Discovery of disease treatment mechanisms through the multiscale interactome

Camilo Ruiz,[1,2] Marinka Zitnik,[3] Jure Leskovec[1,4,‡]
[1] Computer Science Department, Stanford University, Stanford, CA 94305, USA
[2] Bioengineering Department, Stanford University, Stanford, CA 94305, USA
[3] Biomedical Informatics Department, Harvard University, Boston, MA 02115, USA
[4] Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

‡Corresponding author. Email: jure@cs.stanford.edu

Most diseases disrupt multiple genes, and drugs treat such diseases by restoring the functions of the disrupted genes. How drugs restore these functions, however, is often unknown as a drug's therapeutic effects are not limited only to the genes that the drug directly targets. Here, we develop the multiscale interactome, a powerful approach for the discovery of disease treatment mechanisms. We integrate disease-perturbed genes, protein targets, and functional pathways into a multiscale interactome network, which contains 478,728 interactions between 1,661 drugs, 840 diseases, 17,660 proteins, and 9,798 functional pathways. We find that a drug's effectiveness can often be attributed to targeting genes that are distinct from disease-associated genes but that affect the same functional pathways. We develop a random walk-based method that captures how drug effects propagate through functional pathways in a multiscale manner and are coordinated by the protein-protein interaction network in which drugs act. On three key pharmacological tasks, we find that the multiscale interactome predicts what drugs will treat a given disease up to 40% better than prior approaches, reveals treatment mechanisms, and has the unique ability to explain how genetic mutations interfere with treatment mechanisms to cause drug resistance and serious adverse reactions. Our results indicate that molecular-scale interactomes (i.e., protein-protein interaction networks) alone are unable to explain the therapeutic effects of drugs as many drugs treat diseases by reinstating the functional pathways disrupted by the disease rather than directly targeting disease proteins or their regulators. We provide the first general framework for accurately identifying treatment mechanisms, even when drugs seem unrelated to the diseases they are recommended for.

Complex diseases, like cancer, disrupt dozens of proteins that interact in underlying biological networks [1–4]. Treating such diseases requires practical means to control the networks that underlie the disease [5–7]. By targeting even a single protein, a drug can affect hundreds of proteins in the underlying biological network. To achieve this effect, the drug relies on physical interactions between proteins. The drug binds a target protein, which physically interacts with dozens of other proteins, which in turn interact with dozens more, eventually reaching the proteins disrupted by the disease [8–10]. Networks capture such interactions and are a powerful paradigm to investigate the intricate effects of disease treatments and how these treatments translate into therapeutic benefits, revealing insights into drug efficacy [10–15], side effects [16], and effective combinatorial therapies for treating the most dreadful diseases, including cancers and infectious diseases [17–19].

However, existing approaches assume that, for a drug to treat a disease, the proteins targeted by the drug need to be *close* to or even need to *coincide* with the disease-perturbed proteins [10–14] (Figure 1). As such, current approaches fail to capture functional pathways, through which target proteins can restore the functions of disease-perturbed proteins and thus treat a disease [20–25]. Moreover, current approaches are "black-boxes:" they predict treatment relationships but provide little biological insight into how treatment occurs.

Here we present the multiscale interactome, a powerful approach for the discovery of disease treatment mechanisms. We integrate disease-perturbed genes, protein targets and functional pathways in a multiscale interactome network. The multiscale interactome uses the physical interaction network between 17,660 proteins, which we augment with 9,798 functional pathways, in order to fully capture the fundamental biological principles of effective treatments across 1,661 drugs and 840 diseases.

To discover how a drug treats a disease, our approach uses biased random walks which model how drug effects spread through functional pathways in a multiscale manner and are coordinated by the protein-protein interaction network in which drugs act. In the multiscale interactome, drugs treat diseases by propagating their effects through a network of physical interactions between proteins and a hierarchy of functional pathways spanning multiple scales, ranging from cells to tissues and organs. For each drug and disease, we learn a diffusion profile, which identifies the key proteins and functional pathways involved in a given treatment. By comparing drug and disease diffusion profiles, the multiscale interactome provides an interpretable basis to identify the proteins and functional pathways that explain successful treatments.

We demonstrate the power of the multiscale interactome on three key tasks in pharmacology. First, we find the multiscale interactome predicts which drugs can treat a given disease up to 40% more accurately than existing methods that rely only on the molecular-scale interactome, as captured by the protein-protein interaction network. This finding indicates that our approach accurately captures the functional pathways through which target proteins restore the functions of disease-perturbed proteins, even when drugs are distant to diseases they are recommended for. The multiscale interactome also improves prediction on entire drug classes, such as hormones, that rely on functional pathways and thus cannot be accurately represented by existing molecular-scale approaches. Second, we find that the multiscale interactome is a "white-box" method with the ability to reveal the mechanism for each predicted treatment. Finally, we find that the multiscale interactome predicts what genetic mutations cause resistance or serious adverse reactions for a given treatment and, unlike existing approaches, explains how these mutations interfere with the treatment mechanism.

Our results indicate that the failure of existing approaches is not due to algorithmic limitations but is instead fundamental. We find that a drug can treat a disease by influencing the behaviors of proteins that are *distant* from the drug's direct targets in the protein-protein interaction network. We find evidence that as long as those proteins participate in pathways that restore the functions disrupted by the disease proteins, the treatment can be successful. Thus, the molecular-scale interactome (i.e., the protein-protein interaction network) alone is unable to explain the therapeutic effects of drugs, and functional pathway information provides an important component for modeling treatment mechanisms. We provide a general framework for accurately identifying treatment mechanisms, even when drugs seem unrelated to the diseases they are recommended for.

Results

The multiscale interactome as a universal representation of the physiological effects of drugs and diseases. The multiscale interactome models drug treatment by integrating both physical interactions between proteins and a multiscale hierarchy of functional pathways. Crucially, most treatment mechanisms depend on functional pathways [20–24]. Existing network approaches, however, only model physical interactions between proteins [10–14], and thus cannot accurately model such treatments (Figure 1a).

Our multiscale interactome captures the fact that drugs and diseases exert their physiological effects through both proteins and functional pathways (Figure 1b). In particular, the multiscale interactome is a network in which 1,661 drugs interact with the proteins they primarily target (8,568 edges) [26, 27] and 840 diseases interact with the proteins they disrupt through genomic alterations, altered expression, or post-translational modification (25,212 edges) [28]. Subsequently, these protein-level effects propagate in two ways. First, 17,660 proteins physically interact with other proteins according to regulatory, metabolic, kinase-substrate, signaling, and binding relationships (387,626 edges) [29–35]. Second, these proteins alter 9,798 functional pathways according to a rich hierarchy spanning multiple scales: functional changes in proteins alter core cellular processes, which in turn affect the functions of tissues and organs, ultimately affecting phenotypes at the level of the whole body (34,777 edges between proteins, 22,545 edges between functional pathways; Gene Ontology) [36, 37]. By modeling the effect of drugs and diseases on both proteins and functional pathways, our multiscale interactome can model the range of drug treatment mechanisms that rely on both [20–24].

Overall, our multiscale interactome provides the largest systematic dataset to study drug-disease treatments. Nearly 6,000 approved treatments (i.e., drug-disease pairs) spanning almost every category of human anatomy are compiled [27, 38, 39], exceeding the largest prior network-based study by 10X [13] (Anatomical Therapeutic Classification; Supplementary Figure 1).

Propagation of the physiological effects of drugs and diseases through the multiscale interactome. To learn how physiological effects of drugs and diseases propagate through proteins and functional pathways, we harnessed network diffusion profiles (Figure 1c). A network diffusion profile propagates the effects of a drug or disease across the multiscale interactome, revealing the most affected proteins and functional pathways. The diffusion profile is computed by biased random walks that start at the drug or disease node. At every step, the walker can restart its walk or

jump to an adjacent node based on optimized edge weights. The diffusion profile $\mathbf{r} \in \mathbb{R}^{|V|}$ measures how often each node in the multiscale interactome is visited, thus encoding the effect of the drug or disease on every protein and functional pathway.

Diffusion profiles contribute three methodological advances. First, diffusion profiles provide a general framework to adaptively integrate proteins and functional pathways at different scales. When continuing its walk, the random walker jumps between proteins and functional pathways at different scales based on optimized edge weights. For drug and disease treatments, these edge weights encode the knowledge that proteins and functional pathways at different scales have different importance in the physiological effects of drugs and diseases [20,21]. By adaptively integrating both proteins and functional pathways across scales, therefore, diffusion profiles model physiological effects that rely on both.

Second, diffusion profiles provide a mathematical formalization of the principles governing how drug and disease effects propagate in a biological network. Drugs and diseases are known to generate their physiological effects by disrupting or binding to proteins which recursively affect other proteins and functional pathways. The effect propagates via two principles [8, 9, 40]. First, proteins and functional pathways closer to the drug or disease are affected more strongly. Similarly in diffusion profiles, proteins and functional pathways closer to the drug or disease are visited more often since the random walker is more likely to visit them after a restart. Second, the net effect of the drug or disease on any given node depends on the net effect on each neighbor. Similarly in diffusion profiles, a random walker can arrive at a given node from any neighbor.

Finally, comparing diffusion profiles provides a rich, interpretable basis to predict pharmacological properties. Traditional random walk approaches predict properties by measuring the proximity of drug and disease nodes [9]. By contrast, we compare drug and disease diffusion profiles to compare their physiological effects on proteins and functional pathways, a richer comparison. Our approach is thus consistent with recent machine learning advances which harness diffusion profiles to represent nodes [41].

The multiscale interactome accurately predicts which drugs treat a disease. By comparing the similarity of drug and disease diffusion profiles, the multiscale interactome predicts what drugs treat a given disease up to 40% more effectively than molecular-scale interactome approaches (AUROC 0.705 vs. 0.620, +13.7%; Average Precision 0.091 vs. 0.065, +40.0%; Recall@50 0.347 vs. 0.264, +31.4%) (Figure 2a, b, Methods). Moreover, the multiscale interactome accurately mod-

els entire classes of drugs that molecular-scale interactome approaches cannot since these drugs rely on functional pathways. Indeed, the top overall performing drug classes (i.e., sex hormones, modulators of the genital system; Supplementary Figure 2) and the top drug classes for which the multiscale interactome outperforms the molecular-scale interactome (i.e., pituitary, hypothalamic hormones and analogues; Figure 2c, Supplementary Figure 3) harness functional pathways beyond the molecular scale. For example, Vasopressin, a pituitary hormone, treats urinary disorders by binding receptors which trigger smooth muscle contraction in the gastrointestinal tract, free water reabsorption in the kidneys, and contraction in the vascular bed [26, 42, 43]. Vasopressin's treatment mechanism, and that of pituitary and hypothalamic hormones more broadly, relies on functional pathways beyond the molecular scale that are modeled by the multiscale interactome.

The multiscale interactome systematically reveals the proteins and functional pathways used by a drug to treat a disease. Existing approaches are "black boxes:" they predict what drug treats a disease but cannot explain how the drug treats the disease through specific proteins and functional pathways [10–15] (Figure 2d). By contrast, drug and disease diffusion profiles highlight the proteins and functional pathways harnessed in treatment (Figure 2e). For a given drug and disease, we define the treatment mechanism as the k most frequently visited nodes in the drug and disease diffusion profiles which correspond to the proteins and functional pathways most affected by the drug and disease.

Gene expression signatures validate our drug-disease treatment mechanisms (Figure 2f). We find that drugs with more similar treatment mechanisms have more similar gene expression signatures (Spearman $\rho = 0.392$, $p = 5.8 \times 10^{-7}$, $n = 152$) [44,45], indicating that our derived treatment mechanisms reflect the proteins and functional pathways harnessed by drugs.

Case studies further validate our network treatment mechanisms. Consider the treatment of Hyperlipoproteinemia Type III by Rosuvastatin (i.e., Crestor). In Hyperlipoproteinemia Type III, defects in apolipoprotein E (APOE) [46–48] and apolipoprotein A-V (APOA5) [49, 50] lead to excess blood cholesterol, eventually leading to the onset of severe arteriosclerosis [47]. Rosuvastatin is known to treat Hyperlipoproteinemia Type III by inhibiting HMG-CoA reductase (HMGCR) and thereby diminishing cholesterol production [51, 52]. Crucially, the network-derived treatment mechanism replicates these key steps (Figure 2g). Notably, there is no direct path of proteins between Hyperlipoproteinemia and Rosuvastatin. Instead, treatment operates through functional pathways (i.e., cholesterol biosynthesis and its regulation). Consistently, the multiscale interac-

tome identifies Rosuvastatin as a treatment for Hyperlipoproteinemia far more effectively than a molecular-scale interactome approach, ranking Rosuvastatin 72nd out of 1,661 drugs rather than 1,207th. The multiscale interactome explains treatment mechanisms that rely on functional pathways, a feat which molecular-scale interactome approaches cannot accomplish.

Similarly, consider the treatment of Cryopyrin-Associated Periodic Syndromes by Anakinra. In Cryopyrin-Associated Periodic Syndromes, mutations in NLRP3 and MME lead to immune-mediated inflammation through the Interleukin-1 beta signaling pathway [53]. Anakinra treats Cryopyrin-Associated Syndromes by binding IL1R1, a receptor which mediates regulation of the Interleukin-1 beta signaling pathway and thus prevents excessive inflammation [26,54]. Again, our network-derived treatment mechanism replicates these key steps (Figure 2h). Crucially, the treatment mechanism involves the regulation of inflammation and immune system signaling, complex functional pathways which are not modelled by molecular-scale interactome approaches. Again, the multiscale interactome identifies Anakinra as a treatment for CAPS far more effectively than a molecular-scale interactome approach, ranking Anakinra 181st out of 1,661 drugs (molecular-scale interactome ranks it 1,192nd).

The multiscale interactome identifies genetic mutations that cause patient-specific drug resistance and adverse reactions. A key goal of precision medicine is to understand how genetic mutations cause patient-specific drug resistance and adverse reactions [55] (Figure 3a). However, current tools of precision medicine cannot explain how mutations interfere with the treatment mechanism [56], and thus the treatment-altering mutations are not well understood.

We find that mutations cause drug resistance and adverse reactions by targeting genes of high importance in the treatment mechanism (Figure 3b). We define the treatment importance of a genetic mutation according to the visitation frequency of the corresponding protein in the drug and disease diffusion profiles (Methods). Mutations that cause drug resistance and adverse reactions exhibit substantially higher treatment importance scores than other mutations (median network importance = 0.912 vs. 0.513; $p = 2.95 \times 10^{-107}$, Mood's median test), indicating that these treatment altering mutations occur at highly visited nodes. We thus provide the first evidence that the topological position of a genetic mutation influences its ability to cause drug resistance or serious adverse reactions.

We find that the network importance of a genetic mutation in the drug and disease diffusion profiles predicts whether that mutation will cause drug resistance and adverse reactions for

that particular treatment (AUROC = 0.79, Average Precision = 0.82) (Figure 3c). Our diffusion profiles thereby provide a systematic approach to identify mutations likely to alter treatment, with implications for patient stratification in clinical trials and personalized therapeutic selection [57].

Finally, we find that when a mutation in a diseased patient causes resistance to one indicated drug but not another, that mutation primarily targets the treatment mechanism of the resistant drug (Figure 3d, e). Overall, 71.0% of the genes known to cause resistance to one indicated drug but not another exhibit higher network importance in the altered treatments than in the unaltered treatment. We thus provide a network formalism explaining how genetic mutations can cause resistance and adverse reactions in only some drugs indicated to treat a disease.

Consider Benazepril and Diltiazem, two drugs indicated to treat Hypertensive Disease (Figure 3f). A mutation in the AGXT gene leads to drug resistance against Benazepril [58–60] but not Diltiazem [60]. Indeed, our approach gives higher treatment importance to AGXT in treatment by Benazepril than in treatment by Diltiazem, ranking AGXT as the 21st most important gene for Benazepril treatment but only the 2,425th most important gene for Diltiazem treatment. Moreover, our approach explains why AGXT causes resistance to Benazepril but not Diltiazem (Figure 3f). Diltiazem primarily operates at a molecular-scale, inhibiting various calcium receptors (CACNA1S, CACNA1C, CACNA2D1, CACNG1) which trigger relaxation of the smooth muscle lining blood vessels and thus lowers blood pressure [26, 61–63]. By contrast, Benazepril operates at a systems-scale: Benazepril binds to ACE which affects the renin-angiotensin system, a systems-level functional pathway that controls blood pressure through hormones [26, 64, 65]. A mutation in AGXT directly affects ACE, thus impairing the ability of Benazepril to influence the renin-angiotensin system and treat Hypertensive Disease. By contrast, AGXT plays no role in the treatment paths used by Diltiazem. The multiscale interactome can thus explain how genetic mutations interfere with treatment mechanisms to generate drug resistance.

Discussion

The multiscale interactome provides a general approach to systematically understand how drugs treat diseases. By integrating physical interactions and functional pathways, the multiscale interactome improves prediction of what drugs will treat a disease by up to 40% over physical interactome approaches [10, 13]. Moreover, the multiscale interactome systematically reveals the mechanism harnessed by a drug to treat a disease. By contrast, existing methods are "black-boxes" which make predictions without providing mechanistic insight. Finally, the multiscale interactome predicts what mutations will cause resistance or severe adverse reactions for drug treatments. Unlike other methods, the multiscale interactome explains how these mutations interfere with treatment mechanisms.

Our study provides three important implications for systems biology. (1) First, we find that integrating functional pathways into the interactome captures new biological phenomena. Historically, the interactome has modeled only physical interactions between proteins [8–10, 13]. We find drug treatments whose mechanisms critically depend on functional pathways. More broadly, however, these functional pathways are known to be crucial in disease pathogenesis [66–69], disease comorbidities [6], and drug combinations [22–24]. Harnessing the multiscale interactome in these settings may help us answer key pharmacological questions.

(2) Second, we find that diffusion profiles provide both predictive power and interpretability in modeling biological phenomena. Diffusion profiles reveal the proteins and functional pathways harnessed in treatment. Diffusion profiles also explain how genetic mutations affect these treatment mechanisms to cause drug resistance and adverse reactions. They, however, apply broadly in pharmacological studies. For example, in disease protein prediction [70], diffusion profiles may both predict new disease proteins and explain how these proteins are implicated in disease pathogenesis. In studying the efficacy [17] and adverse reactions [16] of drug combinations, diffusion profiles may identify both the functional pathways utilized by drug combinations to treat a disease and the functional pathways which generate polypharmacy adverse reactions. In disease comorbidities [6, 29], diffusion profiles may predict new comorbidities and identify the functional pathways in the initial disease which contribute to the development of the comorbidity.

(3) Finally, our study shows how physical interactions and functional pathways propagate the effects of proteins and ultimately translate them to phenotypes. We find that many drugs neither directly target the proteins associated with the disease they treat nor target proximal proteins. Instead, these drugs affect similar functional pathways which ultimately counteract the functional

effects of the disease. This view expands upon the current view of indirect effects embraced in other biological phenomena. In the omnigenic model of complex disease [71, 72], for example, hundreds of genetic variants affect a complex phenotype through indirect effects that propagate through a regulatory network of physical interactions. Our results suggest that the multiscale interactome with functional pathways spanning cells, tissues, organs, and organ systems can capture these indirect effects in complex disease. Altogether, the multiscale interactome provides a general computational paradigm for network medicine.

Data availability. All data used in the paper, including the multiscale interactome, approved drug-disease treatments, drug and disease classifications, gene expression signatures, and pharmacogenomic relationships is available at github.com/snap-stanford/multiscale-interactome.

Code availability. Python implementation of our methodology is available at github.com/snap-stanford/multiscale-interactome. The code is written in Python. Please read the README for information on downloading and running the code.

Author contribution. C.R., M.Z., and J.L. designed research; C.R., M.Z., and J.L. performed research; C.R., M.Z., and J.L. analyzed data; and C.R., M.Z., and J.L. wrote the paper.

Author information. The authors declare no conflict of interest. Correspondence should be addressed to J.L. (jure@cs.stanford.edu).

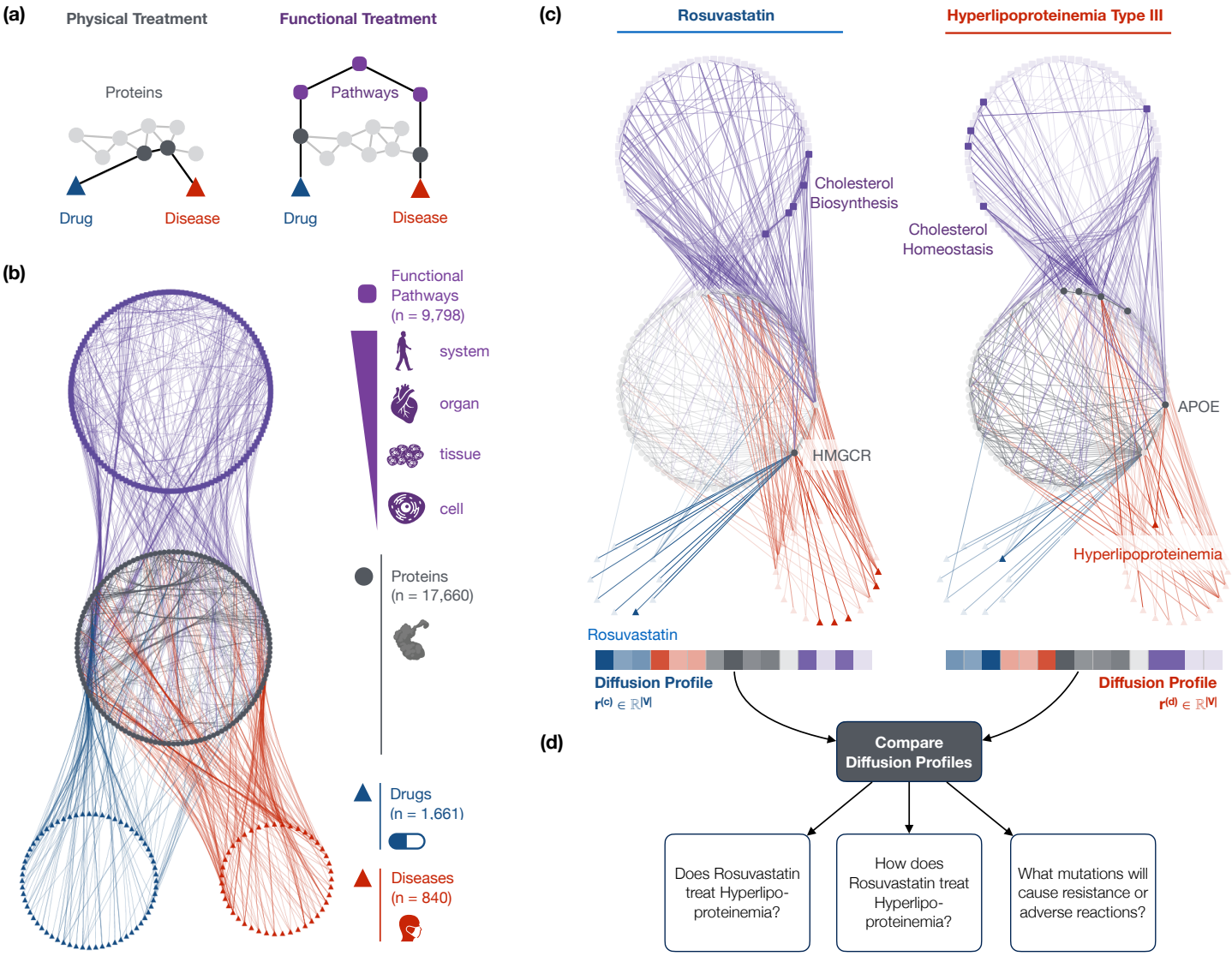

Figure 1: The multiscale interactome models drug treatment through both proteins and functional pathways. (a) Existing network approaches assume that drugs treat diseases by targeting proteins that are proximal to disease proteins in a network of physical interactions [10–14]. However, drugs can also treat diseases by targeting distant proteins that affect the same functional pathways [20–25]. **(b)** The multiscale interactome models drug-disease treatment by integrating both proteins and functional pathways. **(c)** The diffusion profile of a drug or disease captures its effect on every protein and functional pathway. The diffusion profile propagates the physiological effect of the drug or disease via random walks which adaptively explore proteins and functional pathways based on optimized edge weights. Ultimately, the visitation frequency of a node corresponds to the drug or disease's propagated effect on that node (Methods). **(d)** By comparing the diffusion profiles of a drug and disease, we compare their physiological effects on both proteins and functional pathways. Thereby, we predict whether the drug treats the disease (Figure 2a-c), explain how the drug treats the disease (Figure 2d-h), and identify which mutations cause drug resistance or dangerous adverse reactions (Figure 3). For example, Hyperlipoproteinemia Type III's diffusion profile reveals how defects in APOE affect cholesterol homeostasis, a hallmark of the excess blood cholesterol found in patients [46–50]. The diffusion profile of Rovustatin, a treatment for Hyperlipoproteinemia Type III, reveals how binding of HMG-CoA Reductase (HMGCR) reduces the production of excess cholesterol [51,52]. By comparing these diffusion profiles, we thus predict that Rosuvastatin treats Hyperlipoproteinemia Type III, identify the HMGCR and APOE-driven cholesterol metabolic pathways harnessed in treatment, and predict that mutations in APOE and HMGCR may interfere with the treatment mechanism and thus cause drug resistance or dangerous adverse reactions.

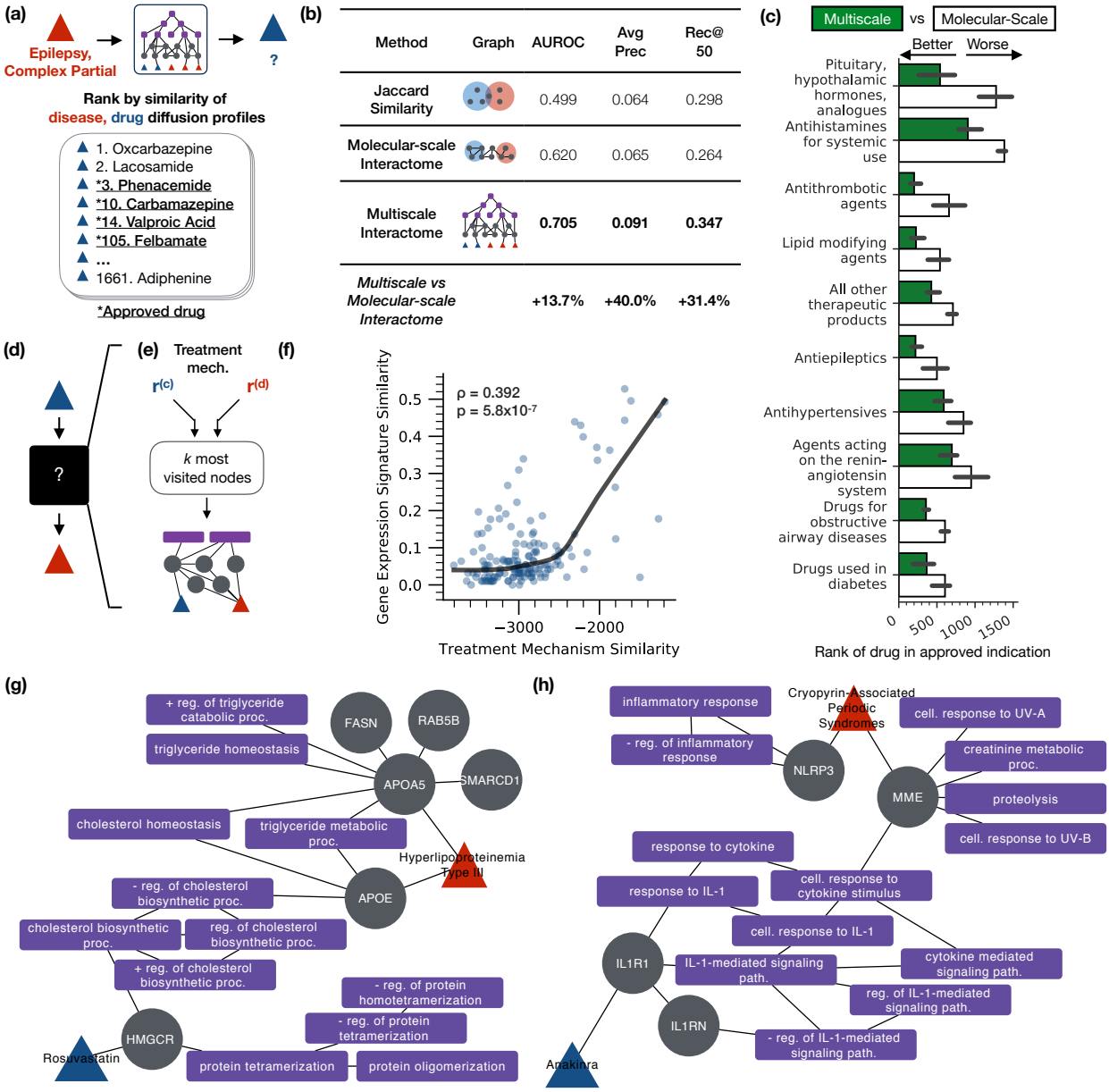

Figure 2: The multiscale interactome accurately predicts what drugs treat a disease and systematically reveals treatment mechanisms. **(a)** To predict whether a drug treats a disease, we compare the drug and disease diffusion profiles according to a correlation distance. **(b)** By incorporating both proteins and functional pathways, the multiscale interactome improves predictions of what drug will treat a given disease by up to 40% over molecular-scale interactome approaches [13]. Reported values are averaged across five-fold cross validation (Methods). **(c)** The multiscale interactome outperforms the molecular-scale interactome most greatly on drug classes that are known to harness functional pathways beyond the protein scale (i.e., pituitary, hypothalamic hormones and analogues; median and 95% CI shown). **(d)** Existing interactome approaches are "black boxes": they predict what drug treats a disease but do not explain how the drug treats the disease through specific biological pathways [10–15]. **(e)** By contrast, the diffusion profiles of a drug and disease reveal the proteins and functional pathways harnessed in treatment. For each drug and disease pair, the treatment mechanism consists of the top k most frequently visited nodes in the drug and disease diffusion profiles. **(f)** Drugs with more similar treatment mechanisms have more similar gene expression signatures (Spearman $\rho = 0.392$, $p = 5.8 \times 10^{-7}$, $n = 152$), suggesting that drug treatment mechanisms capture their physiological effects. **(g)** The multiscale interactome captures treatment mechanisms that molecular-scale interactome approaches cannot faithfully represent. Rosuvastatin treats Hyperlipoproteinemia Type III by binding to HMG CoA reductase (HMGCR) which drives a series of cholesterol biosynthetic pathways affected by Hyperlipoproteinemia Type III [46–52]. **(h)** Anakinra treats Cryopyrin-Associated Periodic Syndromes by binding to IL1R1 which regulates immune-mediated inflammation through the Interleukin-1 beta signaling pathway [26, 54]. Inflammation is a hallmark of Cryopyrin-Associated Periodic Syndromes [53].

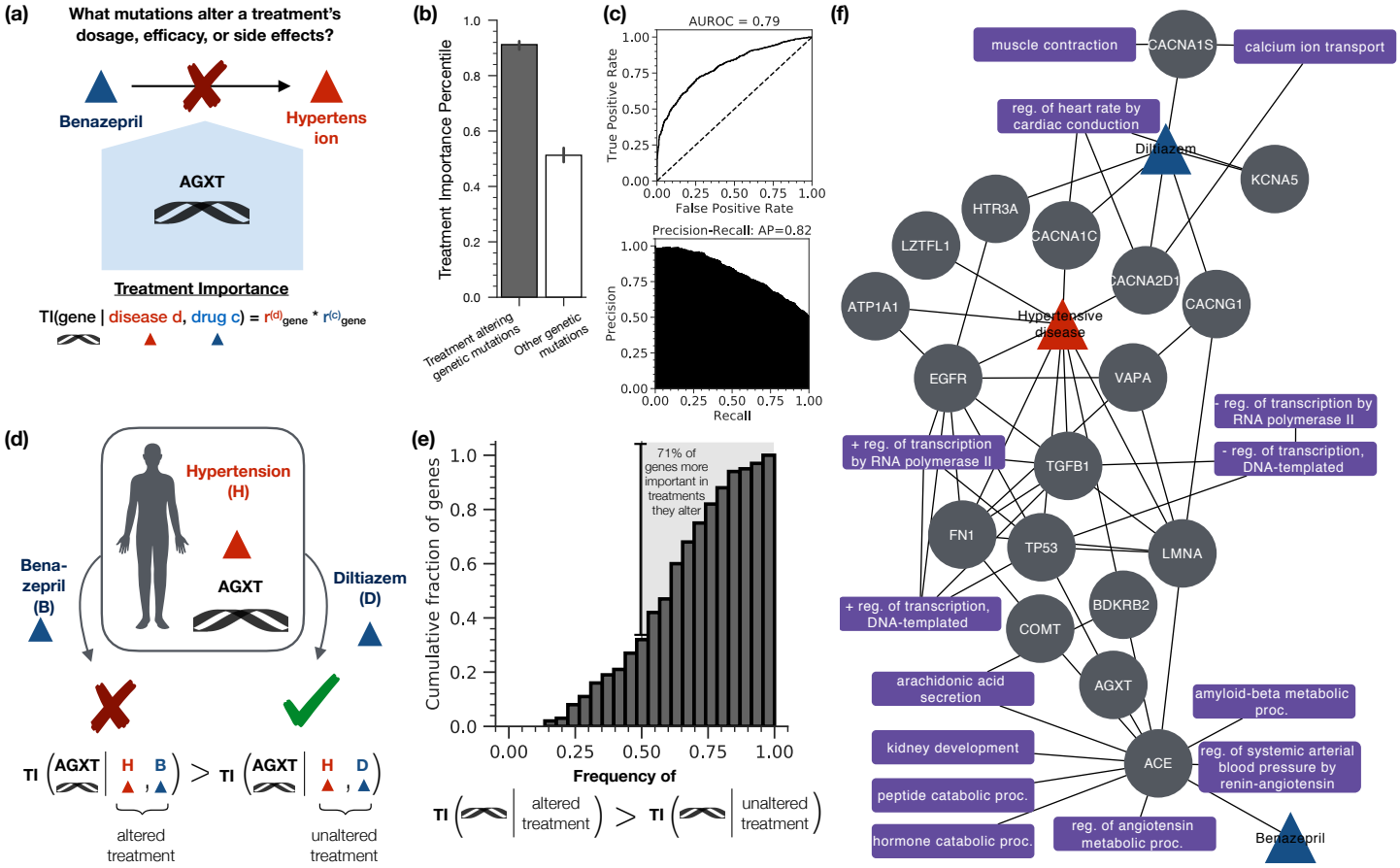

Figure 3: Diffusion profiles identify which genetic mutations cause drug resistance and serious adverse reactions and explain how these mutations interfere with the drug treatment mechanism. **(a)** Genetic mutations cause drug resistance and serious adverse reactions in a range of treatments [57]. A pressing need exists to systematically identify the mutations that cause resistance and serious adverse reactions for a given treatment and explain how these mutations interfere with the treatment mechanism [73]. **(b)** Mutations cause drug resistance and serious adverse reactions by targeting genes of high network importance in the treatment mechanism (median network importance of treatment altering genes = 0.912 vs. 0.513 $p = 2.95 \times 10^{-107}$; Mood's median test; median and 95% CI shown). We define the network treatment importance of a genetic mutation according to its visitation frequency in the drug and disease diffusion profiles (Methods). **(c)** The treatment importance of a mutation in the drug and disease diffusion profiles predicts whether that mutation will cause drug resistance and serious adverse reactions for that particular treatment (AUROC = 0.79, Average Precision = 0.82). **(d)** When a mutation in a diseased patient causes resistance to one indicated drug but not another, that mutation primarily targets the treatment mechanism of the resistant drug. In patients with Hypertensive Disease, a mutation in AGXT generates resistance to Benazepril but not Diltiazem. Indeed, AGXT exhibits a higher treatment importance in Benazepril treatment than in the Diltiazem treatment, ranked as the 21st most important gene rather than the 2,425th most important gene. **(e)** Overall, 71.0% of genes known to cause resistance to one indicated drug but not another exhibit higher treatment importance in the mechanism of the resistant drug. **(f)** Genes uniquely cause resistance in one indicated drug but not another by primarily targeting the resistant drug's treatment mechanism. Shown are the treatment mechanisms for Hypertensive Disease by Benazepril and Diltiazem. AGXT, which uniquely causes resistance in Benazepril, lies along the treatment mechanism of Benazepril but not that of Diltiazem.

Online Methods

The multiscale interactome. The multiscale interactome captures how drugs use both a network of physical interactions and a rich hierarchy of functional pathways to treat diseases. In the multiscale interactome, 1,661 drugs connect to the proteins they target (8,568 edges) [26, 27]. 840 diseases connect to the proteins they disrupt through genomic alterations, altered expression, or post-translational modification (25,212 edges) [28]. 17,660 proteins connect to other proteins based on physical interactions such as regulatory, metabolic, kinase-substrate, signaling, or binding relationships (387,626 edges) [29–35]. Proteins connect to the 9,798 functional pathways they affect (22,545 edges) [36, 37]. Finally, functional pathways connect to each other in a rich hierarchy spanning scales (22,545 edges) [36, 37].

We visualize a representative subset of the multiscale interactome using Cytoscape [74] (Figure 1b).

Drug-protein interactions. We map drugs to their protein targets using DrugBank [26] and the Drug Repurposing Hub [27]. For DrugBank, we map the Uniprot Protein IDs to Entrez IDs using HUGO [75]. For the Drug Repurposing Hub, we map drugs to their DrugBank IDs using the drug names and DrugBank's "drugbank_approved_target_uniprot_links.csv" file. We map protein targets to Entrez IDs using HUGO [75]. We filter drug-target relationships to only include proteins that are represented in the network of physical interactions between proteins (see Methods: Protein-protein interactions). All drug-target interactions are provided in Supplementary Data 1.

Disease-protein interactions. We map diseases to genes they affect through genomic alterations, altered expression, or post-translational modification by using DisGeNet [28]. To ensure high-quality disease-gene associations, we only consider the "curated" set of disease-gene associations provided by DisGeNet which draws from expert-curated repositories: UniProt, the Comparative Toxicogenomics Database, Orphanet, the Clinical Genome Resource (ClinGen), Genomics England PanelApp, the Cancer Genome Interpreter (CGI), and the Psychiatric Disorders Gene Association Network (PsyGeNET). We exclude all disease-gene associations that are inferred, based on orthology relationships from animal models, or based on computational-mining of the literature. Ultimately, diseases are associated with genes they affect via genomic alteration, alteration of expression, or post-translational modification according to the DisGeNet relationship ontology. To avoid circularity in the analysis, we remove disease-gene associations marked as therapeutic. Finally, we filter disease-gene relationships to only consider genes whose protein products were

present in the network of physical interactions between proteins (see Methods: Protein-protein interactions). All disease-protein interactions are provided in Supplementary Data 2.

Protein-protein interactions. We generate a network of 387,626 physical interactions between 17,660 proteins by compiling seven major databases. Across all databases, we only consider human proteins and their interactions; only allow protein-protein interactions with direct experimental evidence; and only allow *physical* interactions between proteins, filtering out genetic and indirect interactions between proteins such as those identified via synthetic lethality experiments. All protein-protein interactions are provided in Supplementary Data 3.

1. *The Biological General Repository for Interaction Datasets [30]* (BioGRID; 309,187 interactions between 16,352 proteins). BioGRID manually curates both physical and genetic interactions between proteins from 71,713 high- and low-throughput publications. We map BioGRID proteins to Entrez IDs by using HUGO [75]. We only include protein-protein interactions from BioGRID that result from experiments indicating a *physical* interaction between the proteins, as described by BioGRID [30], and ignore protein-protein interactions indicating a *genetic* interaction between the proteins. We use the "BIOGRID-ORGANISM-Homo_sapiens-3.5.178.tab" file.

2. *The Database of Interacting Proteins [32]* (DIP; 4,235 interactions between 2,751 proteins). DIP only considers physical protein-protein interactions with experimental evidence and curates these from the literature. We map the UniProt ID of each protein to its Entrez ID by using HUGO [75]. We allow all experimental methods from DIP since they all capture physical interactions [32].

3. *The Human Reference Protein Interactome Mapping Project*. We integrate four protein-protein interaction networks from the Human Reference Protein Interactome Mapping Project that were generated through high-throughput yeast two hybrid assays (HI-I-05 [35]: 2,611 interactions between 1,522 proteins; HI-II-14 [31] 13,426 interactions between 4,228 proteins; Venkatesan-09 [33]: 233 interactions between 229 proteins; Yu-11 [34] 1,126 interactions between 1,126 proteins). Since protein-protein interactions in all four networks result from a yeast two-hybrid system, all protein-protein interactions are physical and experimentally verified. We thus include all protein-protein interactions across these networks. Proteins are already provided with their Entrez ID so no mapping is required.

4. *Menche-2015 [29]* (138,425 interactions between 13,393 proteins). Finally, we integrate the physical protein-protein interaction network compiled by Menche et al. [29]. Menche et al. compiles different types of physical protein-protein interactions from a range of sources. In all cases, protein-protein interactions result from direct experimental evidence. Menche et al. compiles regulatory interactions from the TRANSFAC database; binary interactions from a series of high-throughput yeast-two-hybrid datasets as well as the IntAct and MINT databases; literature curated interactions from IntAct, MINT, BioGRID, and HPRD; metabolic-enzyme coupled interactions from KEGG and BIGG; protein complex interactions from CORUM; kinase-substrate interactions from PhosphositePlus; and signaling interactions from [76]. All proteins are provided in Entrez format and thus do not require further mapping.

Protein-pathway interactions. We map proteins to the functional pathways they affect by using the human version of the Gene Ontology [36, 37] (7,993 proteins; 6,387 functional pathways; 34,777 edges). We only allow experimentally verified associations between genes and functional pathways according to the following IDs: EXP – inferred from experiment, IDA – inferred from direct assay, IMP – inferred from mutant phenotype, IGI – inferred from genetic interaction, HTP – high throughput experiment, HDA – high throughput direct assay, HMP – high throughput mutant phenotype, and HGI – high throughput genetic interaction. We exclude any protein-functional pathway relationships that are inferred from physical interactions to avoid redundancy with the physical network of interacting proteins. We also exclude protein-functional pathway relationships inferred from gene expression patterns since the Gene Ontology states that such interactions are challenging to map to specific proteins [36, 37]. To prevent circularity, we further ignore all associations based on phylogenetically inferred annotations or various computational analyses (sequence or structural similarity, sequence orthology, sequence alignment, sequence modeling, genomic context, reviewed computational analysis). Finally, we ignore associations based on author statements, curator inference, electronic annotations (i.e. automated annotations), and those for which no biological data was available. Some functional pathways in the Gene Ontology have multiple synonymous IDs. For each functional pathway, we use the "master IDs" provided by GOATOOLS [77]. All protein - functional pathway interactions are provided in Supplementary Data 4.

Pathway-pathway interactions. We construct a hierarchy of functional pathways across scales by

using the Gene Ontology's Biological Processes [36, 37]. The Gene Ontology represents a curated hierarchy of functional pathways, where highly specific functional pathways are children of more general functional pathways according to numerous relationship types. For example, "negative regulation of response to interferon-gamma" $\xrightarrow{\text{is a}}$ "negative regulation of innate immune response" $\xrightarrow{\text{is a}}$ "negative regulation of immune response" $\xrightarrow{\text{negatively regulates}}$ "immune response." We allow relationships between functional pathways of the following types: regulates, positively regulates, negatively regulates, part of, and is a. Finally, we only consider functional pathways which are associated with at least one drug target or one disease protein, either directly or implicitly through their children. All functional pathway - functional pathway interactions are provided in Supplementary Data 5.

Constructing dataset of approved drug-disease treatments. We construct a dataset of 5,926 unique, approved drug-disease pairs, exceeding the largest prior network-based study by 10X [13]. We source approved drug-disease pairs from the Drug Repurposing Database [38] ($n_{pairs} = 2,538$; $n_{drugs} = 996$, $n_{diseases} = 463$), the Drug Repurposing Hub [27] ($n_{pairs} = 1,449$; $n_{drugs} = 908$, $n_{diseases} = 265$), and the Drug Indication Database [39] ($n_{pairs} = 3,304$; $n_{drugs} = 1,147$, $n_{diseases} = 615$). In all cases, we filter drug-disease pairs to ensure that only FDA-approved treatment relationships are included.

We extract approved drug-disease pairs from each database as follows. In all cases, drugs are mapped to DrugBank IDs [26] and diseases are mapped to unique identifiers from the National Library of Medicine [78] (NLM UMLS CUIDs: NLM Unified Medical Language System Controlled Unique Identifier):

1. *The Drug Repurposing Database* is a gold-standard database of drug-disease pairs extracted from drug labels and the American Association of Clinical Trials Database [38]. Drugs and diseases in the Drug Repurposing Database are provided with DrugBank IDs and NLM UMLS CUIDs so no additional mapping is required. We extract only the drug and disease pairs designated as "Approved" treatment relationships.

2. *The Broad Institute's Drug Repurposing Hub* is a hand-curated collection of drug-disease pairs compiled from drug labels, DrugBank, the NCATS NCGC Pharmaceutical Collection (NPC), Thomson Reuters Integrity, Thomson Reuters Cortellis, Citeline Pharmaprojects, the FDA Orange Book, ClinicalTrials.gov, and PubMed [27]. We map drugs to DrugBank IDs by comparing their provided names and PubChem IDs to DrugBank's external links

mapping [26]. We map diseases to UMLS CUIDs by using the UMLS Metathesaurus's REST API [78]. Finally, we only include drug-disease pairs with a "Launched" clinical phase attribute, indicating FDA approval.

3. *The Drug Indication Database* provides drug-indications relationships from DailyMed, DrugBank, the Pharmacological Actions sections of the Medical Subject Headings, the National Drug File Reference Terminology, the Physicians' Desk Reference, the Chemical Entities of Biological Interest (ChEBI), the Comparative Toxicogenomics Database, the Therapeutic Claims section of the USP Dictionary of United States Adopted Names and International Drug Names, and the World Health Organization Anatomic-Therapeutic-Chemical classification) [39]. The Drug Indication Database captures both diseases and non-disease medical conditions (i.e. pregnancy) for which a drug is used. Additionally, the Drug Indication Database captures both treatment relationships between drugs and indications as well as prevention, management, and diagnostic relationships. We filter the Drug Indication Database to only include *approved* treatment relationships between drugs and *diseases*.

   We map drugs to DrugBank IDs by using the provided CAS and ChEBI IDs as well as DrugBank's external links mapping [26]. Indications are already provided with UMLS CUIDs.

   We filter indications to only include diseases in two ways. First, we only consider indications with a UMLS semantic type of "B2.2.1.2.1 Disease or Syndrome", "B2.2.1.2 Pathologic Function", or "B2.2.1.2.1.2 Neoplastic Process." Second, we only consider indications present in DisGeNet, a database mapping diseases to their associated genes [28].

   To ensure that drug-disease relationships specifically represent treatment relationships, we filter drug-disease pairs based on the "indication subtype." We remove drug-indication pairs where the indication subtype described is not treatment (i.e. preventative/prophylaxis, diagnosis, adjunct, palliative, reduction, causes/inducing/associated, and mechanism). We additionally remove all drug indication pairs from the Comparative Toxicogenomics Database (CTD). The goal of CTD is to provide broad chemical-disease associations published in the literature [79]. Concurrently, CTD does not subset these chemical-disease associations into drug-disease relationships that represent FDA-approved treatments.

   Finally, we remove overly broad diseases from the Drug Indication Database. We remove disease categories (i.e. diseases with "Diseases" in their name such as "Cardiovascular Dis-

eases" and "Metabolic Diseases). We also remove diseases with more than 130 approved drugs (i.e. Disorder of Eye – 290 approved drugs).

After compiling approved drug-disease treatment pairs, we remove treatments for which drugs rely on binding to non-human proteins (i.e. viral or bacterial proteins) to induce their effect. The multiscale interactome only models human proteins and functional pathways. The multiscale interactome is thus not designed to model treatments which rely on binding to viral or bacterial proteins. To remove such treatments, we map all disease UMLS CUIDs to their corresponding Disease Ontology ID [80]. We then remove diseases corresponding to the "disease by infectious agent category" of the Disease Ontology. The Disease Ontology does not map many UMLS CUIDs to corresponding Disease Ontology IDs. We thus manually curate the final list of diseases to remove additional infectious diseases: malaria, bacterial septicemia, fungal infection, coccidiosis, gonorrhea, gastrointestinal roundworms, shingles, lice, gastrointestinal parasites, tapeworm, syphilis, genital herpes, lungworms, fungicide, fungal keratosis, yeast infection, laryngitis, enterocolitis, protozoan infection, African trypanosomiasis, sepsis, Chagas disease, mites, bacterial vaginosis, scabies, pinworm, equine protozoal myeloencephalitis (EPM), microsporidiosis, and ringworm.

Finally, we filter approved drug-disease treatment pairs to only include drugs with at least one known target in DrugBank [26] or the Drug Repurposing Hub [27] and diseases with at least one associated gene in the curated version of DisGeNet [28] as these are the only drugs and diseases that the multiscale interactome represents (see Methods: Drug-protein interactions, Disease-protein interactions).

Ultimately, we achieve a dataset of 5,926 approved drug-disease pairs, exceeding the largest prior network-based study by 10X [13]. All approved drug-disease pairs are provided in Supplementary Data 6.

Learning drug and disease diffusion profiles. We propagate the effects of each drug and disease across the multiscale interactome by using network diffusion profiles. A drug or disease diffusion profile learns the proteins and functional pathways most affected by each drug or disease. Each drug or disease diffusion profile is computed through biased random walks that start at the drug or disease node. At every step, the random walker can restart its walk or jump to an adjacent node based on optimized edge weights. After many walks, the diffusion profile measures how often every node was visited, thus representing the effect of the drug or disease on that node.

By using optimized edge weights, diffusion profiles learn to adaptively inte-

grate proteins and functional pathways. Diffusion profiles rely on a set of scalar weights which encode the relative importance of different types of nodes: $W = \{w_{\text{drug}}, w_{\text{disease}}, w_{\text{protein}}, w_{\text{functional pathway}}, w_{\text{higher scale functional pathway}}, w_{\text{lower scale functional pathway}}\}$. These weights are hyperparameters which we optimize when predicting the drugs that treat a given disease (see Methods: Model selection and optimization of scalar weights). When a random walker continues its walk, it picks the next node to jump to based on the relative values of these weights. For example, if a random walker is at a protein and has both protein and functional pathway neighbors, it is $\frac{w_{\text{protein}}}{w_{\text{functional pathway}}}$ times more likely to jump to the protein neighbors than the functional pathway neighbors. Notice that proteins connect to drugs, diseases, proteins, and functional pathways, making $\{w_{\text{drug}}, w_{\text{disease}}, w_{\text{protein}}, w_{\text{functional pathway}}\}$ the relevant weights for a random walker currently at a protein. By contrast, functional pathways connect to proteins, higher scale functional pathways, and lower scale functional pathways, making $\{w_{\text{protein}}, w_{\text{higher scale functional pathway}}, w_{\text{lower scale functional pathway}}\}$ the relevant weights for a random walker at a functional pathway. By providing separate weights for higher and lower scale functional pathways, the random walker learns to explore different scales of the hierarchy of functional pathways and integrate them appropriately.

Diffusion profiles represent a general methodology to propagate signals through a heterogeneous biological network. By carefully defining edge weights and the nodes that the random walker restarts to, diffusion profiles can be used in a wide range of biological tasks. Here, we define edge weights for drug, disease, protein, and functional pathway node types, yet more or fewer weights can be used based on the problem of interest. Similarly, here, the random walker jumps to the initial drug or disease node after a restart, but in reality, it can restart to any node or any set of nodes. The edge weights and restart nodes thus make diffusion profiles a flexible approach to propagate signals across a heterogeneous biological network, with applicability to a wide range of problems in systems biology and pharmacology.

Computing drug and disease diffusion profiles through power iteration. Mathematically, we compute diffusion profiles through a matrix formulation with power iteration [81–83]. The diffusion profile computation takes as input:

1. $G = (V, E)$ the unweighted, undirected multiscale interactome with V nodes and E edges.

2. $W = \{w_{\text{drug}}, w_{\text{disease}}, w_{\text{protein}}, w_{\text{functional pathway}}, w_{\text{higher scale functional pathway}}, w_{\text{lower scale functional pathway}}\}$ the set of scalar weights which encode the relative likelihood of the walker jumping from

one node type to another when continuing its walk.

3. α which represents the probability of the walker continuing its walk at a given step rather than restarting.

4. $\mathbf{s} \in \mathbb{R}^{|V|}$ a restart vector which sets the probability the walker will jump to each node after a restart; here, $\mathbf{s}$ is a one-hot vector encoding the drug or disease of interest.

5. ϵ the tolerance allowed for convergence of the power iteration computation.

The diffusion profile computation outputs $\mathbf{r} \in \mathbb{R}^{|V|}$, a drug- or disease-diffusion profile which measures the frequency with which the random walker visits each node. Note that $\sum_i \mathbf{r}_i = 1$.

Before computing the diffusion profile of a drug or disease of interest, we preprocess the multiscale interactome in order to only allow biologically meaningful walks. Diffusion profiles are designed to capture how a drug or disease of interest propagates its effect by recursively affecting proteins and functional pathways. Notice that drugs and diseases do not propagate their effect by using other drugs and diseases as intermediates. Therefore, we disallow paths that have drugs and diseases as intermediate nodes. To accomplish this mathematically, we convert $G = (V, E)$ to a directed graph G' where all previously undirected edges are replaced by edges in both directions. We then make the drug or disease of interest a source node (i.e. no in-edges) and all other drugs and diseases sink nodes (i.e. no out-edges). In G', a random walker starts at the drug or disease of interest and recursively walks to proteins and functional pathways. If the walker reaches any other drug or disease node, it must restart its walk.

Next, we encode G' and the set of scalar weights W into a biased transition matrix $\mathbf{M} \in \mathbb{R}^{|V| \times |V|}$. Each entry $\mathbf{M}_{ij}$ denotes the probability $p_{i \to j}$ a random walker jumps from node i to node j when continuing its walk. Consider a random walker at node i jumping to neighbor j of type t. Let T be the set of all node types adjacent to node i. We compute $p_{i \to j}$ in two steps.

1. First, we compute the probability of the random walker jumping to a node of type t rather than a node of a different type. w_t is the weight of node type t as specified in W:

$$p_t = \frac{w_t}{\sum\limits_{t' \in T} w_{t'}}.$$

2. Second, we compute the probability that the random walker jumps to node j rather than to

another adjacent node of type t. Let n_t be the number of adjacent nodes of type t:

$$\mathbf{M}_{ij} = p_{i \to j} = \frac{p_t}{n_t}.$$

After constructing $\mathbf{M}$, we finally compute the diffusion profile through power iteration as shown in Algorithm 1. The key equation is:

$$\mathbf{r}^{(k+1)} = \overbrace{(1-\alpha)\mathbf{s}}^{\text{Restart walk}} + \overbrace{\alpha(\underbrace{\mathbf{r}^{(k)}\mathbf{M}}_{\text{from node with out-edges}} + \underbrace{\mathbf{s}\sum_{j \in J} \mathbf{r}_j^{(k)}}_{\text{from node without out-edges}})}^{\text{Continue walk...}}.$$

At each step, the random walker can restart its walk at the drug or disease node according to $(1-\alpha)\mathbf{s}$ or continue its walk. If the random walker continues its walk from a node with out-edges, then it jumps to an adjacent node according to $\alpha(\mathbf{r}^{(k)}\mathbf{M})$. If the random walker continues its walk from a node without out-edges (i.e. a sink node), then it restarts its walk according to $\alpha(\mathbf{s}\sum_{j \in J} \mathbf{r}_j^{(k)})$, where J is the set of sink nodes in the graph. At every iteration, $\sum_i \mathbf{r}_i = 1$.

Code for the power iteration implementation is available at github.com/snap-stanford/multiscale-interactome. We use a tolerance of $\epsilon = 1 \times 10^{-6}$.

Algorithm 1 Diffusion profiles through power iteration

% Initialize diffusion profile
$\mathbf{r}_i^{(0)} = \frac{1}{|V|} \forall i$
% While not converged
while $\left\|\mathbf{r}^{(k+1)} - \mathbf{r}^{(k)}\right\|_1 > \epsilon$ **do**
  % Start new walk at drug or disease node or continue walk.
  $\mathbf{r}^{(k+1)} = (1-\alpha)\mathbf{s} + \alpha(\mathbf{r}^{(k)}\mathbf{M} + \mathbf{s}\sum_{j \in J} \mathbf{r}_j^{(k)})$
end while

Predicting what drugs will treat a given disease with diffusion profiles. For a drug to treat a disease, it must affect proteins and functional pathways similar to those disrupted by the disease. The diffusion profiles of the drug $\mathbf{r}^{(c)}$ and the disease $\mathbf{r}^{(d)}$ encode the effect of the drug and the disease on proteins and functional pathways. Therefore, comparing $\mathbf{r}^{(c)}$ and $\mathbf{r}^{(d)}$ allows us to predict what drugs treat a given disease.

For each drug and each disease, we compute the diffusion profile as described above. For each disease, we then rank-order the drugs most likely to treat the disease based on the similarity of the drug and disease diffusion profiles $\text{SIM}(\mathbf{r}^{(c)}, \mathbf{r}^{(d)})$ and a series of baseline methods.

We test five metrics of vector similarity:

1. L2 norm: $\sqrt{\sum_i |\mathbf{r}_i^{(c)} - \mathbf{r}_i^{(d)}|^2}$,
2. L1 norm: $\sum_i |\mathbf{r}_i^{(c)} - \mathbf{r}_i^{(d)}|$,
3. Canberra distance: $\sum_i \frac{|\mathbf{r}_i^{(c)} - \mathbf{r}_i^{(d)}|}{|\mathbf{r}_i^{(c)}| + |\mathbf{r}_i^{(d)}|}$,
4. Cosine similarity: $\frac{\mathbf{r}^{(c)} \cdot \mathbf{r}^{(\mathrm{d})}}{\|\mathbf{r}^{(c)}\|_2 \|\mathbf{r}^{(\mathrm{d})}\|_2}$,
5. Correlation distance: $1 - \frac{(\mathbf{r}^{(c)} - \bar{\mathbf{r}}^{(c)}) \cdot (\mathbf{r}^{(\mathrm{d})} - \bar{\mathbf{r}}^{(\mathrm{d})})}{\|(\mathbf{r}^{(\mathrm{c})} - \bar{\mathbf{r}}^{(\mathrm{c})})\|_2 \|(\mathbf{r}^{(\mathrm{d})} - \bar{\mathbf{r}}^{(\mathrm{d})})\|_2}$.

We additionally test two proximity metrics. In particular, we consider the visitation frequency of the drug node i in the disease diffusion profile as: $\mathbf{r}_i^{(\mathrm{d})}$. We also consider the visitation frequency of the drug node i in the disease diffusion profile multiplied by the visitation frequency of the disease node j in the drug diffusion profile: $\mathbf{r}_i^{(\mathrm{d})} * \mathbf{r}_j^{(\mathrm{c})}$.

We compare to the following baselines. (1) Jaccard similarity between the set of drug targets T and the set of disease proteins S: $\frac{|T \cap S|}{|T \cup S|}$; and (2) the closest distance metric in [10, 13], which is the state-of-the-art proximity metric. Given the set of drug targets T, the set of disease proteins S, and $l(s,t)$ as the shortest path length between nodes s and t, the closest distance metric is: $\frac{1}{|T|} \sum_{t \in T} \min_{s \in S} l(s,t)$.

Evaluating predictions of what drugs will treat a disease. We evaluate how effectively a model ranks the drugs that will treat a disease by using AUROC, Average Precision, and Recall@50. For each disease, a model produces a ranked list of drugs. We identify the drugs approved to treat the disease and, consistent with prior literature, assume that other drugs cannot treat the disease [11–14]. For each disease, we then compute the model's AUROC, Average Precision, and Recall@50 values based on the ranked list of drugs. We report the model's performance across diseases by reporting the median of the AUROC, the mean of the Average Precision, and the mean of the Recall@50 values across diseases.

To ensure robust results, we perform five-fold cross validation. We split the drugs into five folds and create training and held-out sets of the drugs and their corresponding indications. We

compute the above evaluation metrics separately on the training and held-out sets. Ultimately, we report all performance metrics on the held-out set, averaged across folds (Figure 2b).

Model selection and optimization of scalar weights. The diffusion profiles of each drug and disease depend on the scalar weights used to compute them $W = \{w_{\text{drug}}, w_{\text{disease}}, w_{\text{protein}}, w_{\text{functional pathway}}, w_{\text{higher scale functional pathway}}, w_{\text{lower scale functional pathway}}\}$ and the probability α of continuing a walk. Similarly, how effectively diffusion profiles predict what drugs treat a given disease depends on the similarity metric used to compare drug and disease diffusion profiles. We optimize the prediction model across the scalar weights W, the probability of continuing a walk α, and the comparison metrics by performing a sweep and selecting the model with the highest median AUROC on the held-out set, averaged across folds.

After initial coarse explorations for each hyperparameter, we sweep across 486 combinations of hyperparameters sampled linearly within the following ranges $w_{\text{drug}} \in [3, 9], w_{\text{disease}} \in [3, 9], w_{\text{protein}} \in [3, 9], w_{\text{higher scale functional pathway}} \in [1.5, 4.5], w_{\text{lower scale functional pathway}} \in [1.5, 4.5], \alpha \in [0.85, 0.9]$ and set $w_{\text{functional pathway}} = w_{\text{higher scale functional pathway}} + w_{\text{lower scale functional pathway}}$. We also sweep across the seven comparison metrics described above. We repeat this procedure for the both the multiscale interactome and the molecular-scale interactome to identify the best diffusion-based model for both. The optimal weights for the molecular-scale interactome are $w_{\text{drug}} = 4.88, w_{\text{disease}} = 6.83, w_{\text{protein}} = 3.21$ with $\alpha = 0.854$ and use the L1 norm to compare $\mathbf{r}^{(c)}$ and $\mathbf{r}^{(d)}$ (Figure 2c, Supplementary Note 1). The optimal weights for the multiscale interactome are $w_{\text{drug}} = 3.21, w_{\text{disease}} = 3.54, w_{\text{protein}} = 4.40, w_{\text{higher scale functional pathway}} = 2.10, w_{\text{lower scale functional pathway}} = 4.49, w_{\text{functional pathway}} = 6.58$ with $\alpha = 0.860$ and use the correlation distance to compare $\mathbf{r}^{(c)}$ and $\mathbf{r}^{(d)}$ (Figure 2b, c). We utilize these optimal weights for the multiscale interactome for all subsequent sections. Optimized diffusion profiles are provided in Supplementary Data 10.

Evaluating predictions of what drugs will treat a disease by drug category. We analyze the multiscale interactome's predictive performance across drug categories by using the Anatomical Therapeutic Chemical Classification (ATC) [84]. We map all drugs to their ATC class by using DrugBank's XML database "full_database.xml" [26]. We use the second level of the ATC classification and only consider categories with at least 20 drugs. For the drugs in each ATC Level II category, we compute the rank of the drugs for the diseases they are approved to treat. We conduct this analysis twice, first to understand the overall performance of the best multiscale interactome

model (Supplementary Figure 2) and second to understand the differential performance of the best multiscale interactome model compared to the best molecular-scale interactome model using diffusion profiles (Figure 2c; Supplementary Figure 3). The ATC classification for the drugs in our study is provided in Supplementary Data 7.

Learning treatment mechanisms with diffusion profiles. For a given drug-disease pair, diffusion profiles identify the proteins and functional pathways important to treatment. We define the treatment mechanism as the top k proteins and functional pathways in the drug diffusion profile and in the disease diffusion profile. To visualize the treatment mechanism, we induce a subgraph on these nodes and remove any isolated components.

Validation of treatment mechanisms through gene expression signatures. To validate drug treatment mechanisms, we compare drug diffusion profiles to the drug gene expression signatures present in the Broad Connectivity Map [44, 45] (Figure 2f).

We map drugs in the Broad Connectivity Map to DrugBank IDs using PubChem IDs, drug names, and the DrugBank "approved_drug_links.csv" and "drugbank_vocabulary.csv" files [26].

Drugs in the Broad Connectivity Map have multiple gene expression signatures based on the cell line, the drug dose, and the time of exposure. However, drugs only have a single diffusion profile. We thus only consider drugs where activity is consistent across cell lines and select a single representative gene expression signature for each drug. To accomplish this, we follow Broad Connectivity Map guidelines [44, 45] as described next. For drugs:

1. We only consider drugs with similar signatures across cell lines (an inter-cell connectivity score $>= 0.4$) and with activity across many cell lines (an aggregated transcriptional activity score $>= 0.3$).

2. We only consider drugs that are members of the "touchstone" dataset: the drugs that are the most well-annotated and systematically profiled across the Broad's core cell lines at standardized conditions. The Broad Connectivity Map specifically recommends the "touchstone" dataset as a reference.

For gene expression signatures, we utilize the Level 5 Replicate Consensus Signatures provided by the Broad Connectivity Map. Each gene expression signature captures the z-scored change in expression of each gene across replicate experiments ("GSE92742_Broad_LINCS_Level5_COMPZ.MODZ_n473647x12328.gctx"). For these gene ex-

pression signatures:

1. We only consider genes whose expression is measured directly rather than inferred (i.e. "landmark" genes).

2. We only consider signatures that are highly reproducible and distinct (distil_cc_q75 >= 0.2 and (pct_self_rank_q25 <= 0.1).

3. We require that each signature be an "exemplar" signature for the drug as indicated by the Broad Connectivity Map (i.e. a highly reproducible, representative signature).

4. We require that each signature be sufficiently active (i.e. have a transcriptional activity score >= 0.35) and result from at least 3 replicates (distil_n_sample_thresh >= 3).

5. In cases where multiple signatures meet these criteria for a given drug, we select the signature with the highest transcriptional activity score.

The gene expression signatures we ultimately use for each drug are provided in Supplementary Data 8.

Finally, we compare the similarity of drugs based on their diffusion profiles and their gene expression signatures. We compare the similarity of drug diffusion profiles by the Canberra distance, multiplied by -1 so higher values indicate higher similarity. We compare the similarity of drug gene expression signatures based on the overlap in the 25 most upregulated genes U and 25 most downregulated genes D:

$$\frac{1}{2}\left[\frac{|U_{\text{drug1}} \cap U_{\text{drug2}}|}{|U_{\text{drug1}} \cup U_{\text{drug2}}|} + \frac{|D_{\text{drug1}} \cap D_{\text{drug2}}|}{|D_{\text{drug1}} \cup D_{\text{drug2}}|}\right].$$

We use rank transformed gene expression signatures and diffusion profiles. We only allow the comparison of gene expression signatures that are in the same cell, with the same dose, and at the same exposure time. Ultimately, we measure the Spearman Correlation between the similarity of the drugs as described by the drug diffusion profiles and the similarity of the drugs as described the gene expression signatures.

Compiling genetic mutations that alter treatment. We compile genetic mutations that alter treatment by using the Pharmacogenomics Knowledgebase (PharmGKB) [60]. PharmGKB is a gold-standard database mapping the effect of genetic mutations on treatments. PharmGKB is

manually curated from a range of sources, including the published literature, the Allele Frequency Database, the Anatomical Therapeutic Chemical Classification, ChEBI, ClinicalTrials.gov, dbSNP, DrugBank, the European Medicines Agency, Ensembl, FDA Drug Labels at DailyMed, GeneCard, HC-SC, HGNC, HMDB, HumanCyc Gene, LS-SNP, MedDRA, MeSH, NCBI Gene, NDF-RT, PMDA, PubChem Compound, RxNorm, SnoMed Clinical Terminology, and UniProt KB.

We use PharmGKB's "Clinical Annotations" which detail how mutations at the gene level alter treatments. PharmGKB's "clinical_ann_metadata.tsv" file provides triplets of drugs, diseases, and genes known to alter treatment. Treatment alteration occurs when a genetic mutation alters the efficacy, dosage, metabolism, or pharmacokinetics of treatment or otherwise causes toxicity or an adverse drug reaction. We map genes to their Entrez ID using HUGO, drugs to their DrugBank ID using PharmGKB's "drugs.tsv" and "chemicals.tsv" files, and diseases to their UMLS CUIDs by using PharmGKB's "phenotypes.tsv" file. To ensure consistency with the approved drug-disease pairs we previously compiled, we only consider (drug, disease, gene) triplets in which the drug and disease are part of an FDA-approved treatment. Ultimately, we obtain 1,223 drug-disease-gene triplets with 201 drugs, 94 diseases, and 455 genes. All drug-disease-gene triplets are provided in Supplementary Data 9.

Computing treatment importance of a genetic mutation based on diffusion profiles. We define the treatment importance of a mutation in gene i as the product of the visitation frequency of the corresponding protein in the drug and disease diffusion profiles. For a treatment composed of drug compound c and disease d, the treatment importance of a mutation in gene i is:

$$\text{TI}(i|c,d) = \mathbf{r}_i^{(c)} * \mathbf{r}_i^{(d)}.$$

We define the treatment importance percentile as the percentile rank of $\text{TI}(i|c,d)$ compared to mutations in all other genes for the same drug and disease. Intuitively, a mutation in gene i is important to a treatment if the corresponding protein is frequently visited in both the drug and disease diffusion profiles.

Comparing treatment importance of treatment altering genetic mutations vs other genetic mutations. We compare the treatment importance of genetic mutations known to alter a treatment with the treatment importance of other genetic mutations (Figure 3b). In particular, we compare the set of (drug, disease, gene) triplets where the gene is known to alter the drug-disease treatment

with an equivalently sized set of (drug, disease, gene) triplets where the gene is not known to alter treatment. We construct the latter set by sampling drugs, diseases, and genes uniformly at random that are not known to alter treatment from PharmGKB [60]. The drugs and diseases in all triplets correspond to approved drug-disease pairs. Thereby, we construct a distribution of the treatment importance for "treatment altering genetic mutations" and a distribution of the treatment importance for "other genetic mutations" (Figure 3b).

Predicting genes that will alter a treatment based on treatment importance. We evaluate the ability of treatment importance to predict the genes that will alter a given treatment (Figure 3c). For each (drug, disease, gene) triplet, we use the treatment importance of the gene $\text{TI}(i|c, d)$ to predict whether the gene alters treatment or not for that drug-disease pair (i.e. binary classification). We use the set of positive and negative (drug, disease, gene) triplets constructed previously (see Methods: Comparing treatment importance of treatment altering genetic mutations vs other genetic mutations). We assess performance using AUROC and Average Precision (Figure 3c).

Comparing treatment importance of genetic mutations that alter one drug indicated to treat a disease but not another. We analyze how often a genetic mutation has a higher treatment importance in the treatments it alters than in those it does not alter (Figure 3e).

Formally, let i be a gene. Consider a triplet $(d, c_{\text{altered}}, c_{\text{unaltered}})$ of a disease d, a drug c_{altered} approved to treat the disease whose treatment is altered due to a mutation in i, and a drug $c_{\text{unaltered}}$ approved to treat the disease whose treatment is not altered due to a mutation in i. Let n_{triplets} be the total number of such triplets for gene i. For each gene i, we measure the fraction f of triplets $(d, c_{\text{altered}}, c_{\text{unaltered}})$ for which the treatment importance of i is higher in the (c_{altered}, d) treatment than in the $(c_{\text{unaltered}}, d)$ treatment, as shown below. We only consider genes for which $n_{triplets} \geq 100$.

$$f\left[\text{TI}(i|c_{\text{altered}}, d > \text{TI}(i|c_{\text{unaltered}}, d)\right] = \frac{\sum\limits_{\forall(d, c_{\text{altered}}, c_{\text{unaltered}})} \mathbb{1}\{\text{TI}(i|c_{\text{altered}}, d) > \text{TI}(i|c_{\text{unaltered}}, d)\}}{n_{\text{triplets}}}.$$

References

1. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).

2. Creixell, P. *et al.* Pathway and network analysis of cancer genomes. *Nature Methods* **12**, 615 (2015).

3. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics* **16**, 441–458 (2015).

4. Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics* **47**, 106 (2015).

5. Nikolsky, Y., Nikolskaya, T. & Bugrim, A. Biological networks and analysis of experimental data in drug discovery. *Drug Discovery Today* **10**, 653–662 (2005).

6. Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics* **17**, 615 (2016).

7. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Research* **25**, 142–154 (2015).

8. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56–68 (2011).

9. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* **18**, 551 (2017).

10. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature Communications* **9**, 1–12 (2018).

11. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery* **18**, 41–58 (2019).

12. Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J. & Green, J. R. A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics* **19**, 878–892 (2018).

13. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nature Communications* **7**, 10331 (2016).

14. Wang, W., Yang, S., Zhang, X. & Li, J. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* **30**, 2923–2930 (2014).

15. Luo, Y. *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications* **8**, 1–13 (2017).

16. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).

17. Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nature Communications* **10**, 1–11 (2019).

18. Hu, Y. *et al.* Optimal control nodes in disease-perturbed networks as targets for combination therapy. *Nature Communications* **10**, 1–14 (2019).

19. Firestone, A. J. & Settleman, J. A three-drug combination to treat BRAF-mutant cancers. *Nature Medicine* **23**, 913–914 (2017).

20. Zhao, S. & Iyengar, R. Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. *Annual Review of Pharmacology and Toxicology* **52**, 505–521 (2012).

21. Walpole, J., Papin, J. A. & Peirce, S. M. Multiscale computational models of complex biological systems. *Annual Review of Biomedical Engineering* **15**, 137–154 (2013).

22. van Hasselt, J. C. & Iyengar, R. Systems pharmacology: defining the interactions of drug combinations. *Annual Review of Pharmacology and Toxicology* **59**, 21–40 (2019).

23. Han, K. *et al.* Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nature Biotechnology* **35**, 463 (2017).

24. Jia, J. *et al.* Mechanisms of drug combinations: interaction and network perspectives. *Nature Reviews Drug Discovery* **8**, 111–128 (2009).

25. Yu, M. K. *et al.* Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell Systems* **2**, 77–88 (2016).

26. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2017).

27. Corsello, S. M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nature Medicine* **23**, 405 (2017).

28. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* gkw943 (2016).

29. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).

30. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Research* **47**, D529–D541 (2019).

31. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).

32. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* **32**, D449–D451 (2004).

33. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nature Methods* **6**, 83 (2009).

34. Yu, H. *et al.* Next-generation sequencing to generate interactome datasets. *Nature Methods* **8**, 478 (2011).

35. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).

36. Consortium, G. O. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Research* **47**, D330–D338 (2018).

37. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

38. Brown, A. S. & Patel, C. J. A standard database for drug repositioning. *Scientific Data* **4**, 170029 (2017).

39. Sharp, M. E. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *Journal of Biomedical Semantics* **8**, 2 (2017).

40. Klipp, E., Liebermeister, W., Wierling, C. & Kowald, A. *Systems biology: a textbook* (John Wiley & Sons, 2016).

41. Donnat, C., Zitnik, M., Hallac, D. & Leskovec, J. Learning structural node embeddings via diffusion wavelets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1320–1329 (2018).

42. Nielsen, S. *et al.* Vasopressin increases water permeability of kidney collecting duct by inducing translocation of aquaporin-CD water channels to plasma membrane. *Proceedings of the National Academy of Sciences* **92**, 1013–1017 (1995).

43. Holmes, C. L., Landry, D. W. & Granton, J. T. Science review: vasopressin and the cardiovascular system part 1–receptor physiology. *Critical Care* **7**, 427 (2003).

44. Subramanian, A. *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).

45. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).

46. Utermann, G., Jaeschke, M. & Menzel, J. Familial hyperlipoproteinemia type III: Deficiency of a specific apolipoprotein (APO e-iii) in the very-low-density lipoproteins. *FEBS Letters* **56**, 352–355 (1975).

47. Utermann, G. *et al.* Polymorphism of apolipoprotein e: Genetics of hyperlipoproteinemia type III. *Clinical Genetics* **15**, 37–62 (1979).

48. Ghiselli, G., Schaefer, E. J., Gascon, P. & Breser, H. Type III hyperlipoproteinemia associated with apolipoprotein e deficiency. *Science* **214**, 1239–1241 (1981).

49. Wang, J. *et al.* APOA5 genetic variants are markers for classic hyperlipoproteinemia phenotypes and hypertriglyceridemia. *Nature Clinical Practice Cardiovascular Medicine* **5**, 730–737 (2008).

50. Evans, D., Seedorf, U. & Beil, F. Polymorphisms in the apolipoprotein a5 (APOA5) gene and type III hyperlipidemia. *Clinical Genetics* **68**, 369–372 (2005).

51. Moghadasian, M. H. Clinical pharmacology of 3-hydroxy-3-methylglutaryl coenzyme a reductase inhibitors. *Life Sciences* **65**, 1329–1337 (1999).

52. Holdgate, G., Ward, W. & McTaggart, F. Molecular mechanism for inhibition of 3-hydroxy-3-methylglutaryl coa (HMG-CoA) reductase by rosuvastatin (2003).

53. Shinkai, K., McCalmont, T. & Leslie, K. Cryopyrin-associated periodic syndromes and autoinflammation. *Clinical and Experimental Dermatology: Clinical Dermatology* **33**, 1–9 (2008).

54. Kone-Paut, I. & Galeotti, C. Anakinra for cryopyrin-associated periodic syndrome. *Expert Review of Clinical Immunology* **10**, 7–18 (2014).

55. Ashley, E. A. Towards precision medicine. *Nature Reviews Genetics* **17**, 507 (2016).

56. Hansen, N. T., Brunak, S. & Altman, R. Generating genome-scale candidate gene lists for pharmacogenomics. *Clinical Pharmacology & Therapeutics* **86**, 183–189 (2009).

57. Karczewski, K. J., Daneshjou, R. & Altman, R. B. Pharmacogenomics. *PLOS Computational Biology* **8** (2012).

58. Su, X. *et al.* Association between angiotensinogen, angiotensin II receptor genes, and blood pressure response to an angiotensin-converting enzyme inhibitor. *Circulation* **115**, 725–732 (2007).

59. Yu, H. *et al.* A core promoter variant of angiotensinogen gene and interindividual variation in response to angiotensin-converting enzyme inhibitors. *Journal of the Renin-Angiotensin-Aldosterone System* **15**, 540–546 (2014).

60. Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: a worldwide resource for pharmacogenomic information. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **10**, e1417 (2018).

61. Nayler, W. G. & Dillon, J. Calcium antagonists and their mode of action: an historical overview. *British Journal of Clinical Pharmacology* **21**, 97S–107S (1986).

62. Sutton, M. S. J. & Morad, M. Mechanisms of action of diltiazem in isolated human atrial and ventricular myocardium. *Journal of Molecular and Cellular Cardiology* **19**, 497–508 (1987).

63. O'Connor, S. E., Grosset, A. & Janiak, P. The pharmacological basis and pathophysiological significance of the heart rate-lowering property of diltiazem. *Fundamental & Clinical Pharmacology* **13**, 145–153 (1999).

64. Balfour, J. A. & Goa, K. L. Benazepril. *Drugs* **42**, 511–539 (1991).

65. Lavoie, J. L. & Sigmund, C. D. Minireview: overview of the renin-angiotensin system—an endocrine and paracrine system. *Endocrinology* **144**, 2179–2183 (2003).

66. Sanchez-Vega, F. *et al.* Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).

67. Jones, D. Pathways to cancer therapy. *Nature Reviews Drug Discovery* **7**, 875–876 (2008).

68. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).

69. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).

70. Choobdar, S. *et al.* Assessment of network module identification across complex diseases. *Nature Methods* **16**, 843–852 (2019).

71. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

72. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034 (2019).

73. Goldstein, D. B., Tate, S. K. & Sisodiya, S. M. Pharmacogenetics goes genomic. *Nature Reviews Genetics* **4**, 937–947 (2003).

74. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).

75. Yates, B. *et al.* Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Research* gkw1033 (2016).

76. Vinayagam, A. *et al.* A directed protein interaction network for investigating intracellular signal transduction. *Science Signaling* **4**, rs8–rs8 (2011).

77. Klopfenstein, D. *et al.* GOATOOLS: A python library for gene ontology analyses. *Scientific Reports* **8**, 1–17 (2018).

78. Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research* **32**, D267–D270 (2004).

79. Davis, A. P. *et al.* The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Research* **47**, D948–D954 (2019).

80. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research* **47**, D955–D962 (2019).

81. Langville, A. N. & Meyer, C. D. A survey of eigenvector methods for web information retrieval. *SIAM Review* **47**, 135–161 (2005).

82. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank citation ranking: Bringing order to the web. Tech. Rep., Stanford InfoLab (1999).

83. Hagberg, A., Swart, P. & S Chult, D. Exploring network structure, dynamics, and function using NetworkX. Tech. Rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008).

84. Organization, W. H. *et al.* The Anatomical Therapeutic Chemical Classification System with defined daily doses-ATC/DDD (2009).