

MERIT: Mechanism driven model predicts drug outcomes and nominates indications for failed drugs

Koh-Tan, H.H. Caline^{1*}, Meić, Ivan¹, Sarı, Buğra Alptekin¹, Muller, Steve¹ & Richman, Gabriel¹

¹Omic Inc. (USA), Seattle, Washington, USA

*caline@omic.ai

Abstract

Drug development depends on efficacy and safety, but many trial-outcome prediction models incorporate trial design, prior development history or compound identity, enabling compound memorization and inflating apparent performance. We developed MEchanism-Resolved Inference of Trial outcomes (MERIT), a model that predicts trial outcomes from molecular and disease features without using information on similar-compound success. MERIT integrates the disease and drug of interest with large-scale drug-protein, protein-metabolite and immune interaction maps to link a drug's intended and potential off-target effects to tissue-specific efficacy and safety. Across 753 small-molecule drugs and 3,133 trials, MERIT achieved a best-in-class overall AUROC of 0.770 (0.765 for efficacy and 0.784 for safety). MERIT also recovered the eventual approved indications for 83% of failed drugs. Finally, we registered locked, outcome-blind predictions for 55 drug-indication pairs in ongoing Phase III trials, establishing a prospective evaluation cohort.

Introduction

When a drug fails in the clinic, the cause may lie within the molecule or outside (a trial stopped over funding, enrollment or corporate priorities)^{1,2}. Only the former is a verdict on the compound, yet the two are seldom separated. Among genuine biological failures, safety failures arise from off-target pharmacology, metabolic activation and organ-specific toxicity³, whereas efficacy failures, predominant cause of late-stage attrition, reflect a mechanism that does not

engage disease biology⁴. Whether these drug-driven outcomes can be predicted from a compound's biology before development is committed has remained open⁵.

Computational models have predicted trial-related outcomes for years, reporting area-under-receiver-operator-curve (AUCs) of roughly 0.72 to 0.88^{6,7,8,9}. These numbers come from evaluation splits that let the same compound appear in both training and test¹⁰, so a model can reach them by memorizing which drugs tend to succeed without learning why a mechanism fits a disease. It is unknown how much of that reported performance survives once the model can no longer recognize the compound before it. We name and quantify this *memorization gap* -- the accuracy lost when the same compound no longer sits in both training and test, which separates recognition of a known drug from transferable reasoning about mechanisms.

This study introduces the MEchanism-Resolved Inference of Trial outcomes (MERIT) model, which predicts trial outcomes and nominates indications from biology, not compound memorization. We profiled 753 small-molecule drugs across their 3,133 trials from disease context and integrated molecular-mechanism profiling, via a structure-activity-tissue relationship (STAR) framework. Most reported performance did not survive a fair test that scores the model only on novel drugs not found in training dataset. Retrained leading model (HINT) reached 0.626 against MERIT's 0.704, and the highest reported 0.882 (inClinico⁹) held only under a looser evaluation and a reverse-causation trial-count feature. MERIT reached an honest AUC of 0.770, carried by two pre-trial signals: disease complexity and how well the drug's mechanisms match it. Efficacy and safety failures follow opposite logic, a matching problem versus a single-breach problem. MERIT run in reverse nominates the disease a failed drug's mechanism does fit.

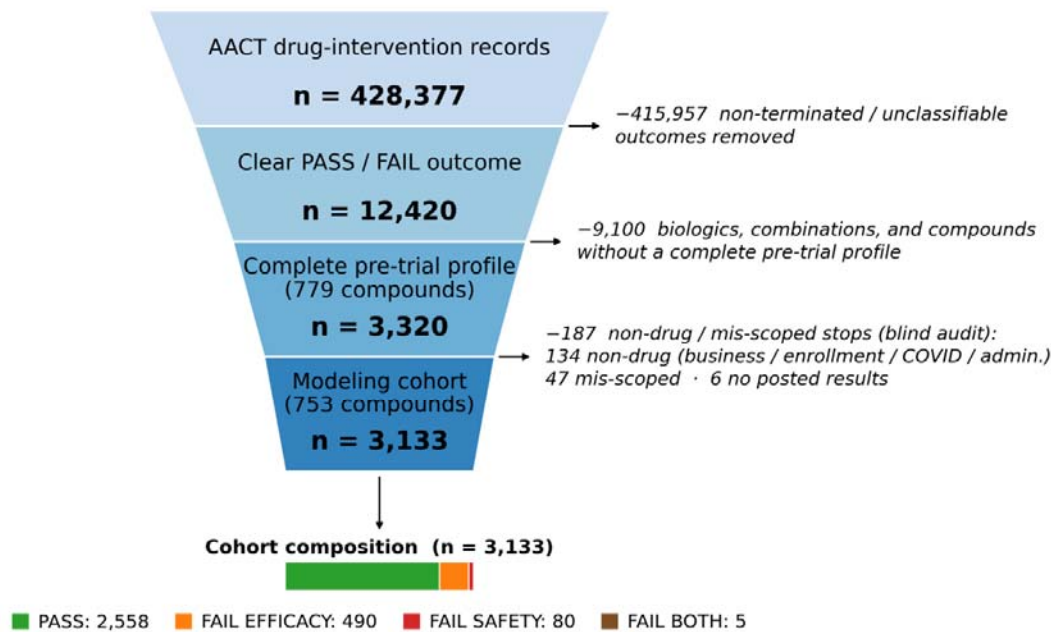
Results

Integrative biological profiling of small-molecule compounds

We assembled 3,133 trials (2,967 distinct; some test one drug in several diseases, counted once each) across 753 small-molecule compounds from ClinicalTrials.gov, classified as PASS, FAIL_EFFICACY, FAIL_SAFETY or FAIL_BOTH by a manual audit of every terminated trial (Figure 1a). PASS denotes Phase III/IV completion, not approval or demonstrated efficacy, a permissive label confounded with trial phase. Restricted to Phase III trials, the efficacy and

overall headlines barely move (AUC 0.753 and 0.748) so the discrimination is not a phase artifact. Each compound was profiled with eight compound-level pre-trial modules: tissue-specific binding (Human Protein Atlas¹¹), pathway engagement (STRING¹² against Open Targets¹³ disease genes), target–disease mechanism and genetics^{13,14}, binding specificity, safety pharmacology, pharmacokinetics (DruMAP), and disease/trial context and complexity, adding a ninth trial-design block during cross-validation (*Figure 1b*). No single module reaches the full model's discrimination (standalone AUC 0.682–0.707; *Figure S1*).

a



b

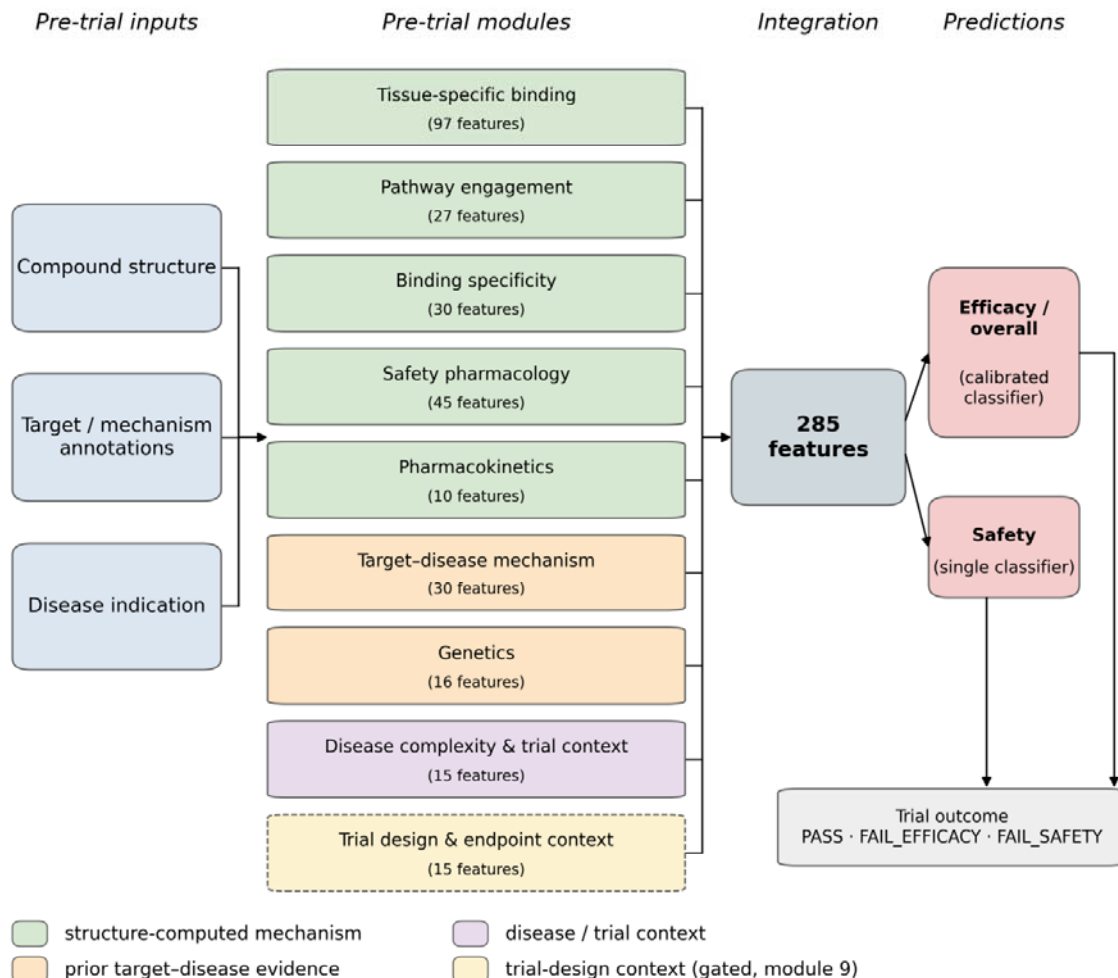


Figure 1. Dataset and computational architecture. (a) Trial-selection funnel: 428,377 AACT records narrow to the modeling cohort of 3,133 trials across 753 compounds after the blind label audit. (b) Computational architecture: pre-trial inputs feed eight core feature modules (1–8) plus a gated ninth endpoint/population/trial-design block (module 9: the leak-safe leverage axes added during cross-validation), combined into 285 features and routed to two task heads (a classifier for efficacy/overall; a single classifier over the six mechanism groups for safety).

Predicting outcomes for unseen drugs

Tested only on drugs whose structure it never encountered in training¹⁰, the full model reaches an overall AUC of 0.770, efficacy 0.765 and safety 0.784 (Figure 2a). All three exclude chance via drug-clustered confidence intervals (CIs), and scrambling the labels collapses performance to chance (0.498 across 50 shuffles; Figure 2b, Table S1). Ranking trials by predicted risk, the 100 flagged most strongly fail at four to six times background (Table S2).

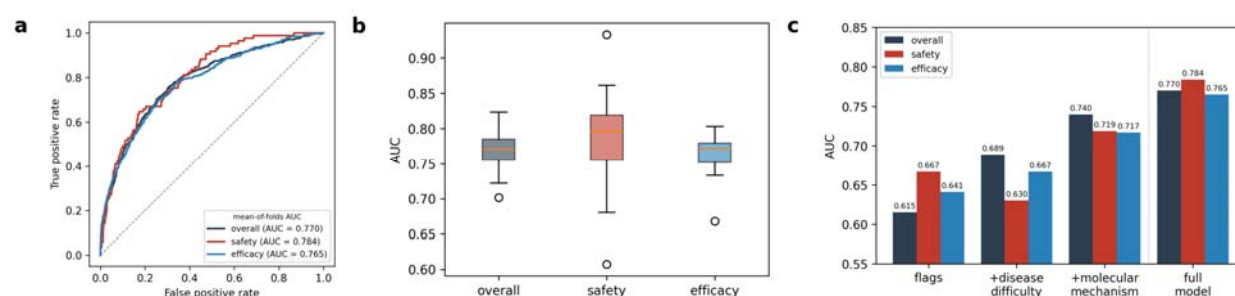


Figure 2. Prediction performance and signal decomposition. (a) ROC curves for the full model under compound-holdout: overall AUC 0.770, safety 0.784, efficacy 0.765. (b) Per-fold AUC across the 25 folds (5 seeds × 5 folds). (c) Signal decomposition (averaged across folds): starting from disease/trial flags, adding the disease-complexity baseline and then the molecular-mechanism features. Mechanism adds the larger share of the efficacy gain (full numbers in Table S1).

MERIT outperforms published methods on the same trials

MERIT was evaluated against HINT⁷ and TrialBench⁸ on the same trials two ways: a leak-free comparison stripping reverse-causation and establishment features from both sides, and each method's own leakier setup. MERIT led under both.

Retrained on this study's cohort under the structure-holdout test (same drugs and splits, restricted to the Phase III trials it targets), HINT reached 0.626 across the 2,787 shared trials against MERIT's 0.704, a 0.078 gap favoring MERIT in 24 of 25 splits (paired t -test $P = 7 \times 10^{-4}$; *Table 1, Figure S2*). Under HINT's own looser split both rise together (0.737 versus 0.654), so the lead is not an evaluation artifact.

On the shared trials (1,142 trials, 412 compounds) TrialBench's features reach 0.794, but that rests on an enrollment count. Removing terminated trials dropped the set to 0.738, leaving mostly establishment metadata: industry sponsorship, a data-monitoring committee and FDA-regulated status each predict approval alone (AUC 0.60 to 0.63), reverse-causation proxies for a well-resourced program. MERIT's pre-trial biology reached 0.768, surpassing their trial-design features by 0.131 (15/15 folds, paired t -test $P = 1 \times 10^{-4}$), and improving their set when added to it (0.807).

Published methods reported higher numbers because the usual evaluation lets a drug's other trials sit in training. Changing only this, every method scored higher, most of all the mechanism-based approach, whose one fixed profile per drug is easiest to memorize (*Figure S2*). Only held-out numbers are reported (overall 0.770; Phase III 0.704 on the HINT-matched cohort).

Table 1. Comparison with published trial-outcome methods on shared trials.

Benchmark	Task	Reported	Honest	MERIT	$\Delta \sim (\text{model-honest}) \sim$	P
HINT ⁷	Phase III completion	0.654	0.626	0.704	+0.078	paired t -test 7×10^{-4} ; Wilcoxon 2×10^{-4} ; DeLong 3×10^{-4}
TrialBench ⁸	Drug approval	0.794	0.637	0.768	+0.131	paired t -test 1×10^{-4} ; Wilcoxon 6×10^{-4}

As a negative control, a classical network-proximity baseline scored efficacy at chance (AUC 0.470)¹⁵, so the signal is MERIT's learned reading of mechanism (efficacy 0.765), not proximity.

Decomposing the pre-trial signal

Adding features one group at a time on identical partitions separated the contribution of each pre-trial input (*Figure 2c*, *Table S1*). Disease and trial flags alone reached an overall AUC of 0.615; adding the indication's historical failure rate (estimated only from training trials) raised it to 0.689. The molecular-mechanism profile added further to 0.740, most by the molecular gain for efficacy (+0.051), exceeding the disease's own difficulty (+0.026). Mechanism alone without the disease-difficulty term reached 0.694, out-predicting disease context (0.667), so mechanism rather than the disease baseline is the larger contributor.

Efficacy failure is a conjunctive matching problem: mechanism and disease must line up, and the gain comes from the features that measure that match (genetic and Mendelian support, Open Targets biology channels¹³, target–disease network links, KEGG co-membership, DepMap dependency). This holds under Phase III restriction, where the molecular gain over disease context barely moved (+0.054 on 2,114 PASS and 326 efficacy failures, versus +0.051 full-cohort).

Safety failure is partly disjunctive: a drug is unsafe if any one of several liabilities is present. After removing the shared disease-difficulty component, the six detectors (promiscuity, hepatic, cardiac, network, tissue, context) behaved largely independently, and the union of their top-decile alerts flagged 26% of failures against 18% for the best single detector (*Note S1*). A single joint classifier over the same six groups, using cross-liability interactions a disjunctive product discarded, discriminated better and more stably than the noisy-odds-ratio (noisy-OR), the rule that flags a drug when any single detector triggers (0.784 vs 0.725, 23/25 folds, $P < 10^{-4}$), so it was used, keeping the detectors as a validated decomposition (*Table S3*). The same split hurt the conjunctive efficacy task (−0.04 to −0.05).

External datasets validate this structure where the liability is structure-predictable: the cardiac detector tracks *in vitro* hERG blockade¹⁶ and clinical cardiac events (SIDER¹⁷), whereas the hepatic detector does not, because trial-halting liver injury is often idiosyncratic (DILrank¹⁸; *Tables S3, S4, Note S2*).

Temporal validation

Trained on pre-cutoff trials and tested on later trials with any already-seen drug removed, so that every test trial is both later and structurally new, MERIT transfers to compounds that did not exist at training (overall 0.704–0.737, efficacy 0.649–0.731 across 2016–2018 cutoffs; safety noisier on 23–31 positives). The drop from 0.770 and 0.765 is expected as the window shrinks. This forward transfer reflects mechanism that generalizes, not a memory of established drugs (*Table S4*).

Why efficacy trials fail

A trial is defined by the population it enrolls and the endpoint it measures, and a mechanistically sound drug can miss on either. Both are interrogable before the trial, unlike whether the effect is large enough. A targeted drug tested in patients not selected for whether that target drives their disease fails far more often (off-target 63% versus on-target 17%). When a primary endpoint is a narrow surrogate in a multi-system disease, even a drug proven for that disease can fail because its mechanism does not move that measure. Across six organ systems, a mismatched endpoint failed 60% of the time against 9% when it matched, a split the model read at chance (AUC 0.46) but a physiological-match score captured well (AUC 0.80). Empagliflozin failed a six-minute-walk endpoint (EMPERIAL¹⁹) that its kidney-and-fluid mechanism does not drive, while succeeding on cardiovascular outcomes (EMPEROR-Reduced²⁰), and tolvaptan showed the reverse in polycystic kidney disease (TEMPO 3:4²¹). Encoded as three pre-trial features that cannot leak the outcome (module 9, applied within Phase III), these re-ranked the trials they apply to without moving the overall AUC. What remains is on-mechanism, right-endpoint trials that failed on effect size alone.

Where the model errs

A blind, web-verified read of every confident miss sorted MERIT's errors into two coherent groups (*Figure 3a*). The safety misses are managed-risk drugs kept safe by dose control (taxanes, PARP and multikinase inhibitors) plus on-target or idiosyncratic injuries shared with passing drugs (*Table S5*). The efficacy misses are right-drug, wrong-setting failures: an approved drug that missed in a particular population, formulation or line (lacosamide, mesalamine). Trial-specific effect size remained, separating by neither placebo response nor endpoint subjectivity (both ≈ 0.73). Neither group could be recovered by any further outcome-blind feature tested

(Table S6, Note S3). MERIT is a detector of risks and mechanisms with a mapped boundary, not an oracle of failure.

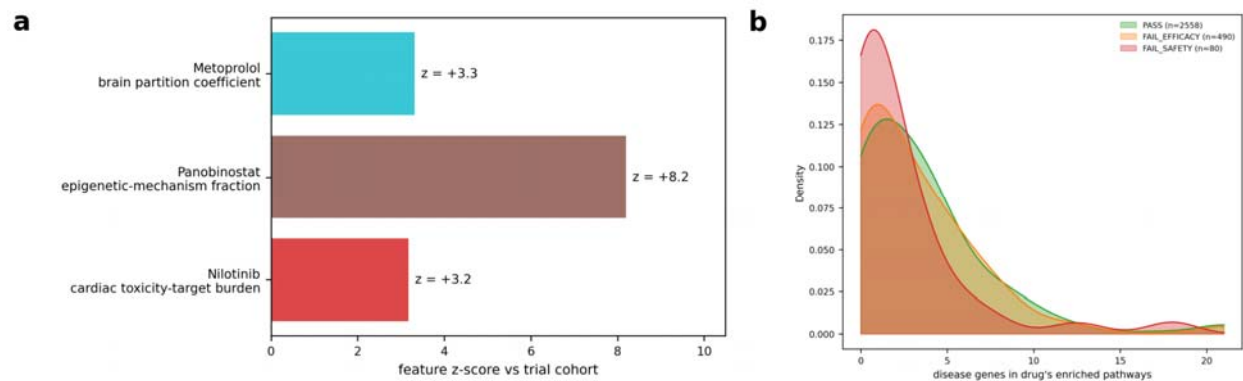


Figure 3. Mechanism of failure: the molecular features as a per-trial pharmacological instrument. (a) Single-feature fingerprints of three confident misses (drug-mean z-score vs the cohort): nilotinib cardiac-target burden ($z = +3.2$), panobinostat epigenetic-mechanism fraction ($z = +8.2$), and metoprolol brain-partition mismatch against COPD ($z = +3.3$). These are strong for the individual drug but rare across the cohort, so they explain single misses without raising population AUC. (b) Disease-pathway overlap by outcome: for each drug, the number of the disease's associated genes that fall in the pathways the drug's targets engage (STRING network against Open Targets disease genes), so a high count means the drug's mechanism is embedded in the disease's gene network and a low count means it acts on pathways outside it. Safety failures engage the fewest (median 1, versus 2 for efficacy failures and 3 for passes; Cliff's $d = -0.25$ safety, -0.05 efficacy, 2000-resample CIs exclude zero): safety-failing drugs act largely on pathways outside the disease's gene network, consistent with off-target toxicity, whereas efficacy failures barely differ from passes. FAIL_BOTH ($n = 5$) is omitted as too few for a density estimate.

A firmer limit sits beneath these cases. A model that reads only the molecule gives the same score to repeat trials of one drug in one disease, but the trials themselves differ: nearly one in five disagree (18%, 95% CI: 14-22%), and most drug-disease pairs that failed once passed on

another attempt. A large part of the efficacy verdict therefore rests on dose, population, endpoint and chance that no molecular feature can carry, and MERIT's residual efficacy error falls largely inside this irreducible band. An efficacy AUC of 0.765 is close to the practical ceiling for a pre-trial molecular model, and closing the gap calls for sharper reading of trial context. Trial-terminating safety failures are too rare among repeated trials to read the same way, so the safety limit lies in what molecular structure can resolve.

The same features also separated safety failures from passes by disease-pathway overlap (Cliff's $d = -0.25$) while efficacy barely differed, explaining rather than predicting on their own (*Figure 3b*).

Repositioning failed drugs by mechanism

Read in reverse, the same logic (efficacy needs a mechanism–disease match) becomes a hypothesis generator: for any compound–disease pairing, one minus the predicted failure probability is a mechanism-fit score. Since MERIT predicts with the compound held out and uses no drug-name feature, the score reflects neither the drug's identity nor any outcome. Three analyses confirmed this (*Figure 4*).

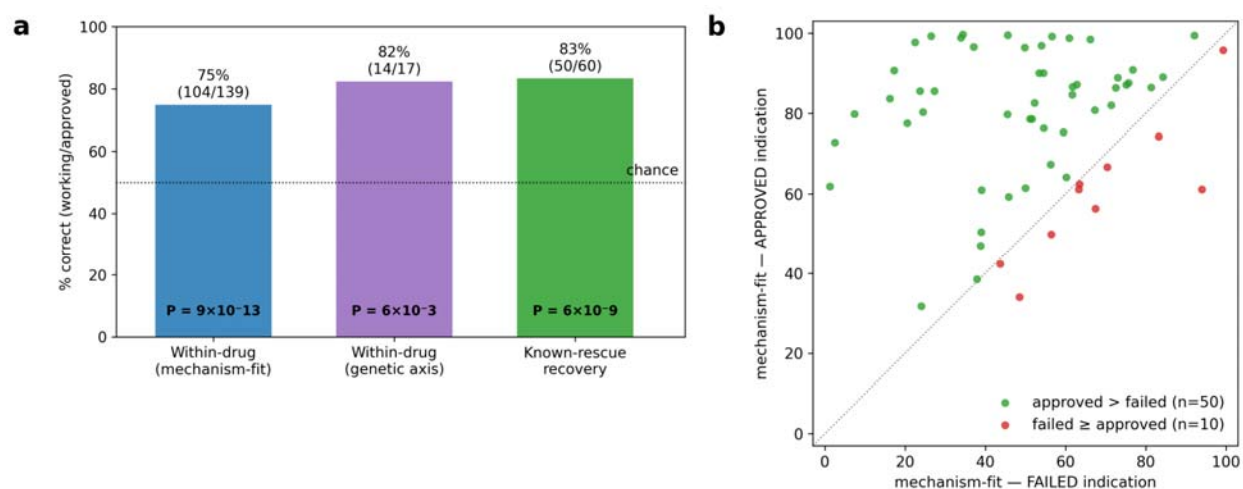


Figure 4. From failure to rescue: repositioning failed drugs by mechanism. (a) Three tests of whether the mechanism score ranks the correct (working/approved) disease highest: within-drug by mechanism-fit (75%, $P = 9 \times 10^{-13}$), within-drug by the independent Open Targets genetic

axis (82% of informative drugs, $P = 6 \times 10^{-3}$), and recovery of known repositionings (83%, $P = 6 \times 10^{-4}$). Dotted line, chance. **(b)** Recovery of known repositionings. For 60 drugs that failed one disease but are approved for another (repoDB), the mechanism-fit score (the failure prediction run in reverse, $\text{fit} = 1 - P(\text{fail})$, computed with the drug held out of training) is plotted for the approved (y) versus failed (x) disease; 83% fall above the diagonal (green), mean fit 0.78 versus 0.51 (paired Wilcoxon $P = 6 \times 10^{-4}$).

First, comparing two diseases for the same molecule removes any cross-drug confound and, because the target is identical, cannot reflect the target's prior clinical track record. Across 139 compounds with both a passing and an efficacy-failing indication, the mechanism-fit score was higher for the working indication in 75% (104/139, $P = 9 \times 10^{-13}$; *Figure 4a*), and the Open Targets genetic association, which cannot leak outcome, agreed wherever it carries signal (14/17 informative compounds, $P = 6 \times 10^{-3}$).

Second, MERIT recovered *known* repositionings. Among 60 cohort compounds that failed one indication but are approved for another (repoDB²²), it ranked the approved indication above the failed one in 50 (83%; mean fit 0.78 versus 0.51; $P = 6 \times 10^{-4}$; *Figure 4b*, *Table S7*), recovering tamoxifen to breast cancer and atorvastatin to hypercholesterolemia. This is not an artifact of prior clinical attention: it used none of the Open Targets known-drug or literature channels and survived removing each disease's base rate (a mechanism-only variant recovers 70%, 42/60), and the top-ranked indication was a documented clinical use in 48 of 54 rescued compounds. Recovery was starkest for imatinib, whose kinase-inhibitor mechanism MERIT mapped to chronic myeloid leukemia and gastrointestinal stromal tumors far above the indications it failed, COVID-19 and loiasis.

Third, applied to the 154 compounds that failed for efficacy with no approved indication, MERIT nominated the indications each target's causal biology most strongly supports, including orphan candidates (*Table S8*). Among the highest-scoring web-verified nominations (146 pairs across 52 compounds; *Table S9*), 88% carried at least preclinical evidence and 67% had a registered trial of the same drug or a same-target agent. MERIT repeatedly pointed at an indication the field independently pursued. Galeterone, a CYP17A1 inhibitor abandoned after a

failed Phase III prostate-cancer trial (ARMOR3-SV), was nominated for congenital adrenal hyperplasia, where the CYP17A1 inhibitor abiraterone is now in trials (*Table S10*). These are pre-trial mechanism hypotheses, not efficacy claims.

Finally, the reverse scoring anticipates repositionings before they are confirmed. Training the efficacy head only on pre-cutoff trials and scoring stranded drugs' later new indications in reverse (*Table S11*), the eventual successes outranked the failures at the drug–disease level (AUC 0.78 at both the 2016 and 2017 cutoffs; Mann–Whitney $P = 0.03$ and 0.02), recovering selinexor→multiple myeloma and siponimod→multiple sclerosis from pre-cutoff mechanisms alone. Power is limited (8 to 12 confirmed successes per cutoff), so this corroborates rather than headlines. One error is systematic: host-directed drugs repurposed against COVID-19 are over-credited (predicted 39% against 82% observed failure, $n = 51$; the immune-target versus viral-driver mismatch behind it is detailed in Methods). This is confined to an emergent disease with no prior therapeutic biology, and non-COVID infectious indications remain calibrated.

Discussion

Two pre-trial inputs account for much of a trial drug's success or failure: difficulty of the indication, and the compound's mechanism--disease fit. The first formalizes a portfolio manager's intuition. The second is a largely overlooked determinant of a drug's fate and the strongest molecular lever on whether a drug works, even if strict evaluation bounds how large it looks.

The asymmetry between the failure modes is biologically coherent. A drug can be tolerable, enter a tractable disease, and still fail because its mechanism never engages, whereas any single liability makes it unsafe, so safety is best modeled as independent detectors and the one-liability split hurts efficacy. This parallels genetic analyses of why trials stop²³ but reads from structure-computed mechanism. The efficacy signal complements gene-level genetic evidence by capturing indirect mechanisms that need no direct target–disease gene overlap (letrozole, azacitidine)^{14,24,25}. Drug-induced expression changes added nothing beyond the structure-, network- and tissue-derived features (446 LINCS L1000 compounds^{26,27}).

Unlike knowledge-graph repositioning models that learn from how often a drug and disease already co-occur²⁸, MERIT's nominations are grounded in target→disease mechanism and come from the model that predicts failure. Recovering known rescues with the drug held out (*Figure 4*) reframes efficacy-failed assets as a search space.

MERIT's overall AUC of 0.770 sits within the 0.70–0.85 range of prior reports^{29, 6, 30}, but those tests do not hold a drug out of its own training data. The single highest, inClinico's 0.882⁹ on Phase II→III transition, comes from a split that holds out neither compound nor target. Reproducing that task, it falls to 0.746 and its signal is trial-registration detail (sponsor, trial size, a count of the program's own later trials), not pre-trial biology (*Figure S3; Note S4*).

Two contributions follow: the *memorization gap* as a measurement instrument -- looser evaluation inflates mechanism-based models, so we urge structure-holdout evaluation as the field's minimum standard; and repositioning, the failure predictor read in reverse recovering known rescues.

The mechanism signal weakens where targets or diseases are poorly annotated in Open Targets (~22% of labels). Added to the full mechanism model rather than as the first block (+0.074; *Table S1*), the disease-difficulty term adds little discrimination (+0.014 AUC) but calibrates predictions toward the base rate, so a well-tolerated compound in a historically successful indication can be over-credited despite a weak match. Among compounds whose target is genetically associated with the disease, those in high-difficulty indications still fail half the time (41–47% across thresholds), compared with 58–80% for off-target compounds. A strong match halves the baseline risk without eliminating it, so confident predictions here remain intrinsically uncertain.

Safety is the hardest task, limited by what the structure can see rather than the number of examples. A blind-audited one-third expansion of the failure set did not improve discrimination, because the added failures are disproportionately invisible to a binding model (*Note S3*). Restricting to Phase III trials gives a safety AUC of ≈0.73, whereas the full-cohort 0.784 counts earlier-phase dose-finding terminations. Most confident safety misses are on-target or class effects whose target is shared with passing drugs, and the head's score is uncorrelated with black-box warning or post-market withdrawal (AUC 0.501 and 0.483), so it predicts trial-termination

risk specifically. A failure differs from its safe counterpart by dose, exposure and population, handled by indication-specific thresholds (*Table S12*). Restricting to small molecules is deliberate, similar to structure-dependent models such as HINT.

All performance is retrospective, so as a forward commitment we registered locked, outcome-blind predictions for 55 novel drug–indication pairs across 54 ongoing Phase III trials (mechanism recomputed per pairing). Each carries a calibrated failure probability fixed by a public timestamp and SHA-256 hash before any trial reports (deposited at <https://doi.org/10.5281/zenodo.21824277>). The *entire* set is scored at readout, by ROC-AUC and calibration, and a confidence stratifier is pre-registered: each prediction is tagged high-confidence or uncertain beforehand, with the hypothesis that the high-confidence tier (27 of 55; 24 PASS, 3 FAIL) is more accurate, the model self-identifying where it can be trusted. The forward-in-time transfer to structurally novel future compounds is the definitive prospective test: brand-new compounds entering Phase I, registered before outcomes are known.

Much of a trial's biological fate is thus set before a single patient is enrolled, early enough to redirect a program before its largest costs. A drug that fails its trial has not always failed its biology, merely directed at the wrong disease, and the same pre-trial mechanism that anticipates the failure can point it at a better one.

References

1. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32**, 40-51 (2014).
2. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273-286 (2019).
3. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203-214 (2010).
4. Hwang, T. J. *et al.* Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern. Med.* **176**, 1826-1833 (2016).
5. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20-33 (2016).
6. Lo, A. W., Siah, K. W. & Wong, C. H. Machine learning with statistical imputation for predicting drug approvals. *Harvard Data Sci. Rev.* **1**, 1 (2019).
7. Fu, T., Huang, K., Xiao, C., Glass, L. M. & Sun, J. HINT: hierarchical interaction network for clinical-trial-outcome predictions. *Patterns* **3**, 100445 (2022).

8. Chen, J. *et al.* TrialBench: multi-modal AI-ready datasets for clinical trial prediction. *Sci. Data* **12**, 1564 (2025).
9. Aliper, A. *et al.* Prediction of clinical trials outcomes based on target choice and clinical trial design with multi-modal artificial intelligence. *Clin. Pharmacol. Ther.* **114**, 972-980 (2023).
10. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463-477 (2019).
11. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
12. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638-D646 (2023).
13. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimaged, redesigned, rebuilt. *Nucleic Acids Res.* **51**, D1353-D1359 (2023).
14. Minikel, E. V. *et al.* Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459-464 (2020).
15. Guney, E., Menche, J., Vidal, M. & Barabási, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331 (2016).
16. Wang, S. *et al.* ADMET evaluation in drug discovery. 16. Predicting hERG blockers by combining multiple pharmacophores and machine learning approaches. *Mol. Pharm.* **13**, 2855-2866 (2016).
17. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075-D1079 (2016).
18. Chen, M. *et al.* DILrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov. Today* **21**, 648-653 (2016).
19. Abraham, W. T. *et al.* Effect of empagliflozin on exercise ability and symptoms in heart failure patients with reduced and preserved ejection fraction, with and without type 2 diabetes. *Eur. Heart J.* **42**, 700-710 (2021).
20. Packer, M. *et al.* Cardiovascular and renal outcomes with empagliflozin in heart failure. *N. Engl. J. Med.* **383**, 1413-1424 (2020).
21. Torres, V. E. *et al.* Tolvaptan in patients with autosomal dominant polycystic kidney disease. *N. Engl. J. Med.* **367**, 2407-2418 (2012).
22. Brown, A. S. & Patel, C. J. A standard database for drug repositioning. *Sci. Data* **4**, 170029 (2017).
23. Razuvayevskaya, O., Lopez, I., Dunham, I. & Ochoa, D. Genetic factors associated with reasons for clinical trial stoppage. *Nat. Genet.* **56**, 1862-1867 (2024).
24. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature* **629**, 624-629 (2024).
25. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856-860 (2015).
26. Subramanian, A. *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437-1452.e17 (2017).
27. Corsello, S. M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405-408 (2017).
28. Huang, K. *et al.* A foundation model for clinician-centered drug repurposing. *Nat. Med.* **30**, 3601-3613 (2024).
29. Gayvert, K. M., Madhukar, N. S. & Elemento, O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem. Biol.* **23**, 1294-1301 (2016).
30. Feijoo, F., Palopoli, M., Bernstein, J., Siddiqui, S. & Albright, T. E. Key indicators of phase transition for clinical trials through machine learning. *Drug Discov. Today* **25**, 414-421 (2020).

31. Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M. & Ide, N. C. The ClinicalTrials.gov results database — update and key issues. *N. Engl. J. Med.* **364**, 852-860 (2011).
32. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373-D1380 (2023).
33. Landrum, G. RDKit: Open-source cheminformatics. <https://www.rdkit.org> (2023).
34. Kawashima, H. *et al.* DruMAP: a novel drug metabolism and pharmacokinetics analysis platform. *J. Med. Chem.* **66**, 9697-9709 (2023).
35. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180-D1192 (2024).
36. Rehm, H. L. *et al.* ClinGen — the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235-2242 (2015).
37. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062-D1067 (2018).
38. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).
39. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576 (2017).
40. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).
41. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785-794 (ACM, 2016).
42. Ke, G. *et al.* LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* **30**, 3149-3157 (2017).
43. INSERM. Orphanet: an online database of rare diseases and orphan drugs. <https://www.orpha.net> (accessed June 2026).
44. Vasilevsky, N. A. *et al.* Mondo: unifying diseases for the world, by the world. Preprint at *medRxiv* (2022).
45. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966-967 (2016).
46. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* **48**, D783-D788 (2020).
47. Whirl-Carrillo, M. *et al.* An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **110**, 563-572 (2021).
48. Mallal, S. *et al.* HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* **358**, 568-579 (2008).

Methods

Data sources and label verification

Clinical trial data were obtained from the Aggregate Analysis of Clinical Trials (AACT) database³¹, comprising 428,377 drug-intervention records. After filtering to drug-type interventions with clear outcomes, 12,420 trials across 2,048 unique drug names were identified. SMILES structures were resolved for 1,886 drugs (92%) through PubChem³² (primary), ChEMBL (secondary) and manual curation; 162 drugs were unresolvable (biologics,

experimental codes, non-drug entries) and explicitly categorized. SMILES were standardized via RDKit³³: salt-stripped (largest fragment retained), neutralized (Uncharger), and deduplicated by the International Chemical Identifier key (InChIKey) connectivity layer (first 14 characters) to IK14, collapsing salt forms while preserving stereoisomers as distinct compounds. 914 unique drug structures remained from 1,158 unique drug identifiers.

Trial outcomes were classified from the why_stopped free-text field using keyword matching with priority ordering (safety > efficacy > enrollment > business), then corrected by the manual review and blind label audit below. All 672 terminated trials with text were manually reviewed, which corrected 34 misclassifications (distinct from the blind audit below that removed non-drug stops from the failure set), most commonly a "Data Safety Monitoring Board" mention triggering the safety keyword when the actual reason was futility. Manual review of the drug list removed 15 non-drug entries (antibody-drug conjugates, peptides, combination regimens, plant supplements) and corrected 4 misidentified compounds.

PASS was defined as Phase III or higher trial completion; Phase I/II completions were labeled IN_PROCESS and excluded as their ultimate outcome is unknown. Disease labels were manually curated where the original AACT conditions were missing or ambiguous; procedural/anesthetic trials were excluded from efficacy evaluation but retained for safety. Anti-pathogen drugs were excluded from efficacy evaluation because the binding pipeline predicts human protein interactions, not drug–pathogen interactions (*Table S13* itemizes the efficacy-evaluation exclusions). A blind, web-verified label audit (below) reclassified non-drug stops out of the failure set. The trial-level modeling cohort comprises **3,133 trials across 753 compounds** (a few trials test a drug in more than one disease and are scored once per disease; 2,967 distinct trials) with a complete molecular-mechanism profile (PASS 2,558; FAIL_EFFICACY 490; FAIL_SAFETY 80; FAIL_BOTH 5). The arm-level robustness cohort (*Table S14*) comprises **3,326 arms across 891 compounds**, with each outcome attributed to the specific investigational arm. Compounds lacking a complete computational profile (predominantly large molecules) are excluded from the modeling cohort.

Computational modules

The candidate feature pool comprises **285 features**: molecular-mechanism features (modules 1–6 below; 255 features, including two leak-safe mechanism-coverage indicators), 14 disease and

trial context features (module 7), the within-fold disease-complexity feature (module 8) added during cross-validation, and 15 features covering the endpoint/population leverage axes together with leakage-free pre-registration trial-design primitives and two further leak-safe trial-context axes, an endpoint mechanism-match for cardiovascular-event endpoints and an as-of-date negative class-precedent flag (module 9). Nested selection retains the top 20 per training fold. The per-module counts below are indicative; the released `feature_list_v8.csv` (Supplementary Information) gives the exact module-family assignment for every candidate feature.

1. **Tissue interaction** (97 features): Predicted binding scores from the proprietary binding pipeline weighted by tissue-specific expression from the Human Protein Atlas¹¹. Thirteen organ systems profiled (heart, brain, liver, kidney, immune, respiratory, endocrine, gastrointestinal, reproductive, musculoskeletal, skin, adipose, other).
2. **Network pathway enrichment** (27 features): Drug binding targets mapped onto STRING¹² protein–protein interaction (PPI) networks. Pathway enrichment (false discovery rate, FDR < 0.05) quantified against Open Targets¹³ disease targets (top 50 genes per disease). Features include pathway-disease overlap (count and fraction), PPI network reachability (fraction of disease genes reachable through drug's interaction network), mechanism classification (immune, apoptosis, cell cycle, epigenetic, DNA damage), and off-target pathway activation.
3. **Binding specificity** (30 features): Distribution statistics of predicted binding scores across the proteome (percentiles, concentration, entropy) and direct-target engagement summaries.
4. **Safety pharmacology** (45 features): Predicted binding to known toxicity targets (cardiac ion channels including hERG, hepatic enzymes, renal transporters, DNA damage machinery) and essential housekeeping genes. Per-organ vital gene binding burden.
5. **Pharmacokinetic predictions** (10 features): DruMAP³⁴-predicted intrinsic clearance, fraction unbound (human and rat), brain penetration (K_p , brain), volume of distribution, Caco-2 permeability, and CYP substrate classification (1A2, 2C9, 2D6, 3A4).

6. **Structured target–disease mechanism and genetic support** (46 features): how well engaging the compound's primary target is expected to modify the indication, computed from target→disease biology alone (Open Targets¹³ biology channels (clinical/known-drug and literature excluded), interaction-network topology, Mendelian and ClinGen/ClinVar genetics, KEGG membership, DepMap dependency, each as a within-disease rank; detailed below under "Structured target–disease mechanism features"). No trial outcome, approval or drug identifier enters these features, which quantify mechanism–disease match independently of binding breadth.
7. **Disease/trial context** (14 features): Binary disease flags (oncology, infectious, CNS, cardiac, autoimmune, metastatic, transplant, severe), trial drug and target counts, and combination flag.
8. **Disease complexity** (1 feature): the indication's historical failure rate (Bayesian-smoothed), computed within each CV training fold. Unseen diseases receive the global rate. Fold-internal computation prevents test-set leakage.
9. **Endpoint, population and trial-design context** (15 features): Five leverage axes, computed without any trial outcome and not subject to outcome leakage, encoding whether a drug's effect matches what the trial demands: an *endpoint-physiology* score (whether the compound's mechanism moves the physiological process the primary endpoint reads out), a *population-leverage* flag (whether a targeted drug's trial enrolled the population its target drives), an *endpoint-difficulty* tier (surrogate/functional/clinical-outcome), an *endpoint mechanism-match* for cardiovascular-event endpoints (whether the target lies on a validated atherothrombotic/cardiometabolic event-reduction axis when the primary endpoint is a discrete cardiovascular event: major adverse cardiovascular events (MACE), recurrent stroke or venous thromboembolism), and an as-of-date *negative class-precedent* flag (whether a same-target-class agent already failed Phase III in the same indication before this trial began, a binary, prior-only signal). Pre-registration trial-design primitives read from the ClinicalTrials.gov design module before enrollment (number of arms, randomization, masking, single-group flag, primary/secondary endpoint counts, primary-endpoint timeframe, eligibility-criteria count, comparator type, maximum age) are added. The leverage axes are scored from pre-trial tables of (endpoint,

disease)→demanded process, target→controlled process and drug-class→responsive population that use neither drug identity nor any outcome. The design primitives use no outcome, are available before enrollment, and are leak-checked (per-feature availability $AUC \approx 0.50$, well below 0.58; the placebo-arm indicator and enrollment are excluded). Comparator type and endpoint type are likewise inadmissible: they are the only design variables that separate the confident efficacy misses (*Table S6*), but both are confounded with how efficacy failures are detected (single-arm trials cannot register an efficacy failure; placebo-controlled trials with posted statistics are preferentially detected). They encode the trial's evidentiary bar (an active-comparator non-inferiority design demands less than a placebo-controlled superiority design) and add a small aggregate efficacy signal ($\sim +0.02$ overall, $\sim +0.016$ efficacy) rather than rescuing individual misses. The module is gated against availability leakage and protected from feature selection for the efficacy/overall heads only (not the safety head).

Structured target–disease mechanism features

The mechanism axis (module 6) is computed entirely from public structured-biology resources without any trial outcome: every feature is a function of (primary target, disease) biology, and none takes a trial outcome, approval status, disease base rate, NCT identifier or drug name as input. The primary target is the factual mechanism-of-action annotation (ChEMBL³⁵ drug-mechanism plus a factual target-identity supplement), resolved on a named pass; all downstream features are computed on the (target, disease) pair, so the axis reflects target→disease fit rather than drug identity. The 44 core features (46 with the two coverage indicators) comprise: Open Targets¹³ target–disease association restricted to its biology evidence channels (genetic association, genetic literature, somatic mutation, RNA expression, animal model, affected pathway), with the clinical (known-drug) channel and the general drug–disease literature co-mention channel excluded by design so no channel can encode prior clinical attention (the retained genetic-literature channel is literature-reported *genetic association*, distinct from the excluded drug–disease co-mention channel); STRING¹² interaction-network topology between the compound's targets and the disease gene module; Mendelian and ClinGen³⁶/ClinVar³⁷ germline-genetic support; KEGG³⁸ pathway co-membership; DepMap³⁹ dependency of disease models on the target; and a within-disease rank of each. Since every channel is a public database

value rather than a learned score, the axis is reproducible from the released feature matrix without any proprietary or language-model component; its leakage properties are inherited from the source data (the genetic-association and Mendelian channels are disease-risk evidence assembled independently of trial outcomes; each channel's availability is complete for the cohort, so feature presence is not an outcome proxy; and the axis is only weakly related to the disease base rate). Restoring the excluded clinical and literature channels changes efficacy discrimination by under 0.01 AUC (0.717 to 0.723), so the whole model, including its headline 0.770 overall and 0.765 efficacy, is free of clinical-precedent signal, not just the repositioning analysis.

Model training and evaluation

Gradient-boosted decision trees were used (500 trees, maximum depth 3, learning rate 0.05, subsample fraction 0.8). The overall task and each per-mechanism safety detector use a single scikit-learn gradient-boosting classifier⁴⁰; the efficacy task uses an equal-weight average of three gradient-boosting implementations (scikit-learn⁴⁰, XGBoost⁴¹ and LightGBM⁴²) fitted with balanced class weights, which was marginally more stable for the rarer efficacy-failure label (random seeds 42, 123, 456, 789, 2024). StratifiedGroupKFold cross-validation with SMILES as the grouping variable ensures zero compound overlap between train and test sets, repeated across 5 random seeds × 5 folds for 25 evaluations. Nested feature selection (top-20 by univariate AUC within each training fold) and median imputation (fitted on training data only) prevent information leakage. Three task models were trained: overall (PASS vs any failure) as a single gradient-boosting classifier, efficacy (PASS vs efficacy failure) as the gradient-boosting ensemble above (a single integrated, non-decomposed predictor), and safety (PASS vs safety failure) as a single gradient-boosting classifier over the union of the six mechanism feature groups (noisy-OR over independent per-mechanism detectors, retained for interpretability, discriminates less well). The primary analysis is at the trial level. It is repeated at the individual investigational arm under the same compound-holdout protocol (*Table S14*), confirming the signal is not an artifact of attributing a combination outcome to the wrong compound. All AUC values are reported as the mean ± SD across the 25 folds; because the 5 seeds reshuffle the same trials, these 25 evaluations are not fully independent and the SD understates between-sample variability. Drug-clustered bootstrap confidence intervals on the headline AUCs are also

computed (described below). Events-per-variable on the 20 selected features is ≈ 28 (overall), 22 (efficacy) and 4 (safety), so safety is power-limited. The analysis uses scikit-learn 1.6.1, XGBoost 3.0.3 and LightGBM 4.6.0 under Python 3.12 and completes the full 5×5 cross-validation in ~ 6 min on CPU; the boosting libraries request a GPU but fall back to CPU, and the reported metrics are CPU-reproducible (the model is regenerated and asserted against the released metrics in a provided notebook).

Probability calibration. Per-fold isotonic regression was applied to raw out-of-fold (OOF) probabilities using a 3-fold nested split within each outer training fold (no label leakage across the outer compound-holdout boundary). Isotonic calibration reduced the pooled-OOF expected calibration error (ECE) from $0.097 \rightarrow 0.035$ (overall) and $0.078 \rightarrow 0.032$ (efficacy), with the fold-averaged AUC essentially unchanged. Safety is reported as a ranker rather than a calibrated probability (its raw score is monotonic in risk but not itself a calibrated probability). Calibrated probabilities are used for overall and efficacy and the raw ranker for safety; the resulting overall-probability calibration is shown in *Table S15*.

Precision at top- k . To express discrimination as an operational triage yield, trials were ranked by predicted failure probability (out-of-fold, averaged over the 5 seeds per trial) and computed precision at k , the fraction of true failures among the k trials ranked most likely to fail, for $k = 10$ to 500 (*Table S2*). Each task is ranked within its own cohort (PASS versus that task's failure label), so the reference point is that task's own cohort-wide failure fraction: 18.4% overall, 17.5% efficacy and 3.2% safety. As ranking is invariant under the monotone isotonic transform, precision at k is identical for calibrated and raw scores and the safety ranker is directly comparable (scripts/strengthening/compute_risk_zones_and_pk.py, which also emits the calibrated-probability risk zones).

Baseline models (random forest, logistic regression) were trained with balanced class weights on the same feature set under identical cross-validation. Label permutation testing (50 shuffles) preserved the compound-grouping structure to test whether apparent signal arises from chance fold assignment.

Benchmark comparisons. MERIT was positioned against published methods on shared data in three ways. HINT is a deep-learning model (molecular graph, disease codes, eligibility-criteria

text), and TrialBench is a multi-modal benchmark of trial-design features. (i) *TrialBench feature comparison*: on the trial-identifier intersection of this study's cohort with their trial-approval task (1,142 trials, 412 compounds), TrialBench's approval label was predicted under compound-holdout (IK14 grouping; 3 seeds $\times$ 5 folds; gradient boosting), comparing their tabular trial-design features, MERIT's pre-trial biology features and the union. For the fair comparison both sides are held to the same standard: MERIT's feature set excludes trial-design and trial-context features (module 9, drug-count and combination flags), and theirs is reduced to legitimate trial-design by removing the enrollment feature (whose value differs systematically between approved and terminated trials) and the establishment metadata (industry sponsor, data-monitoring committee, FDA-regulated status, data-sharing) that behaves as a reverse-causation proxy; their full set and the union are also reported on their own terms (scripts/benchmark/benchmark_final_v2.py). (ii) *HINT model comparison*: we retrained the authors' implementation (message-passing molecular encoder, GRAM embedding of disease ICD codes, BioBERT embedding of eligibility-criteria text, ADMET (absorption, distribution, metabolism, excretion, toxicity) pretraining) on this study's cohort, formatted to its required input, restricted to Phase III and Phase II–III trials, under the structure-based compound-holdout folds shared with the matched MERIT (2,787 unique-NCT, IK14-grouped folds, each NCT in one fold; validation split carved trial-disjoint from training; 5 epochs per fold after the shipped ADMET pretraining). To confirm the implementation is faithful rather than under-trained, HINT was reproduced on its own published benchmark (TOP Phase III split, 1,146 test trials): trained from scratch it reached ROC-AUC 0.715 single-pass and 0.730 bootstrap, matching the authors' shipped predictions (0.721), and a 10-epoch run gave the same result (0.710) (results/benchmark/hint_top_reproduction.json). HINT ships its eligibility-criteria embeddings only as an empty placeholder, so they were regenerated with a faithful BioBERT substitute (dmis-lab/biobert-base-cased-v1.1, mean-pooled). Training and testing HINT on the same regenerated embeddings avoids the ~ 0.04 AUC penalty that appears only when shipped weights are scored with substitute embeddings. This study's cohort resolves to ICD codes for 55% of trials (100% on native TOP); on the ICD-complete subset MERIT leads by a similar margin (Δ AUC +0.11 versus +0.12 on the full cohort; single-seed pooled out-of-fold, not directly comparable to the 5 $\times$ 5 mean-of-folds headline HINT 0.626 versus MERIT's 0.704) (results/benchmark/hint_star_icd_subset.json). Mean test ROC-AUC is reported over 5 seeds $\times$ 5

folds on identical trials and folds. (iii) *Network-medicine baseline*: closest-distance network proximity between each drug's protein targets and the disease gene module over the STRING v12 interactome (combined-score ≥ 400 ; 19,488 nodes), following Guney et al.¹⁵. Drug targets were the union of ChEMBL mechanism-of-action genes and the pipeline target list (matched by name and IK14); the disease module was the top-100 Open Targets associated targets per indication (membership only; association scores are not a predictor, so this is the classical topology baseline, not the genetic feature axis). Proximity d_c was z-scored against 50 degree-preserving random gene-set pairs (log-degree bins) and scored directly as an efficacy predictor (higher z = targets farther from the module $\rightarrow$ failure) on the trials with both annotations (1,385 trials, 223 failures), giving a drug-clustered AUC of 0.470 (95% CI 0.406–0.538), no better than chance (scripts/network_proximity_baseline.py). For the HINT comparison the per-fold gap was tested on the 25 paired AUCs (shared fold assignments, so shared test sets): a paired *t*-test and Wilcoxon rank test across the 25 folds, plus a *t*-test on the 5 per-seed mean differences as the conservative unit (the 5 seeds reshuffle the same trials, so the 25 folds are not independent).

The memorization gap (evaluation-inflation analysis). To quantify how much reported performance depends on the evaluation split, each feature set was scored under two cross-validation schemes differing *only* in fold assignment, holding cohort, label, classifier, imputation, feature selection, seeds and fold count constant; because everything but the fold-grouping is fixed, the protocol is model-agnostic and returns any compound-level model's recognition-driven inflation. The *structure-holdout* scheme groups folds by the IK14 (StratifiedGroupKFold), so no compound appears in both train and test; the *structure-blind* scheme assigns the same trials to folds at random (StratifiedKFold, or random NCT for the STAR arm), letting a compound's other trials remain in training (verified to span the boundary in every blind fold). The inflation is the blind-minus-holdout fold-averaged AUC. The MERIT full-feature arm ran the full production model on the Phase III cohort (2,942 trial-indication rows over the 2,787 unique NCTs of the matched HINT comparison, PASS label); the feature-set arm ran a single gradient-boosting classifier on the TrialBench intersection (1,142 trials, 412 compounds, TrialBench label). One caveat bounds the structure-holdout guarantee: it removes the test compound from the downstream classifier, but the upstream structure-to-target binding predictor is separately trained, so a compound with abundant public bioactivity data could receive a more informative binding profile than a novel one. Three facts bound this. First, that predictor is a pretrained drug–

target *affinity* model: no trial outcome, approval or clinical annotation enters its training or objective, and it is applied as a fixed scorer, not retrained on this cohort, so the most a well-studied compound gains is a more accurate binding profile, not a label. Second, decisively for the efficacy and overall heads, a public-only feature set that never touches the binding layer recovers most of the efficacy increment. Third, the binding-derived safety head's score is uncorrelated with black-box-warning or post-market-withdrawal status (AUC 0.50 and 0.48), so it is not re-detecting already-flagged toxic drugs. The safety task nonetheless leans entirely on the binding layer and is the weakest and least independently reproducible of the three. All *Table 1* benchmark numbers and their paired-significance tests are locked in `benchmark_final_v2.py` and reproduced in `notebooks/08_table1_benchmark_significance.ipynb` (scripts in *Table S16*). The fair comparison holds both sides to a leak-free feature scope (MERIT's biology excludes trial-design and trial-context features; the comparator's design excludes its enrollment leak and establishment metadata).

Safety head (single classifier over the mechanism groups; six-detector interpretability layer). The safety classifier is a single gradient-boosted model over the union of the six per-mechanism feature groups (promiscuity, hepatic, cardiac, network, tissue, context), trained under the identical compound-holdout folds, with a dose-scaled exposure term protected. It was compared against a parameter-free noisy-OR over six independent per-group detectors ($P(\text{fail}) = 1 - \prod_m (1 - p_m)$): the joint model discriminates better and more stably on identical features and folds because the disjunctive product cannot use cross-liability interactions and collapses toward chance when a fold's positives are unfamiliar. The arm-level robustness cohort likewise uses the single classifier (*Table S14*; 0.688). The six detectors are retained as an interpretable, externally-validated decomposition (*Table S3*) and for setting per-mechanism, indication-specific safety thresholds (*Table S12*), but prediction uses the joint model. Efficacy uses the gradient-boosting ensemble and overall a single classifier; the disjunctive per-mechanism decomposition applied to efficacy reduces AUC (-0.04 to -0.05) and is not used. The safety head remains the weakest and most power-limited of the three (85 positives; Discussion): even with the joint model an occasional fold still collapses when its positives fall in the molecular-invisible half of the failure set (on-target class effects and idiosyncratic liver injury), a limit of what structure can see rather than of the aggregation rule.

External validation of the safety mechanism axes (*Table S3*). To assess whether the detectors capture organ-specific biology rather than general toxicity, the cardiac and hepatic panels were validated against external ground truth (matched on IK14). For the cardiac axis, hERG channel-blockade labels¹⁶ (105 cohort compounds, 65 blockers) were predicted from the four cardiac toxicity-target binding features by five-fold cross-validated logistic regression (30 repeats), against a baseline of general promiscuity (proteome-wide bound-target count) and lipophilicity (Caco-2 permeability and brain partition coefficient); organ-specificity was confirmed by residualizing the cardiac features on those confounds and bootstrapping the residualized AUC over 2,000 compound resamples. Clinical concordance used SIDER¹⁷ counts of MedDRA Cardiac-disorders adverse-event terms per compound (n = 438), with hepatic adverse-event counts as a specificity control. For the hepatic axis, the hepatic binding panel was tested against DILIrank¹⁸ drug-induced-liver-injury concern (142 Most/Less-concern hepatotoxins versus No-concern drugs; n = 213). The clinical mechanism of each genuine in-trial safety failure was classified, blind to model output, from its audit rationale (Methods, "Label audit") into an organ liability, establishing the heterogeneity of safety-failure mechanisms reported in Results (*Table S17*).

Signal decomposition. To attribute predictive performance to feature blocks, nested single-GBM (single gradient-boosting model) blocks were trained adding one block at a time on identical seeds and folds: (i) disease/trial binary flags; (ii) plus the within-fold disease-complexity feature; (iii) plus the molecular-mechanism profile through nested top-20 selection. Each block's marginal contribution is the paired fold-averaged AUC increment over the preceding model. The decomposition isolates the disease-complexity feature (+0.074 overall) from the molecular-mechanism profile (+0.051 overall, +0.051 efficacy); the single-GBM molecular-block endpoints (overall 0.740, efficacy 0.717) approach the integrated full model (0.770 / 0.765), the remaining gap reflecting the endpoint/population leverage and pre-registration trial-design context (module 9; ~+0.02 overall), the cytotoxic drug-class axis (module 6; a leak-safe Anatomical Therapeutic Chemical (ATC) class flag protected for the efficacy/overall heads, +0.009 efficacy), the ensemble efficacy head, calibration and the single-classifier safety head, confirming the attribution.

Temporal (leave-future-out) validation. To approximate prospective use we trained on trials with start year before a cutoff and tested on trials in or after it, additionally removing from the test set any compound (by SMILES) present in training, so every test trial is both temporally future and structurally novel. Cohorts, labels, exclusions and the per-mechanism safety protect-set mirror the full model exactly; each head (single-GBM overall, calibrated-ensemble efficacy, single-classifier safety) was fit on the training era and scored on the held-out future era, and the mean test AUC is reported across the 5 seeds at cutoffs 2016/2017/2018 (*Table S4*). Median imputation, nested feature selection and the within-fold disease-complexity feature are all fit on the pre-cutoff training era only, so no post-cutoff trial informs any fitted quantity (prepare_fold). The structured mechanism features (module 6) are a single current snapshot of public databases; because they are computed from target→disease biology rather than trial outcomes, and the forward transfer is carried by the binding-, topology- and genetics-derived layers that are outcome-independent, they do not encode post-cutoff trial results.

Prediction confidence gate (prospective registration). The confidence gate is a pre-registered *stratifier*, not a selection filter: every novel pair is registered and scored, and the gate only labels each high-confidence (inside a held-out reliability band) or uncertain (the middle), so the whole forward distribution is tested and the high-versus-uncertain contrast is falsifiable. The band edges derive from the canonical model's *own* out-of-fold predictions (the same model the lock fits on all cohort rows), so the bands are self-consistent with the scores they gate; on held-out data the banded subset is 90% accurate against 63% in the uncertain middle. The bands are asymmetric by necessity: the PASS band ($\approx$ base rate) cleanly reaches $\geq 90\%$ held-out precision ($P_{\text{fail}} \leq 0.20$), but the against-base-rate FAIL band cannot reach 90% at a usefully inclusive threshold, so it is set at the honest ceiling of $\sim 85\%$ ($P_{\text{fail}} \geq 0.51$) reported here. Forward pairs are additionally out-of-distribution (the raw failure score saturates near 1), so confident-FAIL bets are gated on outcome-blind *evidence* (the number of training trials informing the indication and direction-aware corroboration by the efficacy/safety sub-models), abstaining a confident-FAIL that is both thin-support (< 10 indication trials) and uncorroborated (neither sub-model ≥ 0.5 ; e.g. fenofibrate→diabetic retinopathy). A small committed override (prereg_C_confident_overrides_v1.csv) additionally abstains a bet the threshold gate cannot see is mis-framed: metoprolol→Duchenne muscular dystrophy, a confident FAIL on disease-level mismatch whose endpoint is LVEF change, where beta-blockers are cardioprotective, so

MERIT's disease-level pessimism overrides its own correct endpoint-physiology match. A per-prediction confidence score and tier are recorded (scripts/benchmark/prereg_C_confidence_gate.py).

Prospective registration. As a forward, outcome-blind commitment model predictions were frozen for ongoing trials whose outcomes are not yet known. From the ClinicalTrials.gov v2 API we retrieved every ongoing Phase III trial (recruiting, active-not-recruiting, enrolling-by-invitation or not-yet-recruiting) of the cohort compounds and retained the **novel drug-indication pairs**, a cohort compound tested in an indication it does not hold in the training cohort. Master, platform and basket protocols were excluded and any trial listing more than one condition (so each pair is a single, unambiguous indication; without this guard a histology-agnostic basket fans out into one spurious pair per tumor type), and applied the cohort's efficacy-scope exclusions (anti-pathogen, endogenous-ligand, and procedural anesthesia/sedation/pharmacokinetic-booster agents). The drug-level molecular features (207) are reused from the compound and the indication-context features (10) from the indication; the signal that discriminates a novel pairing, the target→disease mechanism fit, is **recomputed for each pair** from public biology (topology, KEGG co-membership, Open-Targets biology channels, ClinGen, Mendelian/DepMap, direct-target engagement and within-indication ranks; module 6), as are the trial-context features (endpoint-physiology, endpoint-difficulty, population-leverage, cardiovascular-event match, negative class-precedent) from each trial's own protocol and design module, all with no outcome or new trial data. No feature is median-imputed in the forward head; a pair whose disease has no Open-Targets gene module, or whose drug acts without a protein target (a cytotoxic captured by the cytotoxic-class flag, or a metabolic/cofactor agent), is flagged rather than scored on imputed biology. The overall PASS-versus-failure head was fit on all modeling-cohort rows using the identical published model (median imputation, top-*k* univariate selection, within-train Bayesian disease target-encoding, seed-averaged gradient boosting, nested isotonic calibration; the frozen artifact was not modified). To keep the readout attributable to the cohort compound rather than a co-administered agent, the **primary** set restricts to trials where the compound is the investigational focus by a deterministic rule: named in the brief title, the experimental *differentiator* (in an experimental arm but absent from the comparator, so not a shared backbone), and primary completion on or after the lock date. This gives **111 predictions across 105 trials of 76 compounds** (7 of 111 predicted to fail at 0.5, near

base rate). After a novelty filter (excluding approved uses and broad cytotoxic monotherapies), a combination-attribution filter (excluding junior-partner or comparator-arm pairs) and an outcome-blindness filter (excluding trials that had publicly reported a primary readout before the lock date, a web-verified list, since ClinicalTrials.gov fields do not reliably reflect press/conference readouts), **55 genuinely novel predictions across 54 trials of 37 compounds** remain. The entire set of 55 is the primary registered artifact, each registered with its calibrated failure probability and scored at readout by ROC-AUC and calibration (compound-clustered CI), with no post-hoc selection. On top of it the confidence stratifier is pre-registered (Methods, "Prediction confidence gate"), with the hypothesis that the **high-confidence subset (27 of 55; 24 PASS, 3 FAIL)** is the more accurate. Scoring the whole distribution rather than only the confident calls removes any selective-reporting concern and tests the stronger claim that the model knows where it is reliable. The full novel-pair set (711 predictions across 469 trials) is retained as a broader, attribution-caveated artifact under the same protocol. The predictions, their SHA-256 hash and generating commit are deposited at <https://doi.org/10.5281/zenodo.21824277> and <https://github.com/OmicMD/MERIT>, and fixed before any trial reports; at readout they are scored against realized PASS/FAIL under the same label-audit protocol by a frozen, deposited scoring script so the metric cannot be tuned after outcomes are known (build and scoring scripts in *Table S16*; artifacts in prediction/).

Phase-matched efficacy analysis. Because PASS requires Phase III completion while terminated efficacy failures span phases (many efficacy failures terminate before Phase III, where this cohort contains no PASS), trial phase is correlated with the efficacy label. To remove this confound the decomposition was repeated restricted to Phase III trials only (2,114 PASS, 326 efficacy failures in the efficacy cohort), where both classes are present, under the identical compound-holdout protocol. The molecular increment over disease context is essentially unchanged under phase restriction: +0.054 (Phase III only) against +0.051 on the full cohort, so the increment reflects mechanism rather than trial maturity. The absolute headline AUCs are likewise robust to the same restriction (phase_stratified_auc.py, mean-of-folds): efficacy 0.765→0.753 and overall 0.770→0.748, confirming the discrimination is not a phase artifact. Safety, whose failures are predominantly earlier-phase terminations (66 of 85 in Phase II), is the exception, falling from 0.784 to 0.731.

Label audit. Every terminated-trial label entering the failure set was assessed by a blind, web-verified protocol that read each trial's ClinicalTrials.gov record and classified whether the termination reflected a genuine drug-attributable safety or efficacy failure versus a non-drug stop (business/strategic, enrollment-only, COVID-19/operational, administrative, or toxicity/efficacy attributable to a combination partner). The audit was blind to MERIT's prediction. The resulting noise rate was balanced across MERIT's confidence strata (so correction neither inflates nor depends on model performance). The protocol-establishing single-pass audit reclassified 23 unambiguous non-drug terminations (12 safety, 11 efficacy) out of the failure set and re-coded one further efficacy label as safety; the cumulative audit over all label-correction passes excludes 187 trials in total (134 non-drug stops, 47 mis-scoped outcomes, 6 with no posted results; *Table S18*; the exclusion breakdown is shown in *Figure 1a*). The non-drug stops include two trials whose DSMB/subgroup-mortality and business-closure terminations rested on non-credible drug-safety attributions (thiamine and bortezomib), re-adjudicated in a later pass. All reported metrics use the corrected labels. This protocol-establishing pass was a single rater against an explicit, record-traceable criterion. To quantify its reliability, a second annotator independently re-classified a blinded, stratified sample of 68 terminated-trial labels under the same frozen rubric, blind to both the first rater's decision and MERIT output, giving Cohen's $\kappa = 0.70$ on the genuine-failure-versus-non-drug-stop decision (86.6% raw agreement; $n = 67$ after excluding one trial the second rater judged uncertain; the full six-category κ was 0.78; `scripts/benchmark/build_second_rater_kit.py`, `scripts/benchmark/score_second_rater.py`).

Bootstrap stability and feature-CI protocols. PCA stability across the 6-axis framework was assessed via 100 drug-resamples with replacement (578 PASS-cohort drugs, drug-level means, scaled features), refitting a 6-PC PCA on each resample, and matching each bootstrap PC to the canonical PC via Hungarian assignment on $|\text{Pearson } r|$ of the loadings vector (*Note S5*). Per-feature Cliff's d confidence intervals for the pathway-overlap features were computed via trial-level resample with replacement, $N = 2000$, within each outcome cohort, with the 95% CI defined as the 2.5–97.5 percentile of the resample distribution. Both were computed on the modeling cohort. **Headline-AUC confidence intervals** were obtained by a compound-clustered bootstrap that resamples drugs (by SMILES) with replacement ($N = 2000$) and recomputes the same fold-averaged AUC as the headline, weighting each fold's contribution by the resampled compound multiplicity; the unit-weight estimate reproduces the headline. This gives 95% CIs of

0.745–0.796 (overall), 0.742–0.790 (efficacy) and 0.737–0.831 (safety); all three exclude chance, and the wider safety interval reflects its 85 positives.

Repositioning analysis (indication selection and rescue nomination)

The repositioning analyses (*Figure 4*) run the canonical efficacy model in reverse: for a (compound, indication) the out-of-fold predicted failure probability (mean over 5 seeds) defines a mechanism-fit score, $\text{fit} = 1 - P(\text{fail})$; because each prediction is made with the compound held out and uses no drug-name feature, the fit never uses the drug's identity or any trial outcome. The mechanism features exclude the Open Targets known-drug and literature channels (which would encode approval), and a mechanism-only variant with the disease base rate removed still recovers 70% of known repositionings (42 of 60), so the recovery reflects target–disease biology rather than clinical precedent. The Open Targets¹³ genetic-association axis serves as a further outcome-blind cross-check. (i) *Within-drug indication selection*: for compounds with both a PASS and a FAIL_EFFICACY indication, the mechanism-fit was compared between the two indications for the same molecule (paired Wilcoxon), removing cross-drug confounds; the genetic cross-check used the per-compound Open Targets genetic-association score across that compound's indications. (ii) *Recovery of known repositionings*: cohort compounds that failed one indication for efficacy but are Approved for another in repoDB²² (matched by drug name) define a known-rescue set. Recovery is the fraction with the approved indication scored above the failed one (scripts/phase1/repositioning_model_reverse.py). For imatinib, for instance, MERIT's fit scores are 0.99 (chronic myeloid leukemia) and 0.98 (gastrointestinal stromal tumors) against 0.006 (COVID-19) and 0.67 (leishmaniasis). (iii) *Novel nominations*: for compounds that failed for efficacy with no approved indication, target genes were resolved through ChEMBL³⁵ and scored against all diseases by Open Targets genetic association; the highest-supported indications a compound was not trialed for are reported as hypotheses, with Orphanet⁴³/MONDO⁴⁴ rare-disease terms flagged. Nominations are pre-trial mechanism hypotheses, not efficacy predictions. (iv) *Time-split validation* (scripts/repositioning_timesplit.py; *Table S11*): as a prospective-flavored check, the efficacy head was retrained only on trials that began before a cutoff year (2016/2017/2018) and run in reverse on "stranded" compounds (a pre-cutoff efficacy failure and no pre-cutoff success) scored against their realized post-cutoff new indications (the indication absent from the compound's pre-cutoff trials, so held out by construction); the realized trial's own PASS/FAIL is the ground truth. Metrics are computed at the drug–disease nomination unit, with

COVID/coronavirus condition strings normalized to one indication so a single pandemic-era repurposing target is not counted as many trials, and the trial level retained as a sensitivity line; the canonical model is unchanged. A thin-training-support abstention guardrail was tested and rejected because it over-abstains, discarding rare-but-grounded successes; the COVID-19 over-crediting category (host-directed drugs; a 43-point calibration gap on the forward cohort, versus 4 points for non-COVID infectious indications; *Table S11*) is reported as a boundary rather than corrected: these drugs are over-credited because their immune targets overlap the disease's inflammatory module while its actual driver, viral replication, lies outside their mechanism.

Transcriptomic perturbation analysis (evaluated, not adopted)

To test whether drug-induced transcriptional response adds signal beyond the structure-, network- and tissue-derived features, LINCS L1000 consensus perturbation signatures were used²⁶ (up/down gene sets per compound), bridged to structures via the Drug Repurposing Hub²⁷ and matched on IK14 (446 profiled compounds). Because signature availability is itself outcome-correlated (more-studied compounds over-represented), all analyses were evaluated within the jointly profiled subset. Two formulations were tested: (i) *connectivity reversal*, the degree to which a compound's signature opposes its indication's disease signature (from public GEO collections), with a positive control on each compound's own disease signature; and (ii) *mechanistic reconstruction*, each compound's top predicted targets propagated by random walk with restart (restart 0.3/0.6; top-5/10/30 seeds) over STRING¹² (combined score ≥ 400 ; 19,488 nodes) and the directed, signed OmniPath network⁴⁵ (85,217 edges; downstream, out-degree-normalized), scored for enrichment of the compound's measured differential genes (ROC-AUC over the graph, seed genes excluded) against four nulls (mismatched compound, seed-shuffle, node-degree ranking, and generic-core versus compound-specific, the generic core being genes differential in $>25\%$ of compounds). Across every seed threshold, restart rate and both networks, matched and mismatched AUCs were indistinguishable (specificity percentile ≈ 0.50 ; $\Delta\text{AUC} \approx 0.000$), node degree alone matched or exceeded the propagated AUC, and the only reconstructable component was a generic stress/proliferation program shared across $\sim$ half of compounds. No transcriptomic feature was incorporated into MERIT.

Codes and script generation

All authors used Claude Code intermittently (began in Mar 2026) to generate python scripts for data preprocessing, data analysis and generating figure / tables from analysed outputs. The STAR pipeline was created completely without use of any AI-assisted technologies. The scripts and outputs were reviewed, curated and verified by all authors, who take full responsibility for the content of the publication.

Data availability

The clinical-trial data underlying this study are publicly available from the Aggregate Analysis of ClinicalTrials.gov (AACT; <https://aact.ctti-clinicaltrials.org>); the AACT extract used was snapshotted in June 2026. Molecular and biological annotations derive from public resources cited in Methods (PubChem, ChEMBL, Human Protein Atlas, STRING, Open Targets, DruMAP, SIDER, DILrank, repoDB and LINCS L1000). The processed trial- and arm-level modeling cohorts, the curated outcome labels with per-label provenance, the frozen per-compound molecular-mechanism feature matrix, the drug–HLA association and population-allele-frequency tables underlying *Note S3*, and the per-fold out-of-fold predictions underlying every reported metric and figure are released at <https://github.com/OmicMD/MERIT> and <https://doi.org/10.5281/zenodo.21824276>. The locked prospective-registration predictions across the 54 ongoing Phase III trials of novel drug–indication pairs, with their SHA-256 commitment are deposited at <https://doi.org/10.5281/zenodo.21824277> and their generating codes are deposited at <https://github.com/OmicMD/MERIT>.

Code availability

All analysis code is available at <https://github.com/OmicMD/MERIT>, comprising the data-provenance notebook (every transformation from raw AACT records to model predictions, including STRING network queries, Open Targets disease-target retrieval and DruMAP pharmacokinetic prediction) and the model-training, signal-decomposition, public-feature-decomposition, temporal leave-future-out, single-classifier safety-head, external safety-axis validation (hERG/DILrank/SIDER), the published-model, network-proximity and Phase II→III transition (*Note S4*) benchmarking scripts, repositioning (MERIT-in-reverse generator) and

transcriptomic perturbation-analysis scripts, the confident-error residual-probe scripts (target→adverse-event-class priors, reactive-metabolite structural alerts, finer-indication and within-trial design context; *Note S3*) and the prospective HLA×haptentation enrichment tool, and the figure-generation scripts. One component is proprietary: the structure-to-target binding predictions are produced by the STAR framework, which is not released. To keep the study reproducible despite this, we release the pipeline's **outputs** [~3TB of data] upon request, the frozen per-compound predicted binding profiles, as static artifacts, so that all downstream feature construction, model training, evaluation, and every reported result are fully reproducible from the released data without access to the proprietary core; only generating binding profiles for new compounds requires the proprietary pipeline, which is available from the authors for non-commercial academic use on reasonable request. The efficacy and repositioning results in particular do not depend on the proprietary layer: a public-only feature set (structured target–disease mechanism plus STRING/Open Targets pathway overlap) recovers most of the efficacy increment (*Methods*), and the structured mechanism features are released as a frozen per-compound artifact.

Ethics declarations

This study used only publicly available, aggregate and de-identified clinical-trial records and public molecular and biological databases. No human participants, identifiable individual patient data, or animal subjects were involved; no ethical approval or informed consent was required.

Competing interests

GR is the founder of Omic Inc. All other authors are employees of Omic Inc. and declare no other competing interests.

Funding

This study was funded by Omic Inc.

Author contributions

GR conceived and designed the study. CKT designed the STAR pipeline. IM and BAS implemented and optimised the pipeline. SM managed the data and analysis infrastructure. All

authors were involved in the analyses of data. CKT and GR wrote the manuscript draft, and all authors contributed to the manuscript revisions.

Acknowledgments

Not applicable.