Retrospective drug-target benchmarks measure protein druggability rather than disease biology

Abstract

Computational and language-model systems for therapeutic target discovery are commonly evaluated retrospectively: freeze the evidence available before a cutoff date, rank candidate targets for a disease, and check which candidates later entered clinical development. We build such a benchmark from 6 archived Open Targets Platform releases and 599,765 ClinicalTrials.gov studies, with labels anchored on trial start dates, and show that the pooled score this design produces is a mixture of two orthogonal quantities, dominated by the one no system claims to predict.

A target-disease benchmark has a between-target axis (which protein, given the disease) and a within-target axis (which indication, given the protein). On the trial-entry label essentially all discrimination lies between targets: a count of how many other diseases a target was already trialled in reaches AUROC 0.972 [0.966, 0.977], and a disease-independent structural druggability annotation containing no clinical or disease-specific information reaches 0.813. On the within-target axis, where that count is mathematically inert (0.500 by construction), aggregated biological evidence reaches 0.412 [0.374, 0.448] and human genetics 0.498, both at or below chance across 306 targets.

We then evaluate a frontier language model on 55 diseases. Given gene symbols and no evidence it reaches 0.779 [0.723, 0.827]. Supplying the pre-cutoff evidence does not improve it. Masking gene identities while keeping the same evidence gives 0.54, which is the information ceiling of that evidence (0.55 for a logistic regression on the identical features), so the blinded model is at ceiling and the evidence is nearly empty once the confound is matched out. The remaining entity-bound advantage is at least partly recall of the future: asked closed-book whether a matched pair entered trials after the cutoff, the model answers at balanced accuracy 0.633 [0.600, 0.668]. Withholding the confound from the prompt does not remove it either, since the model reconstructs prior clinical development from a gene symbol alone at Spearman 0.705 and AUROC 0.893 [0.861, 0.921].

1 Introduction

Retrospective temporal evaluation is the standard proposal for testing whether a computational system can prioritize therapeutic targets. Restrict the system to evidence available before a cutoff T, ask it to rank candidate targets for a disease, and check which highly ranked targets subsequently entered clinical development. The design appears to avoid the central weakness of static benchmarks, that a model may already know the answer, and to measure something a discovery organisation would pay for.

We built this benchmark from archived database releases rather than reconstructions, and then asked what its scores are made of. The answer is that the pooled score is a mixture of two orthogonal quantities. One is a property of the protein: is it druggable, and does it already have drugs. The other is a property of the target-disease pair: is this protein causally implicated in this disease. Systems in this area claim the second. The benchmark measures the first.

Separating them requires no new data, only conditioning. Prior clinical development is a target-level constant, so it cannot discriminate among a single target's candidate diseases. Conditioning on the target therefore removes it, along with every other target-level property, exactly and without any matching tolerance. What survives that conditioning is the disease-specific signal, and there is almost none of it.

This matters for language-model evaluation in a way that is not obvious. The natural control for a confound is to withhold it: do not tell the model which targets already have drugs. We show that control fails, because the model reconstructs the confound from the gene symbol alone.

Our contributions:

1. A two-axis decomposition of retrospective target-disease benchmarks, showing that pooled AUROC in this literature mixes a between-target axis carrying nearly all the discrimination with a within-target axis carrying almost none (Section 4).
2. Identification of the mechanism as druggability rather than promiscuity: a disease-independent structural tractability annotation, with every clinically derived category removed, reaches 0.813 (Section 5).
3. An evaluation of a frontier language model under four information conditions, using the existing blinding protocol of Cuccarese (2026) as an instrument, that isolates how much of its performance is entity recall rather than evidence use (Section 6).
4. A measurement of shortcut recoverability: how much of the benchmark's dominant confound a pretrained model reconstructs from parametric memory given no inputs, and why this defeats the standard control (Section 8).
5. Label-integrity measurements that any user of these benchmarks needs: retrospective curation invalidates 58.4% of snapshot-difference positives, and only 16.2% of positives under the permissive label are industry-sponsored (Section 9).

2 Related work

Temporal biomedical evaluation. Temporal holdout is established practice here and we claim no novelty for the protocol. Sybrandt et al. (2020) validated hypothesis generation by predicting connections first appearing after 2015 from earlier data; Breit et al. (2020) ship a time-slice split covering gene-disease edges; Narganes-Carlón et al. (2023) trained language models on literature frozen at successive dates to rank prospective drug targets; Siu et al. (2026) predict Phase II to Phase III advancement on a temporal knowledge graph. The OpenBioLink time-slice split defines positives by membership of a later curated release, which is the construction we measure and find wrong for most newly appearing pairs (Section 9). Falaguera et al. (2025) timestamped roughly 28 million Open Targets evidence items and report a median 11-year gap between primary publication and curation, which bounds how large that effect can be.

Shortcut baselines. Zietz et al. (2024) showed that node degree reproduces much of the apparent performance of network-based biomedical prediction, and Breit et al. (2020) raised related concerns for link prediction. Our between-target baselines are domain-specific analogues. The decomposition adds what a degree baseline cannot supply: degree correction leaves the between-target axis intact, and disease-independent literature volume reaches only 0.494 here, so correcting for popularity would not have revealed the problem.

Genetic support and clinical success. Nelson et al. (2015), revised by King et al. (2019) and Minikel et al. (2024), report that mechanisms with human genetic support progress more often. These condition on drug programmes that already exist and ask about progression given a programme; we ask which pairs enter development at all over a pool of 29,251 targets. The estimands differ, and our within-target null does not contradict them.

Proxy labels. That an observable label can diverge from the construct it stands for is established outside this domain: Obermeyer et al. (2019) showed substituting healthcare cost for healthcare need reversed an algorithm's conclusions, Guerdan et al. (2023) give a causal framework, and Dehghani et al. (2021) show method rankings are partly an artefact of benchmark choice. Our second-label result (Section 10) is an instance rather than a new phenomenon.

Identifier blinding. Replacing entity identifiers with anonymous codes before prompting, and comparing against an unblinded control, is an existing protocol: Cuccarese (2026) introduce epistemic blinding for exactly this purpose and apply it to LLM-assisted oncology target prioritization across four cancer types, reporting that blinding changes 16% of top-20 predictions while preserving recovery of validated targets. We use their protocol rather than propose it. Our masked condition differs in what it is measured against: a temporally anchored label with the dominant confound matched out, on which blinding does not preserve discrimination but removes it (Section 6). We also measure the complementary quantity they do not, namely how much of the confound the model reconstructs when identifiers are present (Section 8).

Target prioritization systems. Supervised prioritization (Ferrero et al., 2017), network propagation (Himmelstein et al., 2017), and graph foundation models (Huang et al., 2024) are typically evaluated against a present-day snapshot under random or cluster splits. Agent architectures that build typed evidence graphs for biomedical hypotheses already exist (rar, 2025). None reports a between/within decomposition or a druggability control.

3 Benchmark construction

3.1 Sources and task

We use 6 archived Open Targets Platform releases (Ochoa et al., 2021; Buniello et al., 2025), namely 21.06 through 26.06, as released parquet datasets. These are genuine historical database states, so the June 2021 release contains what Open Targets held in June 2021 and nothing later. We additionally retrieve 599,765 studies from the ClinicalTrials.gov v2 API with start dates, phases, study types and sponsor classes.

Fix $T = 30$ June 2021. For disease d, the candidate pool is every target with any recorded pre-T evidence for d in the 21.06 release: 2,538,652 pairs over 29,251 targets, with 21 typed evidence datasources (Appendix A). The `chembl` datasource encodes known-drug evidence and is the label source, so it is excluded from all feature sets.

3.2 Labels anchored on trial start dates

The Open Targets known-drug table links a drug, target, disease and the ClinicalTrials.gov identifiers supporting that link. We extract these across the four archived releases containing the table, giving 84,451 distinct trials, and join to start dates; 0.20% of lookups fail. For pair (g, d) let $\tau(g, d)$ be the start date of the earliest interventional trial linking that target to that disease. The label is

$$y(g, d) = \mathbf{1}[\, T < \tau(g, d) \leq T + 24 \text{ months} \,],$$

excluding pairs already in development at T. The horizon closes 12 months before the June 2024 release from which labels are read, so a trial starting shortly before that release is not scored as a negative merely because it had not yet been curated. This yields 1,373 positives; evaluation retains diseases with at least 5 positives and 300 candidates, excluding Human Phenotype Ontology terms, giving 57 diseases and 521 positives.

Table 1: Two-axis decomposition on the trial-entry label. Between-target: fix a disease, rank targets (57 diseases). Within-target: fix a target, rank its candidate diseases (306 targets, 510 positives). Intervals are cluster bootstraps. Target-level signals are exactly 0.500 within-target by construction.

Ranking signal	Between-target AUROC [95% CI]	Within-target AUROC [95% CI]
Prior clinical development	0.972 [0.966, 0.977]	0.500 [0.500, 0.500]
Aggregated biological evidence	0.541 [0.487, 0.596]	0.412 [0.374, 0.448]
Literature co-occurrence	0.578 [0.528, 0.625]	0.465 [0.434, 0.497]
Human genetics	0.485 [0.471, 0.497]	0.498 [0.487, 0.509]
RNA expression	0.473 [0.457, 0.487]	0.458 [0.440, 0.475]
Somatic mutation	0.540 [0.517, 0.566]	0.513 [0.501, 0.526]
Target literature volume	0.494 [0.442, 0.546]	0.500 [0.500, 0.500]
Random	0.515 [0.490, 0.541]	0.527 [0.495, 0.559]

3.3 Evaluation and statistics

We report per-disease AUROC with disease-level cluster bootstrap intervals (2000 resamples), and enrichment over the disease base rate, which is 0.63% on average. All AUROC values use mid-rank tie handling. Comparisons are paired on identical diseases and tested with two-sided Wilcoxon signed-rank, Holm-corrected across the 14 tests reported. With 57 paired diseases the design has 80% power to detect a mean AUROC difference of 0.045, and we report that alongside every null.

4 The benchmark has two axes, and only one carries information

A target-disease benchmark supports two conditionings. The *between-target* axis fixes a disease and ranks candidate targets. The *within-target* axis fixes a target and ranks candidate diseases, which is the indication-selection question. Prior clinical development, target popularity, tractability and essentiality are all properties of the target, so they are constant along the within-target axis and cannot discriminate there. Conditioning on the target removes all of them at once, exactly, with no caliper and no modelling assumption.

Table 1 and Figure 1 report both axes. Between targets, prior clinical development reaches 0.972 [0.966, 0.977] while aggregated biological evidence reaches 0.541 [0.487, 0.596]. Within targets, across 306 targets carrying 510 of the 521 positives, prior clinical development is 0.500 and target popularity 0.500, both exactly chance as the algebra requires, and aggregated biological evidence is 0.412 [0.374, 0.448], with genetics at 0.498 and expression at 0.458.

The pooled numbers reported in this literature are a mixture of these two. The axis that carries the discrimination is a property of the protein. The axis that corresponds to the question a prioritization system is asked carries none.

The within-target statistic is a mean over unequal strata, so we calibrate it against a permutation null rather than against 0.5: permuting labels within targets 2000 times gives a null centred at 0.500 with standard deviation 0.016. Against that null, aggregated biological evidence is significantly below chance ($z = -5.328$, $p = 5.0e{-}4$), as are expression ($z = -4.283$) and literature ($z = -2.143$); genetics is indistinguishable from the null ($z = -0.446$, $p = 0.666$) and somatic mutation is slightly above it ($z = 2.326$). We report the anti-predictive direction because it is calibrated, but we do not claim a mechanism for it. One candidate, that for a given protein the disease with the strongest evidence is often the one where the protein is least tractable, is not tested here.

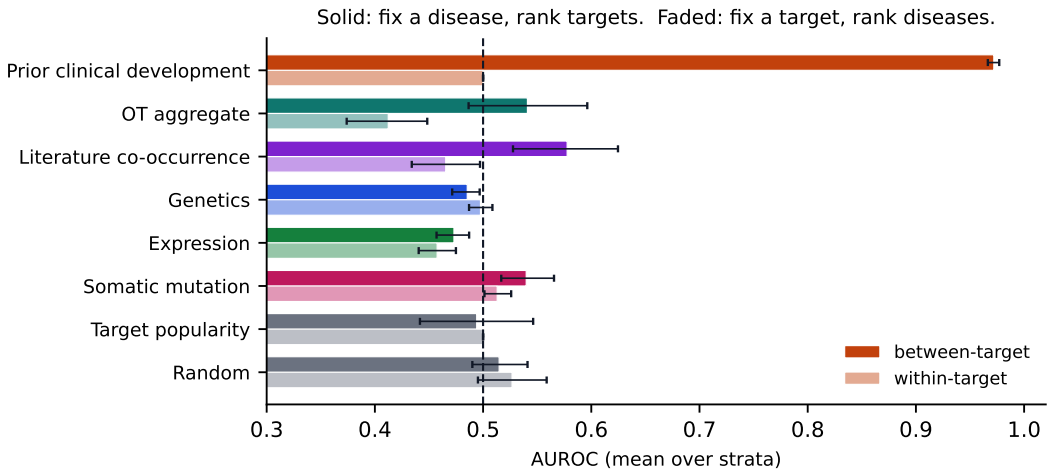

Figure 1: Two-axis decomposition on the trial-entry label. Solid bars fix a disease and rank targets; faded bars fix a target and rank its candidate diseases. Target-level signals are exactly 0.500 on the within-target axis by construction. Error bars are cluster bootstrap intervals.

5 The mechanism is druggability

If the between-target axis measures a property of the protein, the natural candidate is druggability. The 21.06 target table carries tractability annotations derived from protein structure and homology alongside categories derived from clinical precedent. Excluding every clinically derived category, leaving only structure and homology based predictions, this disease-independent annotation reaches AUROC 0.813 on the trial-entry label. Including the clinical-precedence categories raises it to 0.903. Aggregated disease-specific biological evidence reaches 0.541.

An annotation that contains no disease information whatsoever, and no record of any trial, predicts which targets enter trials for a disease far better than all of the evidence linking targets to that disease.

This also explains the benchmark's most striking negatives. The pairs with the strongest disease-specific evidence in the whole benchmark are TP53, BRCA1, BRCA2, PTEN, CDKN2A and DMPK in their canonical diseases, and all are negatives, because tumour suppressors and loss-of-function mechanisms are largely not druggable by conventional modalities. We tested whether biological evidence is specifically anti-predictive among intractable targets and did not find it: splitting at the median clinically blind tractability score gives 0.520 [0.466, 0.571] in the tractable stratum against 0.573 [0.449, 0.690] in the intractable one, with overlapping intervals (Appendix C).

6 What a language model does on this benchmark

We evaluate a frontier language model (Claude Opus 4.5, accessed through the Claude Agent SDK, one call per disease, tool use instructed off) on 55 diseases under four conditions. Candidate sets are either *unmatched*, where negatives are sampled at random and prior clinical development reaches AUROC 0.968, or *matched*, where each positive is paired with negatives of near-identical prior clinical development so that the confound is neutralised by construction (1,840 candidates, 368 positives; the matched sets are described in Appendix B).

Three results follow (Figure 2). First, the model ranks well with no evidence at all: gene symbols alone give 0.779 [0.723, 0.827]. Second, removing the confound by matching costs about 0.1 AUROC

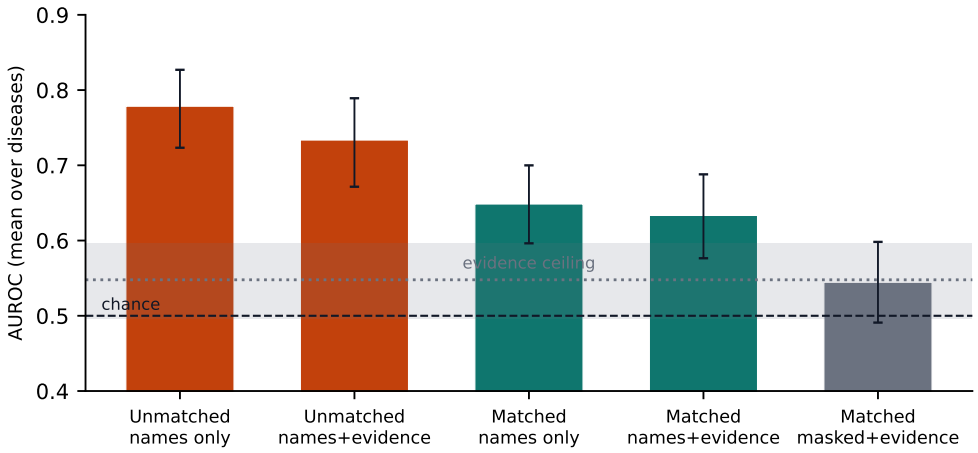

Figure 2: Frontier language model across information conditions on 55 diseases. The dotted line and shaded band are the information ceiling of the supplied evidence, a logistic regression fitted on the identical rendered features. The blinded condition sits at that ceiling.

Table 2: Frontier language model on 55 diseases, per-disease AUROC with cluster bootstrap intervals. Statistical baselines on the same matched sets are in Appendix B: every evidence channel is within noise of chance there, so the matched conditions have a low ceiling.

Condition	AUROC [95% CI]
Unmatched, gene names only	0.779 [0.723, 0.827]
Unmatched, gene names + evidence	0.734 [0.672, 0.789]
Matched, gene names only	0.649 [0.596, 0.700]
Matched, gene names + evidence	0.633 [0.576, 0.688]
Matched, masked identity + evidence	0.545 [0.491, 0.598]
Matched, ceiling: logistic regression on identical features	0.548 [0.497, 0.596]

(paired, $p < 0.001$), and the model's ranking tracks the confound directly: Spearman correlation with prior clinical development is 0.450 [0.380, 0.511] unmatched, falling to -0.089 once matched. Third, supplying the pre-cutoff evidence does not improve the ranking (paired difference -0.0312, $p = 0.0863$).

Interpreting the masked condition requires knowing what the evidence can support. We therefore report every matched condition against an information ceiling: a logistic regression fitted on the identical rendered feature vector handed to the model, evaluated leave-one-disease-out. On the matched sets that ceiling is 0.55 [0.50, 0.60].

Under a continuous evidence rendering, masking gene identities following the blinding protocol of Cuccarese (2026) gives 0.54 [0.49, 0.60], which is at that ceiling. The blinded model is therefore not failing to reason over the evidence; it extracts about as much as a fitted linear model does, and the evidence itself supports very little once the confound is matched out. With gene identities visible the same model reaches 0.63 [0.58, 0.69], a paired gain of 0.052 ($p = 8.8e{-}7$) over the blinded condition and above what the supplied evidence supports. That excess is entity-bound knowledge rather than evidence use.

An earlier version of this experiment used a coarse evidence rendering in which 88.7% of candidates shared a tied profile; the masked arm scored 0.491 there and would have supported a stronger and wrong conclusion that the model cannot use evidence at all. We report the continuous

rendering.

We do not resolve whether the entity-bound excess is legitimate biological knowledge or recall of post-cutoff events; Section 7 bounds it. We tested one model family, so these numbers describe that model rather than language models in general.

7 The cutoff is fictional for a pretrained model

Section 6 leaves the entity-bound excess unexplained. We bound how much of it is recall of post-cutoff events. On the matched sets, closed-book and with no evidence or candidate list, we asked the model directly whether an interventional trial of a drug against target g for disease d started between July 2021 and June 2023, over 736 pairs balanced 50/50 between pairs that did and did not.

The model answers above chance: balanced accuracy 0.633 [0.600, 0.668], and AUROC 0.666 [0.626, 0.704] using signed confidence. It is conservative rather than yes-biased, answering positive for 36.4% of pairs against a true rate of 50%, with specificity 0.769 and sensitivity 0.497.

Because these pairs are matched on prior clinical development, this is not the confound reasserting itself: it is direct knowledge of post-cutoff outcomes. Its magnitude is close to the model's named-condition ranking score on the same sets (0.63), which is consistent with the entity-bound excess over the evidence ceiling being substantially recall rather than reasoning, though our design does not prove that attribution. For a model whose pretraining postdates the cutoff, a temporal split of this kind does not create the counterfactual it appears to create.

8 Withholding the shortcut does not remove it

The standard control for a confound is to keep it out of the model's inputs. We measured whether that works. For 400 targets, stratified across the prior-development distribution, we asked the model closed-book, with no evidence and no candidate list, to estimate how many distinct diseases had a drug against that target in interventional trials as of June 2021.

The model reproduces the benchmark's dominant confound from the gene symbol alone: Spearman 0.705 against the true count, and AUROC 0.893 [0.861, 0.921] for whether a target has any prior clinical development. Its separate druggability rating correlates with prior development at 0.712.

A control that withholds prior clinical development from the prompt therefore does not withhold it from the model. This generalises beyond this benchmark: wherever labels derive from a historical process that a pretrained model has read about, input ablation is not a valid control, and the confound must be removed from the evaluation set by construction rather than from the prompt.

9 Label integrity

Retrospective curation. Of the 8,009 pairs appearing for the first time in the June 2024 release, 58.4% correspond to trials that began before the cutoff (Figure 3). The extreme case is NCT00000482, a trial that started in 1965. A snapshot-difference label is wrong for most of its positives, which is why we anchor on start dates.

What the label counts. The permissive label is any interventional trial, and that is not drug development. Of positives, 75.8% are sponsored by academic or hospital investigators, 16.2% by industry, and 15.0% are Phase 4; only 13.5% are industry-sponsored Phase 1 or 2. Restricting to

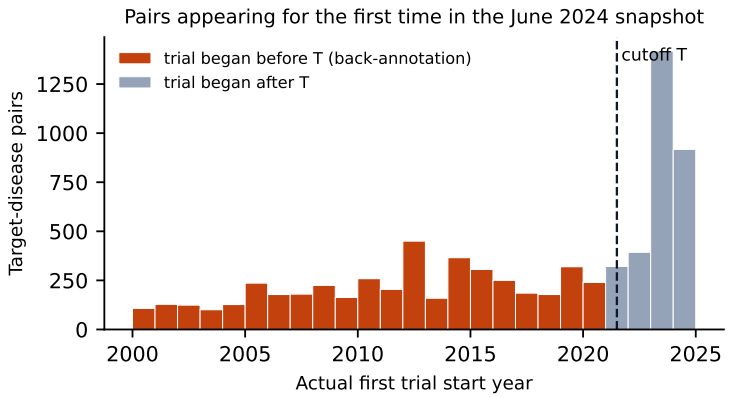

Figure 3: Retrospective curation. For target-disease pairs appearing for the first time in the June 2024 Open Targets release, the actual start year of the earliest linked trial. Most began before the June 2021 cutoff, so treating first appearance in a snapshot as a new event mislabels them.

that strict definition leaves 47 positives across 7 diseases, where prior clinical development still reaches 0.951 [0.925, 0.976] against 0.524 for aggregated biology, and 100.0% of positives remain already-drugged targets. That is a robustness check with wide intervals, not a second primary analysis.

Novel targets are absent. Only 1,520 of 29,251 candidate targets can ever receive a positive label, because the label requires a curated drug mechanism, and 0 targets entered interventional development for the first time during the window. Because our own results say curation lags reality, we re-tested this against the June 2026 release, which has had three further years to curate: the count remains 0.

10 A second label inverts the ordering

Scoring the same signals on whether a pair acquired new human genetic support between the 21.06 and 24.06 releases, restricted to datasources present in both, prior clinical development falls to 0.499 [0.476, 0.525] and aggregated biology rises to 0.660 [0.618, 0.701]. Which signal wins depends on which proxy is chosen for the same latent construct.

We report this as a robustness observation rather than a primary result, because the label is contaminated: using the per-year evidence attribution in the 26.06 release, of the 146 positives whose date could be resolved, out of 616, 39.0% have first genetic evidence dated at or before 2021.

11 Limitations

We evaluate one model family, so Sections 6 and 8 describe that model. Tool use was instructed off rather than enforced; the workflow logs record tool calls, and we report the model's stated rationales alongside its scores, but we cannot prove no retrieval occurred.

The primary cutoff is June 2021, with a replication at June 2022 (Appendix D); the two share most underlying trials, so this is not an independent replication.

The candidate pool is restricted to targets with some pre-cutoff evidence for the disease, which makes the task easier than open discovery, and pool membership is itself an artefact of 2021 cataloguing.

The within-target axis has fewer positives per target than the between-target axis has per disease, and the strict-label analysis rests on 47 positives. Absence of a detectable effect is not absence of an effect: the design cannot detect mean AUROC differences below 0.045.

The genetic-support label is contaminated at a rate we could bound only on a minority of positives.

12 Conclusion

Retrospective target-disease benchmarks contain two orthogonal axes. Nearly all of the discrimination lies on the between-target axis and is explained by properties of the protein, prior clinical development at 0.972 and disease-independent structural druggability at 0.813. On the within-target axis, which is the question prioritization systems are asked, disease-specific biological evidence is at or below chance. A frontier language model scores 0.779 from gene symbols with no evidence, gains nothing from the evidence supplied to it, and falls to chance when gene identities are hidden. Because the model reconstructs prior clinical development from the gene symbol alone at AUROC 0.893, withholding that confound from the prompt does not control it.

Reporting pooled AUROC on these benchmarks conflates the two axes. We recommend reporting both axes separately, including a clinically blind druggability baseline alongside the majority-class baseline, and constructing evaluation sets in which the confound is removed by design rather than omitted from the prompt.

Data and code availability. The benchmark, the matched and within-target splits, the trial-date anchoring procedure, the model evaluation harness and all analysis code are in the accompanying repository. Every number in this paper is generated from committed artifacts by `code/consolidate_results.py`, and `code/check_numbers.py` fails the build if any reported value drifts.

References

Rareagent: Self-evolving reasoning for drug repurposing in rare diseases, 2025. Preprint, 7 October 2025.

Anna Breit, Simon Ott, Asan Agibetov, and Matthias Samwald. Openbiolink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*, 36:4097–4098, 2020. doi: 10.1093/bioinformatics/btaa274.

Annalisa Buniello, Daniel Suveges, Carlos Cruz-Castillo, Maria Bernardo Llinares, Helena Cornu, Irene Lopez, Kirill Tsukanov, Juan Maria Roldán-Romero, Chintan Mehta, Luca Fumis, Graham McNeill, James Hayhurst, et al. Open targets platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Research*, 53:D1467–D1475, 2025. doi: 10.1093/nar/gkae1128.

Michael Cuccarese. Epistemic blinding: An inference-time protocol for auditing prior contamination in llm-assisted analysis, 2026. Preprint, 7 April 2026.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery, 2021.

Maria J. Falaguera, Ellen M. McDonagh, David Ochoa, Polina V. Rusina, Juan Maria Roldan-Romero, David G. Hulcoop, Andrew R. Leach, and Ian Dunham. Temporal trends in evidence

supporting novel drug target discovery. *Nature Communications*, 17:492, 2025. doi: 10.1038/s41467-025-67180-y.

Enrico Ferrero, Ian Dunham, and Philippe Sanseau. In silico prediction of novel therapeutic targets using gene-disease association data. *Journal of Translational Medicine*, 15:182, 2017. doi: 10.1186/s12967-017-1285-6.

Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. Ground(less) truth: A causal framework for proxy labels in human-algorithm decision-making, 2023.

Daniel S. Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L. Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017. doi: 10.7554/eLife.26726.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N. Nadkarni, Benjamin S. Glicksberg, Nils Gehlenborg, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30:3601–3613, 2024. doi: 10.1038/s41591-024-03233-x.

Emily A. King, J. Wade Davis, and Jacob F. Degner. Are drug targets with genetic support twice as likely to be approved? revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genetics*, 15:e1008489, 2019. doi: 10.1371/journal.pgen.1008489.

Eric Vallabh Minikel, Jeffery L. Painter, Coco Chengliang Dong, and Matthew R. Nelson. Refining the impact of genetic evidence on clinical success. *Nature*, 629:624–629, 2024. doi: 10.1038/s41586-024-07316-0.

David Narganes-Carlón, Daniel J. Crowther, and Ewan R. Pearson. A publication-wide association study (pwas), historical language models to prioritise novel therapeutic drug targets. *Scientific Reports*, 13:8366, 2023. doi: 10.1038/s41598-023-35597-4.

Matthew R. Nelson, Hannah Tipney, Jeffery L. Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, Lon R. Cardon, John C. Whittaker, and Philippe Sanseau. The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47:856–860, 2015. doi: 10.1038/ng.3314.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453, 2019. doi: 10.1126/science.aax2342.

David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Asier Gonzalez-Uriarte, Cinzia Malangone, Alfredo Miranda, Luca Fumis, Denise Carvalho-Silva, Michaela Spitzer, et al. Open targets platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research*, 49:D1302–D1310, 2021. doi: 10.1093/nar/gkaa1027.

Pui Chung Siu, Claudia Cabrera, Mani Mudaliar, and Arkaitz Zubiaga. Thbkg: A temporal biomedical knowledge graph for decision-aligned clinical advancement prediction, 2026. Preprint, submitted 6 August 2026.

Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. Agatha: Automatic graph-mining and transformer based hypothesis generation approach, 2020. Preprint, 13 February 2020.

Michael Zietz, Daniel S. Himmelstein, Kyle Kloster, Christopher Williams, Michael W. Nagle, and Casey S. Greene. The probability of edge existence due to node degree: a baseline for network-based predictions. *GigaScience*, 13:giae001, 2024. doi: 10.1093/gigascience/giae001.

A Evidence datasources at the cutoff

The June 2021 release contains 21 direct-association datasources in six types: literature (`europepmc`); genetic association (`ot_genetics_portal`, `phewas_catalog`, `eva`, `genomics_england`, `orphanet`, `uniprot_variants`, `uniprot_literature`, `gene2phenotype`, `clingen`); animal model (`phenodigm`); RNA expression (`expression_atlas`); somatic mutation (`cancer_gene_census`, `intogen`, `eva_somatic`); affected pathway (`reactome`, `slapenrich`, `crispr`, `progeny`, `sysbio`); and known drug (`chembl`, excluded as the label source). The 24.06 release adds `gene_burden`, `impc`, `crispr_screen` and `cancer_biomarkers` and drops `phenodigm` and `phewas_catalog`, so cross-release comparisons use only datasources present in both.

B Matched candidate sets and their ceiling

Each positive is matched to up to four negatives from the same disease whose prior clinical development is within 15%, giving 55 diseases, 1,840 candidates and 368 positives, with median prior indications 78 among positives and 75 among negatives. On these sets prior clinical development falls to 0.521 [0.503, 0.540]. Every statistical evidence channel is within noise of chance: aggregated biology 0.541 [0.486, 0.594], literature 0.550, genetics 0.497, random 0.488. The matched language-model conditions therefore have a low achievable ceiling, which is why we do not read 0.491 as a pure model failure.

C Robustness

Tie handling. A rank-based AUROC breaking ties by input order overstates zero-inflated signals: for the genetic aggregate it gives 0.521 against 0.485 with mid-rank handling. All reported values use mid-rank handling; the precedent baseline is unaffected (0.972 versus 0.972).

Ontological near-duplication. 73.3% of positives involve a target with a pre-cutoff trial for an ontologically related disease. Counting only prior diseases that are neither ancestors nor descendants of the target disease gives AUROC 0.967 [0.956, 0.975] against 0.972 unrestricted.

Tractability stratification. Splitting at the median clinically blind tractability score, aggregated biology reaches 0.520 [0.466, 0.571] among tractable targets and 0.573 [0.449, 0.690] among intractable ones. The intervals overlap and we do not claim a difference.

Oncology. Prior clinical development reaches 0.974 [0.969, 0.978] in oncology (37 diseases, 324 positives) and 0.969 outside it (20 diseases, 197 positives). Aggregated biology reaches 0.620 in oncology and 0.394 [0.298, 0.479] outside it.

D Replication at a second cutoff

Repeating the construction on the June 2022 release with a 12-month horizon gives 23 diseases and 354 positives. Prior clinical development reaches 0.975 [0.967, 0.982] against 0.395 for aggregated biology, 100.0% of positives involve an already-drugged target, the median positive target had 208.0 prior indications, and 0 targets were first-in-class.

E Supervised models

Logistic regression with balanced class weights on log-transformed features, leave-one-disease-out, gives 0.569 [0.525, 0.612] on biological features alone, 0.972 on prior clinical development alone, 0.962 on the two combined, and 0.971 on precedent plus literature volume. Adding 42 biological features to a single integer does not improve performance.