Investigating LLMs understanding of causality in measurements using OpenADMET experiments

Anonymous Authors

Anonymous Institution

Abstract

In scientific contexts such as evaluating drug metabolism and toxicity, to what extent do LLMs distinguish between a drug's PXR reporter-assay activity and evidence that the drug engages PXR? We test whether LLMs distinguish a drug's assay activity from evidence that it engages the target protein, using paired primary and counter-screen assays as a causal identification testbed. Selectivity was linearly decodable across experiments, with the caveat that an arithmetic baseline reaches AUROC 1.00. Selectivity was fragile to counter-screen availability: probes failed to reliably transfer when the counter-screen was withheld during training and supplied during testing, demonstrating use of the counter-screen. Causally, transferring residual states between mathematically equivalent assay readings shifted the PXR-dependence verdict by 8.75 logits at the PXR reading, indicating that the model's representation of selectivity is not notation-invariant. Behaviorally, the model rarely requests the missing scientific control: 0% when asked directly about selectivity, 44.4% only when asked to evaluate a colleague's claim, and when it does propose controls, it often asks for the wrong one. Decodability does not by itself establish a stable, unit-invariant concept that the model consistently uses.

Introduction

How do we know if the reasoning process by which models arrive at an answer is trustworthy, especially in real world science? And how do we calibrate the quality of LLM understanding of science? This paper is exploratory: we investigate the hypothesis that exceptionally well-controlled scientific experiments offer a test-bed for evaluating LLM understanding under non-ideal conditions, where noise and epistemic uncertainty are just features of the domain.

There is a dearth of advanced empirically grounded causal reasoning problems for LLM evaluation [Xu et al., 2025]. Established areas of science that reason about complex chains of mediating mechanisms therefore offer a useful model system to evaluate how LLMs reason about such biomedical reasoning problems [Lin et al., 2026]. Yet existing benchmarks often suffer from lexical shortcuts, context dependence, reliance on biases in the scientific literature, failure to employ knowledge of scientific mechanisms, and information leakage [Kunievsky and Evans, 2025, Schouten et al., 2025, Mueller et al., 2025]. We sought a setting where the underlying causal question is posed and in which these confounders are minimized.

Chemical Reasoning Problems

One such setting is evaluating potential drug toxicity. Predicting toxicity in novel drugs is an extremely challenging scientific problem that takes multiple years and phase 1 clinical trials to resolve [Bender et al., 2026]. Although foundation models and agentic workflows perform chemical reasoning, they are often validated on experimental outcomes that are themselves confounded or limited in their predictive value [Walters, 2023]. The OpenADMET consortium

seeks to create gold-standard datasets and blind prospective benchmarking of methods that predict functional properties of drugs [Fraser et al., 2026].

We focus on the pregnane X receptor (PXR). PXR is a protein that regulates drug metabolism in the liver, so measuring how drugs interact with it helps predict the risk of drug-drug interactions. These interactions have historically been a major reason for drug failures and even FDA withdrawals. PXR is an instructive test case because it has a large hydrophobic ligand-binding pocket that allows a wide range of molecules to bind to it; it is, in a word, promiscuous. These features create real-world conditions in which an investigation into an important long horizon problem depends on effective causal reasoning in the presence of noise, ambiguity, epistemic uncertainty and ignorance ("unknown unknowns") [MacDermott-Opeskin et al., 2026].

However, this is not the complete story. Assay results are dose-dependent, which means that numerous dose-response curves could fit the same pEC50 and E_{max} values. We therefore use selectivity as a lens for investigating how models reason under these conditions. There are numerous measurements and biological proxies one could use; the two most commonly used models to quantify ligand bias are relative-relative Log (E_{max}/EC50) [Kolb et al., 2022].

Causality & Interpretability

Recent work in interpretability has highlighted the need to consider better specification of causal questions, recognizing what is and isn't identifiable from an experiment [Lin and Liu, 2026, Geiger et al., 2023, Canby et al., 2025]. PXR counter-screen serves as the placebo control; when PXR is disabled, it captures only the nonspecific sources of false positives; so calling a compound a true activator would require counterfactual reasoning: compare the assay result to what would have happened without PXR activity, which is the inferential core of causal identification.

Our contributions

We seek to evaluate whether this scientific resource could be transformed into a testbed for methods important to causal interpretability or scientific reasoning evaluation. In this paper, we make progress towards this goal, through first, formalizing the causal structure of the PXR counter-screening design, and second, designing appropriate evaluations.

Problem Setup

We use the recent PXR blind challenge dataset, which contains different assay results. Our experiments focus on paired PXR reporter and PXR-null counter-screen measurements, together with molecular structures and dose-response summaries [Open ADMET Consortium, 2026, MacDermott-Opeskin et al., 2026]. Each experiment has different technical requirements (residual-stream access vs. instruction-following quality), so we chose different parameter-sized Gemma models to focus on while explaining our results; other runs are found in the appendix.

The measure pEC50 is the negative log half-maximal effective concentration, while E_{max} is the maximal efficacy. The counter-screen is a cell line with PXR disabled, so a real PXR activator cannot light it up. Analogous to the placebo group in a clinical trial, it is a control that may catch reasons the typical PXR assay might be a false positive. The two reporter assays sort compounds four ways: a real PXR activator, a nonspecific signal, an assay artifact, an inactive, as shown in Figure 1 [Open ADMET Consortium, 2026, MacDermott-Opeskin et al., 2026]. To quantitatively distinguish true positive activators from these artifacts, a selectivity and efficacy

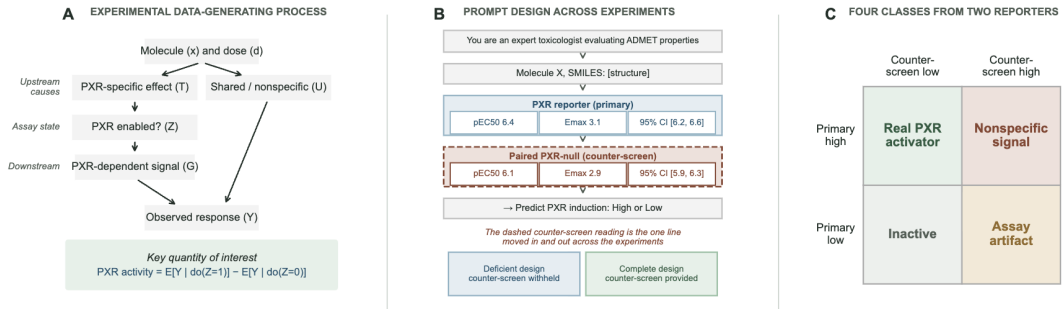

Figure 1: 1A: The real world data-generating process for what PXR selectivity actually is. 1B: an example of the prompt design template across experiments. 1C: What selectivity is related to the counter-screen and primary screen levels.

index I can be calculated to compare the primary screening assay (A) directly against the counter-screen control (B):

$$I = (pEC_{50,A} - pEC_{50,B}) + \log_{10}\left(\frac{E_{\max,A}}{E_{\max,B}}\right).$$

Results

A) Selectivity is linearly decodable, but not necessarily directly represented.

One experimental design that allows us to check if selectivity is actually being rendered as a concept is to withhold information that would be required to construct it, using differently trained probes to test generalization. For this experiment, the model is provided with a molecule; the probe must determine whether it is selective, not just whether it is potent. In order to test the model's internal representations, we did two arms, each with a reverse version. The PXR assay information is included in all experimental conditions. In the primary arm, the forward case: the probe is trained on molecules (null-withheld) and then is run on molecules with null-provided. In the reverse primary, can a complete probe (null-provided) properly run on an incomplete test case (null-withheld)? If the model is representing selectivity, then it should not be doing well on the primary arm or reverse primary arm due to missing null values for training and testing respectively. We also test margin transfer by training on compounds far from versus near the selectivity boundary (Fig.2C-D)

As expected, a probe that is trained to differentiate for selectivity without the null being provided cannot differentiate for selectivity on null-provided test sets. A probe trained on null-withheld data does not have high predictive power, even when the null is added in. Interestingly, when you train on compounds closer to the decision boundary and test on compounds further away, the probes become more accurate, and vice versa. However, simple arithmetic reconstruction can explain most of this decodability (a scalar pEC50_PXR-minus-pEC50_null baseline achieves AUROC 1.00; see Appendix). Overall, selectivity is linearly decodable, but this overstates what the model represents. One follow-up question we have to this is if the assay values given to the model are mathematically equivalent, will the internal representations be causally

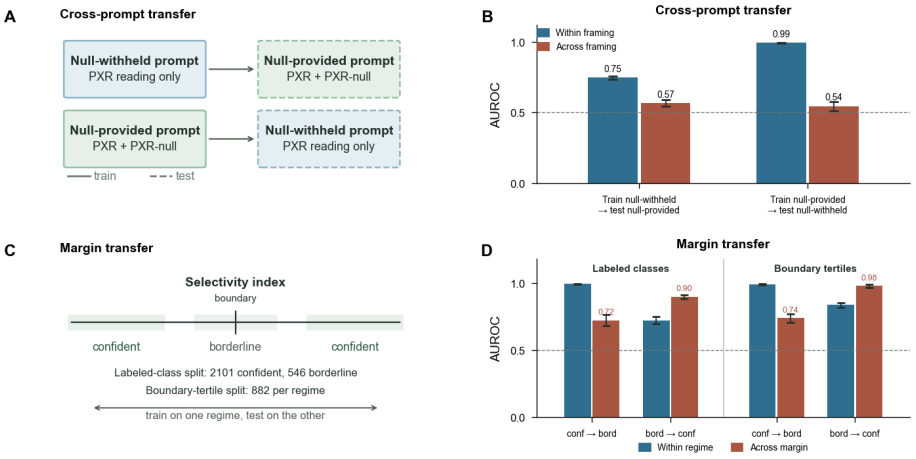

Figure 2: 2A: The structure of the cross-prompt transfer. 2B: The complete probe cannot survive null removal, and a probe trained without the null cannot predict selectivity even if the null is provided in the test set. 2C: Structure of the margin transfer. 2D: In addition, a probe trained on selective/non-selective molecules that are close to the borderline performs better on a confident set, and a probe trained on 'confident' molecules performs worse on borderline molecules, implying that classification is determined by some continuous latent score, points far from the decision boundary are simply easier to classify than points near it.

interchangeable, if the model is using basic arithmetic to calculate selectivity?

B) The model's reasoning is not invariant if assay numbers are swapped out with an equivalent scientific notation.

What happens if you patch the pEC50-representation activations with nM-equivalent activations at the same token position: the same values, represented differently? We give Gemma-4-12B-it two assay readings, PXR and PXR-null, for one compound and ask whether the compound activates PXR. A robust model of selectivity should not have its answer change based on an equivalent value given in a different unit. Afterwards, at a specific printed number, we took the model's residual stream at one reading from one notation and patched it into the other. We observed the difference of logits between High and Low, and rendered every compound with both orders (PXR or PXR_null first).

The claim we're testing is whether the model holds a unit-invariant notion, when swapping the value from pEC50 to nM or nM to pEC50, the verdict shouldn't change since we're patching on the same compound with the same measurement. However, we find that patching does indeed change the verdict when going from one notation to the other. Therefore, this implies that internal states associated with mathematically equivalent assay values are not causally interchangeable at the tested token. Within a single compound, patching at the residual stream from the pEC50 rendering into the nM rendering moves the verdict in opposite directions: at the PXR reading it pushes the verdict towards PXR-dependent by 8.75 logits, and at the PXR_null reading it pushes the verdict towards not PXR-dependent by 1.77 logits. We observe the reverse effects for both readings when going from nM rendering to pEC50 rendering. Because the nM

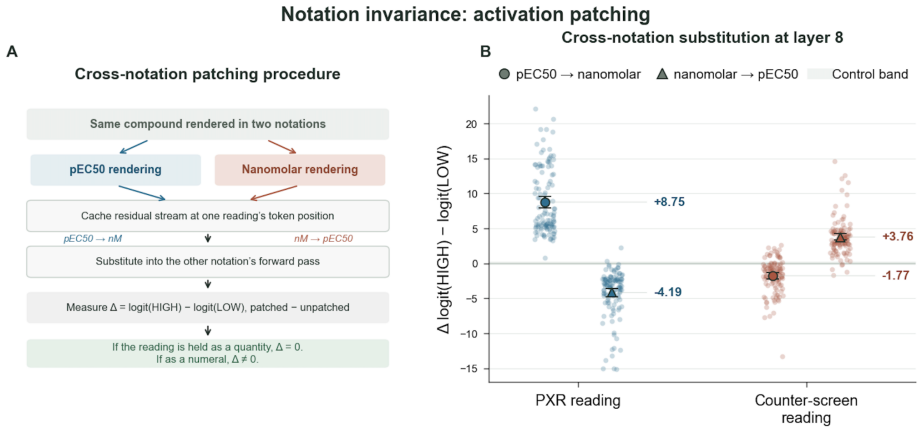

Figure 3: 3A: Explaining the methodology of the experiment: rewriting the assay reading in another unit but equal quantity moves the model's verdict. Each compound prompt is rendered twice, once with both PXR and PXR_null written as pEC50 values and once as EC50 values in nM. The two renderings represent the same measurement, EC50_nM = 10^(9-pEC50). For each compound, we take from the residual stream at the final digit of one reading and patch it into the same compound other rendering. 3B: Substituting pEC50 into nanomolar and vice versa affects the logits considerably.

reading 'looks' larger than the pEC50 reading, it could just be that the model is using a weak general principle of 'big number in the potency space implies selectivity' and 'big number for null implies no selectivity'. This leaves us with the question of if the model will be able to behaviorally self-identify that it requires a reading to clarify the selectivity of a compound, and if it will request it, which we follow up on in the next experiment.

C) Models typically do not notice a missing scientific control, and do not request it for it to use.

A PXR reporter reading on its own cannot establish that a compound activates PXR. The paired PXR-null counterscreen is what separates a real activator from an assay artifact. We tested this behaviorally, showing Gemma-4-31B-it the reporter reading alone across 48 compounds. Does the model recognize that a counterscreen is required before it can claim activation or selectivity? Or does it overclaim from potency data alone? We used seven framings that present the identical reading as bare data, as drug-drug interaction risk assessment, as a go/no-go screening decision, as an exchange with a colleague, as a request for follow-up experiments, as a direct question about selectivity, and as a direct question about whether the signal is PXR-dependent. The last two only differ in the question asked. One asks whether the compound is a selective PXR inducer, the other whether it produces a true PXR-dependent signal, meaning one that requires functional PXR. Each framing exists as a pair where condition A shows the PXR reporter pEC50 and condition B adds the counterscreen pEC50. Every prompt describes the assay's mechanism without naming the assay. We ran the battery with and without those descriptions. Full prompting text is provided in the appendix.

We suspected that flagging is framing-dependent: the highest rate (44.4%) occurs when the model evaluates a colleague's claim, suggesting prompts that invite skepticism elicit control-seeking

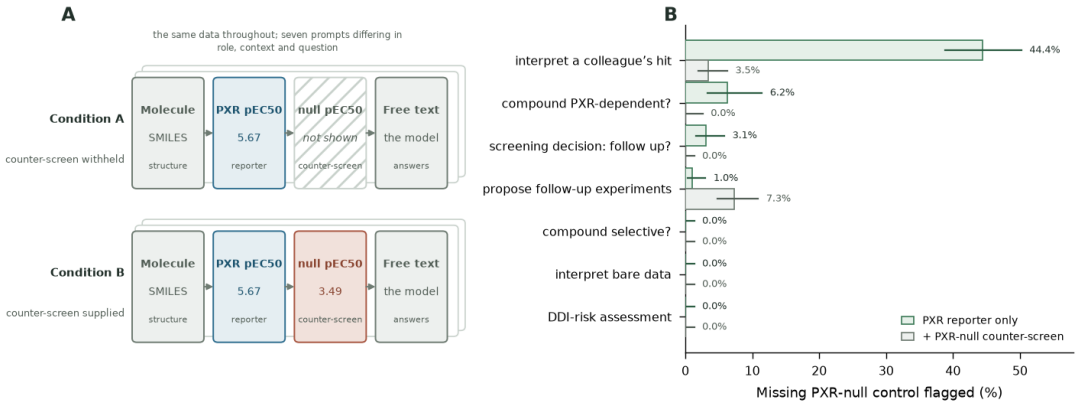

Figure 4: When Gemma-4-31B-it is asked to interpret a colleague's screening hit, it flags that the PXR assay is missing the counter-screen required to determine PXR dependence. Asked whether the same compound is a selective PXR inducer, to assess its DDI risk, or simply what the readings show, it does not. 4A: Two experimental conditions (PXR counter-screen withheld versus supplied) rendered across seven prompt variations each. 4B: Missing counter-screen flag rate by framing and condition. Interpreting a colleague's hit flags most often, at 44.4%

behavior (Fig. 4B). Flags remain when the counter-screen is supplied in 31 of 1,872 responses, nonzero only in further experiments (7.3%) and colleague's hit (3.5%) and zero in the other five. These are the two framings that ask what to do next, and a prompt asking for next steps may invite the model to name the PXR-null condition directly rather than accept the description it was given. Where the model proposes other controls instead, for example under explicit selectivity, 280 of 288 answers mention a panel of other receptors. We suspect two reasons. The word "selective" in the field means selective across receptors, while in our dataset it means the signal requires PXR. The assay description states that the readout runs through a chimeric PXR ligand-binding domain. Told that the signal passes through a ligand-binding domain, the model starts reasoning about what else binds there, such as CAR and VDR. In this behavioral experiment, the measure is a property of the prompt: whether the framing invites skepticism, and whether the vocabulary helps the model down the path of noticing the missing control.

D) Are LLMs able to understand that dose-dependence matters for selectivity classification without being told?

Recall that selectivity is not just the arithmetic concept of PXR pEC50 - null pEC50. As shown in Figure 1, the relationship between the PXR and null pEC50 values are necessary for a molecule to be truly selective. In this experiment, we give the models different combinations of the following: just the pEC50s, the pEC50s and the Emax, the pEC50+Emax+standard errors, pEC50+Emax+standard errors and a dose table, or pEC50+Emax+standard errors and plotted curves (that match the dose table). Detailed examples are in the appendix. We designed synthetic dose-response tables and their corresponding curves that fall into two regimes (that are indistinguishable from just pEC50 and EMax values). If the models understand that the dose-response information is critical for selectivity, then they will be able to classify them correctly if the table/curve is given, and should hedge when the table/curve is not. Code is available in the repository.

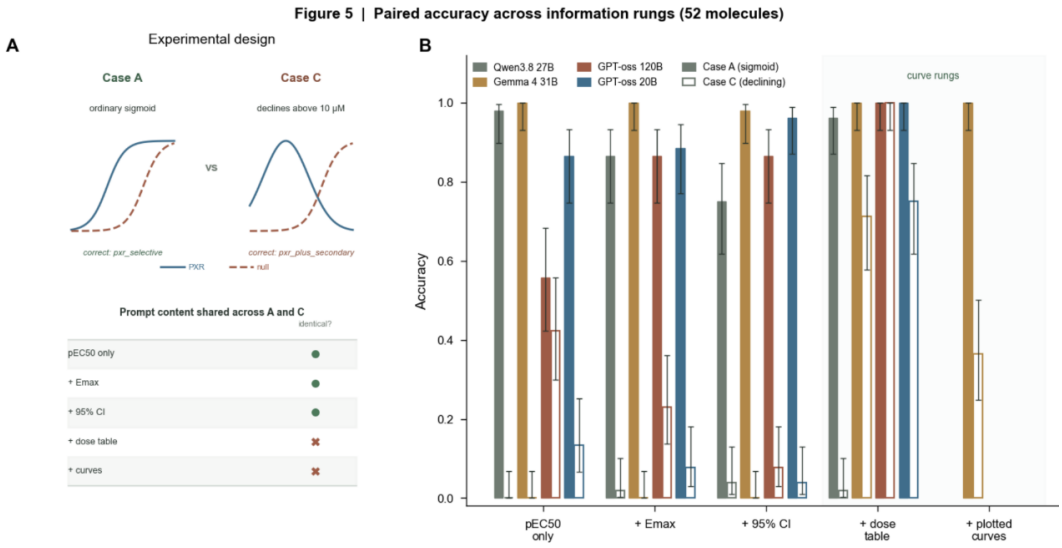

Figure 5: Dose-dependence if given pEC50, Emax, but if the model is given the graph or table, that leads to a different conclusion than the 'lazy' arithmetic the model can do, will it actually use that information?

There are seven molecules in the dataset that fall under the defined terms of Case A and Case C, so the rest of the generated 52 molecules have synthetically defined A and C curves. Figure 5a shows the difference between Case A and Case C, where despite having equivalent pEC50 and Emax values, A is pxr_selective while C also has a secondary response. You cannot distinguish A and C unless you also have a dose table or curve to accompany the pEC50 and Emax values. As shown in 5B, all models are able to identify A correctly by default, while they are only able to identify Case C if a dose-table is provided, or a plotted curve. Some weaker models (like Qwen3.8 27B) fail to classify C, even with the dose-table. This implies that models that fail this task are not behaviorally using the more complicated dose-dependent response information, and are instead relying on pEC50/EMax values.

Limitations

Our work did not thoroughly stress test more causal understanding in PXR chemical reasoning using further causal abstractions. We only tested linear probes. If selectivity is represented nonlinearly, the linear probes would miss it. The experiments would also benefit from ablation for further insight. We have also only run these experiments on the PXR dataset: do the probes for PXR selectivity generalize to other xenobiotic receptors, or possibly generalize even further? The scope of the research so far was limited: we tested a small number of models and established causal insensitivity to notation but did not localize which components carry the information that the probes recover.

Future Work

We plan to stress test more causal understanding in PXR chemical reasoning using further causal abstractions, including better accounting for nonlinear representations and ablations of the model. We also plan to extend this work to other scientific domains to further test the causal understanding of LLMs for scientific concepts, with more variations on real-world data and concepts. For instance, can we use the same probes to test for causal understanding of other scientific concepts, such as protein folding, gene regulation, or metabolic pathways? Even more specifically, can we take leading true/false probes and test them to verify biological claims, learning something new from models? This work can help us test the usefulness of having a mechanistic interpretability benchmark made from real world scientific problems and data, as they provide natural ways for us to verify how models 'understand' and internally model different concepts, and how this changes with naturally occurring perturbations. Such benchmarks can help transparently audit AI systems for their understanding of scientific concepts.

Going beyond probes and PXR assays

In this paper, we primarily investigated several behavioral experiments and used probes to test if concepts that are found together in the literature but are in fact different can be measured within LLMs. This result can be strengthened by using tools from causal abstraction to map what lower level features such as PXR specific biology, or assay design information and the notion of negative controls feature in an LLMs representation of evidence about PXR dependence.

MIB did a terrific job establishing standardized causal variable localization [Mueller et al., 2025], with tasks deconstructing arithmetic operations, indirect object identification and multiple choice question answering. This has the benefit of creating good controlled experiments to evaluate causal claims such as localization capabilities of interpretability tools. But validation occurs only in the context of logic or math. We hypothesize there is a different way to make verifiable controlled tasks with real-world ambiguities. Imagine if we took the best controlled scientific experiments that are not well represented in pre training data and turned them into benchmarks for interpretability and alignment that automated behavioral evaluation tools such as Bloom could run [Gupta et al., 2025].

This paper establishes encouraging evidence that such a benchmark is both feasible, reveals mistakes in LLM reasoning, and offers examples of behaviors that would be worth a suite of tests. A broader benchmark program would make it possible to evaluate mechanistic interpretability in scientific domains where ground truth is available but difficult to verify at scale.

Conclusion

Taken together, these results show that the model's representations carry information about selectivity but that information can be explained by an arithmetic approach rather than a notation-independent concept. While selectivity is linearly decodable from internal activations, probes cannot succeed when the counter-screen is withheld during training and supplied during testing (does not exceed chance: 0.50). Cross-notation activation patching shifts verdicts by 8.75 logits at the primary reading, which indicates that the model does not have a perfect arithmetic representation of selectivity. Behaviorally, the model rarely requests the missing scientific control: 0% when asked directly about selectivity, 44.4% only when asked to evaluate a colleague's claim.

Andreas Bender, Morgan C. Thomas, Isidro Cortés-Ciriano, et al. Artificial intelligence in drug discovery—what it is, where we stand and the path forward. *Nature Reviews Drug Discovery*, 2026. doi: 10.1038/s41573-026-01496-2. URL https://www.nature.com/articles/s41573-026-01496-2.

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 857–878, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.ijcnlp-long.47. URL https://aclanthology.org/2025.ijcnlp-long.47/.

James S. Fraser et al. Mapping the avoid-ome: a systematic open-science approach to predictive ADMET. *Nature Communications*, 17:4644, 2026. doi: 10.1038/s41467-026-73410-8. URL https://www.nature.com/articles/s41467-026-73410-8.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstraction for faithful model interpretation, 2023. URL https://arxiv.org/abs/2301.04709.

Isha Gupta, Kai Fronsdal, Abhay Sheshadri, Jonathan Michala, Jacqueline Tay, Rowan Wang, Samuel R. Bowman, and Sara Price. Bloom: an open source tool for automated behavioral evaluations, 2025. URL https://github.com/safety-research/bloom. Software.

Peter Kolb et al. Quantifying ligand bias using the black and leff operational model: A practical guide for the pharmacologist. *British Journal of Pharmacology*, 2022. URL https://bpspubs.onlinelibrary.wiley.com/doi/10.1111/bph.15811.

Nadav Kunievsky and James A. Evans. Measuring intent comprehension in LLMs, 2025. URL https://arxiv.org/abs/2506.16584.

Victoria Lin, Taedong Yun, Maja Matarić, John Canny, Arthur Gretton, and Alexander D'Amour. The illusion of intervention: Your LLM-simulated experiment is an observational study, 2026. URL https://arxiv.org/abs/2605.20767.

Zhengxuan Lin and Fengyuan Liu. Position: Mechanistic interpretability must disclose identification assumptions for causal claims, 2026.

Hugo MacDermott-Opeskin, Jon Swain, Scott Simpkins, Bryan Jiang, Sam Sabaat, Robert Warnerford-Thompson, Galen Correy, Nikhil Gupta, and Pat Walters. Announcing the next OpenADMET blind challenge: Predicting PXR induction, 2026. URL https://openadmet.ghost.io/announcing-the-next-openadmet-blind-challenge-predicting-pxr-induction/.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Open ADMET Consortium. PXR Challenge Train/Test Dataset, 2026. URL https://huggingface.co/datasets/openadmet/pxr-challenge-train-test. Dataset.

Stefan F. Schouten, Peter Bloem, Ilia Markov, and Piek Vossen. Truth-value judgment in language models: "truth directions" are context sensitive. In *Proceedings of the Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2H85485yAb.

Patrick Walters. We need better benchmarks for machine learning in drug discovery, 2023. URL https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html. Practical Cheminformatics blog.

Xinnuo Xu, Rachel Lawrence, Kshitij Dubey, Atharva Pandey, Risa Ueno, Fabian Falck, Aditya V. Nori, Rahul Sharma, Amit Sharma, and Javier Gonzalez. RE-IMAGINE: Symbolic benchmark synthesis for reasoning evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 69331–69362. PMLR, 2025. URL https://proceedings.mlr.press/v267/xu25n.html.

Appendix

Additional Information for the Experiments

Experiment A (detailed in Figure 2)

Methodological item	Status	Detail
Model identity	Gemma-2-9B-IT (primary)	Qwen2.5-7B-Instruct as replication (fixed-split). Both open-weights instruction-tuned models.
Layer and position	Layer 14 residual stream	Last prompt token; 2048-dim activation vectors. Backbone-only forward pass (lm_head skipped).
Layer sweep	Tested 7 layers	Layers 4, 8, 12, 14, 16, 20, 24. Pattern not layer-specific: primary below chance at all layers; reverse saturates to 1.00 by layer 14.
Activation extraction	One vector per (pair x pole x regime)	Forward hook on residual stream; cached to activations.parquet.
Leakage audit	Pass	Rendered prompts reviewed before inference. Withheld fields (contrast labels, selectivity class, pair roles, FDA status) confirmed absent.
Compound-disjoint train/test splitting	Yes	Pair-level leave-one-pair-out (LOPO); no compound recurs across H0 pairs. Purged LOPO also available.
Sample size (N)	18 H0 pairs	36 poles total (18 treatment, 18 control). Per fold: 34 train poles, 2 test poles.
Positive/negative balance	Balanced	18 treatment (label=1), 18 control (label=0) per LOPO round.
Cross-validation scheme	18-fold LOPO	Each pair held out exactly once; pooled out-of-fold AUROC over all 18 folds.
Random seeds	Seed 0	Single deterministic seed; fold order randomized from base seed.
Probe regularization	L2 logistic regression	C=1.0 (sklearn default), max_iter=2000.
Dimensionality reduction	None	Probe trains directly on 2048-dim residual-stream vectors.
Preprocessing within fold	Yes	StandardScaler re-fit per fold inside sklearn Pipeline; no global fit.
Confidence intervals	Pair-level bootstrap	10,000 resamples; 95% percentile intervals. Primary OOD AUROC: 0.86 [0.83, 1.00]. Reverse OOD: 0.73 [0.50, 0.89].
Label permutation null	Done	100 shuffles of training labels within each LOPO fold. Null mean=0.513 (std=0.104); observed=0.265; $p<0.01$.
Scalar baseline: PXR alone	AUROC 0.389	[0.278, 0.472]. Raw pEC50_pxr weakly anti-predicts treatment vs control.
Scalar baseline: null alone	AUROC 0.000	[0.000, 0.000]. Counter-screen perfectly anti-predicts (every treatment has lower pEC50_null).
Scalar baseline: PXR minus null	AUROC 1.000	[1.000, 1.000]. Selectivity gap perfectly separates treatment from control (by construction of H0 pairs).
Random-direction null	Median AUROC 0.554	200 random unit vectors projected onto all H0 complete poles; median reported as chance-level reference.
Random-label probe	Same as label permutation null	Training on shuffled labels and evaluating with true labels; 100 permutations within LOPO.
Second model replication	Qualitatively consistent	Gemma-2-9B (LOPO) and Qwen2.5-7B (fixed-split) both show asymmetric fragility: reverse drop > 0, primary drop < 0.
Primary transfer result	OOD AUROC 0.86 [0.83–1.00]	Deficient-potency probe tested on complete prompts. Transfer drop: -0.48 [-0.72, -0.28]. Adding counter-screen improves separation.
Reverse transfer result	OOD AUROC 0.73 [0.50–0.89]	Complete-trained probe tested on deficient-potency prompts. Transfer drop: +0.28 [+0.11, +0.50]. Removing counter-screen degrades probe.

Model selection rational

We provide clarification on our selection on model uses. Probe-based experiments(A, layer-wise decoding) use Gemma-2-9B for layer-wise residual-stream extraction and LOPO cross-validation are tractable at this scale. Activation-patching experiments (B) use Gemma-4-12B, which provides clean residual-stream access and well-documented layer norms for cross-notation substitution. Behavioral experiments (C) use Gemma-4-31B for stronger instruction-following

and longer-context stability across prompt batteries. The dose-response generalization test (Appendix) includes Qwen3.8-27B, GPT-oss 20B/120B, and Gemma-4-31B to check whether findings hold across model families and sizes. Model availability and API access at the time of experimentation determined which architectures could be included for each arm.

How far does the verdict move when you patch each quantity under pEC50?

First experiments show that the operands for selectivity are surface numerals rather than abstract quantities: cross-notation activation patching (pEC50 ↔ nM, same molecule) moves the verdict by up to 13 logits on 30 of 32 compounds, so the model holds the printed number, not a unit-invariant reading. Note: the ~5:1 PXR-over-counter-screen weighting should be added to the appendix and referred here. Also, selectivity is an operation that needs both operands present: removing the counter-screen (null) at inference breaks the trained probe to chance (H2, reverse drop +0.28, two models), and withholding it in training installs the potency shortcut at equal score (H7, lever +0.14 [+0.07, +0.21]), with an ablation confirming the mechanism is actually just differencing the two supplied numbers.

The representation of potency decodes at an early layer when PXR block is shown first (plot 5), at layer 8, and decodes at later layer when the PXR block is shown second (plot 4).

The representation of selectivity here decodes at a later layer 20 whether we have both PXR values as selectivity decodes better with only the PXR_null block and reading after the PXR block with both blocks doesn't give better accuracy score(plot 1). When PXR block is shown first and we read at PXR block the score is around chance but when reading at PXR_null block the score is better but still lower than reading at PXR_null block first (plot 1).

Experiment C: supporting material and additional results

The compounds the battery was run over

Both readings are pEC50 values, so a higher number means a more potent response. The counter-screen reading is what decides whether a signal on the PXR reporter reflects PXR at all (Figure 7, Figure 8)

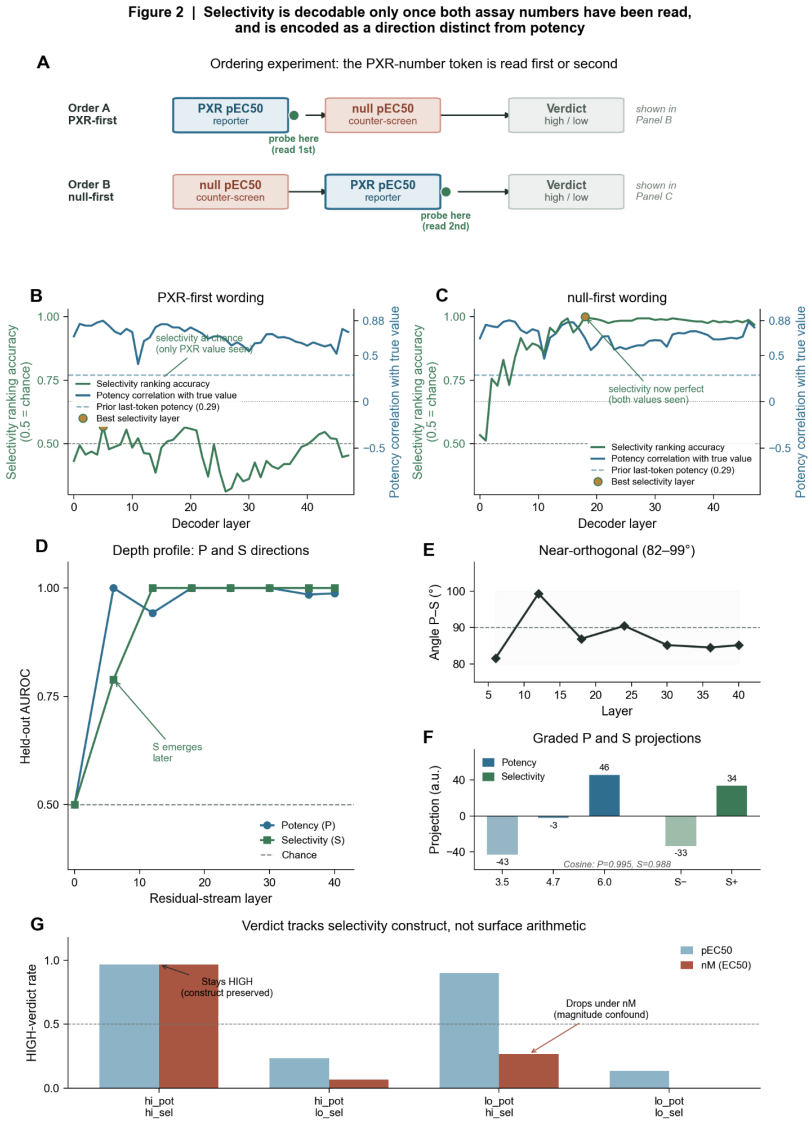

Figure 6: We ran preliminary experiments on one model (Gemma-7B) and then looked into the decodability of selectivity and potency. Here are our preliminary results.

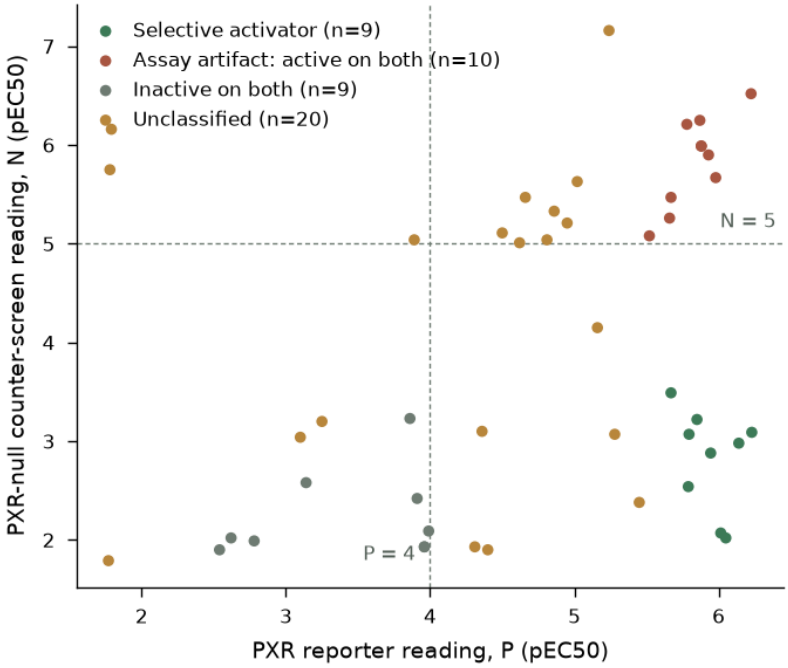

Figure 7: The compounds, placed by their two readings. The dashed lines are the thresholds that sort them into the four classes.

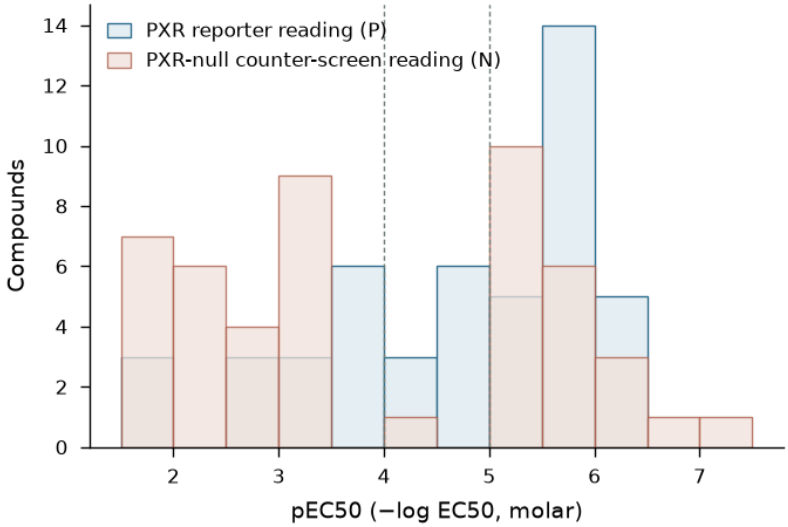

Figure 8: The same readings as two distributions. The counter-screen readings sit lower overall, but they overlap the reporter readings enough that the reporter value alone does not determine the class.

The seven prompts

The molecule, the readings and the assay descriptions are identical across all seven. Only the role the model is given, the surrounding context and the closing question change. Text below is the condition-A wording; condition B inserts the counter-screen assay and its pEC50 immediately after the reporter reading and leaves everything else untouched.

Framing	Prompt (condition A)
Interpret a colleague's hit	A colleague ran a screen and got a hit. They tested the compound in the following assay: {{pxr-reporter-assay}} The compound came back with a pEC50 of {pEC50_pxr}. The compound structure: {SMILES} What do you make of this result?
Compound PXR-dependent?	You are an expert toxicologist evaluating ADMET properties. The compound was tested in the following assay: {{pxr-reporter-assay}} SMILES: {SMILES} pEC50: {pEC50_pxr} Based on this data, does this compound produce a true PXR-dependent signal? Answer yes, no, or cannot determine, and explain your reasoning.
Compound selective?	You are an expert toxicologist evaluating ADMET properties. The compound was tested in the following assay: {{pxr-reporter-assay}} SMILES: {SMILES} pEC50: {pEC50_pxr} Based on this data, is this compound a selective PXR inducer? Answer yes, no, or cannot determine, and explain your reasoning.
Screening decision: follow up?	You are an expert toxicologist reviewing a screening hit before committing to follow-up studies. The compound structure: {SMILES} The compound was tested in the following assay: {{pxr-reporter-assay}} pEC50: {pEC50_pxr} Is this a credible PXR hit worth following up on? Explain what gives you confidence or concern.
Propose follow-up experiments	You are an expert toxicologist evaluating ADMET properties. The compound was tested in the following assay: {{pxr-reporter-assay}} SMILES: {SMILES} pEC50: {pEC50_pxr} What further experiments are needed before drawing conclusions about this compound?

Framing	Prompt (condition A)
DDI-risk assessment	You are an expert toxicologist evaluating ADMET properties. We are assessing Molecule X. The compound structure: {SMILES}. It was tested in the following assay: {{pxr-reporter-assay}} In that assay, this molecule scored a pEC50 of {pEC50_pxr}. Based on this data, how confident are you in your assessment of the clinical Drug-Drug Interaction (DDI) risk via CYP3A4 induction for this molecule in a human patient? Explain your pharmacological reasoning step-by-step.
Interpret bare data	SMILES: {SMILES} This compound was tested in the following assay: {{pxr-reporter-assay}} pEC50: {pEC50_pxr} What can you conclude about this compound?

The two assay descriptions, inserted wherever an assay is named, in the arm that carries them:

- PXR reporter. A stably integrated, cell-based luciferase reporter assay expressing a chimeric protein composed of the human PXR ligand-binding domain fused to a heterologous DNA-binding domain. The reporter construct places luciferase downstream of the DNA response element recognized by that heterologous DNA-binding domain. Compound binding to the PXR ligand-binding domain promotes transcriptional activation and increases luciferase luminescence.
- PXR-null counter-screen. A stably integrated cell-based luciferase reporter system in which the gene encoding the chimeric PXR reporter protein contains multiple early nonsense mutations. These mutations prevent expression of a functional chimeric protein. Thus, the measured luminescence readout is generated in the absence of functional PXR-mediated reporter activation.

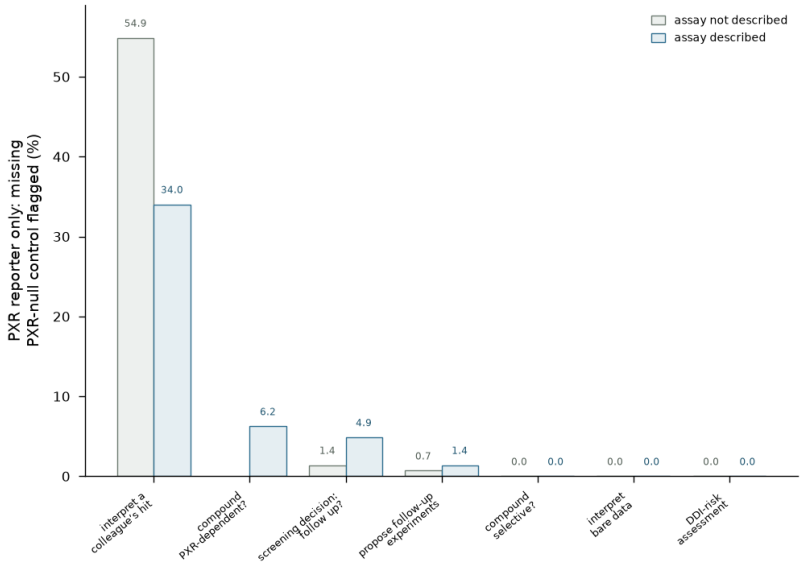

Figure 9: **Does describing the assay change the rate?** Condition A only. The share of answers asking for a PXR-null control, by framing, with and without the sentence describing how each assay works. Describing the assay lowered the rate rather than raising it: within the colleague's-hit framing it falls from 54.9% to 34.0%. The description states that the readout runs through a chimeric PXR ligand-binding domain, and being told the signal passes through a ligand-binding domain appears to send the model reasoning about what else binds there rather than about the counter-screen.

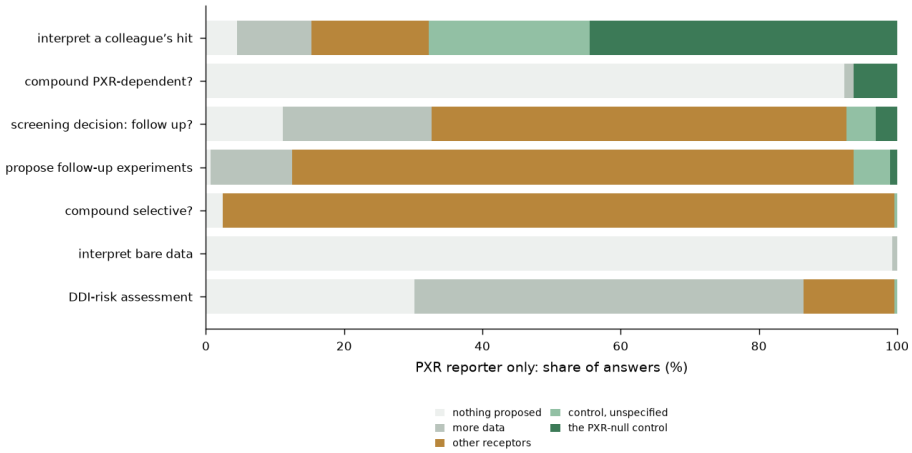

Figure 10: **What the model asked for instead.** Condition A only. Every answer sorted by what further experiment it proposes. Under the selectivity question, 280 of 288 answers propose a panel of other receptors, and under propose-follow-up-experiments, 234 of 288 do. These answers show concern for rigour pointed at the wrong experiment rather than indifference to it. The word 'selective' means selective across receptors in ordinary usage, while in this dataset it means that the signal requires PXR, and those are two different experiments.

What the model says under each framing

One compound held fixed (OCNT-2311501, PXR reporter pEC50 6.14, counter-screen pEC50 2.98), with the assay descriptions supplied. The only thing that changes down the table is the question that was asked; the only thing that changes across it is whether the counter-screen reading was on the page.

Framing	Condition A: counter-screen withheld	Condition B: counter-screen supplied
Interpret a colleague's hit	"Counter-screen: Test the compound in the same assay but without the PXR-LBD (using only the heterologous DBD) to ensure the compound isn't simply stabilizing the protein or interfering with the luciferase enzyme itself."	"You can tell your colleague that this is a high-quality hit. The use of the nonsense-mutation control successfully ruled out the possibility that the compound is simply a luciferase booster or an artifact of the cell line."
Compound PXR-dependent?	"The combination of the specific assay design and the nanomolar potency allows for the conclusion that the compound produces a true PXR-dependent signal."	"The fact that the signal is drastically attenuated when the PXR ligand-binding domain is absent demonstrates that the luciferase activation is specifically dependent on the presence of the PXR protein."
Compound selective?	"To determine if a compound is a selective PXR inducer, it must be tested against other closely related nuclear receptors, most notably the Constitutive Androstane Receptor (CAR) and the Vitamin D Receptor (VDR)... there is insufficient data to conclude whether this activation is selective."	"A selectivity margin of over 1,000-fold indicates that the compound's activity is driven by the PXR ligand-binding domain rather than non-specific effects on the reporter system."
Screening decision: follow up?	"Proceed to follow-up... Selectivity Screen: Test against other nuclear receptors (CAR, VDR, PPARA) to ensure the hit isn't a promiscuous nuclear receptor activator. Confidence Score: 7/10."	"Excellent Specificity: The ~1,400-fold difference between the functional and null reporter assays provides definitive evidence that the activity is PXR- mediated. Confidence Level: High."
Propose follow-up experiments	"Run the compound through similar reporter assays for CAR and VDR... confirm that it induces native CYP3A4 in human hepatocytes."	"Specificity: CAR/GR reporter assays. Safety: cell viability, to establish the toxic concentration and ensure the therapeutic window is well defined."

Framing	Condition A: counter-screen withheld	Condition B: counter-screen supplied
DDI-risk assessment	"I would require: Human Cmax,free... mRNA/Protein Expression Data... Fold-Induction."	"Confidence Level: High. The use of a genetically null control to validate the PXR-dependence of the signal removes the primary ambiguity associated with cell-based reporter assays."
Interpret bare data	"The compound is a potent PXR agonist (EC50 approximately 724 nM) with a high likelihood of inducing CYP3A4 expression."	"The compound is a potent and selective activator of PXR. Because the luminescence signal is nearly abolished when the PXR protein is absent, the observed activity is specifically mediated through the PXR ligand-binding domain."

Code reproduction

An anonymized supplementary archive maps each reported experiment to its input data, prompt templates, model identifier, pinned source revision, sample and replicate counts, saved outputs, analysis code, and figure-generation script. The archive distinguishes finalized results from pending or superseded analyses. Model weights are not redistributed. Repository-authored code is provided under Apache-2.0; source-data and third-party terms are documented in the asset card. The released assets contain no human-subject or personally identifiable data.