

CausalDiffChem: Causal Network Restoration for Therapeutic Prioritization and Repurposing Across Diseases

Francis Armah

`francis.armah@ucc.edu.gh`

University of Cape Coast

Evans Abor

University of Cincinnati

Pius Sam

University of Cape Coast, Department of Biomedical Sciences

Emmanuel Sarpong

University of Cape Coast, Department of Biomedical Sciences

George Ghartey-Kwansah

University of Cape Coast, Department of Biomedical Sciences

Article

Keywords:

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11272130/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

CausalDiffChem: Causal Network Restoration for Therapeutic Prioritization and Repurposing Across Diseases

Evans Kwabena Abor^{1,2,*}, aborek@mail.uc.edu, Pius Kwasi Sam¹, piuskwasiam@gmail.com, Emmanuel Adu Sarpong¹, emmanuel.sarpong@stu.ucc.edu.gh, George Ghartey-Kwansah¹, george.ghartey-kwansah@ucc.edu.gh, Francis Ackah Armah^{1, *}, francis.armah@ucc.edu.gh

¹ Department of Biomedical Sciences, School of Allied Health Sciences, University of Cape Coast, Cape Coast, Ghana.

² Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA.

³ Department of Biochemistry and Biotechnology, College of Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.

* Corresponding Author

Abstract

Computational drug repurposing commonly relies on transcriptomic signature matching to identify compounds that reverse disease-associated gene expression, but such approaches do not distinguish putative causal drivers from downstream consequences of disease. Here, we present CausalDiffChem, a causal-network-guided framework that reconstructs disease-specific structural causal models from transcriptomic data using the NOTEARS algorithm and prioritizes compounds according to *net_correction*, a measure of network restoration under a simulated intervention propagated through the inferred linear structural causal model. We applied CausalDiffChem to multiple sclerosis, sarcoidosis, leishmaniasis and Alzheimer's disease, comprising 408 disease samples and 374 healthy controls. Independently reconstructed networks recovered biologically coherent disease-associated architecture, including a dominant apolipoprotein D interaction in Alzheimer's disease and a macrophage-associated regulatory axis in sarcoidosis. Screening FDA-approved, clinical-stage and preclinical compound libraries identified disease-specific candidates and recurrent prioritization across independently constructed networks. Cisplatin ranked first in both the FDA-approved and clinical-stage sarcoidosis libraries and recurred among the top-ranked candidates across multiple diseases, whereas memantine hydrochloride ranked first across all three compound libraries in leishmaniasis with identical restoration scores. Gene-level analyses further showed that high aggregate restoration could coexist with limited correction of dominant network interactions or discordant global alignment, demonstrating that network restoration, breadth and global transcriptional similarity capture complementary dimensions of therapeutic response. CausalDiffChem therefore provides an interpretable framework for prioritizing therapeutics according to their predicted effects on disease-specific regulatory structure, without requiring curated pathway annotations, and generates experimentally testable hypotheses for drug repurposing across heterogeneous diseases.

Main

Computational drug repurposing has expanded the search for therapeutic opportunities by identifying existing compounds with potential activity beyond their original indications.[1,2] A dominant strategy, exemplified by the Connectivity Map and related approaches, uses transcriptomic signature reversal to prioritize compounds whose molecular effects oppose disease-associated expression profiles.[1,2] Although this paradigm has identified clinically relevant candidates, it remains fundamentally correlational: it does not distinguish genes that contribute causally to disease from those whose altered expression arises downstream of established pathology. Moreover, the reproducibility of transcriptional signature-based predictions can vary across independent datasets.[3]

Structural causal models (SCMs) provide a framework for moving beyond static expression similarity by representing relationships among molecular variables as directed graphs.[4] Within a causal model, simulated interventions can be used to estimate how a perturbation propagates through the broader network. Pearl's do-calculus provides the formal foundation for intervention-based causal inference, distinguishing the consequences of an intervention from those inferred from observational association alone.[5] Recent studies suggest that causal networks inferred from molecular data can provide information about intervention effects and may support therapeutic prioritization beyond conventional correlation-based approaches.[6,7]

Here we introduce CausalDiffChem, a causal network restoration framework for therapeutic prioritization and drug repurposing across diseases. The framework reconstructs disease-specific SCMs directly from transcriptomic data using the NOTEARS continuous optimisation algorithm, without requiring curated pathway annotations, protein-interaction networks, or manually specified regulatory relationships.[8] Candidate compounds are represented by molecular fingerprints and a drug-likeness scale to generate perturbation vectors, which are introduced into each independently reconstructed disease network through a simulated intervention. The resulting post-intervention state is propagated through the inferred linear SCM and compared with the corresponding healthy reference state. We summarize this restoration using `net_correction`, which quantifies the mean direction-aware correction of network nodes toward the healthy state, complemented by the percentage of corrected nodes (%H) and the cosine similarity between the intervention-induced and disease-to-healthy expression shifts (R).

Figure 1. CausalDiffChem framework for causal-network-guided therapeutic prioritization.

For each disease, transcriptomic data comprising disease and healthy samples were reduced to the 20 most variable genes and used to reconstruct a disease-specific structural causal model (SCM) using NOTEARS. In parallel, each candidate compound was represented by molecular fingerprints and a drug-likeness score to generate a perturbation vector. The perturbation was propagated through the inferred linear SCM using a simulated intervention, $do(T)$, to obtain an intervention state. Network restoration was quantified using `net_correction`, the percentage of corrected nodes (%H), and global alignment with the healthy-state transition vector (R). This procedure was applied independently across four disease modules and FDA-approved, clinical-

stage and preclinical compound libraries. Gene-level restoration profiles and cross-disease and cross-library convergence were subsequently evaluated.

We applied CausalDiffChem to transcriptomic datasets comprising 408 disease samples and 374 healthy controls across four disease modules representing distinct pathological contexts: multiple sclerosis, sarcoidosis, leishmaniasis, and Alzheimer's disease. Each disease network was reconstructed independently, with no parameters, regulatory relationships, or pathway information shared between modules. FDA-approved, clinical-stage, and preclinical compound libraries were subsequently scored against each disease-specific network using the same intervention and restoration framework. This design enabled evaluation of disease-specific therapeutic prioritization while testing whether independently reconstructed networks and independently screened compound libraries repeatedly identify convergent candidates.

Results

Disease-specific structural causal models recover biologically coherent regulatory architecture

We reconstructed independent structural causal models (SCMs) for four disease modules spanning neuroinflammatory, systemic granulomatous, parasitic, and neurodegenerative pathology: multiple sclerosis (MS, blood; 99 disease and 45 control samples), sarcoidosis (39/35), leishmaniasis (25/10), and Alzheimer's disease (AD; 245/284, ADNI). Across 408 disease samples and 374 healthy controls (782 total), NOTEARS inferred compact directed acyclic graphs containing 48–61 retained interactions among the twenty most variable genes in each disease, with no shared parameters, curated pathway information, or manually specified regulatory relationships across modules (Figure 2).

Positive interactions predominated in all four networks, although their relative proportions and overall topologies differed substantially between diseases (Figure 2). AD showed the highest proportion of positive edges (87%), indicating a strongly co-directional inferred architecture. Sarcoidosis (76% positive) and leishmaniasis (67% positive) showed intermediate proportions, whereas MS had the most balanced distribution, with 58% positive and 42% negative edges. Thus, despite identical network-construction procedures, the inferred architectures differed in both interaction density and the balance between positive and negative relationships.

The strongest interaction in the AD network was $APOD_1 \rightarrow APOD$ ($w=0.932$), connecting two measurements of apolipoprotein D, a gene implicated in Alzheimer's disease through roles in lipid transport and neuronal membrane maintenance.[9] The strongest inhibitory interaction was $ARRDC3 \rightarrow GLOD5$ ($w=-0.167$). In sarcoidosis, the strongest positive edge was $IFI44L \rightarrow IFIT1$ ($w=0.890$), connecting two interferon-stimulated genes. The network also recovered an inferred macrophage-associated axis involving $ANXA3 \rightarrow ARG1$ and $ARG1 \rightarrow IL1R2$, while $IL1R2 \rightarrow MAFG$ ($w=-0.554$) was the strongest inhibitory interaction. The ARG1-centered relationships are notable in the context of macrophage polarization and granulomatous disease biology, although the inferred network does not by itself establish the direction or functional consequences of these interactions in vivo.[10]

In MS, the strongest positive interaction was DEFA3→DEFA4 ($w=0.851$), connecting two alpha-defensin genes associated with neutrophil antimicrobial responses, whereas SMAP2→RPL21 ($w=-0.462$) was the strongest inhibitory interaction. In leishmaniasis, HLA-DRB5→HLA-DRB1 ($w=0.737$) was the strongest positive edge, connecting two MHC class II antigen-presentation genes. A second prominent relationship was CREB1→RAMP2-AS1, linking a transcriptional regulator with a long non-coding RNA, while XPNPEP3→CNFN ($w=-0.608$) represented the strongest inhibitory interaction. The presence of antigen-presentation, macrophage-associated, and tissue-related genes within the inferred leishmaniasis network is consistent with the biological context of the cutaneous disease cohort, although these associations remain computationally inferred rather than experimentally validated causal relationships. Full edge-weight distributions for all four networks, ranked by absolute magnitude, are shown alongside the network diagrams in Figure 2.

Disease cohort sizes varied substantially, from 25 disease samples in leishmaniasis to 245 in AD. This variation represents an important consideration when interpreting the inferred networks, because smaller cohorts provide less information for distinguishing stable regulatory relationships from sampling variation. Nevertheless, the same sparsity penalty, optimization procedure, edge-pruning threshold, and network dimensionality were applied uniformly across all four diseases. Leishmaniasis produced the largest inferred network, with 61 retained interactions, compared with 48–60 interactions in the other modules. Network size alone, however, cannot establish greater biological robustness; assessment of network stability across independent replication cohorts will be required to determine the reproducibility of individual inferred edges.

Figure 2. Disease-specific structural causal models reveal distinct regulatory architectures.

Structural causal models reconstructed independently from the 20 most variable genes in multiple sclerosis, sarcoidosis, leishmaniasis and Alzheimer's disease. Directed edges represent inferred causal relationships, with edge direction indicated by arrows and edge magnitude represented by the corresponding edge weight. Positive and negative interactions are shown separately. The strongest interactions identified in each disease included APOD_1→APOD in Alzheimer's disease, IFI44L→IFIT1 in sarcoidosis, DEFA3→DEFA4 in multiple sclerosis and HLA-DRB5→HLA-DRB1 in leishmaniasis. Edge-weight distributions are shown alongside each network to illustrate differences in the magnitude and polarity of inferred interactions across disease modules.

Causal network restoration prioritizes disease-specific FDA-approved therapeutics

Ranking FDA-approved compounds by **net_correction**, defined as the mean direction-aware fractional restoration of SCM nodes toward the healthy reference state under simulated intervention, identified a distinct leading candidate for each disease (Figure 3; Supplementary Tables S3 and S4). Cyclizine hydrochloride, an antihistaminic antiemetic, ranked first in MS ($\text{net_correction} = 0.5236$; 80% of nodes corrected; $R = 0.6589$), with the strongest restoration observed for IFIT2, HERC5, and MT1A. Cisplatin, a platinum-based chemotherapeutic, ranked

first in sarcoidosis (0.4454; 55% breadth; $R = -0.7212$), with prominent correction of ARG1, BPI, and CRISP3, including genes within the macrophage-associated regulatory axis identified in the reconstructed network. Memantine hydrochloride, an NMDA-receptor antagonist approved for moderate-to-severe AD, ranked first in leishmaniasis (0.5064; 70% breadth; $R = 0.2270$), with strongest restoration of CREB1, RNA28S5, and S100A4. In AD, mecamylamine hydrochloride, a non-selective nicotinic antagonist, ranked first (0.5355; 80% breadth; $R = -0.3091$), most strongly restoring LRRC40, TSPAN13, and C2ORF69.

Convergence was evident beyond the individual top-ranked compounds. Cisplatin ranked second in MS (0.5120) while ranking first in sarcoidosis, and carbamazepine ranked third in the sarcoidosis FDA-approved library (0.4185) while independently ranking first in its preclinical library. Sarcoidosis also contained memantine as a separate catalogue entry from memantine hydrochloride (rank 6; 0.4113), while leishmaniasis contained memantine at rank 2 (0.4978) immediately below memantine hydrochloride. In AD, mecamylamine hydrochloride showed the largest first-to-second ranking gap of the four diseases, with a net_correction of 0.5355 compared with 0.4957 for abametapir. Abametapir also appeared among the top FDA-approved candidates for leishmaniasis (rank 5; 0.4734), providing an additional example of cross-disease recurrence. Thus, convergence was not restricted to the single highest-ranked compound but was also evident across the broader ranked candidate sets.

Two compounds initially ranked first in the sarcoidosis and leishmaniasis FDA-approved libraries, respectively—halothane and sevoflurane—were excluded from top-ranked reporting because their established use as inhalation anesthetics does not provide a plausible basis for chronic therapeutic repurposing in these disease settings. After this predefined viability exclusion, cisplatin and memantine hydrochloride became the reported leading candidates for sarcoidosis and leishmaniasis, respectively.

Cisplatin recurred within the FDA-approved top ten for three of the four diseases: MS (rank 2; 0.5120), sarcoidosis (rank 1; 0.4454), and leishmaniasis (rank 9; 0.4552). Importantly, its restoration profile differed across these disease networks. In MS, cisplatin corrected 75% of nodes and showed positive global alignment with the healthy reference ($R = 0.1031$), whereas in sarcoidosis it corrected 55% of nodes but showed strongly negative global alignment ($R = -0.7212$). Abametapir similarly recurred in leishmaniasis (rank 5; 0.4731; 85% breadth; $R = 0.3853$) and AD (rank 2; 0.4957; 75% breadth; $R = -0.2506$), with opposite global-alignment patterns despite substantial restoration in both networks. These findings indicate that recurrence of a compound across independently reconstructed disease networks does not imply an identical restoration profile, reinforcing the importance of evaluating magnitude, breadth, and global transcriptional alignment separately.

Magnitude and breadth represent complementary dimensions of network restoration

The percentage of nodes corrected (%H) provided a complementary measure of restoration breadth and did not necessarily track with net_correction magnitude (Figure 3). Cyclizine hydrochloride in MS and mecamylamine hydrochloride in AD produced the broadest restoration

among the four leading FDA-approved compounds, with 80% of nodes corrected, followed by memantine hydrochloride in leishmaniasis (70%) and cisplatin in sarcoidosis (55%).

The difference between magnitude and breadth was particularly evident in sarcoidosis. Cisplatin achieved a *net_correction* of 0.4454 despite correcting only 55% of nodes, indicating that its overall score was driven by relatively strong restoration among a more restricted subset of genes. By contrast, cyclizine hydrochloride combined the highest breadth (80%) with a higher *net_correction* of 0.5236, reflecting more distributed restoration across the network. These results demonstrate that *net_correction* captures the magnitude of directionally correct restoration, whereas %H captures its breadth across the network; neither metric alone fully describes the resulting therapeutic response. The global similarity measure *R* provides a third, distinct dimension, as illustrated by compounds such as cisplatin and mecamlamine hydrochloride, which achieved substantial node-level restoration despite negative global alignment with the healthy reference.

Figure 3. Causal-network restoration prioritizes disease-specific FDA-approved therapeutics.

FDA-approved compounds were ranked independently for each disease according to *net_correction*. The leading viable candidates were cyclizine hydrochloride for multiple sclerosis, cisplatin for sarcoidosis, memantine hydrochloride for leishmaniasis and mecamlamine hydrochloride for Alzheimer's disease. For each leading compound, *net_correction*, percentage of corrected nodes (%H), and global alignment (*R*) are shown. Full rankings are provided in Supplementary Table S3, with the complete screened library in Supplementary Table S4. Halothane and sevoflurane were excluded from top-ranked reporting as inhalation anesthetics lacking a plausible role as chronic-disease repurposing candidates.

Causal network restoration generalizes across preclinical and clinical-stage compounds

Extending *net_correction* scoring to the preclinical and clinical-stage libraries identified distinct leading compounds across diseases and development stages (Figure 4; Supplementary Tables S5–S8). In MS, dizocilpine led the preclinical library (*net_correction* = 0.4223; 70% breadth; *R* = 0.2038), whereas cisplatin led the clinical-stage library (0.5120; 75% breadth; *R* = 0.1031), with neither coinciding with the FDA-approved leader, cyclizine hydrochloride. In AD, GS967 led the preclinical library (0.3767; 45% breadth; *R* = -0.7915), whereas mecamlamine hydrochloride again ranked first in the clinical-stage library (0.5355; 80% breadth; *R* = -0.3091), reproducing its FDA-approved result exactly. Sarcoidosis was led by carbamazepine in the preclinical library (0.4185; 60% breadth; *R* = -0.6358) and cisplatin in the clinical-stage library (0.4454; 55% breadth; *R* = -0.7212). Leishmaniasis showed the strongest within-disease consistency: memantine hydrochloride ranked first in all three libraries, with an identical *net_correction* of 0.5064, 70% breadth, and *R* = 0.2270.

The broader rankings revealed additional recurrence and salt-form effects. In MS, MPEP hydrochloride and MPEP occupied the second and third preclinical positions (0.418 and 0.395,

respectively), representing two catalogue entries for the same underlying compound. In sarcoidosis, memantine hydrochloride ranked second in the preclinical library (0.362), whereas in leishmaniasis it ranked first, with carbamazepine ranking third (0.451). Conversely, carbamazepine ranked first in sarcoidosis, illustrating that the relative ordering of the same candidates changed when evaluated against independently reconstructed disease networks. In leishmaniasis, RITA ranked second in both the preclinical and clinical-stage libraries with numerically identical scores of approximately 0.498. The clinical-stage library also contained amantadine hydrochloride and amantadine at ranks three and four (0.467 and 0.462), respectively. In AD, mecamlamine hydrochloride showed the largest separation between the first- and second-ranked clinical-stage candidates, with a `net_correction` of 0.535 compared with 0.445 for isoflavone.

Cross-disease recurrence was particularly evident in the preclinical library (Figure 4). Pinosylvin appeared within the top ten for all four diseases, although its restoration profile differed substantially across networks. Its highest `net_correction` occurred in leishmaniasis (0.4236; 85% breadth; $R = 0.2007$), whereas its breadth was 50% in both sarcoidosis and AD, accompanied by negative global similarity ($R = -0.627$ and -0.6945 , respectively). RITA similarly showed a positive global alignment in leishmaniasis (0.4976; 80% breadth; $R = 0.2743$) but negative values in sarcoidosis (0.3255; 60%; $R = -0.5619$) and AD (0.3572; 75%; $R = -0.6046$). GS967 appeared in the top ten for sarcoidosis, leishmaniasis, and AD, with `net_correction` values of 0.3390, 0.3704, and 0.3767, respectively, while breadth declined from 65% and 55% to 45% and global similarity became increasingly negative.

Figure 4. Causal-network restoration extends across preclinical and clinical-stage compound libraries.

Candidate compounds were independently ranked using the same `net_correction` framework across preclinical and clinical-stage libraries for each disease. Top-ranked compounds and the corresponding restoration magnitude, breadth (%H), and global alignment (R) are shown for each disease–library combination. Cross-disease recurrence among the top-ranked candidates is also shown, highlighting compounds that repeatedly emerged from independently reconstructed disease networks. Memantine hydrochloride showed identical top-ranked restoration across all three libraries in leishmaniasis, whereas cisplatin independently led both the FDA-approved and clinical-stage sarcoidosis rankings.

Additional cross-disease recurrence was observed for MPEP hydrochloride, (Z)-SMI-4a, and NQ301, which each appeared in the preclinical top ten for both MS and AD. Carbamazepine and memantine hydrochloride likewise appeared in the top ten for both sarcoidosis and leishmaniasis. These recurring compounds did not produce uniform restoration profiles across diseases, demonstrating that recurrence reflects repeated prioritization by independent networks rather than identical network responses.

The clinical-stage library showed a similarly notable pattern, with cisplatin appearing within the top ten for MS (rank 1), sarcoidosis (rank 1), and leishmaniasis (rank 5). Thus, compounds identified independently across development-stage libraries frequently recurred across disease modules, despite each disease network having been reconstructed independently and scored without shared causal parameters.

Gene-level restoration further distinguished the mechanisms underlying these rankings. In AD, GS967 most strongly restored THEMIS, ZEB1, and TSPAN13, whereas mecamylamine hydrochloride preferentially restored LRRC40, TSPAN13, and C2ORF69. These distinct restoration profiles demonstrate that compounds with different development-stage positions can achieve network correction through different subsets of genes, motivating the gene-level analysis of the leading compounds presented in Figure 5.

Gene-level restoration reveals the regulatory mechanisms underlying therapeutic prioritization

Having established the leading compound in each disease–library combination, we next examined the signed, uncapped gene-level restoration profiles of these compounds (Figure 5). This analysis provides a mechanistic view of the node-level changes underlying net_correction and helps explain instances in which node-level restoration and global transcriptional similarity (R) diverged. Because net_correction caps individual-node restoration at 1, the gene-level profiles additionally reveal restoration magnitudes that are otherwise obscured by the aggregate score.

In AD, mecamylamine hydrochloride strongly restored LRRC40, TSPAN13, and C2ORF69, with C2ORF69 reaching +2.05 in uncapped units. In contrast, APOD and APOD_1, which formed the strongest interaction in the AD network, remained among the least-restored nodes (−1.99 and −0.66, respectively). Thus, substantial correction across individual nodes occurred despite incomplete restoration of the dominant APOD/APOD_1 interaction, coinciding with a negative global similarity ($R = -0.3091$). The preclinical leader GS967 produced even greater restoration of C2ORF69 (+5.28), together with strong correction of THEMIS and ARRDC3, but APOD_1 remained negatively displaced (−0.81). Across the AD leaders, therefore, the dominant APOD/APOD_1 axis was not fully restored despite substantial correction elsewhere in the network.

Sarcoidosis showed the largest single-node restoration observed across the study. Cisplatin produced an uncapped IL1R2 correction of +19.77, more than fivefold greater than its next-largest positive effect, BPI (+3.80). Other prominent positive effects included ARG1 and CRISP3, whereas OLFM4, CEACAM8, and MMP8 moved further from the healthy state. This combination of strong positive restoration in selected nodes and opposing movement in others was accompanied by a negative global similarity ($R = -0.7212$). Carbamazepine, the preclinical leader in the same network, produced a distinct profile: ARG1 was its strongest positive correction (+5.54), whereas IL1R2 moved in the opposite direction (−3.30) and SLPI showed the most negative value in the panel (−4.49). Despite these different node-level profiles, cisplatin and carbamazepine achieved similar aggregate net_correction values (0.4454 and 0.4185, respectively).

Figure 5. Gene-level restoration reveals distinct mechanisms underlying compound prioritization.

Signed, uncapped gene-level restoration profiles for the leading compound from each disease and compound library. Positive values indicate movement toward the healthy reference state, whereas negative values indicate movement away from the healthy state. Unlike the capped `net_correction` metric, gene-level values retain restoration magnitudes exceeding the healthy-state reference gap. Distinct gene-level mechanisms underlie compounds with similar `net_correction` values: cisplatin strongly restored IL1R2 in sarcoidosis while producing opposite-direction changes in neutrophil-associated genes, whereas carbamazepine produced a distinct restoration profile on the same network. In Alzheimer's disease, mecamylamine hydrochloride strongly corrected LRRC40, TSPAN13 and C2ORF69 but did not restore the APOD/APOD_1 interaction that constituted the strongest inferred network edge. These profiles illustrate why network-level correction and global transcriptional alignment can diverge.

MS similarly demonstrated substantial variation in gene-level restoration across development stages. Cisplatin, the clinical-stage leader, produced the largest positive correction observed across the three MS leaders, at SMAP2 (+12.32). The same node was minimally displaced by cyclizine hydrochloride (−0.26), the FDA-approved leader, and strongly displaced in the diseaseward direction by dizocilpine (−8.08), the preclinical leader. Thus, SMAP2 alone illustrates how compounds with similar prioritization within the same disease network can produce markedly different node-level responses.

Leishmaniasis provided a contrasting profile. Memantine hydrochloride, which ranked first across all three libraries with identical aggregate scores, produced a positive global similarity ($R = 0.2270$) together with 70% node-level breadth. Its strongest gene-level effect was observed for RNA28S5 (+6.51), providing a prominent contributor to the aggregate restoration profile. MS's FDA-approved leader, cyclizine hydrochloride, showed the highest positive global similarity among the leading compounds ($R = 0.6589$), together with 80% breadth. Together, these profiles illustrate that broad node-level restoration and positive global transcriptional alignment can coincide, whereas other compounds can achieve substantial `net_correction` despite discordant global alignment.

The gene-level analysis does demonstrate that `net_correction` summarizes the extent of directionally correct restoration but does not specify which regulatory nodes account for that score or whether the resulting network state is globally aligned with health. The uncapped profiles therefore provide complementary mechanistic information, revealing both exceptionally strong restoration of individual nodes and persistent diseaseward displacement of other components of the same network.

Convergent therapeutic prioritization across independent compound libraries

The independent compound libraries and disease-specific networks nevertheless produced repeated convergence in therapeutic prioritization (Figure 4; Supplementary Tables S5–S8). Leishmaniasis showed the strongest within-disease convergence: memantine hydrochloride ranked first in the FDA-approved, preclinical, and clinical-stage libraries, with identical *net_correction* (0.5064), breadth (70%), and global similarity ($R = 0.2270$) in each library. Sarcoidosis likewise showed cross-library convergence around cisplatin, which ranked first in both the FDA-approved and clinical-stage libraries, while carbamazepine independently ranked first in the preclinical library and produced restoration involving the same ARG1/BPI/CRISP3 region highlighted for cisplatin.

Cross-disease convergence was also evident. Cisplatin appeared within the top ten FDA-approved or clinical-stage candidates for MS, sarcoidosis, and leishmaniasis, while MPEP hydrochloride, (Z)-SMI-4a, and NQ301 each appeared within the preclinical top ten for both MS and AD. Carbamazepine and memantine hydrochloride each recurred within the preclinical top ten for sarcoidosis and leishmaniasis. The preclinical library showed particularly broad recurrence: pinosylvin appeared in the top ten for all four diseases, while GS967 and RITA each appeared in the top ten for three diseases.

Importantly, recurrence did not imply an identical response across disease networks. Pinosylvin, for example, showed its strongest and broadest restoration in leishmaniasis (*net_correction* = 0.4236; 85% breadth; $R = 0.2007$), whereas its appearances in sarcoidosis and AD were characterized by narrower breadth and negative global similarity. RITA similarly showed positive global alignment in leishmaniasis but negative alignment in sarcoidosis and AD. Thus, the repeated appearance of a compound reflects convergence in **prioritization**, whereas its gene-level restoration profile remains dependent on the disease-specific causal network.

Taken together, these results show that independently reconstructed disease networks and independently scored compound libraries repeatedly identify overlapping therapeutic candidates. At the same time, the distinct gene-level and global-response profiles observed for recurring compounds indicate that this convergence is not simply a consequence of compounds producing a uniform transcriptional response across diseases. Rather, the same compound can be prioritized by different networks while acting on different subsets of nodes and producing different degrees of global restoration. This distinction provides the basis for interpreting therapeutic convergence alongside the positive-control analysis of known clinical therapies.

Known-therapeutic recovery as a positive-control check

To determine whether the FDA-approved ranking recovered established therapeutic signal beyond the specific candidates highlighted above, we examined the full, unfiltered *net_correction* ranking of 2,831 FDA-approved compounds for each disease. Small-molecule therapeutics with established clinical use for the corresponding disease were identified independently, and their positions within the complete ranking were recorded. This analysis was intended as a positive-control assessment rather than a comparison with an alternative computational method: it asked whether clinically established treatments received favorable *net_correction* scores, irrespective of whether they were the top-ranked compounds identified by CausalDiffChem.

In Alzheimer's disease, two mechanistically distinct approved therapeutic classes ranked within the upper portion of the library. Donepezil, a cholinesterase inhibitor, ranked 170th of 2,831 compounds (top 6.0%), whereas memantine hydrochloride, an NMDA-receptor antagonist, ranked 252nd (top 8.9%). Neither was the overall top-ranked compound, which was mecamlamine hydrochloride. Tacrine, an earlier-generation cholinesterase inhibitor that is now largely discontinued because of hepatotoxicity, ranked 781st (top 27.6%). Rivastigmine ranked substantially lower, within the bottom third of the library (72–94th percentile depending on salt form), and this result was retained without adjustment.

The remaining disease modules showed more heterogeneous recovery. In multiple sclerosis, siponimod and ozanimod, both sphingosine-1-phosphate receptor modulators, ranked within the upper portion of the library (top 17.9% and 21.0%, respectively), whereas fingolimod ranked lower (top 63–67%). In sarcoidosis, chloroquine and leflunomide ranked within the upper two-fifths of the library (top 38.9% and 39.2%, respectively), while prednisone and related corticosteroid formulations ranked substantially lower (81st–95th percentile). In leishmaniasis, miltefosine ranked within the upper quartile (top 23.2%), whereas pentavalent antimonials and amphotericin B ranked near the lower end of the library (79th–99th percentile).

Overall, the positive-control analysis showed that established therapeutics can receive favorable net_correction scores, most notably in Alzheimer's disease, where two mechanistically distinct approved treatments ranked within the top 10% of the full FDA-approved library. However, recovery was heterogeneous across diseases and therapeutic classes. A favorable net_correction rank therefore should not be interpreted as evidence of established clinical efficacy, nor should a lower rank be interpreted as evidence against clinical effectiveness. Rather, these results indicate that the twenty-gene causal networks capture aspects of therapeutic biology that overlap with established treatments in some disease contexts, while incompletely representing broader, systemic, or route-dependent mechanisms of action in others.

Discussion

We developed CausalDiffChem, a causal-network-guided framework for computational drug repurposing that prioritizes compounds according to their predicted ability to restore disease-associated molecular states toward a healthy reference. Across four biologically distinct diseases, independently reconstructed structural causal models generated disease-specific therapeutic rankings while repeatedly identifying compounds across diseases and compound-development stages. Cisplatin, for example, ranked first in both the FDA-approved and clinical-stage sarcoidosis libraries and recurred among the top-ranked candidates in three of four diseases. Memantine hydrochloride showed an even stronger within-disease convergence, ranking first across the FDA-approved, clinical-stage and preclinical libraries for leishmaniasis with numerically identical scores. Because these rankings were generated from independently reconstructed networks without shared disease-specific parameters or curated pathway relationships, their recurrence provides evidence that causal-network restoration can identify therapeutic signals that extend beyond a single disease transcriptomic signature or compound library.

The gene-level analyses further show why `net_correction` should not be interpreted independently of restoration breadth or global transcriptional alignment. These metrics capture complementary properties of the simulated intervention: `net_correction` measures direction-aware restoration across individual network nodes, `%H` measures the breadth of nodes moving toward the healthy state, and `R` measures the overall alignment of the induced displacement with the disease-to-healthy direction. In Alzheimer's disease, mecamlamine hydrochloride produced substantial node-level restoration but left `APOD` and `APOD_1` among the least-corrected genes, including the network's strongest inferred interaction. Similarly, cisplatin in sarcoidosis strongly restored the `ARG1/BPI/CRISP3` axis while several neutrophil-associated genes moved away from the healthy reference, resulting in a strongly negative `R` despite substantial `net_correction`. By contrast, the leading compounds in MS and leishmaniasis showed greater concordance between node-level restoration and global alignment. Thus, the framework distinguishes compounds that broadly realign a network from those that produce strong but spatially heterogeneous restoration, a distinction that would be obscured by a single transcriptomic similarity metric.

The positive-control analysis provides additional evidence that established therapeutic biology is captured by the inferred networks, although recovery was heterogeneous across diseases. In Alzheimer's disease, donepezil and memantine hydrochloride both ranked within the top 10% of the complete FDA-approved library despite neither being the overall CausalDiffChem leader. Several established treatments in MS, sarcoidosis and leishmaniasis also ranked favorably, whereas corticosteroids and several injectable antileishmanial therapies ranked substantially lower. These findings emphasize that favorable computational ranking and established clinical efficacy are related but not equivalent. CausalDiffChem evaluates restoration of a compact transcriptional network and may therefore capture only a subset of the mechanisms underlying therapies whose effects are systemic, broad-spectrum, tissue-specific or strongly dependent on pharmacokinetics and route of administration.

Several limitations should be considered. First, the disease cohorts differed substantially in sample size, particularly the relatively small leishmaniasis cohort, which may reduce the stability of inferred network relationships. Replication in larger independent cohorts will be important to establish the robustness of individual edges and therapeutic rankings. Second, the current framework does not intrinsically distinguish causal-network compatibility from clinical feasibility. The initial appearance of halothane and sevoflurane as top-ranked compounds in two FDA-approved libraries illustrates this limitation and supports incorporation of explicit pharmacological viability criteria before final candidate prioritization. Third, compound perturbations were approximated from molecular fingerprints and physicochemical properties rather than experimentally measured drug-induced transcriptional responses. The resulting gene-level effects should therefore be considered mechanistic hypotheses rather than experimentally established drug mechanisms. Finally, the linear, acyclic SCM formulation provides a tractable representation of regulatory structure but does not capture all biological feedback, nonlinear regulation or cell-type-specific interactions.

These limitations also define clear directions for future development. Experimental perturbation of the predicted high-impact genes and convergent compounds will be essential to determine whether the simulated interventions correspond to biological responses. In particular, compounds

recurring across independently reconstructed disease networks provide attractive candidates for systematic validation. Larger replication cohorts, experimentally measured drug-response signatures, cell-type-resolved data and extensions to nonlinear or cyclic causal models could further improve the fidelity of the framework. Incorporating pharmacological viability and translational constraints directly into candidate prioritization would also reduce the occurrence of computationally favorable but clinically impractical compounds.

Overall, these findings establish causal-network restoration as a complementary strategy for drug repurposing that moves beyond simple reversal of disease-associated expression signatures. By linking compound perturbations to inferred disease-specific regulatory structure, CausalDiffChem provides an interpretable framework in which aggregate restoration, breadth, global alignment and gene-level effects can be examined together. The repeated convergence of candidates across independently reconstructed networks and compound libraries supports further experimental investigation of causal-network-guided therapeutic prioritization as a general strategy for identifying repurposing opportunities across heterogeneous diseases.

Methods

Study design

Transcriptomic datasets representing four disease modules were retrieved from the National Center for Biotechnology Information Gene Expression Omnibus (GEO)[11] and the Alzheimer's Disease Neuroimaging Initiative (ADNI).[12] The datasets comprised multiple sclerosis blood transcriptomes (GSE17048),[13] sarcoidosis (GSE37912),[14] cutaneous leishmaniasis (GSE55664),[15] and Alzheimer's disease microarray data from ADNI. Each GEO dataset was downloaded in its deposited, preprocessed form (log₂-normalized microarray intensities or VST-normalized RNA-seq counts, as originally submitted by the respective dataset authors) and processed using the GEOparse interface where appropriate. Binary case/control labels were assigned from the original dataset sample metadata. Each dataset was analyzed independently, and the twenty most variable genes within each disease cohort were selected for structural causal model (SCM) construction.

Structural causal model construction

For each disease, an SCM was reconstructed from z-scored expression values of the twenty most variable genes using the NOTEARS continuous optimization algorithm, which learns a weighted directed acyclic graph W by minimizing a least-squares reconstruction loss subject to the smooth acyclicity constraint

$$h(W) = \text{tr}(e^{W \circ W}) - d = 0,$$

where $d = 20$ is the number of network nodes.[5,8,20] The constraint was enforced using an augmented Lagrangian formulation optimized with L-BFGS-B. An L1 sparsity penalty ($\lambda_1 = 0.05$) was applied throughout. The penalty coefficient ρ was initialized at 1.0 and multiplied by 10, up to

a maximum of 10^{12} , whenever the acyclicity violation failed to decrease by at least 75% between outer iterations. The outer augmented-Lagrangian procedure was run for up to 500 iterations, with convergence defined as $h(W) \leq 10^{-8}$. Each outer iteration used L-BFGS-B for up to 200 inner iterations with a function tolerance of 10^{-14} .

After convergence, edges with $|w| < 0.05$ were pruned; this threshold was also used to define edge counts and edge-weight distributions reported in the Results. A GPU-accelerated implementation of the same objective and acyclicity constraint, using PyTorch autograd to differentiate through the matrix exponential, is included in the released codebase for larger gene sets; the twenty-gene networks reported here were reconstructed using the CPU implementation described above. Non-finite expression values (NaN or infinite), where present, were replaced by the per-gene median calculated from finite observations within the same cohort before network reconstruction. Each disease SCM was reconstructed independently, with no parameters, pathway annotations, or regulatory relationships shared across disease modules.

Expression normalization quality control

Before SCM reconstruction, each disease expression matrix was screened for appropriate z-score normalization by comparing the mean absolute disease-state expression value with a threshold of 1.5. Datasets exceeding this threshold, indicating that the data remained on a $\log_2$ or raw-intensity scale rather than being fully standardized, were re-normalized by per-gene z-scoring across all samples. This quality-control procedure was applied uniformly across all four disease cohorts.

Causal perturbation and net_correction scoring

For each candidate compound, a molecular perturbation vector was constructed from two Extended-Connectivity Fingerprints (ECFP4 and ECFP6), corresponding to Morgan fingerprint radii 2 and 3, respectively.[16] The radius-2 fingerprint was generated directly at the dimensionality $d = 20$ of each disease network. The radius-3 fingerprint was generated at 512 bits and reduced to the same dimensionality by averaging contiguous, non-overlapping 25-bit segments ($\lceil 512/20 \rceil = 25$); the remaining 12 highest-index bits were not included in the folded representation. The two fingerprints were combined as

$$0.6 \times \text{ECFP4} + 0.4 \times \text{folded ECFP6},$$

and the resulting vector was rescaled to the range $[-, 11]$.

A drug-likeness scale was calculated as

$$0.5 \times QED + 0.3 \times \text{clip}\left(\frac{\log P - 1}{3}, 0, 1\right) + 0.2 \times \left(1 - \frac{TPSA}{140}\right),$$

using QED,[17] logP,[18] and TPSA[19]. The compound perturbation vector δ_T was then calculated as the signed fingerprint multiplied by the absolute disease–healthy expression gap and the drug-likeness scale, with an additional gap-dependent term:

$$\delta_T = (\text{signed fingerprint}) \times |\text{disease} - \text{healthy}| \times S + (\text{disease} - \text{healthy}) \times S \times 0.2,$$

where S denotes the drug-likeness scale.

The perturbation was propagated through the reconstructed causal network using the closed-form linear structural-equation intervention solution

$$v_{\text{int}} = (I - W + \varepsilon I)^{-1}(v_{\text{disease}} + \delta_T),$$

where $\varepsilon = 10^{-6}$ was used for numerical stability. This represents a simulated intervention propagated through the inferred linear structural causal model rather than a general application of Pearl's full do-calculus framework.[5]

For each node i , the correction score was calculated as

$$C_i = \min\left(1, \frac{|v_{\text{int},i} - v_{\text{dis},i}|}{|v_{\text{hea},i} - v_{\text{dis},i}| + \varepsilon}\right) \times 1[\text{sign}(v_{\text{int},i} - v_{\text{dis},i}) = \text{sign}(v_{\text{hea},i} - v_{\text{dis},i})].$$

The reported **net_correction** was the mean of C_i across all 20 network nodes. The percentage of nodes corrected (%H) was calculated as the percentage of nodes for which the directional indicator equaled 1, irrespective of correction magnitude. Global transcriptional alignment (R) was calculated as the cosine similarity between the intervention-induced change and the disease-to-healthy change:

$$R = \cos(v_{\text{int}} - v_{\text{dis}}, v_{\text{hea}} - v_{\text{dis}}).$$

Therapeutic ranking, cross-library convergence, and gene-level restoration analysis

Within each disease and compound library, compounds were ranked in descending order of net_correction after application of the exclusion and molecular-weight/QED filtering criteria described below. The highest-ranked compound in each disease–library combination was designated the top-ranked or leading compound.

Cross-disease and cross-library convergence was assessed by tabulating each compound's appearances within the top ten ranked candidates for each disease and library after application of the same exclusion and filtering criteria. A compound appearing within the top ten for more than one disease or more than one library within a disease was considered recurrent or convergent.

As a positive-control analysis of the FDA-approved rankings, known clinically established small-molecule therapeutics for each disease were identified from standard treatment guidelines and located within the corresponding full, unfiltered net_correction ranking. This analysis used all 2,831 screened FDA-approved compounds rather than only the top twenty reported in Supplementary Table S4. Position was expressed as “top X%,” calculated as

$$\text{top } X\% = \frac{\text{rank}}{\text{total compounds screened}} \times 100,$$

where a smaller value indicates a higher-ranked compound. Therapeutics absent from the screened library, principally monoclonal antibodies and other biologics excluded by the molecular-weight criteria, were not evaluated.

Gene-level restoration profiles in Figure 5 represent a signed, uncapped extension of the node-level correction measure. For each node, the uncapped magnitude was calculated as

$$M_i = \frac{|v_{\text{int},i} - v_{\text{dis},i}|}{|v_{\text{hea},i} - v_{\text{dis},i}| + \varepsilon}.$$

The value was assigned a positive sign when the intervention moved the node toward the healthy state and a negative sign when it moved away from the healthy state. Unlike C_i , M_i was not capped at 1, allowing restoration exceeding the original disease–healthy expression gap to be visualized. These uncapped values were used only for gene-level visualization and interpretation in Figure 5; compound ranking and cross-disease comparisons remained based exclusively on net_correction and %H.

Compound libraries and filtering

FDA-approved, clinical-stage, and preclinical compounds were filtered to molecular weight ≤ 900 Da before scoring. FDA-approved and clinical-stage compounds additionally required QED ≥ 0.3 . These criteria excluded peptide biologics and other compounds inconsistent with the small-molecule perturbation model. Several preclinical entries, including rusalatide acetate, AOD9604, terlipressin, cibinetide, and triptorelin, exceeded 1,200 Da and were therefore removed by the molecular-weight filter.

Azobenzene, a photoswitchable reference compound without therapeutic development history, and a compound catalogued as “Alodan,” whose structure corresponds to chlorbicyclen, an organochlorine insecticide rather than a therapeutic, were excluded from top-ranked reporting but retained and flagged in the complete result tables.

Halothane and sevoflurane ranked first by net_correction in the FDA-approved sarcoidosis and leishmaniasis libraries, respectively. Both are inhalation anesthetics without a plausible role as repurposed chronic-disease therapeutics and were therefore excluded from top-ranked reporting. The next-ranked viable compound was reported in their place throughout the Results.

Blood–brain barrier permeability prediction

Blood–brain barrier (BBB) permeability was predicted for top-ranked compounds across all four disease modules using Random Forest classifier and regression models developed from the B3DB benchmark dataset within the study’s own codebase. The classification model was trained on 7,805 labelled compounds that passed RDKit featurization, comprising 4,956 BBB-positive

and 2,851 BBB-negative compounds. The regression model was trained on the 1,058-compound subset with continuous logBB measurements.

Each compound was represented by a 107-dimensional feature vector comprising 43 RDKit physicochemical and topological descriptors, including molecular weight, logP, TPSA, hydrogen-bond donor and acceptor counts, ring counts, Kier/Hall connectivity and shape indices, BertzCT topological complexity, and E-state and partial-charge indices, together with a 64-bit radius-2 Morgan (ECFP4) fingerprint.

Both models used 300 trees, unlimited maximum depth, and a minimum leaf size of 2. Balanced class weighting was applied to the classifier. Five-fold stratified cross-validation (*random_state* = 42) was performed before refitting the models on the full training sets. The classifier achieved a cross-validation ROC-AUC of 0.956 ± 0.005 and accuracy of 0.882 ± 0.010 . The regression model achieved a mean absolute error of 0.344 ± 0.011 and R^2 of 0.565 ± 0.041 . TPSA and logP were among the most important features by Gini importance.

Compounds with predicted logBB > -1 were classified as CNS-penetrant. When the trained model was unavailable for a molecule, a Lipinski-style proxy was applied using logP 1–4, molecular weight <450 Da, TPSA <90 Å², fewer than 8 rotatable bonds, and ≤3 hydrogen-bond donors.

BBB predictions were used solely as an additional translational assessment of top-ranked compounds. Compound ranking was determined exclusively by net_correction; BBB permeability was not used in SCM reconstruction, perturbation generation, net_correction scoring, or therapeutic ranking.

Statistical reporting

net_correction and *R* were reported to four decimal places and %H to the nearest percentage point. No rounding was applied before ranking. No statistical significance testing was applied to compound rankings because net_correction, %H, and *R* are deterministic computational prioritization measures rather than inferential statistical estimates with associated *P* values or confidence intervals.

Software availability and reproducibility

All analyses were performed using a fixed, disease-agnostic codebase applied identically across the four disease modules, without manual disease-specific adjustment of hyperparameters, filtering thresholds, or scoring formulas. SCM optimization was initialized from a zero adjacency matrix, without random initialization. Neither SCM construction nor net_correction scoring for the FDA-approved, clinical-stage, or preclinical libraries involved stochastic procedures, making the reported results reproducible given the same input expression data and compound structures.

A fixed random seed is used elsewhere in the codebase solely for sampling compounds for a separate ZINC15 virtual-screening extension and was not used in the analyses reported here. Molecular descriptors and fingerprints were calculated using RDKit, and causal-network

optimization used the SciPy L-BFGS-B implementation. The released source code includes the scripts used for SCM construction, net_correction scoring, and BBB prediction, together with a requirements file specifying software dependencies. Sample sizes, edge counts, and top-ranked compounds for each disease and compound library are provided in the Supplementary Information to facilitate independent re-analysis.

References

1. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313:1929–1935.
2. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;171:1437–1452.e17.
3. Lim N, Pavlidis P. Evaluation of Connectivity Map shows limited reproducibility in drug repositioning. *Scientific Reports*. 2021;11:17624.
4. Porcu E, Sadler MC, Lepik K, et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nature Communications*. 2021;12:5647.
5. Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press; 2009.
6. Maathuis MH, Kalisch M, Bühlmann P. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*. 2009;37:3133–3164.
7. Domingo-Fernández D, Gadiya Y, Patel A, Mubeen S, Rivas-Barragan D, Diana CW, et al. Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLOS Computational Biology*. 2022;18(2):e1009909.
8. Zheng X, Aragam B, Ravikumar P, Xing EP. DAGs with NO TEARS: Continuous optimization for structure learning. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.
9. Kolanek A, Cemaga R, Maciejczyk M. Role and Diagnostic Significance of Apolipoprotein D in Selected Neurodegenerative Disorders. *Diagnostics*. 2024;14(24):2814.
10. Wilson JL, Mayr HK, Weichhart T. Metabolic Programming of Macrophages: Implications in the Pathogenesis of Granulomatous Disease. *Front Immunol*. 2019;10:2265.
11. Barrett T, et al. NCBI GEO: archive for functional genomics data sets — update. *Nucleic Acids Research*. 2013;41:D991–D995.
12. Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology*. 2010;74(3):201–209.
13. Gandhi KS, McKay FC, Cox M, Riveros C, Armstrong N, Heard RN, et al. The multiple sclerosis whole blood mRNA transcriptome and genetic associations indicate dysregulation of specific T cell pathways in pathogenesis. *Hum Mol Genet*. 2010;19(11):2134–2143.
14. Zhou T, Zhang W, Sweiss NJ, Chen ES, Moller DR, Knox KS, et al. Peripheral blood gene expression as a novel genomic biomarker in complicated sarcoidosis. *PLoS One*. 2012;7(9):e44818.

15. Novais FO, Carvalho LP, Passos S, Roos DS, Carvalho EM, Scott P, et al. Genomic profiling of human *Leishmania braziliensis* lesions identifies transcriptional modules associated with cutaneous immunopathology. *J Invest Dermatol*. 2015;135(1):94–101.
16. Rogers D, Hahn M. Extended-Connectivity Fingerprints. *J Chem Inf Model*. 2010;50(5):742–754.
17. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. *Nat Chem*. 2012;4(2):90–98.
18. Wildman SA, Crippen GM. Prediction of Physicochemical Parameters by Atomic Contributions. *J Chem Inf Comput Sci*. 1999;39(5):868–873.
19. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem*. 2000;43(20):3714–3717.
20. Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B – Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans Math Softw*. 1997;23(4):550–560.

Acknowledgements

The authors thank the ADNI participants and investigators for their contributions of data. The ADNI project was funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering.

Author information

Authors and Affiliations

1. **Department of Biomedical Sciences, School of Allied Health Sciences, University of Cape Coast, Cape Coast, Ghana**

Evans Kwabena Abor, Pius Kwasi Sam, Emmanuel Adu Sarpong, George Ghartey-Kwansah & Francis Ackah Armah

2. **Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA**

Evans Kwabena Abor

Contributions

E.K.A. conceived and designed the study, developed the CausalDiffChem framework, performed the computational analyses, and wrote the manuscript. F.A.A. and G.G.-K. supervised the project and contributed to study design and interpretation. F.A.A. and G.G.-K. critically revised the manuscript. P.K.S. and E.A.S. contributed to the study and critically revised the manuscript.

Corresponding authors

Correspondence to Evans Kwabena Abor or Francis Ackah Armah.

Funding

This work received no specific funding.

Competing interests

The authors declare no competing interests.

Data availability

All transcriptomic datasets used in this study are publicly available. ADNI data are available through the ADNI database (registration required); GEO datasets are available under accessions GSE17048, GSE37912, and GSE55664. Compound libraries were obtained from Selleckchem (L1300) and ZINC15. Full ranked results for all compound libraries across the four disease modules are provided as a single Excel workbook in the Supplementary Information.

Code availability

All source code is released under an MIT licence at <https://github.com/aborek-abor/causaldiffchem> and archived at Zenodo (<https://doi.org/10.5281/zenodo.20602315>).

Supplementary Information

Supplementary Tables

**Disease transcriptomic data
(disease vs. healthy)**

MS · Sarcoidosis · Leishmaniasis · AD

Top-20 most variable genes
selected per disease

NOTEARS structure learning
→ disease-specific SCM ($\mathbf{W}$)
(Figure 2)

Compound libraries

FDA-approved · Preclinical · Clinical-stage

ECFP4 + ECFP6 fingerprint,
drug-likeness scale
→ perturbation vector $\delta\mathbf{T}$

Simulated $\text{do}(\cdot)$ intervention

$$\mathbf{v}_{\text{int}} = (\mathbf{I} - \mathbf{W} + \epsilon\mathbf{I})^{-1}(\mathbf{v}_{\text{dis}} + \delta\mathbf{T})$$

net_correction, %H, R scoring
and ranking
(Figures 3–5)

Per-disease, per-library
compound rankings

Gene-level restoration
profiles (Figure 5) and
cross-disease
convergence analysis

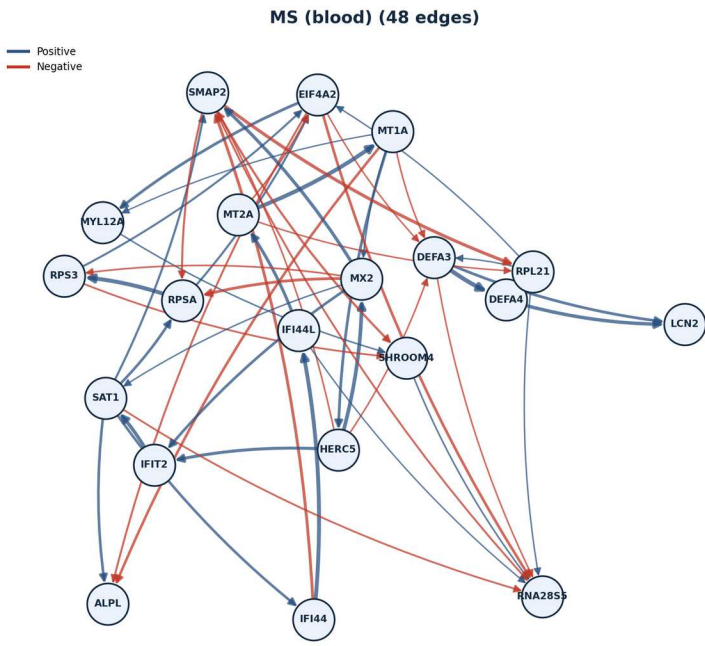
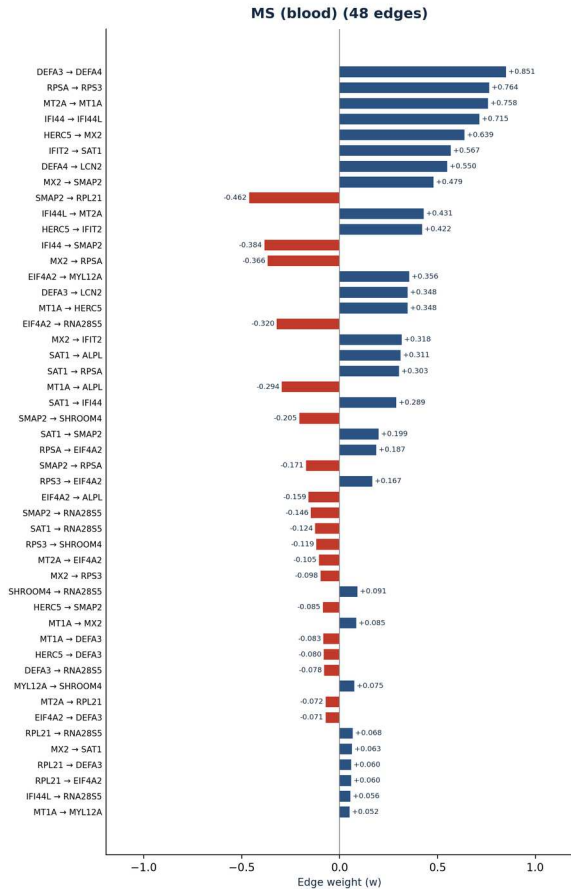
Positive-control check
against known clinical
therapeutics

SCM network structure

Edge weight distribution (ranked)

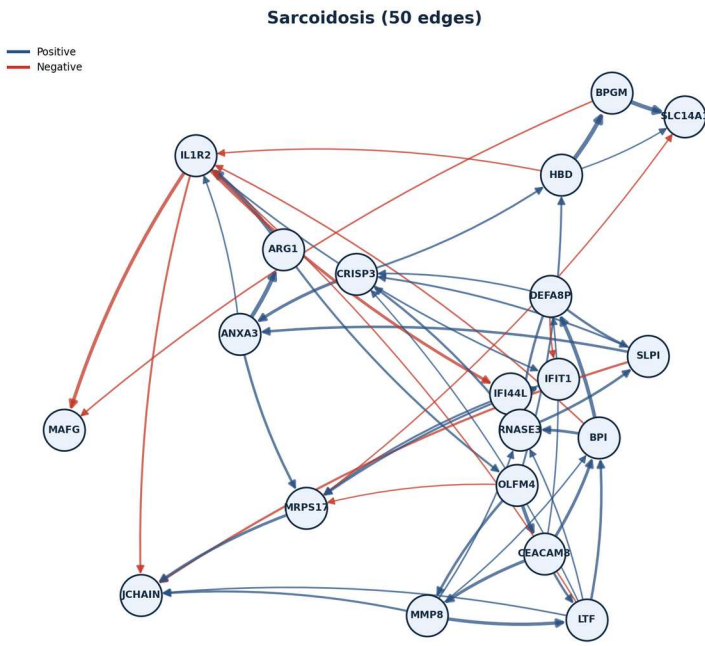
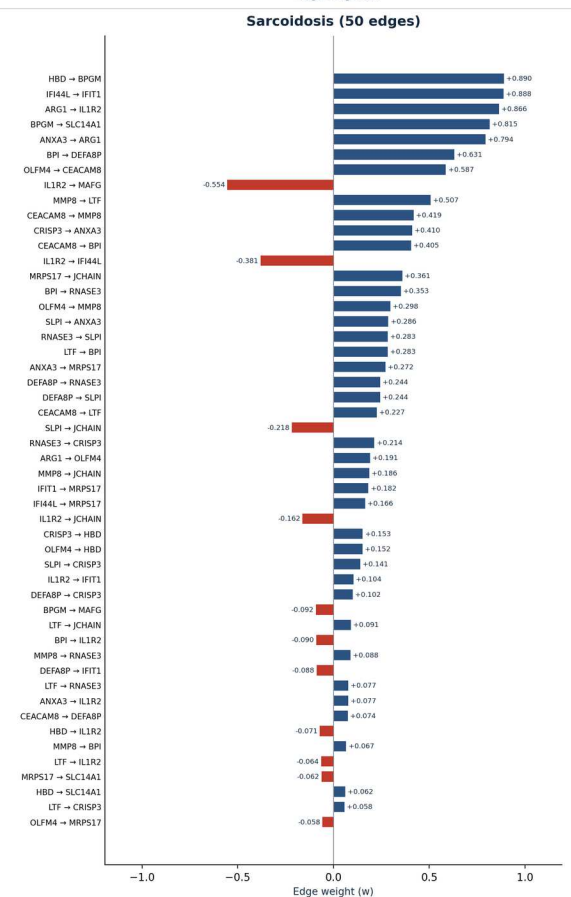
a

MS (blood)



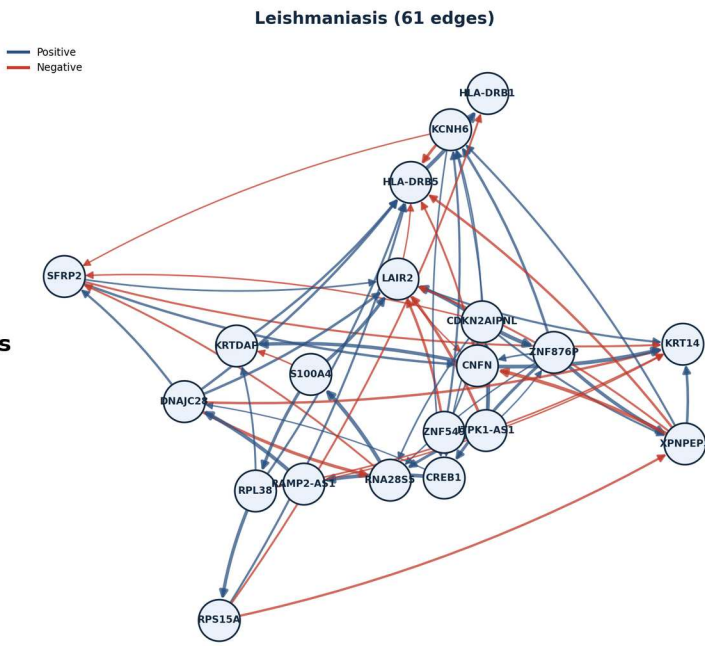
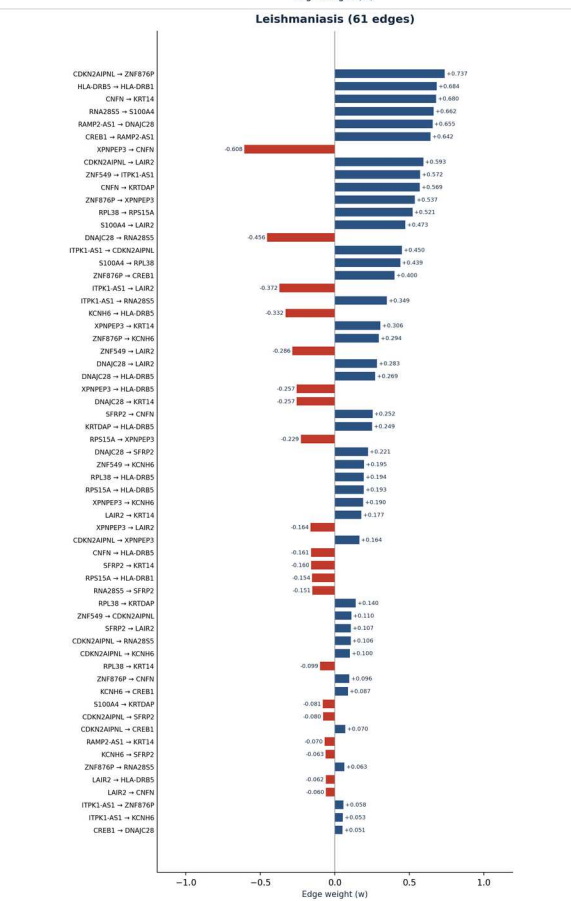
b

Sarcoidosis



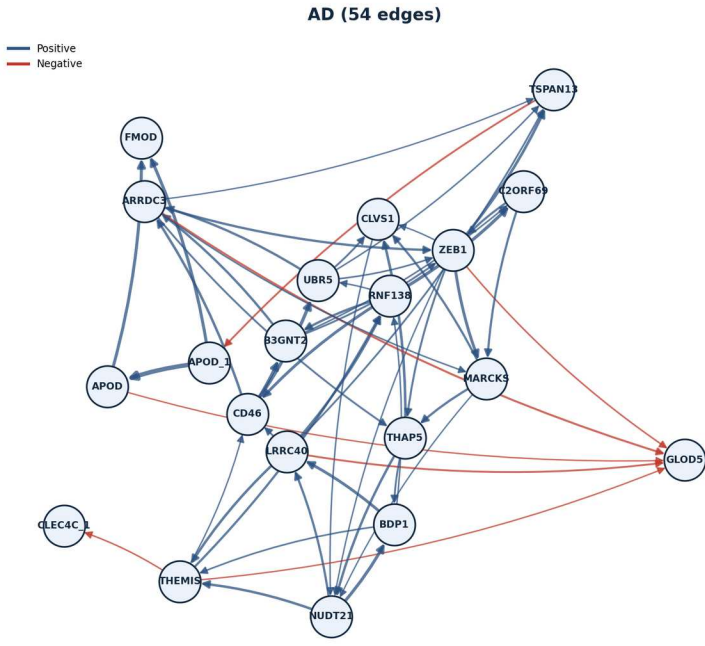
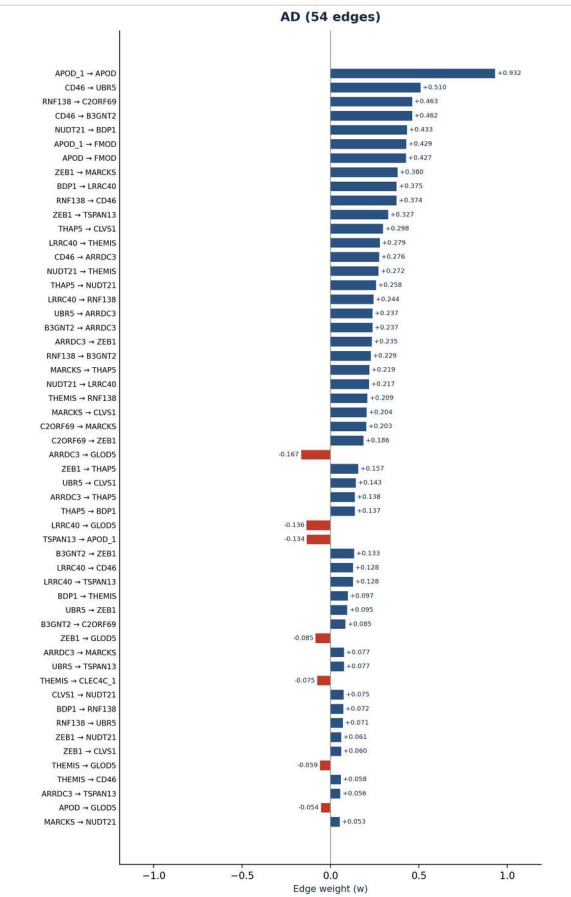
c

Leishmaniasis



d

AD

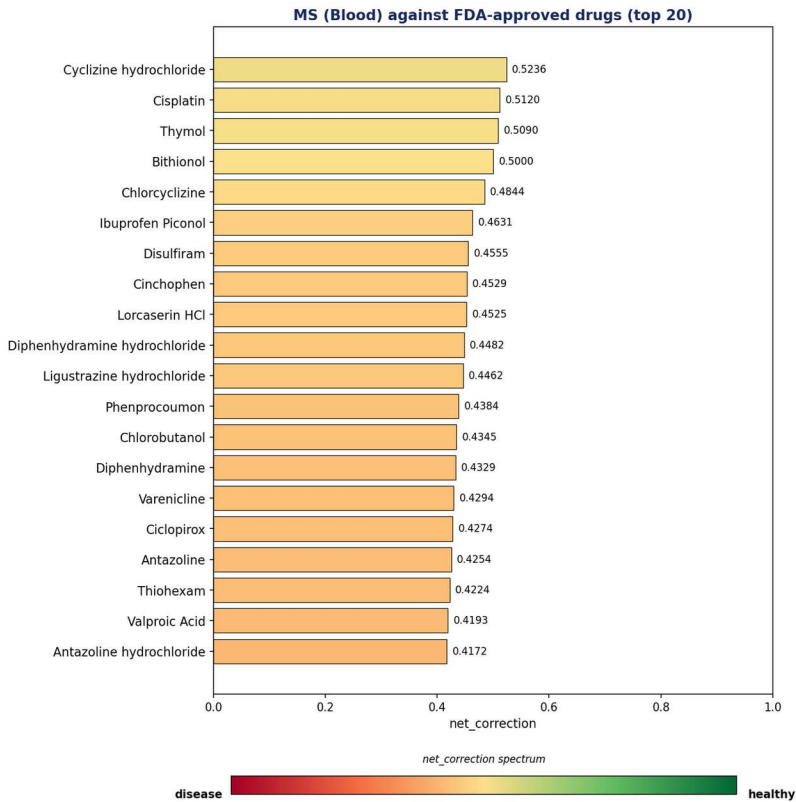
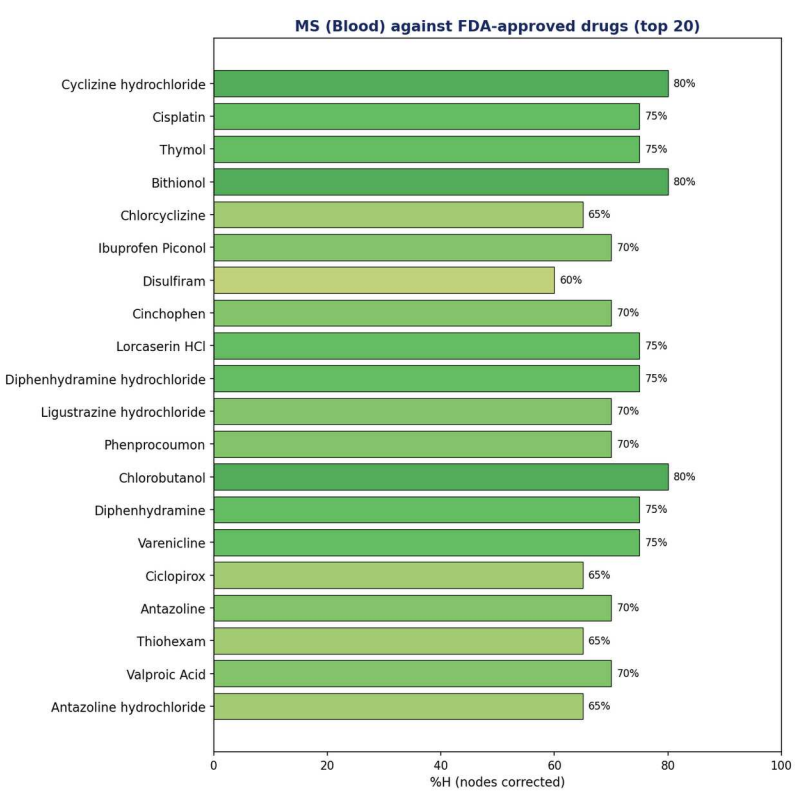


net_correction (magnitude)

%H (breadth)

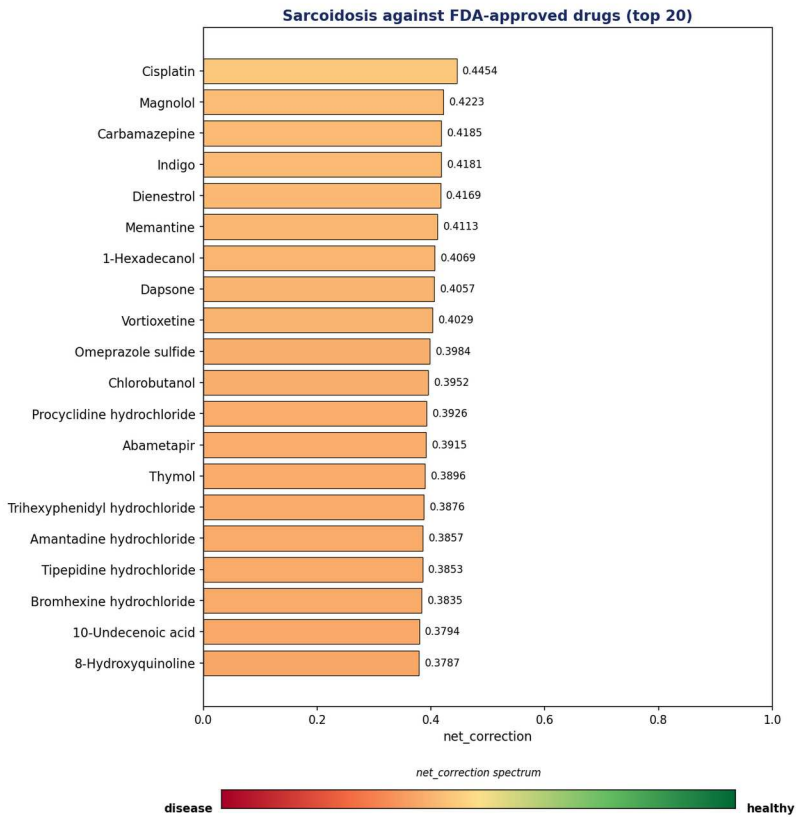
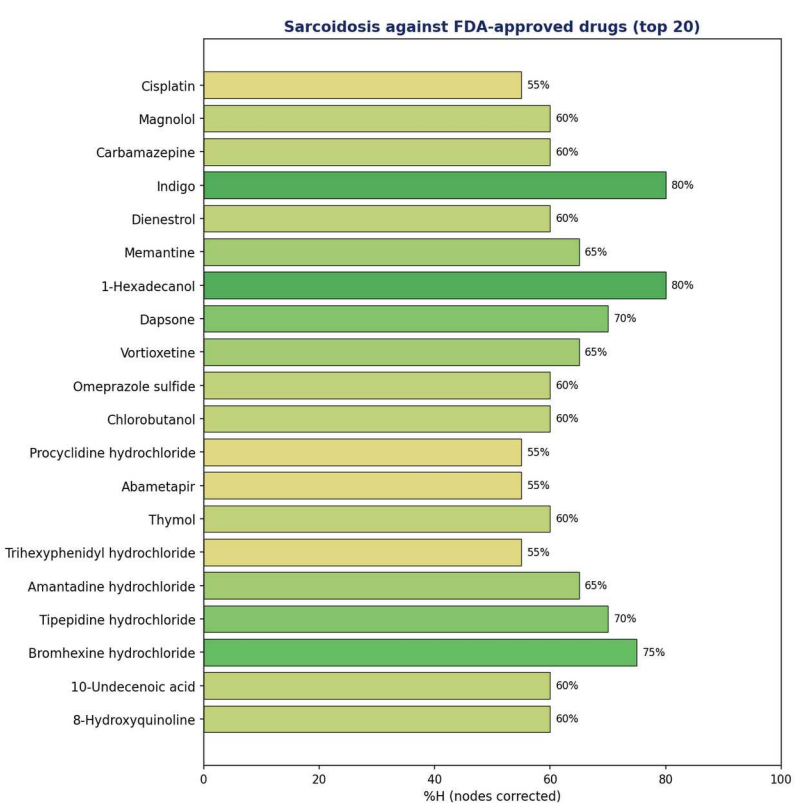
a

MS (blood)



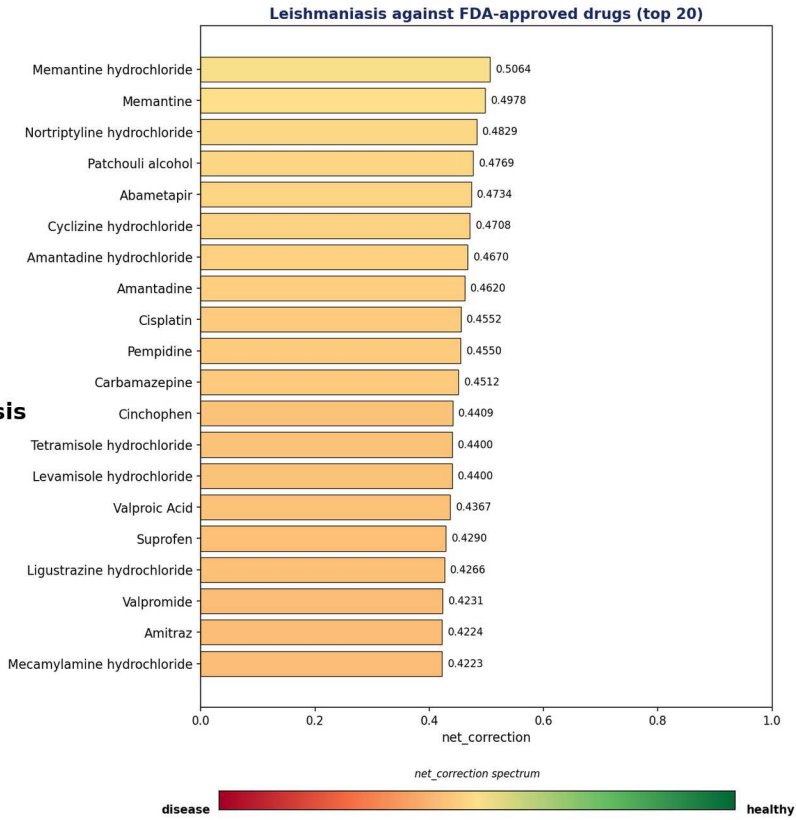
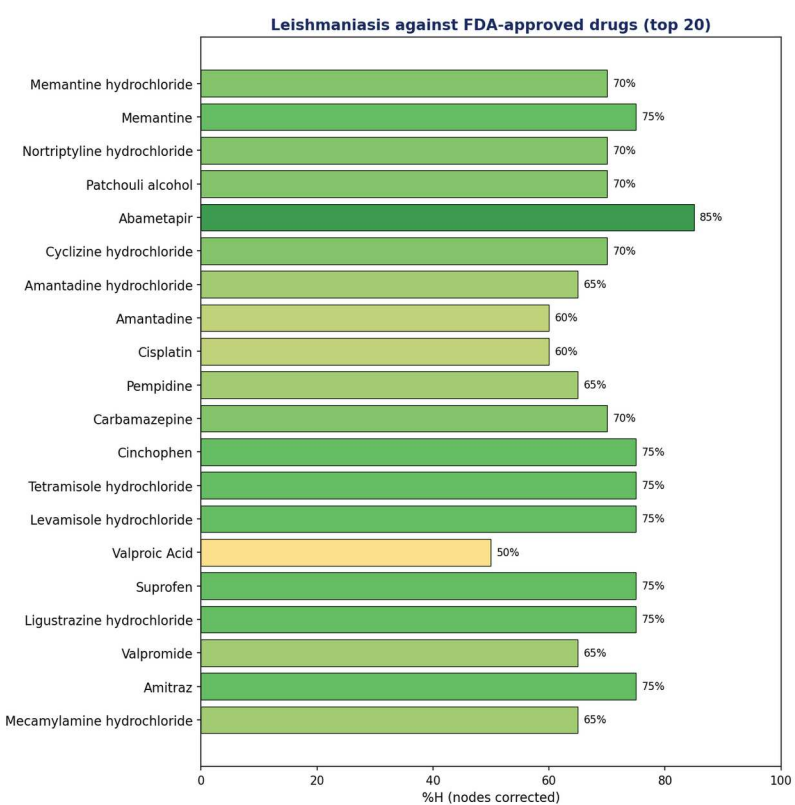
b

Sarcoidosis



c

Leishmaniasis



d

AD

