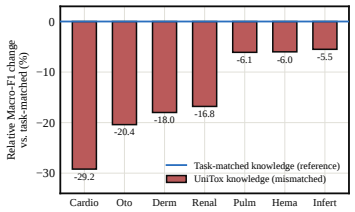


Figure 3: Relative Macro-F1 changes from the task-matched reference (blue line), averaged over swaps with other endpoints' specifications.

For each of the seven endpoints shown in the figure, performance averaged across mismatched specifications was lower than in the task-matched condition, with relative decreases of 5.5–29.2%. Swapping specifications jointly changes the targets of evidence acquisition and the knowledge used for interpretation. With the conditions held fixed, these performance differences support the predictive utility of using property-aligned specifications to organize evidence acquisition and interpretation. Whereas evidence ablation examines the contributions of the acquired evidence layers, this experiment highlights the importance of specifying which factors and relations to investigate through evidence acquisition. Further interpretation of these experiments is provided in the Appendix C.

Table 4: AUROC of ModernBERT fine-tuned on each model's generated rationale, with the SMILES string alone as a baseline.

Model	BBBP	DILI	HIA	Bioav	Overall
Only SMILES	84.54	79.52	90.99	64.54	79.90
General LLMs					
Llama3.1	81.97	75.92	88.54	62.85	77.32
GPT-4o	84.57	80.85	88.25	66.03	79.93
DeepSeek-v4-pro	85.47	80.58	89.58	65.25	80.22
Qwen3.7-max	85.17	78.52	90.13	67.57	80.60
Specialized Molecular LLMs					
Galatica 6.7B	82.48	82.82	85.66	67.26	79.56
Mol-LLaMA	85.31	79.78	74.56	65.33	76.25
EDT-Former	84.85	77.94	78.86	62.58	76.06
Multi-Agents					
TxGemma-Chat-9B	81.99	77.14	88.11	65.15	78.10
TxGemma-Chat-27B	83.42	76.62	86.50	65.91	78.11
Ours					
Ours	**86.23**	**83.44**	**93.05**	**68.02**	**82.69**

4.3 Utility and Assessed Quality of Generated Rationales

Alongside its prediction, PharmLens generates a rationale grounded in the specification and molecule-specific evidence. In the absence of property-specific ground-truth rationales, we assess their value along two dimensions: downstream predictive utility and assessed content quality. A ModernBERT probe evaluates the predictive information contained in the rationales, while LLM judges and experts assess their relative quality and agreement across evaluations.

4.3.1 PREDICTIVE UTILITY

Evaluation setting. Following the approach of MolCap-Arena (Edwards et al., 2024), we concatenate SMILES with each model's rationale in the form "SMILES[CLS]rationale" and fine-tune ModernBERT (Warner et al., 2025) for property prediction, comparing AUROC with a SMILES-only input. Sentences that explicitly state the predicted label are removed using the same procedure for all models' rationales. Here, supervised training of ModernBERT serves as a downstream probe of the informational utility of the generated rationales.

Results. Adding PharmLens's rationales yields the highest AUROC on all four endpoints and consistently outperforms the SMILES-only condition. Mean AUROC increases from 79.90 to 82.69 (Table 4). Rationales from other models underperform SMILES alone on some endpoints, indicating that adding a rationale does not inherently improve prediction. PharmLens's consistent gains support the additional predictive information provided in this probe by rationales that describe property-specific evidence.

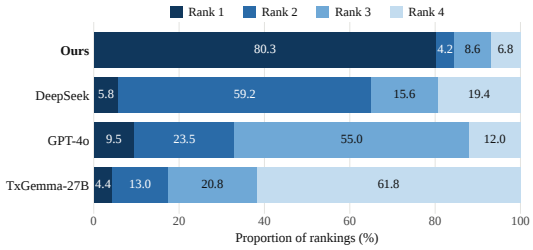


Figure 5: Expert ranking distributions for Bioavailability rationales, with Rank 1 indicating the highest assessed quality.

4.3.2 QUALITY ASSESSMENT BY LLM JUDGES AND EXPERTS.

LLM assessment. GPT-4o, Claude-Sonnet-4.5, and Qwen3.7-max independently evaluate Bioavailability rationales. Each judge assigns scores from 1 to 5 for Specificity (relevance to the particular molecule and property), Fidelity (consistency between the chemical content and the conclusion), Insightfulness (connections among the evidence), and Overall quality. Fidelity is a criterion for evaluating rationale content; the evaluation prompts are provided in the Appendix. PharmLens generally receives high scores for Overall quality, Specificity, and Fidelity, with a consistent advantage in Specificity. Differences in Insightfulness depend on the judge (Fig. 4).

Expert assessment. Experts in drug discovery, medicinal chemistry, and computational chemistry rank the rationales from PharmLens, GPT-4o, DeepSeek-V4-Pro, and TxGemma-27B from first to fourth for 20 Bioavailability molecules. PharmLens ranks first in 70.8% of the aggregated assessments, achieves a mean rank of 1.49, and is never ranked last (Fig. 5). This indicates that its rationales receive the highest relative quality assessment within this set of cases (Zheng et al., 2023).

Converting LLM-judge scores to per-molecule rankings and comparing them with expert rankings yields Kendall's τ of 0.724-0.885 and Spearman's ρ of 0.784-0.924 (Table 5. These results indicate similar judgments by LLM judges and experts regarding the relative quality of the evaluated Bioavailability rationales. Together with the downstream improvements, these evaluations support both the predictive utility of PharmLens's rationales and their high quality as assessed by experts.

Table 5: Agreement between LLM judges and experts in ranking Bioavailability rationales, measured by Kendall's τ and Spearman's ρ.

LLM Judge	Kendall's τ	Spearman ρ
GPT-4o	0.724	0.784
Claude-Sonnet-4.5	0.840	0.884
Qwnen3.7-max	0.885	0.924

5 CONCLUSION

We introduced PharmLens to specify which factors and functional relations to assess for molecular property prediction and how to acquire the necessary evidence. By organizing property-specific mechanistic knowledge into reusable specifications, PharmLens connects assessment-target selection with molecule-specific evidence acquisition and interpretation without requiring pre-existing molecular annotations. Predictive performance on TDC and UniTox, property-cliff discrimination on UniTox, and the downstream utility of generated rationales support the practical value of this approach. Evidence ablations and specification swaps further indicate that both property-aligned assessment targets and molecule-specific relation and interaction evidence contribute to predictive performance. These findings suggest that translating property-specific mechanistic knowledge into explicit molecular assessment procedures is a useful direction for evidence-based molecular property prediction.

AI USE STATEMENT

The authors used generative AI tools to assist with translation and improve manuscript clarity. These tools were also used to obtain feedback on methodology and experimental design and to check the consistency of theoretical analyses and proofs. The theoretical framework was developed by the authors, with AI assistance in this context limited to consistency checks and feedback. AI was not used to originate research concepts or hypotheses, implement the methodology, conduct experiments, interpret results, or draw conclusions. Additionally, LLMs were used to construct mechanistic knowledge specifications and evaluate the model's rationales through an LLM-as-a-judge approach, as described in the main text. All AI-assisted suggestions were reviewed and evaluated by the authors before adoption.

ETHICS STATEMENT

The proposed model is designed to support computational drug discovery research. Its predictions and rationales should not be interpreted as experimentally validated mechanisms or clinical recommendations. Biases and errors in source literature, predictive tools, and large language model (LLM) outputs may affect its assessments. Experiments were conducted using existing public benchmarks.

REPRODUCIBILITY STATEMENT

The source code is publicly available in the GitHub repository https://github.com/anonymousqwerty350/PharmLens linked in the submission.

REFERENCES

Membrane transporters in drug development. *Nature reviews Drug discovery*, 9(3):215–236, 2010.

Gerald T Ankley, Richard S Bennett, Russell J Erickson, Dale J Hoff, Michael W Hornung, Rodney D Johnson, David R Mount, John W Nichols, Christine L Russom, Patricia K Schmieder, et al. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environmental toxicology and chemistry*, 29(3):730–741, 2010.

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 *Proceedings of the 31st international conference on computational linguistics*, pp. 354–379, 2025.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

Nada J Daood, Sean R Carey, Elena Chung, Tong Wang, Anna Kreutz, Mounika Girireddy, Suman Chakravarti, Nicole C Kleinstreuer, Jacqueline B Tiley, Lauren M Aleksunes, et al. Machine learning modeling for abc transporter efflux and inhibition: Data curation, model development, and new compound interaction predictions. *Molecular pharmaceutics*, 22(11):7022, 2025.

Sarah Dawson, Simone Stahl, Nikki Paul, Jane Barber, and J Gerald Kenna. In vitro inhibition of the bile salt export pump correlates with risk of cholestatic drug-induced liver injury in humans. *Drug Metabolism and Disposition*, 40(1):130–138, 2012.

Carl Edwards, Ziqing Lu, Ehsan Hajiramezanali, Tommaso Biancalani, Heng Ji, and Gabriele Scalia. Molcap-arena: A comprehensive captioning benchmark on language-enhanced molecular property prediction. *arXiv preprint arXiv:2411.00737*, 2024.

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. In *International Conference on Learning Representations*, volume 2024, pp. 48221–48251, 2024.

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. *arXiv preprint arXiv:2102.09548*, 2021.

Dongki Kim, Wonbin Lee, and Sung Ju Hwang. Mol-llama: Towards general understanding of molecules in large molecular language model. *Advances in Neural Information Processing Systems*, 38:26921–26960, 2026.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2025 update. *Nucleic acids research*, 53(D1): D1516–D1525, 2025.

Namkyeong Lee, Edward De Brouwer, Ehsan Hajiramezanali, Tommaso Biancalani, Chanyoung Park, and Gabriele Scalia. Rag-enhanced collaborative llm agents for drug discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 561–569, 2026.

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. In *International Conference on Learning Representations*, volume 2024, pp. 17352–17371, 2024.

Zongru Li, Xingsheng Chen, Honggang Wen, Regina Qianru Zhang, Ming Li, Xiaojin Zhang, Hongzhi Yin, Qiang Yang, Kwok-Yan Lam, Pietro Lio, et al. A systematic survey and benchmark of deep learning for molecular property prediction in the foundation model era. *Journal of Chemical Theory and Computation*, 22(10):4866, 2026.

Xuan Lin, Long Chen, and Yile Wang. Attrilens-mol: Attribute guided reinforcement learning for molecular property prediction with large language models. *arXiv preprint arXiv:2508.04748*, 2025.

Jueon Park, Yein Park, Minju Song, Soyon Park, Donghyeon Lee, Seungheun Baek, and Jaewoo Kang. Cotox: Chain-of-thought-based molecular toxicity reasoning and prediction. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4002–4007. IEEE, 2025.

Philipp Schake, Sarah Naomi Bolz, Katja Linnemann, and Michael Schroeder. Plip 2025: introducing protein–protein interactions to the protein–ligand interaction profiler. *Nucleic Acids Research*, 53 (W1):W463–W465, 2025.

Jake Silberg, Kyle Swanson, Elana Simon, Angela Zhang, Zaniar Ghazizadeh, Scott Ogden, Hisham Hamadeh, and James Zou. Unitox: leveraging llms to curate a unified dataset of drug-induced toxicity from fda labels. *Advances in Neural Information Processing Systems*, 37:12078–12093, 2024.

Antonia F Stepan, Daniel P Walker, Jonathan Bauman, David A Price, Thomas A Baillie, Amit S Kalgutkar, and Michael D Aleo. Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the united states. *Chemical research in toxicology*, 24(9):1345–1410, 2011.

Derek Van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling*, 62(23):5938, 2022.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2526–2547, 2025.

Gregory L Warren, C Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H Lambert, Mika Lindvall, Neysa Nevins, Simon F Semus, Stefan Senger, et al. A critical assessment of docking programs and scoring functions. *Journal of medicinal chemistry*, 49(20):5912–5931, 2006.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. Molrag: unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15513–15531, 2025.

Yuejiang Yu, Chun Cai, Jiayue Wang, Zonghua Bo, Zhengdan Zhu, and Hang Zheng. Uni-dock: Gpu-accelerated docking enables ultralarge virtual screening. *Journal of chemical theory and computation*, 19(11):3336–3345, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Jiaxi Zhuang, Yaorui Shi, Jue Hou, Yunong He, Mingwei Ye, Mingjun Xu, Yuming Su, Linfeng Zhang, Ying Qian, Guolin Ke, et al. Reasoning-enhanced large language models for molecular property prediction. *arXiv preprint arXiv:2510.10248*, 2025.

Appendix of PharmLens

CONTENTS

A Implementation Details **14**

A.1 Construction of Task-Specific Knowledge 14

A.2 Measurement Space for Ligand-Centric Evidence 15

A.3 Relation Evaluators for Pathway-Centric Evidence 16

A.4 Protein–Ligand Interaction Evidence 17

B Experimental Details **18**

B.1 Property-Cliff Experiment 18

B.2 LLM-as-a-judge Evaluation 18

B.3 Human Expert Evaluation 19

C Further Analysis **21**

C.1 Discrimination of Property-Cliff Molecules 21

C.2 Functional Role of Property-Aligned Specifications 21

C.3 Robustness to Backbone Memorization 22

D Prompt Templates for Agents **22**

D.1 PhyschemAgent Prompt 22

A Implementation Details

A.1 Construction of Task-Specific Knowledge

$\mathcal{K}_\tau$ consists of a single JSON document that records the structural and physicochemical conditions involved in property expression, as well as the proteins involved in their manifestation, under molecular_context and relevant_proteins, respectively. The former lists the conditions that support and those that exclude the property, along with the mechanism by which each condition leads to the property. The latter contains, for each protein on the panel, its physiological role (`role`), the mechanism contributing to the property (ℓ_p, `logic`), the type of relationship to verify (`relation`), and the structure to be used for docking (`pdb_id`). In addition, it records the causal chain from initiating events to the property (`pathway_rules`), the definition of the property, the question sentence provided to the prediction agent, and the label name.

The document is generated by GPT-4o, which receives only the target property as input, searches PubMed abstracts, and constructs it therefrom; each item is required to specify the identifier and the corresponding sentence from the supporting literature. The prompt used for construction is included in Appendix D. Below, taking hematotoxicity as an example, we confirm that each item of $\mathcal{K}_\tau$ does not merely list proteins and conditions, but also contains the pathways leading to the property.

Construction of relevant_proteins. The relation assessed at all five proteins on the hematological-toxicity panel is the same, namely inhibition. The routes by which the establishment of inhibition leads to toxicity, however, differ from one another, and that difference comes from the mechanistic position that ℓ_p specifies.

Example of the relevant_proteins field of K^{path} for hematological toxicity.

DHFR

ROLE Folate metabolism and de novo nucleotide biosynthesis | RELATION inhibitor | PDB_ID `1KMS`

LOGIC Dihydrofolate reductase reduces dihydrofolate to tetrahydrofolate, the one-carbon carrier required for de novo thymidylate and purine synthesis. Depletion of tetrahydrofolate stalls DNA replication in rapidly dividing bone-marrow hematopoietic progenitors, producing myelosuppression; [...]

Schweitzer BI, Dicker AP, Bertino JR. Dihydrofolate reductase as a therapeutic target. *FASEB J* 1990;4(8):2441-52. PMID 2185970. Howard SC, McCormick J, Pui CH, *et al.* Preventing and managing toxicities of high-dose methotrexate. *Oncologist* 2016;21(12):1471-82. PMID 27496039.

DHODH

ROLE De novo pyrimidine biosynthesis in proliferating cells | RELATION inhibitor | PDB_ID `1D3G`

LOGIC Dihydroorotate dehydrogenase is a flavin enzyme on the inner mitochondrial membrane that oxidizes dihydroorotate to orotate, the rate-limiting step of de novo pyrimidine biosynthesis [...]. Depletion of uridine and cytidine nucleotides limits DNA and RNA synthesis in rapidly dividing bone-marrow progenitors, producing cytopenias; [...]

Reis RAG, Calil FA, Feliciano PR, *et al.* The dihydroorotate dehydrogenases: past and present. *Arch Biochem Biophys* 2017;632:175-91. PMID 28666740. Sykes DB, Kfoury YS, Mercier FE, *et al.* Inhibition of dihydroorotate dehydrogenase overcomes differentiation blockade in acute myeloid leukemia. *Cell* 2016;167(1):171-186.e15. PMID 27641501.

VKORC1

ROLE Vitamin K cycle and coagulation factor synthesis | RELATION inhibitor | PDB_ID `6WV3`

LOGIC Vitamin K epoxide reductase regenerates reduced vitamin K, the cofactor for gamma-carboxylation of coagulation factors II, VII, IX, and X in hepatocytes. Depletion of functional vitamin K-dependent clotting factors prolongs coagulation and produces bleeding; [...]

Rost S, Fregin A, Ivaskevicius V, *et al.* Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature* 2004;427(6974):537-41. PMID 14765194. Li T, Chang CY, Jin DY, *et al.* Identification of the gene for vitamin K epoxide reductase. *Nature* 2004;427(6974):541-4. PMID 14765195. Stafford DW. The vitamin K cycle. *J Thromb Haemost* 2005;3(8):1873-8. PMID 16102054.

PSMB5

ROLE Proteasomal protein turnover in megakaryocytes | RELATION inhibitor | PDB_ID `5LF3`

LOGIC PSMB5 carries the chymotrypsin-like active site that drives the rate-limiting proteolysis of the proteasome, degrading ubiquitinated proteins such as IkB and thereby permitting NF-kB activation. In megakaryocytes this NF-kB activity is required for proplatelet formation and platelet release [...]; its loss lowers circulating platelets. [...]

Shi DS, Smith MCP, Campbell RA, *et al.* Proteasome function is required for platelet production. *J Clin Invest* 2014;124(9):3757-66. PMID 25061876. Murai K, Kowata S, Shimoyama T, *et al.* Bortezomib induces thrombocytopenia by the inhibition of proplatelet formation of megakaryocytes. *Eur J Haematol* 2014;93(4):290-6. PMID 24750292. Lonial S, Waller EK, Richardson PG, *et al.* Risk factors and kinetics of thrombocytopenia associated with bortezomib for relapsed, refractory multiple myeloma. *Blood* 2005;106(12):3777-84. PMID 16099887.

Inhibition of DHFR and TYMS acts at two consecutive points of the folate-dependent pathway, and inhibition of DHODH acts on the de novo pyrimidine pathway; both deplete nucleotides and thereby block DNA synthesis in dividing bone-marrow progenitors. Inhibition of VKORC1 blocks the gamma-carboxylation of coagulation factors and leads to bleeding, while inhibition of PSMB5 blocks platelet release from megakaryocytes and leads to thrombocytopenia. Three distinct mechanisms thus sit under a single property label, and $\mathcal{K}_\tau^{\text{path}}$ decomposes them into the respective proteins and routes. The way in which an identical relation call is mapped onto the property diverges in this manner, and what specifies that arrangement is ℓ_p.

Construction of molecular_context. The same principle applies to $\mathcal{K}_\tau^{\text{lig}}$. Each condition is not merely stated as the presence or absence of a structure; it also carries the route by which that structure leads to the property.

Example of the molecular_context field of K^{lig} for hematological toxicity.

- Aromatic amine in which an amino, hydroxylamino, or nitroso nitrogen is bonded directly to an aromatic ring carbon; oxidative bioactivation to nitrenium and nitroso species covalently modifies marrow progenitors and neutrophils, linked to agranulocytosis and hemolytic anemia.
- Alkylating warhead such as a nitrogen mustard, aziridine, epoxide, or alkyl sulfonate ester that covalently crosslinks DNA and arrests division of hematopoietic progenitors.
- Antifolate pharmacophore built on a 2,4-diaminopteridine or 2,4-diaminopyrimidine that inhibits DHFR or TYMS and blocks folate-dependent nucleotide biosynthesis.
- Coumarin, 4-hydroxycoumarin, or indanedione motif that antagonizes vitamin K-dependent coagulation and increases hemorrhagic risk.

Aromatic amine is not stated merely as the name of a functional group; instead, structural requirements—such as nitrogen being directly bonded to an aromatic ring carbon—are detailed alongside its pathway of undergoing oxidative activation to attack bone marrow progenitor cells. Just as an alkylating warhead, an antifolate pharmacophore, and a coumarin skeleton lead to distinct pathways—DNA cross-linking, inhibition of nucleotide synthesis, and coagulation antagonism, respectively—a shift in the pathway indicated by these conditions changes the measurements that must be verified. Because this description defines the measurement selection for physicochemical agents, measurements at this layer become a procedure for verifying the quantities indicated by the mechanism, rather than an arbitrary set of descriptors.

A.2 Measurement Space for Ligand-Centric Evidence

The space $\mathcal{F}$ from which the physicochemical agent selects measurements is defined as the 208 two-dimensional descriptors exposed by RDKit's `Descriptors.descList`. Of these, 85 are the functional-group substructure counters (`fr_*`) provided by `Chem.Fragments`, and the remaining 123 are physicochemical and topological descriptors.

At the selection step, each item is presented to the agent in the form `name: description`, where the description is taken verbatim from the first line of the RDKit docstring. Table 6 lists the full candidate pool presented to the agent.

Table 6: Candidate measurement pool of molecular descriptors

Family	n	Descriptors
Size / composition	12	`MolWt`, `ExactMolWt`, `HeavyAtomMolWt`, `HeavyAtomCount`, `NumValenceElectrons`, `NumRadicalElectrons`, `NumHeteroatoms`, `RingCount`, `NOCount`, `NHOHCount`, `NumRotatableBonds`, `FractionCSP3`
Polarity / H-bonding	3	`TPSA`, `NumHDonors`, `NumHAcceptors`
Lipophilicity / refractivity / drug-likeness	4	`MolLogP`, `MolMR`, `LabuteASA`, `qed`
Partial-charge extrema	4	`Max/MinPartialCharge`, `Max/MinAbsPartialCharge`
Electrotopological extrema	4	`Max/MinEStateIndex`, `Max/MinAbsEStateIndex`
Topological complexity	4	`BalabanJ`, `BertzCT`, `HallKierAlpha`, `Ipc`
Fingerprint density	3	`FpDensityMorgan1-3`
Connectivity indices	12	`Chi0`, `Chi1`, `Chi{0-1}{n,v}`, `Chi{2-4}{n,v}`
Shape indices	3	`Kappa1-3`
Ring counts	9	`Num{Aromatic,Aliphatic,Saturated}{Rings,Carbocycles,Heterocycles}`
BCUT2D eigenvalues	8	`BCUT2D_{MW,LOGP,MR,CHG}{HI,LOW}`
Surface-area partitions (VSA)	57	`SlogP_VSA1-12`, `SMR_VSA1-10`, `PEOE_VSA1-14`, `EState_VSA1-11`, `VSA_EState1-10`
Functional-group counters	85	`fr_*` — the full set of SMARTS substructure counters in `Chem.Fragments`, e.g. `fr_Al_COO`, `fr_Ar_OH`, `fr_NH0/1/2`, `fr_aniline`, `fr_epoxide`, `fr_nitro`, `fr_quatN`, `fr_sulfonamd`, `fr_thiophene`
Total	**208**	

A.3 Relation Evaluators for Pathway-Centric Evidence

relation evaluator predicts relationships between molecules and proteins—the intermediate representations forming mechanisms—rather than property τ itself. The target relationships consist of the four types defined in Section 3.3 (substrate, inhibitor, antagonist, and agonist), and an evaluator is built for each defined type per protein. These relationship labels are distinct from the final target endpoint property τ (e.g., BBBP, DILI), and the molecules in the relationship dataset do not overlap with those in the TDC and UniTox datasets.

A.3.1 Construction of Training Data

The evaluator for each relationship type is trained on a dedicated training dataset constructed per relationship. For substrates, an author-curated open dataset (Daood et al., 2025) is used for four efflux transporters (P-gp, BCRP, MRP1, MRP2). The remaining three relationships are constructed from ChEMBL records. Each record is organized along two axes: potency and direction. Potency is converted to pChEMBL ($-\log_{10}$ of the molar concentration measurement) to compare records with different metrics and units on a unified scale, while direction is determined from the assay type (Table 7).

Table 7: Labeling criteria for the potency-based relations. Left: potency thresholds on pChEMBL. Right: direction inferred from the assay type.

Potency (pChEMBL)			Direction (assay type)		
pChEMBL	conc.	label	type	measures	direction
≥ 6	$\leq 1\,\mu$M	active	EC_{50}/AC_{50}	activation conc.	agonism
5–6	1–10 μM	discard	IC_{50}	inhibition conc.	antagonism, inhibition
< 5	$> 10\,\mu$M	inactive	K_i/K_d	binding affinity	none (binding only)

On the direction axis, binding constants (K_i, K_d) do not convey direction and are therefore not used as positive evidence for a relationship. For receptors or channels, K_i often indicates occupancy measured by radioligand displacement; treating this as inhibition reverses the direction. For instance, the DRD2 agonist bromocriptine would be misclassified as an antagonist based on high affinity. Consequently, binding constants are used solely as evidence of binding and serve as negative evidence indicating the absence of a relationship only when binding itself is not observed.

Combining the two axes yields the following labeling rules. For all three relationships, records with a pChEMBL of 6 or higher in assays directly measuring the respective relationship—that is, IC_{50} for inhibition and antagonism, and EC_{50} and AC_{50} for agonism—are taken as positive, while records with a pChEMBL under 5 are taken as negative. Binding constants are treated identically across all three relationships, where only records with affinity falling in the inactive range serve as negative evidence. Cell- or organism-level assays, as well as records whose assay descriptions fail to specify the measurement content, are excluded due to the inability to determine direction.

A.3.2 Relation Evaluator Design

The predictor for each target is an ensemble of 12 models representing all combinations of three molecular representations (ECFP6, MACCS, RDKit descriptors) and four algorithms (Random Forest, SVM, XGBoost, DNN). The average of their predicted probabilities, $P_p(m)$, serves as the relationship signal. Combining different representations and different inductive biases complements signals that a single representation might miss. Evaluation is conducted using scaffold split-based 5-fold cross-validation, and performance is reported in Table 8.

Table 8: Cross-validation performance of the relation predictors under scaffold split (mean ± std). CCR denotes balanced accuracy. Values are computed from the consensus probability of the 12-model ensemble.

relation	target	AUPRC	AUROC	CCR	MCC
inhibitor	MEK1	0.987 ± 0.008	0.975 ± 0.013	0.930 ± 0.032	0.851 ± 0.066
	DHODH	0.973 ± 0.015	0.960 ± 0.018	0.887 ± 0.035	0.781 ± 0.082
	mTOR	0.992 ± 0.003	0.966 ± 0.016	0.859 ± 0.061	0.760 ± 0.073
	ABL1	0.975 ± 0.014	0.946 ± 0.026	0.858 ± 0.035	0.711 ± 0.084
	PSMB5	0.975 ± 0.013	0.939 ± 0.031	0.825 ± 0.032	0.675 ± 0.070
	ACE	0.965 ± 0.015	0.918 ± 0.027	0.785 ± 0.076	0.647 ± 0.098
	EGFR	0.975 ± 0.006	0.927 ± 0.010	0.805 ± 0.033	0.642 ± 0.047
	TYMS	0.922 ± 0.051	0.887 ± 0.057	0.815 ± 0.041	0.632 ± 0.090
	BRAF	0.990 ± 0.005	0.947 ± 0.017	0.740 ± 0.048	0.603 ± 0.091
	hERG	0.766 ± 0.051	0.904 ± 0.020	0.710 ± 0.029	0.549 ± 0.050
	CYP19A1	0.959 ± 0.009	0.897 ± 0.015	0.757 ± 0.047	0.544 ± 0.069
	DHFR	0.890 ± 0.034	0.845 ± 0.071	0.760 ± 0.051	0.539 ± 0.085
	CYP17A1	0.987 ± 0.007	0.937 ± 0.038	0.705 ± 0.098	0.531 ± 0.208
	VEGFR2	0.962 ± 0.006	0.881 ± 0.009	0.719 ± 0.020	0.499 ± 0.030
	Nav1.5	0.612 ± 0.111	0.880 ± 0.041	0.635 ± 0.038	0.464 ± 0.084
	COX2	0.858 ± 0.061	0.809 ± 0.061	0.725 ± 0.073	0.464 ± 0.128
	Average	*0.924*	*0.914*	*0.782*	*0.618*
antagonist	ESR1	0.989 ± 0.013	0.966 ± 0.026	0.872 ± 0.028	0.777 ± 0.077
	DRD2	0.863 ± 0.142	0.900 ± 0.132	0.821 ± 0.118	0.700 ± 0.200
	Average	*0.926*	*0.933*	*0.847*	*0.738*
agonist	OPRM1	0.946 ± 0.025	0.967 ± 0.016	0.901 ± 0.028	0.813 ± 0.053
	PXR	0.790 ± 0.058	0.901 ± 0.040	0.798 ± 0.083	0.706 ± 0.115
	Average	*0.868*	*0.934*	*0.849*	*0.759*
substrate	MRP2	0.872 ± 0.040	0.863 ± 0.068	0.796 ± 0.055	0.589 ± 0.119
	MRP1	0.845 ± 0.141	0.844 ± 0.108	0.753 ± 0.136	0.484 ± 0.252
	Pgp	0.815 ± 0.050	0.813 ± 0.057	0.733 ± 0.062	0.471 ± 0.121
	BCRP	0.833 ± 0.044	0.819 ± 0.040	0.709 ± 0.057	0.421 ± 0.109
	Average	*0.841*	*0.835*	*0.748*	*0.491*

However, substrate determination for CYP drug-metabolizing enzymes is performed using CypReact, an established external substrate prediction tool, instead of this QSAR ensemble.

A.4 Protein–Ligand Interaction Evidence

Receptor Preprocessing. A single PDB structure is fixed for each protein. Valence-violating atoms and isolated hydrogens generated during conversion are cleaned up, followed by hydrogen addition and partial charge assignment using unidocktools proteinprep to convert the structure into PDBQT format.

Binding Site Definition. Binding sites are computed directly from the structure without manual coordinate specification. If a co-crystallized ligand is present, its position is utilized: crystallization

additives and buffer components are excluded, and the largest residue among candidates with 5–100 atoms is selected as the primary ligand. The docking box is centered at the geometric center of this ligand, and the edge length along each axis is defined by adding 12 Å to the ligand's span along that axis, capped within the range of 15–25 Å. This follows Vina's default protocol of adding padding to the bound ligand's coordinates and clamping with a lower bound, with the lower and upper bounds adjusted for an automated pipeline where binding site sizes vary substantially across targets. If no valid co-crystallized ligand is available, the same rule is applied to the top-ranked pocket predicted by fpocket.

Docking and Interaction Extraction. Docking is executed using UniDock (search_mode=balance), and subsequent analyses exclusively use the top-scoring pose. This pose is merged with the original protein structure to form a complex, and contacts are extracted using PLIP. Since both the reference ligand and query molecules undergo the exact same procedure, contact pattern comparisons represent a direct contrast between two poses generated in a uniform manner.

B Experimental Details

B.1 Property-Cliff Experiment

Purpose. This experiment is designed as a diagnostic to reveal what predictions are based on. Performance on the full evaluation set cannot distinguish between a model that has learned structure–label correlations and a model that operates through a mechanism. The two approaches diverge where structural similarity fails to predict labels; thus, we isolate pairs of molecules with nearly identical structures but opposite labels and separately evaluate performance on them.

Construction of Property-Cliff Pairs. Structural similarity determination follows the criteria of MoleculeACE (Van Tilborg et al., 2022). A pair is considered structurally similar if any of the following is 0.9 or higher: ECFP Tanimoto (atom-level composition), ECFP Tanimoto of the Bemis–Murcko generic framework (core scaffold), or scaled Levenshtein-based SMILES similarity (textual representation). Since these three metrics capture different types of structural differences, taking their union accounts for instances where a correlation-based predictor might treat two molecules as identical if indicated by any single metric. Only the label axis differs from the original benchmark. While MoleculeACE requires a 10-fold or greater difference in activity values, UniTox uses binary labels; therefore, pairs with opposite labels are defined as cliff pairs, and the molecules involved are defined as the cliff subset. Figure 6 shows cliff pairs drawn from the cardiotoxicity task.

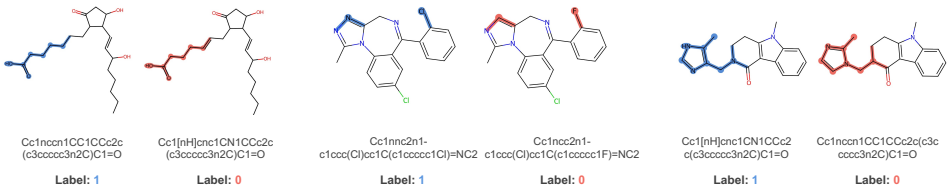


Figure 6: Examples of property-cliff pairs in the cardiotoxicity task.

B.2 LLM-as-a-judge Evaluation

Grading protocol. Grading is performed independently for each evidence unit. Judges view only one piece of evidence at a time, and the generative model's identity, ground truth labels, and evidence from other models are not provided. Grading is conducted once per judge at temperature = 0.

Judge prompt

System

You are an expert chemist evaluating the quality of a rationale produced for a molecular property prediction. For your reference, the SMILES of the molecule and the target property are provided. You are evaluating the quality of the reasoning, NOT whether the final predicted value is correct.

Evaluate the rationale on the following four criteria:

- Overall: How good is this rationale overall?
  1 - Poor, 2 - Below average, 3 - Acceptable, 4 - Good, 5 - Excellent

- Specificity: How well is the reasoning tailored to this specific molecule and property, rather than generic statements that could apply to any molecule?
  1 - Completely generic, 2 - Mostly generic, 3 - Minimally specific, 4 - Moderately specific, 5 - Highly specific to this molecule and property

- Fidelity: Is the reasoning chemically sound, valid, and relevant to the molecule, with the conclusion following logically from the stated grounds?
  1 - Unsound or irrelevant, 2 - Weakly grounded, 3 - Partially sound, 4 - Mostly sound, 5 - Fully sound and well-supported

- Insightfulness: Does the rationale offer meaningful chemical insight useful to an expert, beyond restating obvious facts?
  1 - No insight, 2 - Minimal, 3 - Some, 4 - Good insight, 5 - Deep, expert-level insight

First, provide a brief explanation of your assessment. Then, score each criterion on a scale of 1 to 5, where a higher score indicates better quality. Follow the format below:

Explanation of the evaluation: ...
Final Decision:
- Overall: ...
- Specificity: ...
- Fidelity: ...
- Insightfulness: ...

User

[Molecule Information]
SMILES: **{SMILES}**
Target Property: **{PROPERTY}**

[Rationale]
{RATIONALE}

Results by Endpoint. The main text presents Bioavailability as a representative example. Tables 9, 10, 11, and 12 report all scores assigned by the three judges across the four metrics for each of the four endpoints. Regardless of judge or endpoint, PharmLens consistently ranks near the top, with a gap particularly noticeable in Specificity.

B.3 Human Expert Evaluation

Recruitment. We publicly recruited researchers in drug discovery, medicinal chemistry, computational chemistry, and pharmacokinetics through an online survey. The recruitment call was distributed via relevant research community channels, and eligibility was restricted to individuals holding a bachelor's degree or higher in related fields and currently engaged in research or industry practice.

Participants. A total of 39 participants completed the survey. Table 13 summarizes the demographic composition of the participants. By highest degree attained, 14 participants held a Ph.D. (35.9%), 14 held a Master's degree (35.9%), and 11 held a Bachelor's degree (28.2%). By primary role, industry and research institute researchers constituted the largest group at 26 (66.7%), followed by 9 graduate students (23.1%), 3 professors (7.7%), and 1 postdoctoral researcher. In terms of primary discipline, 12 participants were in pharmacy/medical science, 16 in biology/biotechnology, 5 in computational biology, 3 in chemistry, and 3 in veterinary/animal science. Because the evidence evaluated herein centers on target relationships and binding mechanisms, a pharmacological and biological background forms the core of the evaluation, supplemented by expertise in chemistry and computational fields to

Table 9: LLM-as-a-judge scores (1–5 scale, mean over 127 molecules). Three judges scored independently; arms are ordered by degree of molecular specialization. Ov. = Overall, Sp. = Specificity, Fi. = Fidelity, In. = Insightfulness, Avg. = Average.

Rationale source	Claude-Sonnet-4.5					GPT-4o					Qwen3.7-Max				
	Ov.	Sp.	Fi.	In.	Avg.	Ov.	Sp.	Fi.	In.	Avg.	Ov.	Sp.	Fi.	In.	Avg.
Galactica-6.7B	1.354	1.913	1.126	1.276	1.417	2.165	2.378	2.244	1.984	2.193	1.315	1.654	1.102	1.197	1.317
TxGemma-Chat-27B	2.173	1.969	2.134	1.921	2.049	3.102	3.165	3.693	2.913	3.218	1.858	1.929	1.646	1.480	1.728
Mol-LLaMA	2.386	2.457	2.598	2.039	2.370	3.079	3.346	3.866	2.984	3.319	2.591	3.063	2.685	2.134	2.618
GPT-4o	2.709	2.197	2.913	2.071	2.473	3.142	3.197	3.992	2.976	3.327	2.543	2.441	2.898	1.953	2.459
DeepSeek-V4-Pro	2.882	2.787	3.110	2.315	2.774	3.370	3.559	4.024	2.992	3.486	3.496	3.575	3.772	2.898	3.435
Qwen3.7-Max	3.843	3.929	3.937	3.598	3.827	4.039	4.370	4.472	3.331	4.053	4.677	4.850	4.606	4.606	4.685
TheKnow	**4.756**	**4.772**	**4.803**	**4.756**	**4.772**	**4.118**	**4.913**	**4.780**	**3.354**	**4.291**	**4.928**	**4.917**	**4.964**	**4.973**	**4.946**

Table 10: LLM-as-a-judge evaluation of rationales for HIA

Rationale source	Claude-Sonnet-4.5					GPT-4o					Qwen3.7-Max				
	Ov.	Sp.	Fi.	In.	Avg.	Ov.	Sp.	Fi.	In.	Avg.	Ov.	Sp.	Fi.	In.	Avg.
Galactica-6.7B	1.345	1.755	1.291	1.164	1.389	1.918	2.118	2.100	1.773	1.977	1.300	1.509	1.255	1.073	1.284
TxGemma-Chat-27B	2.336	2.218	2.200	1.873	2.157	2.955	3.000	3.509	2.764	3.057	2.045	2.245	1.855	1.609	1.939
Mol-LLaMA	2.491	2.555	2.691	2.073	2.452	3.036	3.173	3.791	2.882	3.220	2.809	3.273	2.918	2.318	2.843
GPT-4o	2.764	2.409	2.982	2.145	2.575	3.209	3.282	3.982	2.964	3.359	2.773	2.827	2.991	2.173	2.691
DeepSeek-V4-Pro	3.155	2.964	3.645	2.564	3.082	3.482	3.636	4.155	3.036	3.577	3.936	4.100	4.264	3.227	3.882
Qwen3.7-Max	3.955	3.982	4.136	3.582	3.914	4.036	4.255	4.564	3.400	4.066	4.655	4.773	4.582	4.618	4.657
TheKnow	**4.709**	**4.591**	**4.891**	**4.473**	**4.666**	**4.064**	**4.800**	**4.791**	**3.409**	**4.264**	**4.973**	**4.991**	**4.964**	**4.991**	**4.980**

Table 11: LLM-as-a-judge evaluation of rationales for DILI

Rationale source	Claude-Sonnet-4.5					GPT-4o					Qwen3.7-Max				
	Ov.	Sp.	Fi.	In.	Avg.	Ov.	Sp.	Fi.	In.	Avg.	Ov.	Sp.	Fi.	In.	Avg.
Galactica-6.7B	1.511	1.798	1.362	1.426	1.524	1.904	1.989	1.957	1.851	1.926	1.181	1.362	1.053	1.085	1.170
TxGemma-Chat-27B	1.840	1.809	1.915	1.787	1.838	2.883	2.989	3.319	2.670	2.965	1.574	1.617	1.574	1.319	1.521
Mol-LLaMA	2.340	2.277	2.713	2.106	2.359	3.128	3.309	3.819	2.862	3.279	2.681	2.947	2.809	2.138	2.644
GPT-4o	2.521	2.160	2.894	2.149	2.431	3.468	3.596	3.968	2.968	3.500	2.830	2.840	3.160	2.181	2.753
DeepSeek-V4-Pro	3.053	3.106	3.500	2.713	3.138	3.713	3.989	4.287	3.213	3.801	4.330	4.479	4.521	3.957	4.322
Qwen3.7-Max	3.521	3.585	3.681	3.319	3.527	4.000	4.596	4.351	3.394	4.085	4.553	4.564	4.521	4.553	4.548
TheKnow	**4.011**	**4.277**	**4.096**	**3.979**	**4.090**	**4.255**	**4.702**	**4.819**	**3.830**	**4.402**	**4.936**	**4.936**	**4.915**	**4.968**	**4.939**

Table 12: LLM-as-a-judge evaluation of rationales for BBBP

Rationale source	Claude-Sonnet-4.5					GPT-4o					Qwen3.7-Max				
	Ov.	Sp.	Fi.	In.	Avg.	Ov.	Sp.	Fi.	In.	Avg.	Ov.	Sp.	Fi.	In.	Avg.
Galactica-6.7B	1.288	1.705	1.207	1.154	1.339	1.933	2.195	2.027	1.813	1.992	1.219	1.475	1.099	1.086	1.220
TxGemma-Chat-27B	2.267	2.182	2.242	1.961	2.163	3.049	3.086	3.520	2.820	3.119	2.158	2.291	1.985	1.712	2.036
Mol-LLaMA	2.421	2.509	2.561	2.028	2.380	3.101	3.379	3.833	2.911	3.306	2.739	3.190	2.756	2.190	2.719
GPT-4o	2.832	2.432	3.084	2.204	2.638	3.303	3.433	4.000	2.990	3.432	2.793	2.781	3.037	2.089	2.675
DeepSeek-V4-Pro	3.246	3.137	3.572	2.726	3.170	3.690	3.877	4.200	3.116	3.720	4.135	4.276	4.342	3.564	4.079
Qwen3.7-Max	3.846	3.926	3.975	3.533	3.820	4.054	4.387	4.502	**3.554**	4.124	4.648	4.769	4.579	4.589	4.646
TheKnow	**4.547**	**4.551**	**4.723**	**4.274**	**4.524**	**4.079**	**4.860**	**4.840**	3.392	**4.292**	**4.978**	**4.990**	**4.966**	**4.980**	**4.978**

support structure-level assessments. This composition ensures that evaluations are not biased toward the perspective of any single subfield.

Table 13: Demographics of the human evaluation participants (N=39).

(a) Highest degree

	N (%)
Ph.D.	14 (35.9)
M.S.	14 (35.9)
B.S.	11 (28.2)

(b) Position

	N (%)
Researcher	26 (66.7)
Graduate student	9 (23.1)
Faculty	3 (7.7)
Postdoc	1 (2.6)

(c) Field

	N (%)
Pharmaceutical & medical	12 (30.8)
Biology & biotechnology	16 (41.0)
Computational biology	5 (12.8)
Chemistry	3 (7.7)
Veterinary & animal science	3 (7.7)

Evaluation protocol. Each participant was presented with evidence generated by four models (PharmLens, GPT-4o, DeepSeek-V4-Pro, and TxGemma-27B) for 20 molecules and assigned rankings from 1st to 4th based on relative quality under blinded conditions. The presentation order of the evidence was randomized across items to eliminate position bias. All participants ranked the full set of molecules presented, with no missing responses.

C Further Analysis

C.1 Discrimination of Property-Cliff Molecules

PharmLens's relative advantage in the property-cliff evaluation indicates that property-specific evidence acquisition and integration provide useful discriminative information even when structurally similar molecules have opposite labels (Table 2). The pair-level results reinforce this interpretation. PharmLens outperforms the baselines both in the proportion of all cliff pairs for which both molecules are classified correctly and in conditional accuracy among pairs assigned different predicted labels. Its advantage on both metrics means that it successfully distinguishes more molecular pairs and that, when it predicts different labels, those distinctions agree with the true labels more frequently. This supports its ability to jointly distinguish opposite outcomes among similar molecules, beyond average classification performance on individual molecules.

The $\mathbf{E}_{\text{PCE}}$ and IE ablations identify evidence sources that contribute to this discrimination. Removing either source reduces performance on the full set, but the decrease is larger on the cliff subset, with consistent patterns at both task and group aggregation levels. The contributions of property-relevant protein–relation evidence and auxiliary structural evidence are therefore particularly pronounced when distinguishing similar molecules. This observation also relates to the design that separates reuse of a property-level specification from molecule-specific assessment. A shared specification defines the factors and relations to examine for an endpoint, while evidence acquired for each molecule informs the assessment of those items. Taken together, cliff performance and the ablations support the utility of sharing property-specific assessment criteria while capturing molecule-specific differences through relation and structural evidence.

C.2 Functional Role of Property-Aligned Specifications

For every evaluated endpoint, the mean performance across alternative specifications is lower than under the task-matched specification. This indicates the importance of property-aligned knowledge organization for prediction even with identical condition. The knowledge encoded in a specification identifies the molecular factors and protein–relation hypotheses to assess and provides the property-specific direction for interpreting the resulting evidence. The results therefore support the predictive relevance of the knowledge constructed in this study under the evaluated conditions and its utility in guiding evidence acquisition and interpretation aligned with the target property.

Evidence ablation and specification swapping examine different aspects of this design. Ablation shows that relation and structural evidence acquired by the procedure contribute to prediction, whereas swapping shows that specifying which factors and relations require evidence also matters. Together,

the results support the contributions of both molecule-specific evidence use and property-specific assessment-target selection to predictive gains. This clarifies the functional role of a specification as an explicit intermediate representation linking literature knowledge to assessment targets and tool execution. Improvements across multiple backbones further support the utility of this evidence acquisition and integration procedure beyond any single evaluated LLM.

C.3 Robustness to Backbone Memorization

Since the backbone of PharmaLens is DeepSeek-V4-Pro and the labels of UniTox are derived from FDA drug labels, there is a possibility that the backbone encountered the safety records of those drugs during pre-training. That is, the performance might originate from compound-level knowledge memorized by the backbone, rather than mechanism reasoning. To verify this, the molecules that the backbone failed to predict on its own ($n = 615$) are defined as an interval where the backbone lacks usable prior knowledge, and the performance of PharmaLens is measured exclusively on this subset. Because the Macro-F1 of the backbone in this interval is 0.000 by definition, the focus is not on contrasting with the backbone, but on how well PharmaLens maintains its performance on the same backbone.

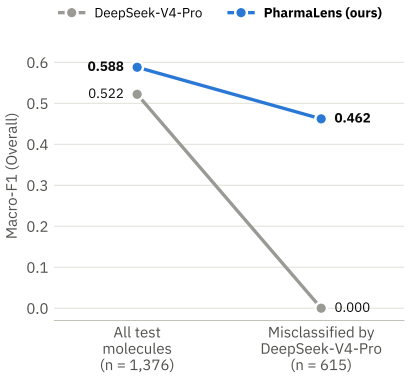


Figure 7: Retention on the backbone's error subset. Macro-F1 on the full test set vs. the subset DeepSeek-V4-Pro answers incorrectly.

PharmaLens achieves 0.462, retaining 79% of its overall dataset performance (0.588). The fact that accurate predictions are maintained even for molecules that the backbone cannot answer from memory indicates that the source of performance is not the replay of memorized labels, but the measured evidence and pathway logic. Had it relied on memory, the performance in this interval would have converged to zero.

D Prompt Templates for Agents

D.1 PhyschemAgent Prompt

PhyschemAgent prompt

Stage 1 — measurement selection $\mathcal{F}_\tau$)

System

You are a medicinal chemist choosing which molecular measurements are important for: **{TASK_DESCRIPTION}**

You do not compute anything. You only decide what is worth computing.

User

Task: **{TASK_DESCRIPTION}**

=== Structural features associated with each class ===
Features associated with the POSITIVE class:
{POSITIVE_FEATURES}

Features associated with the NEGATIVE class:
{NEGATIVE_FEATURES}

=== AVAILABLE DESCRIPTORS ===
{DESCRIPTOR_LIST}

=== AVAILABLE FUNCTIONAL GROUPS ===
{FG_LIST}

=== Your task ===
Below are the RDKit tools available. Select ONLY those whose value would actually change your judgement for this task.

There is no target count. Select every tool that is genuinely diagnostic for this task.

Think carefully before writing your answer.

Respond with a single JSON object and nothing else.
```json
{
  "descriptors": ["exact names from the descriptor list above"],
  "functional_groups": ["exact fr_∗ names from the functional group list above"],
  "rationale": "Why these, tied to the mechanism of this specific task. Name which features you expect to separate the classes and in which direction."
}
```

Stage 2 — assessment of the measured $\mathcal{F}_\tau$ (per molecule)

System

You are a medicinal chemist assessing the ligand-intrinsic evidence for: **{TASK_DESCRIPTION}**

You are given measurements only — RDKit computed them, they are exact, and none of them is pre-labelled as good or bad. Interpreting them is your job.

Judge only what the structure itself supports. A molecule whose physicochemistry is favourable can still be ruled out downstream, and that is fine — say what the structure says, no more.

User

Molecule: **{SMILES}**

=== Measured properties ===
{MEASURED_VALUES}

=== Structural features associated with each class (domain knowledge) ===
POSITIVE class:
{POSITIVE_FEATURES}

NEGATIVE class:
{NEGATIVE_FEATURES}

=== Your task ===
Decide whether the measured structure favours the positive class, the negative class, or neither.

Cite the specific numbers that drive your call — a claim with no number behind it is not evidence.

Reason from two sources, and use both:
1. The domain-knowledge features listed above.
2. Your own medicinal-chemistry knowledge of how these measured properties govern this task.

The list above is a starting point, not a closed set — it does not exhaust what matters here:
- If a measured value is diagnostic for a reason the list never names, say so and use it.
- If your own knowledge contradicts the list, say that too rather than deferring to the list.
- Either way, the number still has to be on the table. An appeal to your own knowledge is held to the same standard as any other claim.

Think carefully before writing your answer.

Respond with a single JSON object and nothing else.
```json
{
  "verdict": "favors_positive" | "favors_negative" | "neutral",
  "confidence": 0.0-1.0,
  "key_evidence": ["each entry names a measured value and what it implies, e.g. 'TPSA=32.7, far below the ~90 ceiling for passive crossing'"],
  "matched_context": {"positive": ["which POSITIVE-class features this molecule actually matches"], "negative": ["which NEGATIVE-class features it actually matches"]},
  "rationale": "2-3 sentences."
}
```

RelationAgent prompt

System

You are auditing the ligand-protein evidence assembled for: **{TASK_DESCRIPTION}**

Your primary task for each protein is to evaluate and determine if the provided evidence is sufficient to establish a definitive relationship with the molecule.

Escalate a protein when:
- The relation-evaluator probability is ambiguous.
- No relation-evaluator exists for it.
- The relation-evaluator call and the docking evidence point in opposite directions.

Escalation is not failure; it routes the protein to a structural comparison against ligands whose relationship to it is known.

User

Molecule: **{SMILES}**

=== Ligand-protein evidence ===
{PROTEIN_EVIDENCE}

Assess each protein listed above. Return one entry per protein, in the same order.

Respond with a single JSON object and nothing else.
```json
{
  "assessments": [
    {
      "protein": "protein name exactly as given",
      "sufficient": true | false,
      "trigger": null | "ambiguity_band" | "no_ml_model" | "ml_docking_conflict",
      "reason": "One sentence. If sufficient, state the relationship the evidence establishes and what settles it. If not, state precisely what is missing."
    }
  ]
}
```

InteractionAgent prompt

System

You determine what relationship a molecule has with one protein, by comparing how it binds against how known ligands of that protein bind. Context: **{TASK_DESCRIPTION}**

You are here because the QSAR evidence could not settle it. Recognition is the CONTACT PATTERN — the specific residues a real ligand must engage, which follow from the protein's binding mechanism.

Otherwise, compare the contacts first and the score second:
- A molecule that reproduces the known actives' contact signature is likely a real ligand even with a mediocre score.
- A molecule that makes none of those contacts is likely not one, however good its score looks.

If the reference set is small or the evidence genuinely does not separate, answer "uncertain".

User

Molecule: **{SMILES}**
Protein: **{PROTEIN}**
Relationship in question: is this molecule a **{RELATION}** of **{PROTEIN}**?

=== What recognition means for this protein (domain knowledge) ===
{RECOGNITION_LOGIC}

=== How THIS molecule binds ===
{QUERY_BINDING}

=== How KNOWN **{RELATION}**S of **{PROTEIN}** bind (reference set, source: **{ACTIVE_SOURCE}**) ===
{ACTIVES_REFERENCE}

=== How KNOWN NON-**{RELATION}**S bind (reference set, source: **{INACTIVE_SOURCE}**) ===
{INACTIVES_REFERENCE}

=== Contact frequency, actives vs non-actives (sorted by the gap) ===
{CONTACT_CONTRAST}

Percentages are the fraction of reference molecules making that contact.
How separated the two halves are overall: **{DISCRIMINABILITY}**

=== Your task ===
Two different score baselines appear above, and they are not interchangeable:
- The query's quartile is cut from the whole task train set.
- The reference p25/p50/p75 are cut from these reference molecules alone.

Compare the query's raw kcal/mol against the reference numbers — never its quartile, which says nothing about where it sits among known ligands.

Compare the contact pattern first, then the score. State which known-active contacts this molecule reproduces and which it misses.

Respond with a single JSON object and nothing else.
```json
{
  "relation": "**{RELATION}**" | "non-**{RELATION}**" | "uncertain",
  "confidence": 0.0-1.0,
  "evidence": "Name the specific contacts reproduced and missed, and the score relative to the reference distribution. Then say what that pattern means mechanistically."
}
```

PredictionAgent prompt

System

You are an expert pharmacologist making the final call on: **{TASK_DESCRIPTION}**

Three specialists have already reported, and their findings are given to you as evidence:

(1) A physicochemical assessment of the ligand itself.
(2) Relationship verdict per protein.
(3) Docking scores and PLIP contacts, for the proteins where that evidence was retained.

Read the evidence honestly.

=== Reasoning through each per-protein finding, then explaining it ===
Each protein has a relationship in question for this task (substrate, agonist, antagonist, inhibitor, and so on). The relationship, and what it means for this task, is fixed by the decision rules and the protein's mechanism below — read it off the rules for each protein; it is not the same across proteins. A finding tells you whether that relationship holds, does not hold, or is uncertain. Reason through it in three steps.

Step 0 — Is this protein on the causal path at all? From the decision rules, decide whether this protein's mechanism acts on the process the question asks about. A mechanism that operates downstream of, or beside, that process does not bear on the answer, and no finding about it — either way — changes your call. Say so and move on to the next protein.

Step 1 — Direction if the relationship holds. From the decision rules and the protein's mechanism, determine which task class it points to when its relationship holds. Derive this from each protein's own mechanism: two proteins sharing a relationship type can point to opposite classes (an efflux substrate and an influx substrate are both substrates yet oppose each other).

Step 2 — Apply the actual finding.
- Holds → contributes in the Step 1 direction, weighted by how strong the evidence is.
- Does not hold → this pathway is not operating for this molecule. That REMOVES its contribution; it does not create one in the opposite direction. Absence of a mechanism is not evidence for the opposite class.
  Then judge how much the absence even tells you: if engagement would have been expected from the structure, its absence is informative about that pathway; if molecules like this one are non-engagers by default, it tells you almost nothing.
- Uncertain → little weight.

=== When the panel says little ===
These proteins are a subset of the pathways that determine the answer, not all of them. If none of them is engaged, that is not a case for the negative class — it means the mechanistic panel is silent for this molecule, and the call rests on the physicochemical evidence and on what the structural features below indicate. Do not convert silence into a verdict in either direction.

In the RATIONALE, for each protein, explain in mechanistic terms whether it bears on this process at all, which class holding its relationship would point to, and what the actual finding changes. Give the reason, not the label.

=== Weighing physicochemistry against mechanism ===
The two carry independent weight; neither overrules the other by default. A molecule with an ideal profile can still be blocked by a transporter, and a molecule with a poor profile can still get in through a carrier.

=== Mechanistic decision rules for this task ===
{PATHWAY_RULES}

=== Structural features associated with each class ===
POSITIVE class:
{POSITIVE_FEATURES}

NEGATIVE class:
{NEGATIVE_FEATURES}

User

Molecule: **{SMILES}**

=== Evidence (1): Physicochemical assessment ===
{PHYSCHEM_EVIDENCE}

=== Evidence (2, 3): Ligand-protein relationships, docking and contacts ===
{PROTEIN_EVIDENCE}

=== Question ===
{PREDICTION_QUESTION}

=== Your task ===
Write your reasoning as connected prose: what the physicochemistry implies and what each protein contributes. For each protein, explain in mechanistic terms whether it bears on this process at all, which class holding its relationship would point to, and what the actual finding changes — remembering that a relationship which does not hold removes that pathway's contribution rather than creating one in the opposite direction. Give the reason, not the label.

Then, on a new line at the very end, give the final answer in exactly the format the question specifies. Write the exact phrase "final answer" nowhere except this last line.