NeoRx: Causal Inference for Automated Drug Target Identification via Pearl's Do-Calculus over Multi-Source Biomedical Knowledge Graphs

Kelyn Paul Njeri
NeoForge Labs (Independent Research)
kelyn@neoforgelabs.tech ORCID: 0009-0000-1068-4512

Abstract

Contemporary drug-discovery pipelines rank putative therapeutic targets by gene–disease association strength—a fundamentally correlational metric that conflates genuine causal drivers with downstream bystanders. We present **NeoRx**, an end-to-end computational platform that applies Pearl's causal inference framework to multi-source biomedical knowledge graphs in order to distinguish *causal* drug targets from *correlational* ones. NeoRx assembles a unified causal knowledge graph from eight heterogeneous databases (Monarch Initiative, Open Targets, KEGG, Reactome, STRING, UniProt, RCSB PDB, and ChEMBL), then applies the backdoor criterion, causal effect estimation, and leave-one-source-out sensitivity analysis to every candidate gene. A composite causal confidence score integrates effect magnitude (30

Keywords: causal inference, drug discovery, do-calculus, knowledge graphs, target identification, Pearl's framework

1 Introduction

1.1 The Correlation Problem in Drug Discovery

Drug target identification is the foundational step in pharmaceutical development. Current computational approaches rely overwhelmingly on *correlational* evidence: genome-wide association studies (GWAS) identify statistical associations between genetic variants and disease phenotypes; transcriptomic analyses flag differentially expressed genes; and network-based methods prioritise highly connected "hub" genes. These approaches share a critical limitation—they cannot distinguish between genes that *cause* a disease and genes that are *caused by* it.

The consequences are severe. Consider HIV: tumour necrosis factor alpha (TNF-α) is strongly associated with HIV progression—plasma TNF-α levels are elevated during active infection, and numerous GWAS studies confirm the association. A correlation-based pipeline would rank TNF-α highly as a therapeutic target. Yet TNF-α elevation is a *consequence* of immune activation, not a cause of infection. Inhibiting TNF-α in HIV patients does not treat the infection—TNF-α elevation is an inflammatory consequence, not a causal driver, of viral pathogenesis [1]. In contrast, CCR5 is causally upstream of HIV-1 infection: it serves as the viral co-receptor, and the loss-of-function allele CCR5-Δ32 confers near-complete resistance to HIV-1 infection [2]. Maraviroc, a CCR5 antagonist, is an approved antiretroviral [3].

This example illustrates a general principle that pervades drug discovery: **correlation is not causation**, and acting on correlational targets wastes resources and can harm patients.

1.2 Pearl's Causal Inference Framework

Judea Pearl's structural causal model (SCM) framework [4] provides the mathematical tools to move beyond correlation. The key constructs are:

- **Directed Acyclic Graphs (DAGs)** encoding causal assumptions about variable relationships
- **The do-operator** $P(Y \mid do(X = x))$, which represents the distribution of outcome Y when treatment X is *intervened upon* (set to x), as opposed to merely *observed* at value x
- **The backdoor criterion**, which identifies valid adjustment sets that block all confounding paths between treatment and outcome, enabling non-parametric identification of causal effects
- **Sensitivity analysis** and refutation tests that assess the robustness of causal estimates to unmeasured confounders

These tools have transformed causal reasoning in epidemiology, economics, and social science [5], but have seen limited adoption in drug target identification. We address this gap.

1.3 Contributions

We present NeoRx, which makes the following contributions:

1. **A causal inference pipeline for drug target identification** that applies the backdoor criterion, causal effect estimation, and sensitivity analysis to biomedical knowledge graphs—the first system, to our knowledge, to apply Pearl's graphical do-calculus framework—as opposed to instrumental-variable methods such as Mendelian randomisation [9]—to automated drug target identification at scale.
2. **A multi-source knowledge graph assembly framework** that integrates eight heterogeneous databases—including ChEMBL for validated drug–target relationships and pathogen-encoded targets—with automatic node merging, edge provenance tracking, and source corroboration scoring.
3. **A pathogen target evaluation pathway** that identifies non-human drug targets (e.g., *Plasmodium* DHFR-TS, HIV Pol) from ChEMBL and scores them by clinical phase, drug diversity, and organism–disease relevance, enabling the system to find targets invisible to human-only knowledge graphs.
4. **A biological intelligence layer** comprising disease-type-aware target classification and tissue-expression boolean gating that prevents biologically implausible targets from advancing.
5. **A disease specificity score** that replaces network centrality, penalising promiscuous hub genes (TP53, AKT1) while rewarding disease-specific targets.
6. **A composite scoring function** that weights causal confidence above binding affinity—inverting the priority of conventional virtual screening.
7. **A validation framework** benchmarked against clinically approved drugs across seven diseases, demonstrating a mean $F_1 = 0.474$.
8. **Open-source release** of the complete platform.

2 Related Work

2.1 Network-Based Target Identification

Several platforms rank targets using network topology. DisGeNET [6] aggregates gene–disease associations from curated databases and GWAS catalogs. Open Targets [7] computes association scores across genetic, functional, and literature evidence. These provide valuable data sources

but rank by *association strength*, not causal evidence. Network propagation methods (e.g., Random Walk with Restart [8]) prioritise topologically central genes, which correlates with but does not guarantee causal involvement.

2.2 Causal Inference in Biomedicine

Mendelian randomisation (MR) uses genetic variants as instrumental variables to estimate causal effects of exposures on outcomes [9]. While principled, MR requires large-scale genotype–phenotype datasets and is limited to exposures with valid instruments. DoWhy [10] provides a general-purpose causal inference library, but its application to drug target identification has been limited to individual case studies. CausalNex [11] enables Bayesian network structure learning but does not integrate multi-source biomedical knowledge.

More recently, Zheng et al. [15] applied Mendelian randomisation across the druggable genome to systematically identify causal therapeutic targets, and the EpiGraphDB platform [16] integrates MR results with pathway and literature evidence. These instrumental-variable approaches require large-scale GWAS summary statistics and are limited to exposures for which valid genetic instruments exist. NeoRx differs in three respects: (i) it operates on curated biomedical knowledge graphs using Pearl's structural causal model rather than genetic instruments; (ii) it requires only a disease name as input, not pre-computed GWAS data; and (iii) it integrates target identification with downstream molecule generation and virtual screening in a single pipeline.

2.3 Virtual Screening Pipelines

Conventional virtual screening pipelines (AutoDock Vina [12], Glide [13]) optimise binding affinity to a given target—they do not question whether the target itself is valid. This decoupling of target validation from compound screening is a fundamental design flaw that NeoRx addresses.

3 Methods

3.1 System Architecture

NeoRx operates as a six-stage pipeline:

1. **Graph Construction** — assemble a causal knowledge graph from eight databases
2. **Causal Target Identification** — apply do-calculus to classify targets
3. **Molecule Generation** — generate novel candidates via VAE
4. **Drug-Likeness Screening** — filter through Lipinski, PAINS, QED
5. **Molecular Docking** — estimate binding affinity via AutoDock Vina
6. **Composite Scoring and Reporting** — rank candidates with causal-weighted scores

The pipeline requires only a disease name as input and produces a ranked list of drug candidates with full causal reasoning for each target.

3.2 Multi-Source Knowledge Graph Assembly

3.2.1 Data Sources

We query eight databases for each disease:
ccc

Source	Data Type	API
Monarch Initiative	Gene–disease associations	Monarch v3 REST

Source	Data Type	API
Open Targets	Gene–disease associations, disease specificity	GraphQL
KEGG	Pathway memberships	REST
Reactome	Pathway memberships	REST
STRING	Protein–protein interactions	REST
UniProt	Protein metadata, druggability	REST
RCSB PDB	3D structures	RCSB Search API
ChEMBL	Validated drug–target relationships	Local SQLite (v36)

Gene–disease associations from Monarch and Open Targets are queried in parallel via a thread pool executor. ChEMBL queries a local SQLite copy of ChEMBL v36 (28 GB), traversing the chain `drug_indication` → `drug_mechanism` → `target_dictionary` → `component_sequences` to identify both human and pathogen targets with validated drug evidence. Per-gene enrichment (KEGG, Reactome, STRING, UniProt, PDB) is then performed for up to $G_{\max} = 20$ genes.

3.2.2 Graph Schema

The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ uses a typed node and edge schema:

Node types $\mathcal{V}_t \in$ {gene, protein, pathway, phenotype, disease, drug, metabolite, pathogen_gene}

The `pathogen_gene` node type represents non-human drug targets from ChEMBL (e.g., *Plasmodium falciparum* DHFR-TS, HIV-1 Pol polyprotein). These nodes use the ID format `pathogen:{organism}:{symbol}` and bypass biological classification and tissue filtering, which only apply to human genes.

Edge types $\mathcal{E}_t \in$ {activates, inhibits, phosphorylates, binds, regulates, participates_in, associated_with, causes, treats, interacts_with, upregulates, downregulates}

Each node carries a confidence score $s \in [0, 1]$, source provenance string, UniProt accession, PDB identifiers, and a metadata dictionary. Each edge carries a weight $w \in [0, 1]$, source database identifier, and evidence text.

3.2.3 Node Merging

Different databases may refer to the same gene using different identifiers. We normalise to gene symbols and merge duplicates by node ID:

- **Score**: $s_{\text{merged}} = \max(s_1, s_2)$
- **UniProt ID**: keep the first non-empty value
- **PDB IDs**: $\text{PDB}_{\text{merged}} = \text{PDB}_1 \cup \text{PDB}_2$
- **Metadata**: dictionary union (later entries overwrite)
- **Source provenance**: concatenated string (e.g., "Monarch, OpenTargets")

Edges are deduplicated by the triple (source_id, target_id, edge_type).

3.2.4 Disease Node

A disease node is automatically created and connected to all gene/protein nodes via `associated_with` edges unless such edges already exist. This ensures the graph has a clear outcome variable for causal analysis.

3.3 Causal Target Identification

For each candidate gene X_i and disease outcome Y, we evaluate the causal relationship through a six-step procedure.

3.3.1 Step 0: Biological Pre-Classification

Before statistical analysis, each gene is classified by a disease-type-aware biological classifier (Section 3.4) and filtered for tissue expression relevance (Section 3.5). Genes classified as `HOST_SYMPTOM` or lacking tissue expression are automatically demoted regardless of subsequent statistical evidence.

3.3.2 Step 1: Graph-Based Causal Analysis

We first seek directed causal pathways from X_i to Y in $\mathcal{G}$:

$$\pi(X_i, Y) = \text{shortest\_directed\_path}(\mathcal{G}, X_i, Y)$$

If no directed path exists, we fall back to undirected paths (through PPI edges), which provide weaker causal evidence.

We then compute the **backdoor adjustment set Z** following Pearl's criterion [4]. A set $\mathbf{Z}$ satisfies the backdoor criterion relative to (X_i, Y) if:

1. No node in $\mathbf{Z}$ is a descendant of X_i
2. $\mathbf{Z}$ d-separates X_i from Y in the mutilated graph $\mathcal{G}_{\overline{X_i}}$ (with arrows into X_i removed)

We verify d-separation using NetworkX's `d_separated()` function and cap the adjustment set at 10 variables for computational tractability.

Identifiability: A target is *identifiable* if $|\pi(X_i, Y)| > 0$ — i.e., there exists at least one path from treatment to outcome.

3.3.3 Step 2: Causal Effect Estimation

We estimate the causal effect as a composite of four evidence components:

$$\hat{\tau}(X_i, Y) = s_{X_i} \cdot \psi(\pi) \cdot \delta(\mathbf{Z}) \cdot (1 + c_{X_i}) \cdot \phi(\mathbf{D}) \cdot \mathbb{1}[\text{direct}]$$

where:

- s_{X_i} is the gene's association score from the knowledge graph
- $\psi(\pi) = \frac{1}{|\pi|} \prod_k w_{e_k}$ is the **path strength** — the product of edge weights along the shortest directed path, discounted by path length. For undirected paths, ψ is halved.
- $\delta(\mathbf{Z})$ is the **d-separation factor**: 1.2 if the adjustment set blocks confounding paths (verified by `d_separated()`), 0.8 if no adjustment set exists.
- c_{X_i} is the **betweenness centrality** of X_i in $\mathcal{G}$
- $\phi(\mathbf{D}) = \min(1.5, 1 + 0.1 \cdot |\mathbf{D}|)$ is the **multi-source corroboration factor**, where $\mathbf{D}$ is the set of distinct databases confirming edges involving X_i
- $\mathbb{1}[\text{direct}] = 1.5$ if any outgoing edge from X_i has type `causes`, otherwise 1.0

A p-value proxy is computed as:

$$p = \max(0.001,\ 0.5 \cdot (1 - \min(1, |\hat{\tau}|)))$$

3.3.4 Step 3: Sensitivity Analysis

We perform three robustness tests and average their scores:

Test 1: Leave-One-Source-Out Stability. For each database d_k contributing edges to X_i, we remove all edges from d_k and recompute the path strength. The stability score is:

$$R_1 = \max(0, \min(1, 1 - \text{CV}(\{\hat{\tau}_{-d_k}\}_k)))$$

where CV is the coefficient of variation of the leave-one-out estimates. For targets supported by a single source, $R_1 = 0.3$.

Test 2: Directed Path Existence. If a directed path exists:

$$R_2 = \min\left(1, \frac{1}{|\pi_{\text{directed}}|}\right)$$

Undirected-only connections yield $R_2 = 0.3$; no connection yields $R_2 = 0$.

Test 3: Multi-Source Connectivity.

$$R_3 = \min\left(1, \frac{\deg^+(X_i) + \deg^-(X_i)}{10}\right)$$

The overall robustness score is:

$$R = \frac{R_1 + R_2 + R_3}{3}$$

3.3.5 Step 4: Topological Evidence

We collect four additional evidence signals:

- **Pathway connections** n_p: number of edges with type `participates_in`
- **Protein interactions** n_{int}: number of edges with type `interacts_with`
- **Source-level scores** $\{s_d\}_{d \in \mathbf{D}}$: per-database association scores
- **Druggability score** $\delta_{\text{drug}} \in [0, 1]$: a heuristic combining structural data availability (+0.2 for PDB structures), protein characterisation (+0.1 for UniProt), Open Targets tractability data (+0.15), UniProt druggability flag (+0.15), and protein family keywords (+0.15), from a base of 0.3

3.3.6 Step 5: Composite Causal Confidence

The causal confidence score $C(X_i)$ is a weighted combination:

$$C(X_i) = 0.30 \cdot \bar{\tau} + 0.25 \cdot R + 0.15 \cdot \mathbb{1}[\text{identifiable}] + 0.10 \cdot \kappa + 0.10 \cdot \sigma + 0.10 \cdot \delta_{\text{drug}}$$

ccc

Component	Weight	Description
$\bar{\tau} = \min(1, \lvert\hat{\tau}\rvert)$	0.30	Normalised causal effect magnitude
R	0.25	Robustness (sensitivity analysis)
$\mathbb{1}[\text{identifiable}]$	0.15	Backdoor identifiability
$\kappa = \min\left(1, \frac{\lvert\mathbf{D}\rvert}{N_{\text{active}}} \cdot \bar{s}\right)$	0.10	Multi-source consensus
$\sigma = \frac{1}{\log_2(n_{\text{diseases}}+2)}$	0.10	Disease specificity
δ_{drug}	0.10	Druggability

where $\bar{s}$ is the mean source score, N_{active} is the number of databases that contributed gene-level data, and n_{diseases} is the number of diseases associated with the gene in Open Targets.

Disease specificity replaces the former network centrality component. Hub genes (TP53, AKT1, EGFR) are associated with thousands of diseases—they are generic, not specific. The specificity formula $\sigma = 1/\log_2(n_{\text{diseases}} + 2)$ penalises promiscuous genes: TP53 (~3000 diseases) receives $\sigma \approx 0.087$, while NPC1 (~3 diseases) receives $\sigma \approx 0.431$. This directly encodes the insight that a gene associated with everything is not a useful drug target for any specific disease.

3.3.7 Step 5b: Bootstrap Confidence Intervals

We compute 95% bootstrap confidence intervals via $B = 200$ bootstrap resamples. In each resample, we perturb:

- Effect: $\hat{\tau}^* \sim \hat{\tau} + \mathcal{N}(0, 0.05)$
- Robustness: $R^* \sim R + \mathcal{N}(0, 0.05)$
- Druggability: $\delta^* \sim \delta + \mathcal{N}(0, 0.03)$
- Source scores: with probability 0.3, drop one source
- Pathway/interaction counts: $n^* = n + U\{-1, 0, 1\}$

The CI is $[q_{2.5\%}, q_{97.5\%}]$ of the bootstrap distribution $\{C^{*(b)}\}_{b=1}^{B}$.

3.3.8 Step 6: Target Classification

Each target is classified by the following rules:

Rule	Classification
`target_type = HOST_SYMPTOM`	**CORRELATIONAL** (biological override)
`tissue_relevant = False`	**CORRELATIONAL** (tissue override)
$C \geq 0.6$ AND $R \geq 0.4$ AND identifiable AND $E \geq 2$	**CAUSAL**
$C < 0.4$ OR $R < 0.3$	**CORRELATIONAL**
$C \geq 0.6$ AND $E < 2$	**INCONCLUSIVE**
Otherwise	**INCONCLUSIVE**

where E is the number of independent evidence streams (Section 3.3.9).

3.3.9 Evidence Stream Counting

We count six categories of independent evidence:

1. **Gene–disease association** — at least one of Monarch or Open Targets
2. **Pathway membership** — $n_p > 0$ (KEGG or Reactome)
3. **Protein interactions** — $n_{\text{int}} > 0$ (STRING)
4. **Structural data** — at least one PDB structure
5. **Functional annotation** — UniProt GO terms or druggability flag
6. **Drug evidence** — ChEMBL validated drug–target relationship

Each category counts at most once. CAUSAL classification requires $E \geq 2$.

3.3.10 Pathogen Target Evaluation

Targets with node type `pathogen_gene` bypass the standard human-gene evaluation pipeline (Steps 0–6) and enter a dedicated pathogen target evaluation path. These targets come from ChEMBL and represent validated drug targets in pathogen organisms (e.g., PfDHFR-TS, HIV-1 Pol). Their confidence is based on drug evidence rather than causal graph analysis:

$$C_{\text{path}}(X_i) = 0.40 \cdot d_{\text{score}} + 0.25 \cdot \delta_{\text{drug}} + 0.15 \cdot \mathbb{1}[\text{identifiable}] + 0.10 \cdot r_{\text{org}} + 0.10 \cdot \sigma$$

where d_{score} is the drug evidence score (60% max clinical phase + 20% drug diversity + 20% mechanism diversity), and r_{org} is the organism–disease relevance score:

- $r_{\text{org}} = 1.0$ if the pathogen organism matches the disease's primary pathogen (e.g., *P. falciparum* for malaria)
- $r_{\text{org}} = 0.3$ for off-target pathogens in infectious diseases (e.g., bacterial targets in an HIV query)
- $r_{\text{org}} = 0.0$ for non-infectious diseases (cancer, Alzheimer's, T2D)—these diseases have no causative pathogen, so any pathogen targets in ChEMBL are from co-prescribed medications and should not be ranked

Robustness is set by clinical phase: Phase 4 $\rightarrow R = 1.0$, Phase 3 $\rightarrow 0.9$, Phase 2 $\rightarrow 0.7$, Phase 1 $\rightarrow 0.5$, preclinical $\rightarrow 0.3$. When $r_{\text{org}} < 0.5$, robustness is penalised by $R \leftarrow R \times 0.3$, ensuring off-target pathogen proteins are demoted to INCONCLUSIVE.

Gene symbol quality gates. ChEMBL's `component_synonyms` table contains gene symbols of variable quality. We apply three quality filters before accepting a pathogen gene symbol: (1) a blocklist of 20 common English words used in enzyme nomenclature ("reverse", "transcriptase", "protease", etc.); (2) a minimum length filter ($|\text{symbol}| \geq 2$); and (3) a no-space filter for UNIPROT fallback synonyms. Targets with no usable gene symbol are skipped entirely.

3.4 Disease-Type-Aware Target Classification

A dedicated classifier prevents biologically implausible targets from advancing. It operates on a taxonomy of nine disease types: `INFECTIOUS_VIRAL`, `INFECTIOUS_PARASITIC`, `INFECTIOUS_BACTERIAL`, `AUTOIMMUNE`, `CANCER`, `METABOLIC`, `NEURODEGENERATIVE`, `GENETIC`, and `OTHER`.

The classifier applies a six-step pipeline (described in full in our companion paper [Paper 3]):

1. **Symptom marker blocklist** — 119+ genes across 9 protein families (GABA receptors, sodium channels, potassium channels, glutamate receptors, dopamine receptors, serotonin receptors, coagulation factors, acute phase proteins, neuronal structural proteins) are blocked for infectious diseases
2. **Inflammatory cytokine check** — TNF, IL6, IL1B, CXCL8, IL10 are classified as `HOST_SYMPTOM` in infectious contexts, as cytokine elevation reflects immune activation, not disease mechanism [17]
3. **Immune receptor check** — TLR, NOD, HLA, MHC, cytokine, and chemokine genes are classified as `HOST_IMMUNE` for infections, `HOST_INVASION` for autoimmune diseases
4. **Known invasion target database** — 14 curated host–pathogen interaction targets (e.g., CCR5 for HIV, ACE2 for COVID-19, glycophorins for malaria)
5. **Evidence quality gate** — Targets with no source, no metadata, and zero score are `CORRELATIONAL`
6. **Default** — Cancer targets default to `HOST_INVASION`; others with evidence default to `HOST_INVASION`

Note: The cytokine check (step 2) is ordered *before* the immune receptor check (step 3), ensuring that inflammatory cytokines such as TNF and IL6 are correctly classified as `HOST_SYMPTOM` rather than `HOST_IMMUNE` for infectious diseases. This prevents the critical error of promoting inflammatory bystanders as drug targets in malaria and Ebola.

3.5 Tissue Expression Filtering

The tissue filter checks whether each candidate gene is expressed in disease-relevant tissue using the Human Protein Atlas [14]. It operates as a **boolean gate**:

- **Pass**: The gene is expressed in at least one disease-relevant tissue (or expression data is unavailable — "unknown = pass")

- **Fail**: The gene is expressed only in tissues unrelated to the disease → automatic demotion to CORRELATIONAL

For each of 12 supported diseases, we define relevant tissue sets (e.g., malaria → blood, liver, spleen, bone marrow; Alzheimer's → brain, cerebral cortex, hippocampus). A curated tissue ontology mapping (`_TISSUE_SYNONYMS`) normalises ~40 HPA tissue names to canonical organ-level forms (e.g., "cerebral cortex" → "brain", "lymphoid tissue" → "lymph node"). Tissue comparison happens at the canonical level.

Design decision: The tissue gate never modifies `causal_confidence`. Confidence remains a pure measure of causal evidence quality. The gate provides an independent biological criterion: if a gene is only expressed in irrelevant tissues, it cannot be a valid drug target for this disease, regardless of how strong the statistical evidence appears.

3.6 Composite Drug Candidate Scoring

Once causal targets are identified, NeoRx generates candidate molecules (via VAE; see companion paper [Paper 4]) and scores them across six dimensions:

$$S = \sum_{i=1}^{6} w_i \cdot \hat{x}_i$$

ccc

Dimension	Weight w_i	Normalisation
Causal confidence	**0.30**	Clamped to $[0, 1]$
Binding affinity	0.25	$\frac{a-0}{-12-0}$, where $a \in [-12, 0]$ kcal/mol
QED (drug-likeness)	0.15	Already $[0, 1]$
Synthetic accessibility	0.10	$\frac{10-\text{SA}}{10-1}$, where SA $\in [1, 10]$
ADMET	0.10	Already $[0, 1]$
Novelty	0.10	Already $[0, 1]$

The critical design decision: causal confidence receives the highest weight (0.30), exceeding even binding affinity (0.25). A moderately-binding molecule aimed at a validated causal target is ranked above a tightly-binding molecule aimed at a correlational bystander. This inverts the priority of conventional virtual screening.

Weights are configurable via a JSON environment variable `NEORX_WEIGHTS` or function argument, and are automatically renormalised to sum to 1.0.

3.7 Validation Framework

We validate NeoRx against known clinically approved drug targets across seven diseases:
ccc

Disease	Known Targets	Known False Positives
Malaria	DHFR, DHPS, GYPA, GYPB, CR1, BSG, DARC	GABRD, GABRA1, SCN2A, TNF, IL6
HIV	CCR5, CXCR4, CD4, POL	TNF, IL6, CRP
Type 2 Diabetes	GLP1R, SLC5A2, DPP4, PPARG, INSR, INS	CRP, TNF
Lung Cancer	EGFR, ALK, KRAS, PIK3CA, ERBB2, TP53, BRAF	—
Breast Cancer	ERBB2, ESR1, PIK3CA, BRCA1, BRCA2, CDK4	—

Disease	Known Targets	Known False Positives
Alzheimer's	BACE1, APP, PSEN1, MAPT, ACHE	—
Ebola	NPC1, GP	TNF, IL6

We report precision, recall, and F1 score, with quality grades: A ($F_1 > 0.7$), B ($F_1 > 0.4$), C ($F_1 > 0.2$), F ($F_1 \leq 0.2$).

4 Results

4.1 Case Study: HIV

The ChEMBL pathogen pipeline identifies three HIV-encoded targets—POL ($C = 0.990$, robustness = 1.000), ENV ($C = 0.955$, robustness = 1.000), and GAG ($C = 0.887$, robustness = 0.900)—all classified as CAUSAL with `PATHOGEN_DIRECT` type. POL encodes the Pol polyprotein (protease/RT/integrase), the target of 26+ FDA-approved antiretrovirals including efavirenz, dolutegravir, and darunavir. ENV encodes the envelope glycoprotein gp160/gp120, the target of ibalizumab. GAG encodes the capsid and matrix proteins targeted by lenacapavir.

Among human targets, CCR5 ($C = 0.706$, robustness = 0.805, 5 evidence streams) ranks #11 overall, correctly classified as `HOST_INVASION`—it encodes the viral co-receptor required for HIV-1 cell entry and is the target of maraviroc. CD4 ($C = 0.625$, robustness = 0.863, 5 evidence streams) is classified as `HOST_IMMUNE`, ranking #14.

Bacterial ribosomal targets (RPOA, RPSA) from co-prescribed antibiotics appear in the ChEMBL results but are correctly demoted to INCONCLUSIVE (robustness = 0.300, below the 0.4 threshold) because their organism does not match HIV.

Overall validation: $P = 0.429$, $R = 0.750$, $F_1 = 0.545$ (Grade B). CXCR4, the alternative co-receptor, falls outside the top 20 candidates.

> **"POL has a causal confidence of 0.990, robustness 1.000. Target type PATHOGEN_DIRECT: this is a validated drug target with Phase 4 approved drugs (efavirenz, dolutegravir, darunavir). Pol polyprotein encodes protease, reverse transcriptase, and integrase—the targets of 26+ FDA-approved antiretrovirals."**

4.2 Case Study: Malaria

Malaria demonstrates the transformative impact of ChEMBL integration. The pathogen pipeline identifies DHFR-TS ($C = 0.896$, robustness = 1.000, `PATHOGEN_DIRECT`) as the top target—the *Plasmodium falciparum* bifunctional dihydrofolate reductase-thymidylate synthase, targeted by pyrimethamine. PPPK-DHPS ($C = 0.825$, robustness = 0.700) is the second true positive—the dihydropteroate synthase targeted by sulfadoxine. Additional pathogen targets include MT-CYB (cytochrome b, the target of atovaquone) and DXR (the target of fosmidomycin).

The biological intelligence layer correctly demotes SCN10A ($C = 0.552$) as `HOST_SYMPTOM`—a neuronal sodium channel associated with seizure symptoms rather than parasite biology—while the correlation-only baseline ranks SCN10A and SCN2A among its top 10.

Bacterial off-target genes (FOLP, RPSA, DACB) from co-prescribed antibiotics appear in ChEMBL results but are correctly demoted to INCONCLUSIVE (robustness ≤ 0.300) via the organism–disease relevance penalty.

Overall: $P = 0.400$, $R = 0.286$, $F_1 = 0.333$ (Grade C). The recall limitation reflects that 5 of 7 ground truth targets (BSG, CR1, DARC, GYPA, GYPB) are host RBC invasion receptors that fall outside the top 20 candidates from Open Targets—a candidate pool ceiling (see Section 5.2).

"DHFR-TS has a causal confidence of 0.896, robustness 1.000. Target type PATHOGEN_DIRECT: Plasmodium falciparum bifunctional dihydrofolate reductase-thymidylate synthase, validated by pyrimethamine (Phase 4)."

4.3 Case Study: Alzheimer's Disease

APP ($C = 0.723$, robustness = 0.888, 6 evidence streams) and PSEN1 ($C = 0.711$, robustness = 0.840, 5 evidence streams) are the top-ranked human targets, both correctly classified as `HOST_INVASION` CAUSAL targets. APP encodes the amyloid precursor protein, the substrate for amyloid-β production, while PSEN1 encodes presenilin-1, the catalytic subunit of γ-secretase. APOE ($C = 0.626$) and PSEN2 ($C = 0.668$) also achieve CAUSAL status.

The organism–disease relevance mechanism correctly handles Alzheimer's as a non-infectious disease: bacterial ribosomal targets RPOA and RPSA from co-prescribed antibiotics appear in ChEMBL results but receive $r_{\text{org}} = 0.0$, resulting in robustness = 0.210 and classification as INCONCLUSIVE. Without this mechanism, these irrelevant bacterial targets would rank #1–2.

BACE1 (β-secretase, the target of verubecestat development) and MAPT (tau protein) are not captured in the top 20, representing a recall limitation. ACHE (acetylcholinesterase, the target of donepezil) ranks #9 but at $C = 0.592$, just below the 0.6 CAUSAL threshold. Overall: $P = 0.333$, $R = 0.400$, $F_1 = 0.364$ (Grade C).

4.4 Cross-Disease Validation

Table 2. Cross-disease validation against clinically approved drug targets (top $N = 20$ candidates per disease).
ccccccccc

Disease		Fou
HIV	20	
Malaria	20	
Type 2 Diabetes	20	
Alzheimer's	20	
Lung Cancer	20	
Breast Cancer	20	
Ebola	20	
Mean	—	

TP = true positives (known approved targets recovered as CAUSAL); FP = known false positives that passed as CAUSAL; Missed = known targets not recovered in top 20. Grades: A ($F_1 > 0.7$), B ($F_1 > 0.4$), C ($F_1 > 0.2$), F ($F_1 \leq 0.2$).

Performance by disease category. NeoRx achieves Grade A on lung cancer (100% recall of all 7 validated targets), Grade B on HIV, type 2 diabetes, and breast cancer, and Grade C on malaria, Alzheimer's, and Ebola. The addition of ChEMBL as an eighth data source with a dedicated pathogen target pipeline transforms infectious disease performance: malaria improves from Grade F ($F_1 = 0.000$) to Grade C ($F_1 = 0.333$) with DHFR-TS and PPPK-DHPS identified as CAUSAL; Ebola improves from Grade F ($F_1 = 0.000$) to Grade C ($F_1 = 0.400$) with the GP envelope glycoprotein identified; and HIV improves from Grade C to Grade B. Zero known false positives are promoted to CAUSAL across all 7 diseases.

Table 3. NeoRx vs. correlation-only baseline. The correlation-only baseline ranks targets by raw association score without causal analysis, classification, or tissue filtering.
ccccccc

Disease	Method		P	
HIV	Corr-only	0.100		0.500
	NeoRx	**0.429**		**0.750**
Malaria	Corr-only	0.050		0.143
	NeoRx	**0.400**		**0.286**
Type 2 Diabetes	Corr-only	0.250		0.833
	NeoRx	**0.417**		**0.833**
Alzheimer's	Corr-only	0.150		0.600
	NeoRx	**0.333**		**0.400**
Lung Cancer	Corr-only	0.350		1.000
	NeoRx	**0.538**		**1.000**
Breast Cancer	Corr-only	0.300		1.000
	NeoRx	0.308		0.667
Ebola	Corr-only	0.000		0.000
	NeoRx	**0.333**		**0.500**

NeoRx outperforms the correlation-only baseline in 6 of 7 diseases. The improvements are most dramatic for infectious diseases: HIV F_1 improves from 0.167 to 0.545 (+227%), malaria from 0.074 to 0.333 (+350%), and Ebola from 0.000 to 0.400 (∞%). For lung cancer, NeoRx also outperforms correlation-only ($F_1 = 0.700$ vs. 0.519, +35%), recovering all 7 validated targets with higher precision. Breast cancer is the only disease where the correlation-only F_1 (0.462) exceeds NeoRx's (0.421), because BRCA1 ($C = 0.599$) and BRCA2 ($C = 0.587$) fall marginally below the 0.6 CAUSAL threshold—these are loss-of-function tumor suppressors where the actual drug target (PARP1, which IS found as CAUSAL) operates via synthetic lethality.

4.5 Ablation Studies

We ablate the biological intelligence layer by removing (a) the target classifier, (b) the tissue expression filter, and (c) both layers simultaneously.

Table 4. Ablation results (F_1 scores across diseases).
ccccc

Disease	Full NeoRx
HIV	**0.545**
Malaria	**0.333**
Type 2 Diabetes	**0.556**
Alzheimer's	**0.364**
Lung Cancer	**0.700**
Breast Cancer	0.421
Ebola	**0.400**

Key observations:

1. **The causal inference engine drives the primary improvement** over correlation-only (mean F_1 improvement of +80% across 6 of 7 diseases). The ChEMBL pathogen pipeline and disease specificity score account for most of the gains, especially on infectious diseases.
2. **The classifier adds value for malaria.** Removing the classifier reduces malaria F_1 from 0.333 to 0.267 because symptom markers (SCN10A) re-enter the CAUSAL set. The classifier's impact is concentrated on infectious diseases where symptom markers are most problematic.

3. **The tissue filter contributes to breast cancer precision.** The boolean tissue gate correctly demotes tissue-irrelevant genes without the over-sensitivity seen in the earlier modifier-based approach.
4. **Most diseases are robust to ablation** because the core confidence formula (with disease specificity replacing centrality) and the ChEMBL pathogen pipeline provide the dominant signal. The biological intelligence layer acts as a safety net against specific failure modes rather than driving scores broadly.

4.6 Computational Cost

Table 5. Wall-clock execution times on a single-threaded consumer workstation (Apple M-series, 16 GB RAM).
ccc

Phase	Uncached	Cached
Knowledge graph construction	129–180 s per disease	$<$ 1 s (API response caching)
Causal target identification	44–335 s per disease	44–335 s (no caching)
Total (7-disease benchmark)	~25 min	~25 min

Graph construction is dominated by REST API calls to eight external databases (Monarch, Open Targets, KEGG, Reactome, STRING, UniProt, PDB, ChEMBL) and is fully parallelised across data sources via a thread pool executor. ChEMBL queries run against a local SQLite database and complete in $<$ 1 second. The causal identification phase—backdoor analysis, sensitivity testing, bootstrap confidence intervals (200 resamples), and HPA tissue expression queries—runs in 44–335 seconds per disease depending on graph density. Ebola is fastest (44 s, sparse graph) while breast cancer is slowest (335 s, dense graph with many targets requiring HPA lookup). A full 7-disease benchmark completes in under 25 minutes from cold start.

5 Discussion

5.1 The Causation Gap

The cross-disease benchmark reveals the transformative impact of integrating validated drug–target evidence (ChEMBL) alongside causal graph analysis.

Where the system excels. Lung cancer achieves Grade A ($F_1 = 0.700$) with 100% recall—all 7 validated targets (EGFR, ALK, KRAS, PIK3CA, ERBB2, TP53, BRAF) are correctly identified as CAUSAL. For type 2 diabetes ($F_1 = 0.556$, Grade B), 5 of 6 ground truth targets are found, including DPP4 (sitagliptin), GLP1R (semaglutide), and SLC5A2 (dapagliflozin). These diseases benefit from strong multi-source evidence in human knowledge graphs and well-characterised host-intrinsic biology.

Where ChEMBL transforms performance. The addition of ChEMBL as an eighth data source with a dedicated pathogen target pipeline produces the largest gains on infectious diseases:

- **HIV**: F_1 improves from 0.167 (correlation-only) to 0.545 (Grade B). The pathogen pipeline identifies POL ($C = 0.990$), ENV ($C = 0.955$), and GAG ($C = 0.887$)—all Phase 4 targets with approved drugs—while the human-gene pipeline independently recovers CCR5 and CD4.
- **Malaria**: F_1 improves from 0.074 to 0.333 (Grade C). DHFR-TS and PPPK-DHPS—the targets of pyrimethamine and sulfadoxine, respectively—are identified as `PATHOGEN_DIRECT` with high confidence, despite being invisible to human-only knowledge graphs.

- **Ebola**: F_1 improves from 0.000 to 0.400 (Grade C). The GP envelope glycoprotein ($C = 0.970$)—the target of monoclonal antibodies (REGN-EB3)—is correctly identified as the top target.

The organism–disease relevance mechanism. A critical challenge in ChEMBL integration is that the database contains drugs prescribed to patients with a given disease, not only drugs that treat the disease mechanism. For non-infectious diseases, this means bacterial/fungal targets from co-prescribed antibiotics appear alongside legitimate targets. The organism–disease relevance score (r_{org}) addresses this: non-infectious diseases receive $r_{\text{org}} = 0.0$ for all pathogen targets, which, combined with the robustness penalty ($R \leftarrow R \times 0.3$), demotes bacterial targets (RPOA, RPSA) to INCONCLUSIVE. For infectious diseases, only pathogens matching the disease's causative organism receive $r_{\text{org}} = 1.0$.

The false-positive demotion benefit. NeoRx achieves zero known false positives across all 7 diseases. The biological intelligence layer demotes SCN10A and SCN2A (sodium channel symptom markers) in malaria, while the organism relevance mechanism demotes bacterial ribosomal proteins in Alzheimer's and breast cancer. This type-I error reduction has direct economic value in drug discovery.

5.2 Limitations

1. **Candidate pool ceiling (`max_genes=20`)**: The graph builder selects the top 20 genes by Open Targets association score for enrichment. For infectious diseases, host invasion receptors (BSG, CR1, DARC for malaria; NPC1 for Ebola) often rank below position 20 in Open Targets and are therefore never evaluated by the causal inference engine. ChEMBL's pathogen pipeline bypasses this limitation for pathogen-encoded targets but not for human host targets of pathogens.
2. **ChEMBL gene symbol quality**: Pathogen gene symbols in ChEMBL's `component_synonyms` table vary widely in quality. Common English words ("reverse", "protease"), single-character symbols ("P"), and multi-word protein names appear as `GENE_SYMBOL` entries. We apply three quality filters (bad-word blocklist, minimum length, no-space UNIPROT filter), but novel edge cases may emerge for under-studied organisms.
3. **Causal direction from observational data**: Without interventional data (e.g., CRISPR knockouts, Mendelian randomisation instruments), causal direction is inferred from graph topology and edge semantics. This is an assumption, not a proof.
4. **Simplified causal model**: The current implementation uses graph-based proxies for causal effect estimation rather than full DoWhy analysis on individual-level data (e.g., TCGA expression matrices). With real patient data, the framework could employ proper instrumental variable or regression discontinuity designs.
5. **Breast cancer BRCA1/BRCA2 threshold miss**: BRCA1 ($C = 0.599$) and BRCA2 ($C = 0.587$) fall marginally below the 0.6 CAUSAL threshold. These are loss-of-function tumor suppressors—not drug targets themselves—but their synthetic lethality partner PARP1 IS correctly identified as CAUSAL (rank #13). This represents a valid scoring outcome rather than a system failure.
6. **Druggability heuristics**: The druggability score is based on structural and functional annotations rather than experimental druggability assays.
7. **Precision ceiling**: Even for the best-performing disease (lung cancer, $P = 0.538$), nearly half of CAUSAL targets are not known approved drugs. Many may be genuinely causal but not yet validated clinically, making precision an imperfect metric for pipeline quality.

5.3 Future Work

- **Selectivity scoring for pathogen targets**: Adding selectivity ratios (pathogen target IC_{50} / human ortholog IC_{50}) from ChEMBL bioactivity data to penalise targets with high

human cross-reactivity

- **Adaptive candidate pool sizing**: Making `max_genes` disease-type-dependent (e.g., 40 for infectious diseases, 20 for cancer) to improve recall of host invasion receptors
- Integration with individual-level data sources (UK Biobank, TCGA, GEO) for full DoWhy causal effect estimation
- Extension to multi-target combination therapies using the graphical framework
- Temporal causal models incorporating longitudinal expression data
- Integration of Mendelian randomisation as an additional evidence stream

Figures

Six-stage pipeline from disease name input to ranked drug candidates. Stage 1: Multi-source knowledge graph assembly from eight databases (Monarch Initiative, Open Targets, KEGG, Reactome, STRING, UniProt, PDB, ChEMBL) with parallel API queries, automatic node merging, and pathogen target extraction. Stage 2: Causal target identification via backdoor criterion, causal effect estimation, sensitivity analysis, and bootstrap confidence intervals, with a dedicated pathogen target evaluation path for non-human targets. Stage 3: Biological intelligence layer (disease-type-aware classification and tissue expression boolean gating). Stage 4: Molecule generation via variational autoencoder. Stage 5: Drug-likeness screening (Lipinski, PAINS, QED) and molecular docking (AutoDock Vina). Stage 6: Composite scoring with causal confidence weighted above binding affinity.

Figure 1: NeoRx pipeline architecture.

Subgraph showing the dual-pipeline architecture. Pathogen targets: POL (top-ranked, $C = 0.990$) connects to the HIV disease node via ‘PATHOGEN$_D IRECT$‘$classificationwithPhase4drugevidence(efavirenz, dolutegravir, darunavir); ENV$(C = 0.955)$and$GAG(C = 0.887)similarlyconnectviavalidateddrugmechanisms.Humantargets : CCR5(C = 0.706,$‘$HOST_INVASION$‘$)withdirectedcausalpathtothediseasenodevia$‘$causes$‘$edges, supportedby5evia targetpathogenproteins(RPOA, RPSA)areshownwithdashedconnectionsandINCONCLUSIVEclassificationvia

Figure 2: HIV causal knowledge graph (excerpt).

Bar chart comparing NeoRx (blue) vs. correlation-only baseline (grey) for each disease. NeoRx outperforms correlation-only in 6 of 7 diseases. The most dramatic improvements are on infectious diseases: HIV F_1 from 0.167 to 0.545 (+227%), malaria from 0.074 to 0.333 (+350%), Ebola from 0.000 to 0.400. Lung cancer achieves Grade A ($F_1 = 0.700$, 100% recall). The only marginal loss is on breast cancer (0.462 vs. 0.421), where BRCA1/BRCA2 fall just below the CAUSAL threshold.

Figure 3: Precision–recall comparison across seven diseases.

6 Conclusion

NeoRx demonstrates that combining Pearl's causal inference framework with validated drug–target evidence from ChEMBL produces dramatically improved target identification compared to correlation-based methods. Across a 7-disease benchmark, NeoRx achieves a mean $F_1 = 0.474$ and outperforms the correlation-only baseline in 6 of 7 diseases. The strongest result is lung cancer (Grade A, $F_1 = 0.700$, 100% recall of all 7 validated targets). The addition of ChEMBL as an eighth data source with a dedicated pathogen target evaluation pathway transforms infectious

disease performance: HIV improves from Grade C to Grade B ($F_1 = 0.545$), malaria from Grade F to Grade C ($F_1 = 0.333$) with DHFR-TS and PPPK-DHPS identified as CAUSAL, and Ebola from Grade F to Grade C ($F_1 = 0.400$) with the GP envelope glycoprotein identified. The disease specificity score—replacing the former network centrality component—penalises promiscuous hub genes and improves precision across all diseases. The biological intelligence layer successfully blocks symptom markers and bacterial off-target proteins, achieving zero known false positives across all 7 diseases. The platform is disease-agnostic, fully automated, requires only a disease name as input, and is released as open-source software.

References

[1] Appay, V. & Sauce, D. "Immune Activation and Inflammation in HIV-1 Infection: Causes and Consequences." *J. Pathol.* 214, 231–241 (2008).

[2] Liu, R. et al. "Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection." *Cell* 86, 367–377 (1996).

[3] Dorr, P. et al. "Maraviroc (UK-427,857), a Potent, Orally Bioavailable, and Selective Small-Molecule Inhibitor of Chemokine Receptor CCR5." *Antimicrob. Agents Chemother.* 49, 4721–4732 (2005).

[4] Pearl, J. *Causality: Models, Reasoning, and Inference.* Cambridge University Press (2000).

[5] Hernán, M. A. & Robins, J. M. *Causal Inference: What If.* Chapman & Hall/CRC (2020).

[6] Piñero, J. et al. "DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants." *Nucleic Acids Res.* 45, D833–D839 (2017).

[7] Ochoa, D. et al. "Open Targets Platform: Supporting Systematic Drug–Target Identification and Prioritisation." *Nucleic Acids Res.* 49, D1302–D1310 (2021).

[8] Valdeolivas, A. et al. "Random Walk with Restart on Multiplex and Heterogeneous Biological Networks." *Bioinformatics* 35, 497–505 (2019).

[9] Davey Smith, G. & Hemani, G. "Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies." *Hum. Mol. Genet.* 23, R89–R98 (2014).

[10] Sharma, A. & Kiciman, E. "DoWhy: An End-to-End Library for Causal Inference." *arXiv:2011.04216* (2020).

[11] Beaumont, P. et al. "CausalNex." https://causalnex.readthedocs.io/ (2021).

[12] Trott, O. & Olson, A. J. "AutoDock Vina: Improving the Speed and Accuracy of Docking." *J. Comput. Chem.* 31, 455–461 (2010).

[13] Friesner, R. A. et al. "Glide: A New Approach for Rapid, Accurate Docking and Scoring." *J. Med. Chem.* 47, 1739–1749 (2004).

[14] Uhlén, M. et al. "Tissue-Based Map of the Human Proteome." *Science* 347, 1260419 (2015).

[15] Zheng, J. et al. "Phenome-Wide Mendelian Randomization Mapping the Influence of the Plasma Proteome on Complex Diseases." *Nat. Genet.* 52, 1122–1131 (2020).

[16] Liu, Y. et al. "EpiGraphDB: A Database and Data Mining Platform for Health Data Science." *Bioinformatics* 37, 1304–1311 (2021).

[17] Wallis, R. S. et al. "Tuberculosis and TNF: A Review of the Evidence." *Nat. Rev. Immunol.* 8, 866–874 (2008).

Appendix A: Software Availability

NeoRx is released under a non-commercial source-available licence at https://github.com/cod3smith/neorx.git. Free for personal, academic, and research use; commercial use requires a separate licence (contact kelyn@neoforgelabs.tech). The platform requires Python $\geq$ 3.13 and can be installed via `pip install neorx`. CLI usage:

```
neorx run "Alzheimer's" --top-n 10 --candidates 50
neorx identify HIV --top-n 5
neorx graph "Type 2 Diabetes"
```

Appendix B: Reproducibility

All experiments use a fixed random seed (`NEORX_SEED=42`). The bootstrap CI computation uses 200 resamples. Knowledge graph construction caches API responses with a configurable TTL (default 7 days). ChEMBL queries run against a local SQLite copy of ChEMBL v36. The complete test suite (23+ tests covering ChEMBL integration, critical fixes, and generic mocks) is included in the repository.

Funding

This research received no external funding.

Declaration of Interest

The author declares no competing interests. This is independent research conducted under NeoForge Labs.

Data Availability

All source code, trained models, and experimental scripts are available at https://github.com/cod3smith/neorx.git under a non-commercial source-available licence (free for personal, academic, and research use; commercial use requires a separate licence). ChEMBL v36 data is publicly available from the European Bioinformatics Institute. All external databases (Monarch Initiative, Open Targets, KEGG, Reactome, STRING, UniProt, RCSB PDB) are publicly accessible via their respective APIs.