Do LLMs reason about the causal validity of real-world measurements?

Sanskriti Shindadkar[1], Nice Cailie Ineza[2], Laurent Hau-Kai Vong[2], Afnan Alabdulwahab[2], Suho Lee[3], and Manjari Narayan[4]

[1] University of California, Berkeley [2] PRISM AI Safety Research [3] New York University [4] Nyaya Labs

Abstract

As more R&D, including AI safety research, is delegated to LLMs and agentic systems, we need to evaluate the capacity of AI to improve scientific methodology itself. In the life sciences, perturbing a complex biological system and attributing changes to a specific component cause within that system is dependent on a number of mechanistic assumptions about both the intervention and measurements of the outcome. Much progress in science comes from testing and revising such assumptions. For instance, when evaluating drug toxicity, can LLMs distinguish between an imperfect assay's measurement of a drug engaging PXR from the causal attribution that the drug engages PXR? (See Figure 1). We investigate such questions using a series of counterfactual behavioral experiments, activation-based probes, and complementary interpretability tools. We utilize highly-specific PXR reporter assays and other placebo-like control assays to decode gold standard concepts of PXR selectivity. However, upon reparametrising the same assay pairs into different yet equivalent units, the decoding performance breaks. Models demonstrate asymmetric behavior with respect to the assay manipulated: reparametrizing the primary assay induces 6x larger perturbations than reparametrizing the control assay. Additionally, when numerical summary statistics and full dose-response curves conflict, several models make choices inconsistent with the full nonlinear dose-responses from the assay, These preliminary experiments demonstrate that LLMs differ in their utilization of measurement mechanisms in interpreting evidence about PXR engagement. This study was inspired by the PXR blind challenge by the OpenADMET consortium. We hope exceptional scientific experiments with rich sets of orthogonal controls are systematically used to generate realistic counterfactual evaluations of model reasoning. Such an endeavor would alleviate the need for full reverse engineering of model internals to evaluate LLM causal reasoning.

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].

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 model details are in Section B. All experiments were run on Gemma-4-12B-it for a comparative baseline.

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

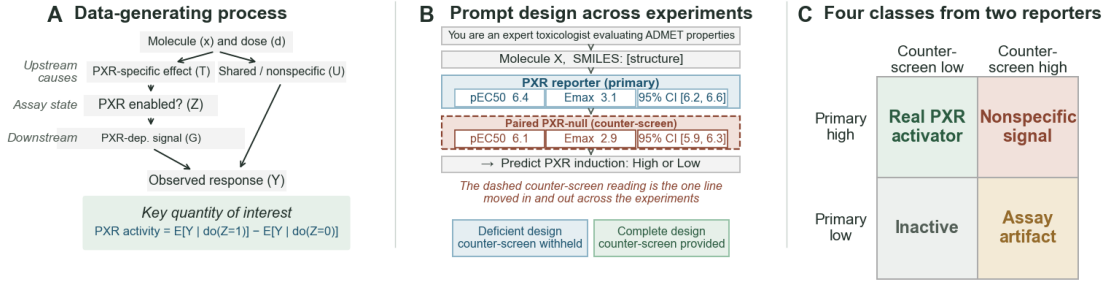


Figure 1: 1A: The real world data-generating process for what PXR selectivity actually is. 1B: an example of the prompt design template across experiments. 1C: Selectivity is a joint function of the primary-screen and counter-screen levels; the four quadrants sort compounds into real activators, nonspecific signals, assay artifacts, and inactives.

inactive, as shown in Figure 1. 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. To quantitatively distinguish true positive activators from these artifacts, a selectivity and efficacy index I (Equation 1) can be calculated to compare the primary screening assay (A) directly against the counter-screen control (B):

$$I = (pEC_{50,A} - pEC_{50,B}) + \log_{10}\left(\frac{E_{\max,A}}{E_{\max,B}}\right). \quad (1)$$

Results

A) Selectivity Is Linearly Decodable, but the Decodability Is Arithmetic.

We are checking if selectivity is being used as a concept through running different experiments in which we train different probes, and compare performance in conditions where selectivity is either withheld (in testing and/or training), and in conditions where probes are trained on pEC50 and E_{max} data and tested out of 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 (Figure 2).

A probe trained to distinguish selective from non-selective compounds when the null assay is withheld achieves 0.75 AUROC on null-withheld prompts, but only 0.57 AUROC when tested on prompts that include the null. The 0.75 result may partly reflect potency: many selective compounds have higher-than-average PXR-positive assay values, so the probe may use that association rather than information about the contrast between assays. In the reverse direction,

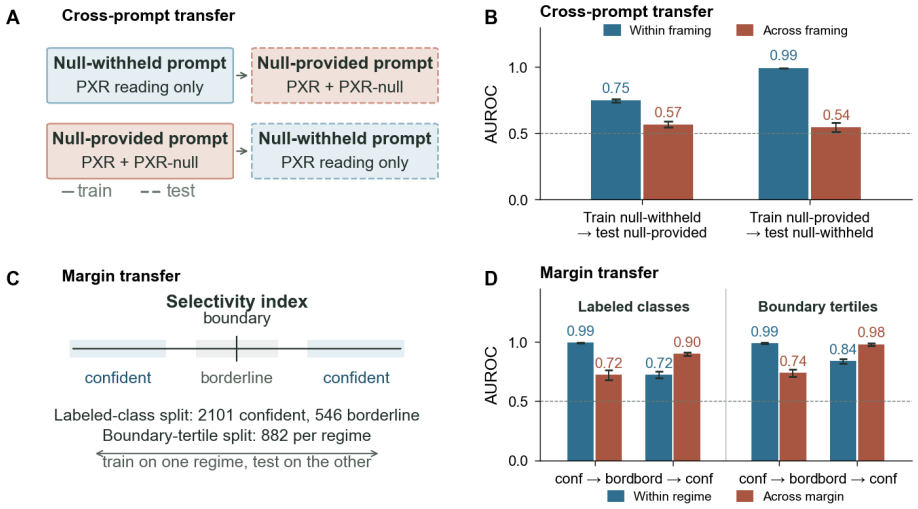


Figure 2: 2A: The structure of the cross-prompt transfer. 2B: The complete probe cannot survive null removal (0.99 to 0.54 AUROC), and a probe trained without the null performs near chance when evaluated on null-provided prompts (0.75 to 0.57 AUROC). 2C: Structure of the margin transfer. 2D: In addition, 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.

a probe trained with the null assay achieves 0.99 AUROC on prompts that include it, but only 0.54 AUROC when tested on null-withheld prompts. This asymmetry shows that information decodable from one prompt condition does not necessarily transfer to the other. More generally, a probe's success shows that information is present in the model's activations in a form the probe can read; it does not establish that the model uses that information to make its answer.

In the margin-transfer analysis, probes trained on compounds closer to the decision boundary become more accurate when tested on compounds farther from it, while probes trained on confident compounds perform worse on borderline compounds. Simple arithmetic reconstruction can explain much of the selectivity decodability: a scalar pEC50_PXR-minus-pEC50_null baseline achieves AUROC 1.00 (see Appendix Section A.1). Selectivity being linearly decodable does not show that the model causally uses it. These results motivate a follow-up question: if assay values are mathematically equivalent, will the model's internal representations be causally interchangeable?

Although this is linearly decodable using pxr_positive - pxr_null due to many assay results being naturally clear, we ran another experiment where the correct selectivity conclusion required use of the full selectivity index, `(pEC50_PXR - pEC50_null) + log10(Emax_PXR / Emax_null)`, and the pattern still held (as seen in Appendix Figure 5).

B) 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.

The reason that this experiment is one behavioural indicator of causal reasoning is that a good reasoning agent here must flag that it is missing vital information in order to reach a scientifically sound conclusion about PXR-dependence, while a weak one might not notice. For PXR, the positive assay on its own can be riddled with confounding factors, and a null is necessary in order to determine PXR-dependence. A model that asks for controls constantly would still be unhelpful, but what we're evaluating here is whether models recognize the missing evidence and respond appropriately. A scientifically appropriate response would be to flag the missing null which limits a high-confidence causal claim about PXR dependence, and optionally make a reasonable estimate of reporter activity using the information provided.

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 Table 1.

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 behavior (Figure 3). Flags remain when the counter-screen is supplied in 31 of 1,344 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 (Figure 8, Figure 9).

C) 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 is necessary for a

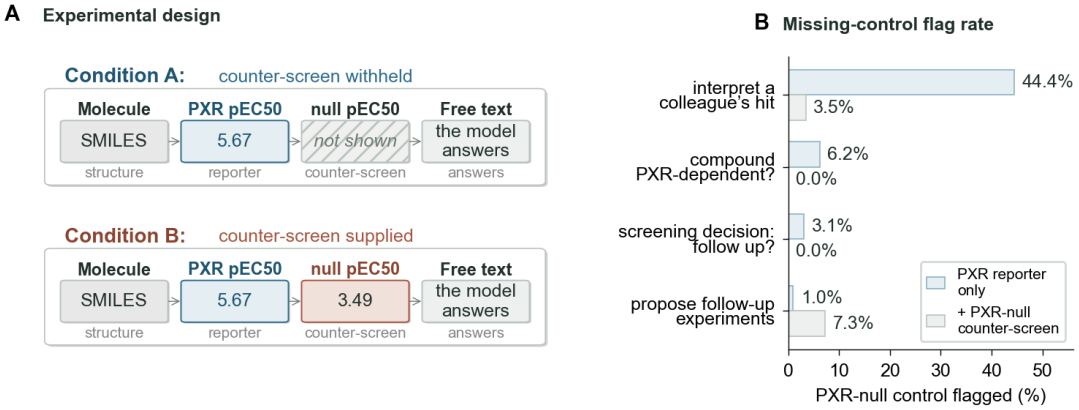


Figure 3: When Gemma-4-31B-it is asked to interpret a colleague's screening hit, it flags that the PXR assay is missing the counter-screen required to determine PXR dependence. Asked whether the same compound is a selective PXR inducer, to assess its DDI risk, or simply what the readings show, it does not. 3A: Two experimental conditions (PXR counter-screen withheld versus supplied) rendered across seven prompt variations each. 3B: Missing counter-screen flag rate by framing and condition (Gemma-4-31B-it; Gemma-4-12B-it produces fewer than 1% flags across all framings). Interpreting a colleague's hit flags most often, at 44.4%.

molecule to be truly selective. In this experiment, we give the models different combinations of the following: just the pEC50s, the pEC50s and the E_{max}, the pEC50+E_{max}+standard errors, pEC50+E_{max}+standard errors and a dose table, or pEC50+E_{max}+standard errors and plotted curves (that match the dose table). See Figure 4 for examples. We designed synthetic dose-response tables and their corresponding curves that fall into two regimes (that are indistinguishable from just pEC50 and E_{max} 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.

Seven molecules in the dataset naturally fall in the Case A/Case C ambiguity, so 45 additional molecules have synthetically defined values for the A and C curves. Figure 4 shows the difference between Case A and Case C, where despite having equivalent pEC50 and E_{max} 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 E_{max} values. As shown in Figure 4, case A was generally easier for the models to classify than Case C under the summary-only conditions, while case C accuracy generally improved when a dose table was provided, although the improvement varied considerably across models. 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/E_{max} values.

Cross-Assay Replications

We repeated the dose-curve and missing-null prompt experiments using the AR dataset, using a subset of 37 compounds, over 7 models of various size+capabilities. These investigations are preliminary, yet suggest that the trends we're seeing in the PXR dataset are not specific to just

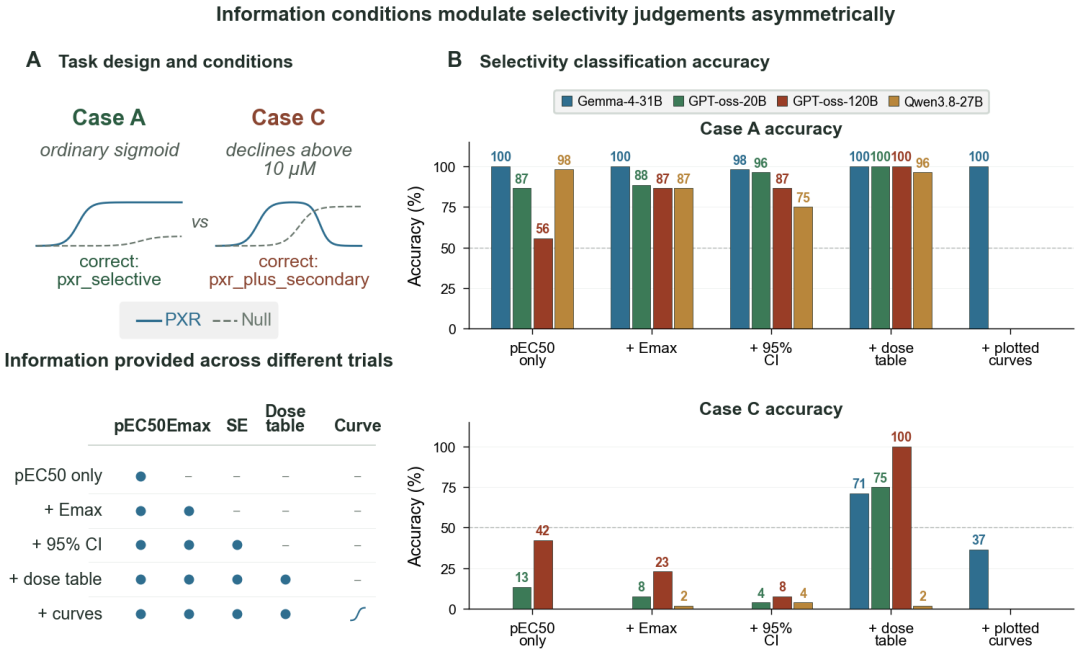


Figure 4: Models usually resort to naive arithmetic, and rarely reason using the more nuanced interpretation of selectivity dependent on PXR. 4A: Visual representation of Case A and Case C, which are PXR selective and PXR+secondary response respectively, but have identical pEC50 and E_{max} values at a concentration. Also included is a summary table of which information is provided across different conditions. 4B: Comparing classification accuracy for Case A and Case C (supplied pEC50 and E_{max} values are the same).

that dataset. The trend in the dose-curve responses of most models not being able to reliably identify case C (with the more nuanced dose-curve response) even when given the table or curves holds, as does the fact that most models rarely inquire about the missing null. These findings can be found in the Appendix, in Figure 13, Figure 14, and Figure 15.

Limitations

We only used linear probes; if selectivity is represented in models nonlinearly, our experiments could not have detected this. The experiments would benefit from ablations to localize where the decodable information resides. Most experiments use the PXR dataset; the preliminary AR assay replication in Figure 15 provides an initial cross-assay check. Whether the probes generalize to other xenobiotic receptors or more broadly remains open. The scope of the research so far was limited: we tested a small number of models and established causal sensitivity to notation but did not localize which components carry the information that the probes recover, nor whether the ~5:1 overweighting of the primary assay observed in patching is implemented in specific attention heads or distributed across layers.

Future Work

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 LLM's representation of evidence about PXR dependence. We also aim to test how well these methods generalize, both to related datasets, such as the EDKB Androgen Receptor Binding Dataset, and to other scientific domains, including protein folding, gene regulation, and metabolic pathways. 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. One impactful future research direction is to take high-quality controlled scientific experiments (not well represented in pre-training data) and turn them into benchmarks for interpretability and alignment that automated behavioral evaluation tools such as Bloom could run [Gupta et al., 2025]. Together, these directions could form the basis for mechanistic interpretability benchmarks grounded in real-world scientific problems, where naturally occurring perturbations provide verifiable tests of how models represent and reason about scientific concepts.

Conclusion

Taken together, these results suggest that the model's behavior is consistent with interpreting gross assay summaries rather than a notation-independent causal concept. A scalar pEC50 difference alone separates the classes perfectly. Probe AUROC is higher on the confident set, whose molecules lie farther from the selective/non-selective boundary, than on the borderline set. Cross-notation activation patching produces an aggregate shift of +8.75 logits at the primary reading, with substantial compound-level variation, indicating that equivalent assay values do not produce fully interchangeable model readouts in this test. Behaviorally, the model rarely requests the missing scientific control: 0-1% for most prompts across most open-weight models we tested. Therefore, models behave as if selectivity is a property of the numbers printed in the prompt, not of the causal relationship those numbers are meant to establish.

As Vafa et al. ask whether foundation models trained on orbital trajectories develop inductive biases consistent with Newtonian mechanics [Vafa et al., 2025], we ask whether models trained on public scientific data account for the causal structure of measurements when evaluating molecular function. This project takes a step toward establishing well-controlled experiments as a novel, hitherto-unexploited source of scientific reasoning evaluations.

Acknowledgments

This research was conducted as part of the PRISM AI Safety Research Fellowship, a program of Women Who Do Data (W2D2), with support from the Long-Term Future Fund (LTFF).

Computational resources for this project were supported by a Bluedot Grant.

References

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.

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.

National Center for Biotechnology Information. qhts assay to identify small molecule agonists of the androgen receptor (ar) signaling pathway using the mda cell line, 2013. URL https://pubchem.ncbi.nlm.nih.gov/bioassay/743040. PubChem BioAssay AID 743040; source: Tox21.

National Center for Biotechnology Information. qhts assay to identify small molecule agonists of the androgen receptor (ar) signaling pathway in the presence of an antagonist - cell viability counter screen, 2018a. URL https://pubchem.ncbi.nlm.nih.gov/bioassay/1259379. PubChem BioAssay AID 1259379; source: Tox21.

National Center for Biotechnology Information. qhts assay to identify small molecule agonists of the androgen receptor (ar) signaling pathway in the presence of an antagonist, 2018b. URL https://pubchem.ncbi.nlm.nih.gov/bioassay/1259381. PubChem BioAssay AID 1259381; source: Tox21.

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.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 60727–60747, 2025. URL https://proceedings.mlr.press/v267/vafa25a.html.

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

A Additional Information for the Experiments

A.1 Experiment A (Detailed in Figure 2)

Methodological item	Status	Detail
Model identity	Gemma-4-12B-it	Open-weights instruction-tuned model. Qwen2.5-7B-Instruct as independent replication (same fixed-split protocol).
Layer and position	Layer 12 (fwd) / 30 (rev)	Last prompt token; residual-stream vectors. Per-direction layer selection from candidate sweep. Backbone-only forward pass (lm_head skipped).
Layer sweep	7 candidate layers	Layers 6, 12, 18, 24, 30, 36, 42. Forward direction peaks at layer 12 (AUROC 0.74); reverse saturates by layer 30 (AUROC 0.99).
Activation extraction	One vector per (compound × regime)	Forward hook on residual stream at last prompt token; cached to activations.parquet.
Leakage audit	Pass	Rendered prompts reviewed before inference. Withheld fields (contrast labels, selectivity class, pair roles, FDA status) confirmed absent from null-withheld prompts.
Train/test splitting	Fixed 50/50 random split	Compound-disjoint: 1323 train, 1324 test. No compound appears in both splits.
Sample size (N)	2647 compounds	Full selectivity-annotated compound set with binary selectivity labels (boundary SI = 1.21).
Probe regularization	L2 logistic regression	C=1.0 (sklearn default), max_iter=2000. StandardScaler re-fit per split inside sklearn Pipeline; no global fit.
Random seeds	Seed 0	Single deterministic base seed; 20 repeated random train/test splits for CI estimation.
Confidence intervals	Bootstrap (2000 resamples)	95% percentile intervals over 20 repeated splits.
Scalar baseline	PXR−null = AUROC 1.000	Arithmetic subtraction of the two pEC50 values perfectly separates selective from non-selective compounds, establishing that decodability alone does not require a learned concept.
Forward transfer result	Within 0.75 [0.73, 0.76]	Null-withheld probe tested on null-provided prompts. Cross-regime AUROC: 0.57 [0.54, 0.59]. Gap: 0.18.
Reverse transfer result	Within 0.99 [0.99, 0.99]	Null-provided probe tested on null-withheld prompts. Cross-regime AUROC: 0.54 [0.51, 0.58]. Gap: 0.45.
Margin transfer (categorical)	Confident / borderline	Forward: within 0.99 [0.99, 0.99], cross 0.72 [0.68, 0.76]. Reverse: within 0.72 [0.70, 0.75], cross 0.90 [0.88, 0.91]. n = 2101 / 546.
Margin transfer (tertile)	Top / bottom tertile	Forward: within 0.99 [0.98, 0.99], cross 0.74 [0.71, 0.77]. Reverse: within 0.84 [0.82, 0.85], cross 0.98 [0.97, 0.99]. n = 882 / 882.
Second model replication	Qualitatively consistent	Qwen2.5-7B-Instruct (fixed-split, layer 14) shows the same asymmetric fragility: reverse transfer degrades, forward transfer improves.

We were wondering if a more complex arithmetic representation of selectivity could be recovered through the same probe methods. As such, we synthesized Emax and pEC50 values such that pxr_positive-pxr_null would be insufficient for decoding selectivity. The trends still held: a probe trained without the null still could not perform with a high AUROC on the test set with the null provided.

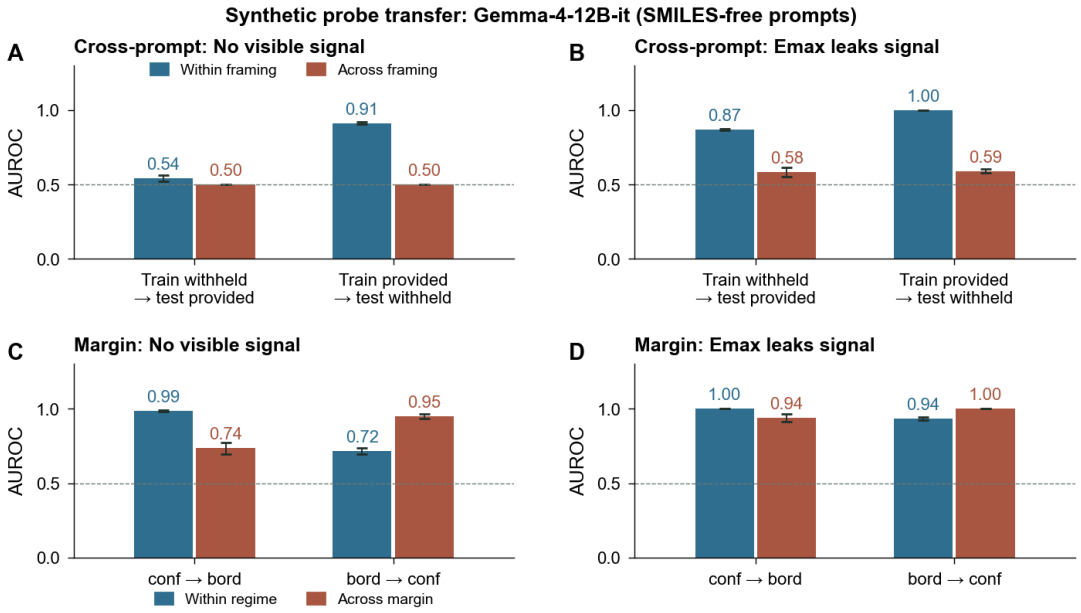


Figure 5: Synthetic probe transfer: Gemma-4-12B-it (SMILES-free prompts)

A.2 Experiment B: 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 6, Figure 7).

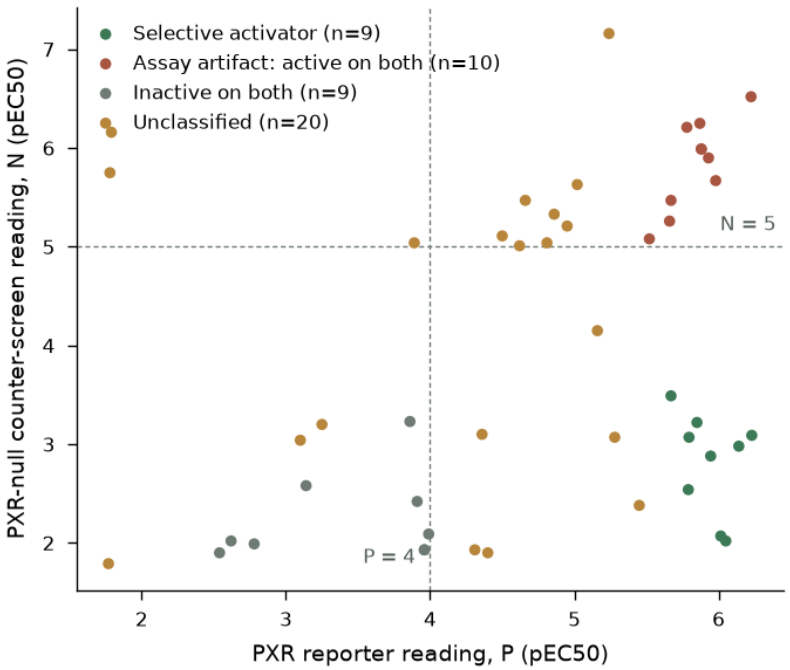


Figure 6: The compounds, placed by their two readings. The dashed lines are the thresholds that sort them into the four classes.

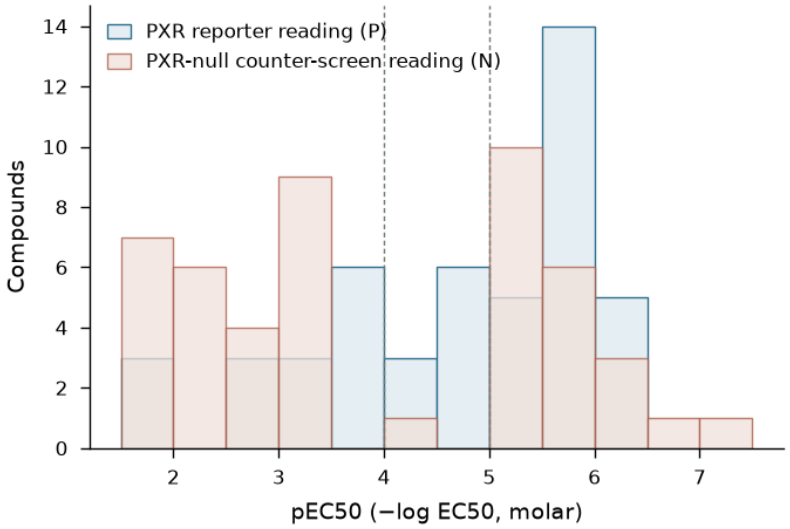


Figure 7: The same readings as two distributions. The counter-screen readings sit lower overall, but they overlap the reporter readings enough that the reporter value alone does not determine the class.

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.

Table 1: Full prompting text.

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.

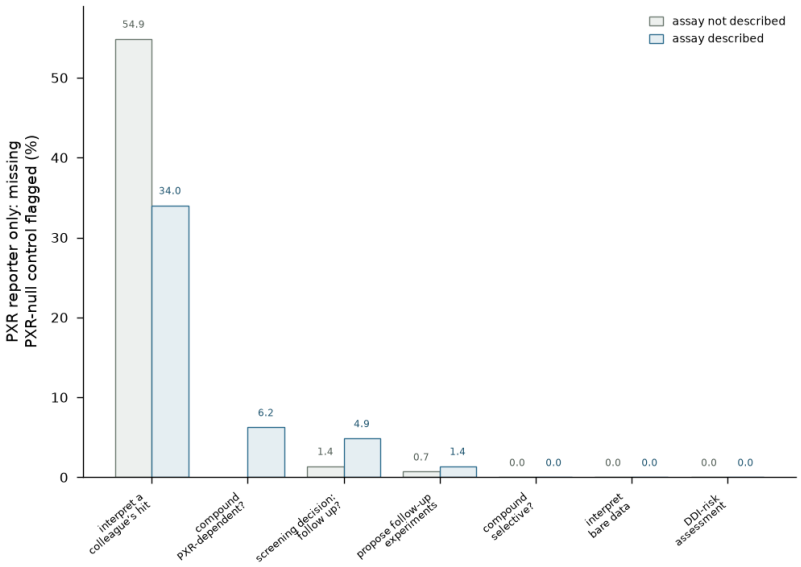


Figure 8: **Does describing the assay change the rate?** Condition A only. The share of answers asking for a PXR-null control, by framing, with and without the sentence describing how each assay works. Describing the assay lowered the rate rather than raising it: within the colleague's-hit framing it 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.

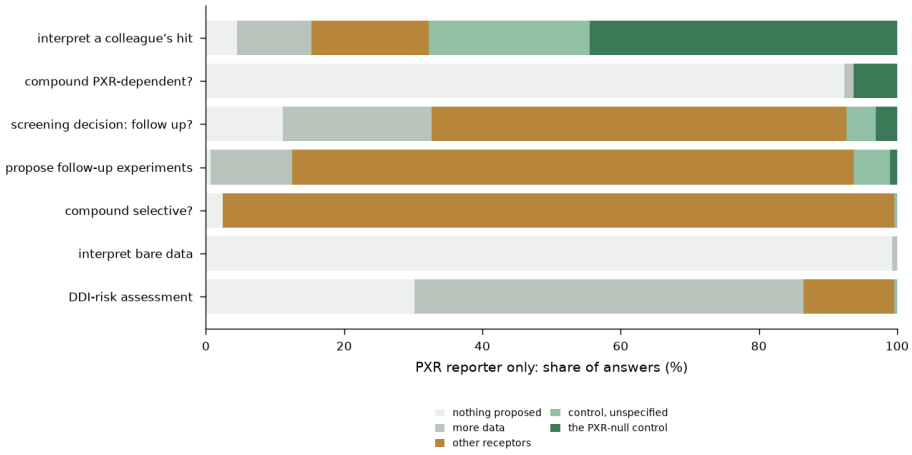


Figure 9: **What the model asked for instead.** Condition A only. Every answer sorted by what further experiment it proposes. Under the selectivity question, 280 of 288 answers propose a panel of other receptors, and under propose-follow-up-experiments, 234 of 288 do. These answers show concern for rigour 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."

B Model Selection Rationale

We provide clarification on our selection of models. Probe-based experiments (A, layer-wise decoding) use Gemma-4-12B-it with a fixed 50/50 compound-disjoint split over 2647 compounds. Activation-patching experiments (Appendix) use Gemma-4-12B-it, which provides clean residual-stream access and well-documented layer norms for cross-notation substitution. Behavioral experiments (B) 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.

C The Model's Reasoning Is Not Invariant If Assay Numbers Are Swapped Out with an Equivalent Scientific Notation (pEC50 and nM)

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. Thus, at the tested layer and

token position, the residual states associated with mathematically equivalent assay values were not causally interchangeable. 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 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.

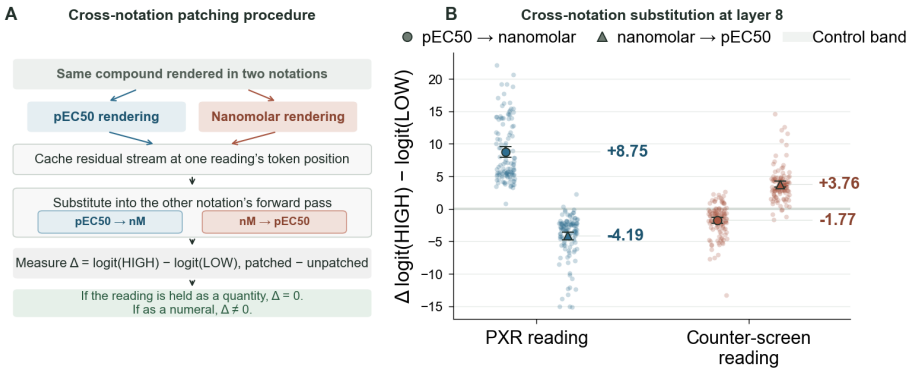


Figure 10: 5A: Explaining the methodology of the experiment: rewriting the assay reading in another unit but equal quantity moves the model's verdict. Each compound prompt is rendered twice, once with both PXR and PXR-null written as pEC50 values and once as EC50 values in nM. The two renderings represent 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 [layer 8]. 5B: Substituting pEC50 into nanomolar and vice versa affects the logits considerably.

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. This preliminary run used 32 compounds with multi-token patching (all four digit tokens); Figure 10 reports a larger run (118 compounds) with single-token patching, where effects range from -15 to +21 logits across individual compounds. Selectivity is an operation that needs both operands present: removing the counter-screen (null) at inference breaks the trained probe to chance (reverse-transfer drop +0.28, two models), and withholding it in training installs the potency shortcut at equal score (lever +0.14 [+0.07, +0.21]), with an ablation confirming the mechanism is actually just differencing the two supplied numbers.

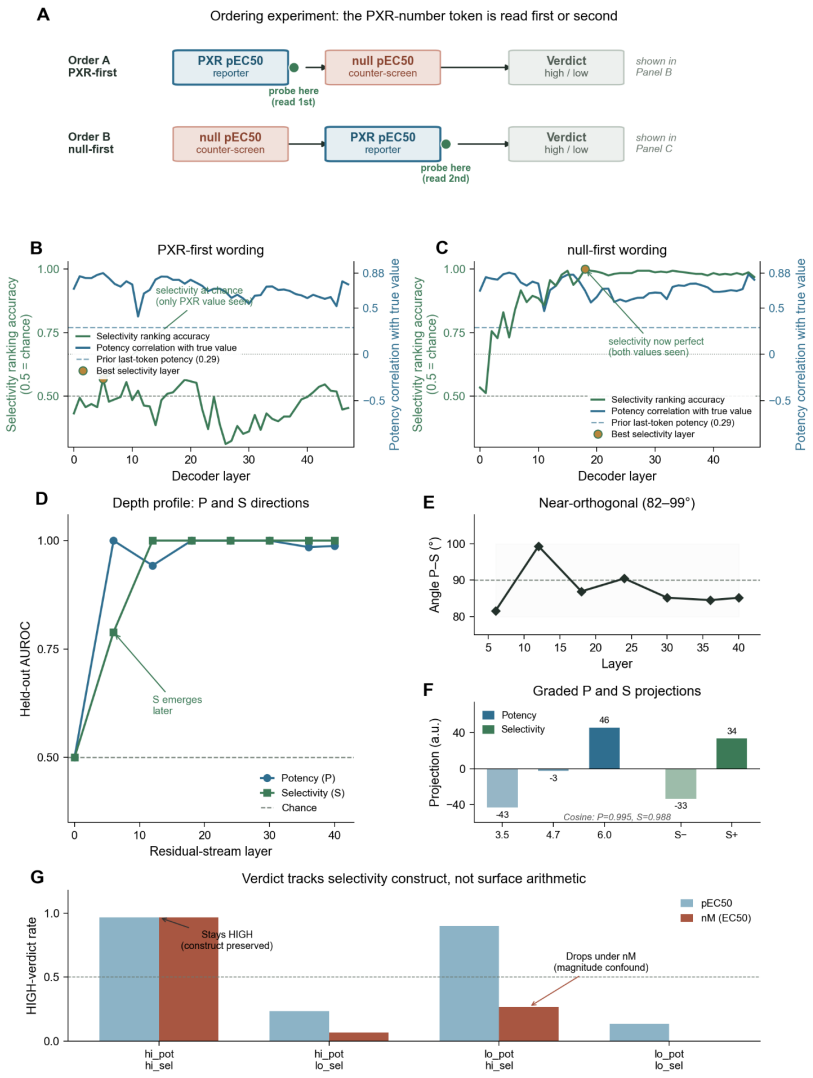


Figure 11: We ran preliminary experiments on one model (Gemma-4-12B-it) and then looked into the decodability of selectivity and potency. Here are our preliminary results.

The representation of potency decodes at an early layer when PXR block is shown first (plot 5), at layer 8, and decodes at a later layer when the PXR block is shown second (plot 4) (Figure 11, Figure 12).

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).

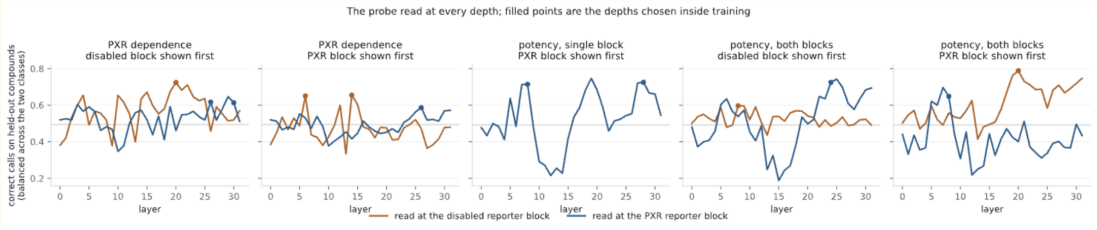


Figure 12: The probe read at every depth; filled points are the depths chosen inside training

D Generalizing to AR (Androgenen Receptor) Datasets

For the AR dataset, there are three assays.

1. Baseline AR agonist assay [National Center for Biotechnology Information, 2013]
2. AR agonist assay + nilutamide [National Center for Biotechnology Information, 2018b]
3. AR + nilutamide viability counter-screen [National Center for Biotechnology Information, 2018a]

The model is given a baseline AR-responsive reporter assay, the same reporter assay under pharmacological AR blockade with nilutamide, and a parallel cell-viability counter-screen. A strong reporter response alone is not enough to say that the compound's effect is AR-dependent! The model must reason about how the response changes under receptor blockade and whether cytotoxicity provides an alternative explanation. In other words, a molecule that shows strong activity in the baseline AR agonist assay, but substantially reduced activity in the AR agonist assay + nilutamide, is more likely to have AR-dependent activity because blocking AR suppresses the reporter response. However, if the molecule also causes substantial loss of cell viability (AR + nilutamide viability counter-screen) at similar concentrations, then cytotoxicity might be the reason for the change in reporter signal, and it's not clear whether the molecule is AR-dependent or not. Therefore, the model must take into account assay values *from all three assays* in order to make a statamtent about AR-dependency. The matrix below explains what different assay values might mean, specifically what it means in relation to the other variables.

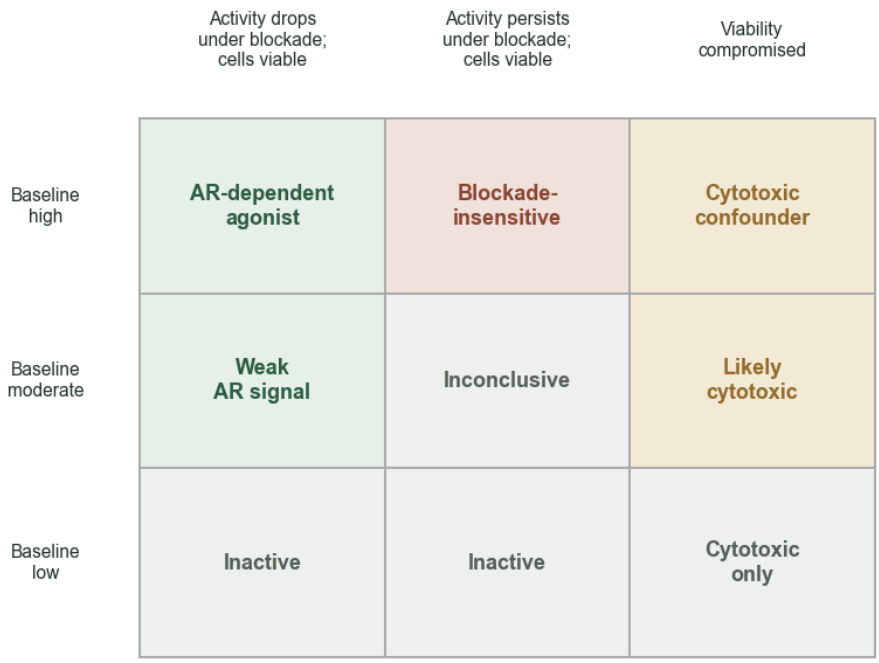


Figure 13: AR decision matrix: what different assay values might mean.

In order to run the following experiments, we first had to define ground truth for what defines whether something is an AR activator/dependent, then run prompts, and then run the experiments on different models. The models include deepseek-v4, gemma-4-26b, gpt-5.4, qwen3.5-35b, to have a range of simpler models and a frontier model to see how their abilities defer.

D.1 Defining Ground Truth for AR

The ground truth for AR is a little bit tricky, and I'm still working through adjusting the definitions to be scientifically correct, and also looking into the literature. The way these are defined might change for future experiments.

Here are the categories:

1. **Baseline_inactive**: Baseline AR activity is not "Active" in the PubChem outcome field. (6,814 compounds.)
2. **Likely_cytotoxic**: Baseline AR is active, but the viability counter-screen also flags active, suggesting cell death rather than AR-specific activity. (9 compounds.)
3. **AR_dependent**: Nilutamide blockade substantially suppressed the response, by any of three criteria: delta_pAC50 > 1.0 (potency shift exceeding one log unit upon blockade),

delta_Emax > 30 (efficacy drop exceeding 30 percentage points), or baseline active but blocked inactive. (150 compounds.)

4. **Blockade_insensitive**: Baseline active, still active under blockade, and |delta_pAC50| < 0.5 (potency barely shifts). These are "fake-outs" — compounds whose high baseline activity is NOT AR-mediated. (50 compounds.)

5. **Inconclusive**: Does not meet criteria for any of the above. (13 compounds.)

For the behavioral battery, the ground truth mapping for AR-specific is: AR_dependent → HIGH, blockade_insensitive → LOW, baseline_inactive → LOW. Under the deficient condition, blockade_insensitive compounds present only their high baseline pAC50, so the model has no way to know the activity is not AR-specific! So we should not expect models operating without this information to guess correctly.

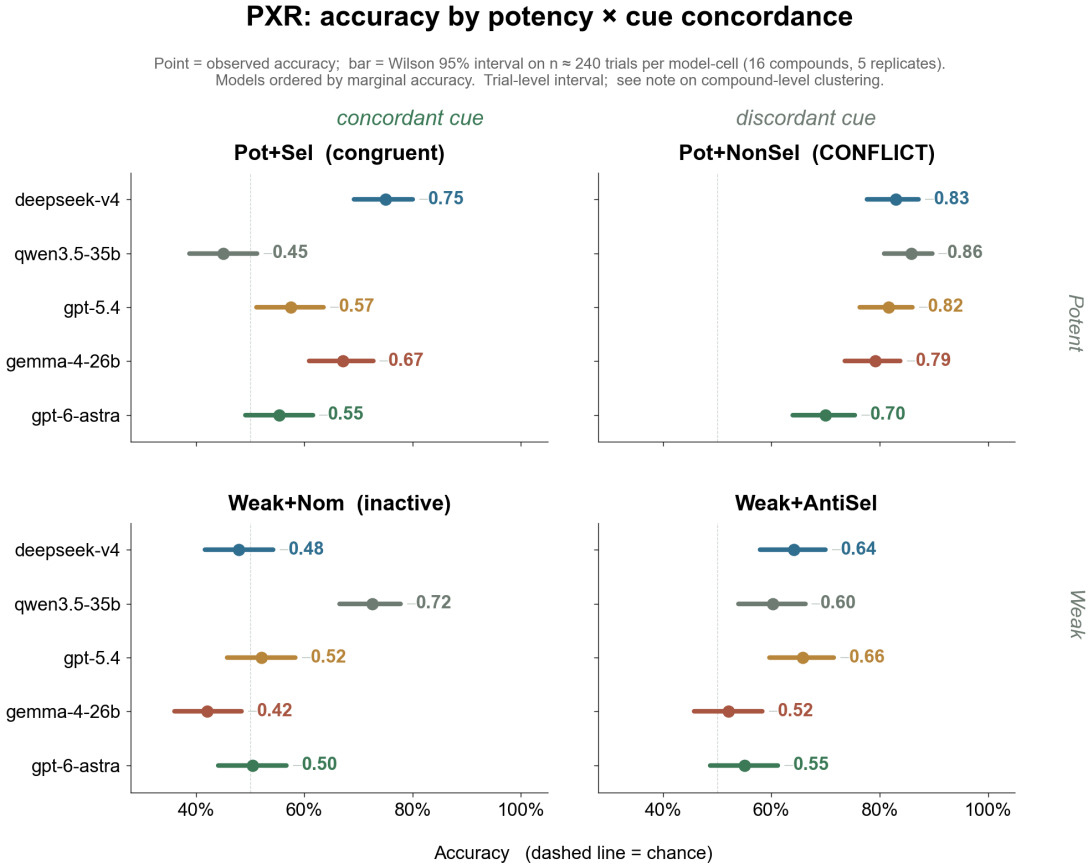


Figure 14: PXR: accuracy by potency × cue concordance | Wilson 95% CI | 16 compounds, 4800 trials, 5 models.

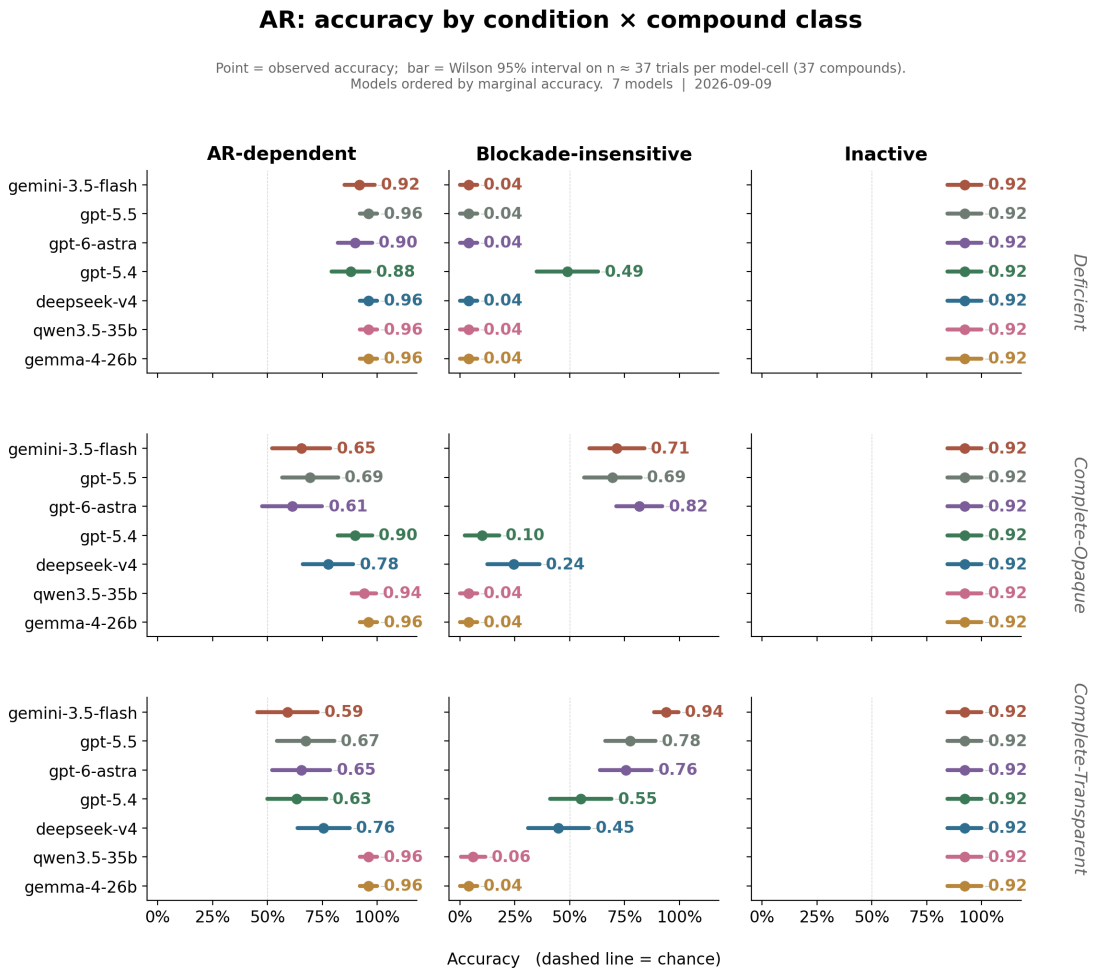


Figure 15: AR: accuracy by condition × compound class | Wilson 95% CI | 37 compounds, 7 models.

E 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.