Open access • Posted Content • DOI:10.1101/2020.04.30.069690

Identification of disease treatment mechanisms through the multiscale interactome

Camilo Ruiz, Marinka Zitnik, Jure Leskovec

Institutions: Stanford University, Harvard University

Published on: 28 Oct 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Interactome

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-21770-8 OPEN

Identification of disease treatment mechanisms through the multiscale interactome

Camilo Ruiz [1,2], Marinka Zitnik[3] & Jure Leskovec [1,4]✉

Most diseases disrupt multiple proteins, and drugs treat such diseases by restoring the functions of the disrupted proteins. How drugs restore these functions, however, is often unknown as a drug's therapeutic effects are not limited to the proteins that the drug directly targets. Here, we develop the multiscale interactome, a powerful approach to explain disease treatment. We integrate disease-perturbed proteins, drug targets, and biological functions into a multiscale interactome network. We then develop a random walk-based method that captures how drug effects propagate through a hierarchy of biological functions and physical protein-protein interactions. On three key pharmacological tasks, the multiscale interactome predicts drug-disease treatment, identifies proteins and biological functions related to treatment, and predicts genes that alter a treatment's efficacy and adverse reactions. Our results indicate that physical interactions between proteins alone cannot explain treatment since many drugs treat diseases by affecting the biological functions disrupted by the disease rather than directly targeting disease proteins or their regulators. We provide a general framework for explaining treatment, even when drugs seem unrelated to the diseases they are recommended for.

[1] Computer Science Department, Stanford University, Stanford, CA, USA. [2] Bioengineering Department, Stanford University, Stanford, CA, USA. [3] Biomedical Informatics Department, Harvard University, Boston, MA, USA. [4] Chan Zuckerberg Biohub, San Francisco, CA, USA. ✉email: jure@cs.stanford.edu

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 systematic 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] (Fig. 1). As such, current approaches fail to capture biological functions, through which target proteins can restore the functions of disease-perturbed proteins and thus treat a disease[20–25] (Supplementary Fig. 3). Moreover, current systematic approaches are black-boxes: they predict treatment relationships but provide little biological insight into how treatment occurs. This suggests an opportunity for a systematic, explanatory approach. Indeed for particular drugs and diseases, custom networks have demonstrated that incorporating specific biological functions can help explain treatment[26–29].

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

To identify how a drug treats a disease, our approach uses biased random walks which model how drug effects spread through a hierarchy of biological functions 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 biological functions. For each drug and disease, we learn a diffusion profile, which identifies the key proteins and biological functions involved in a given treatment. By comparing drug and disease diffusion profiles, the multiscale interactome provides an interpretable basis to identify the proteins and biological functions 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 more accurately than existing methods that rely on physical interactions between proteins (i.e., a molecular-scale interactome). This finding indicates that our approach accurately captures the biological functions through which target proteins affect 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 biological functions and thus cannot be accurately represented by approaches which only consider physical interactions between proteins. Second, we find that the multiscale interactome is a white-box method with the ability to identify proteins and biological functions relevant in treatment. Finally, we find that the multiscale interactome predicts what genes alter drug efficacy or cause serious adverse reactions for a given treatment and identifies biological functions that help explain how these genes interfere with treatment.

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 affect the same biological functions disrupted by the disease proteins, the treatment can be successful. Thus, physical interactions between proteins alone are unable to explain the therapeutic effects of drugs, and functional information provides an important component for modeling treatment mechanisms. We provide a general framework for identifying proteins and biological functions relevant in treatment, even when drugs seem unrelated to the diseases they are recommended for.

Results

The multiscale interactome represents the effects of drugs and diseases on proteins and biological functions. The multiscale interactome models drug treatment by integrating both physical interactions between proteins and a multiscale hierarchy of biological functions. Crucially, many treatments depend on biological functions (Supplementary Fig. 3)[20–24]. Existing systematic network approaches, however, primarily model physical interactions between proteins[10–14], and thus cannot accurately model such treatments (Fig. 1a, Supplementary Fig. 1).

Our multiscale interactome captures the fact that drugs and diseases exert their effects through both proteins and biological functions (Fig. 1b). In particular, the multiscale interactome is a network in which 1,661 drugs interact with the human proteins they primarily target (8,568 edges)[30,31] and 840 diseases interact with the human proteins they disrupt through effects like genomic alterations, altered expression, or post-translational modification (25,212 edges)[32]. 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)[33–39]. Second, these proteins alter 9,798 biological functions according to a rich hierarchy ranging from specific processes (i.e., embryonic heart tube elongation) to broad processes (i.e., heart development). Biological functions can describe processes involving molecules (i.e., DNA demethylation), cells (i.e., the mitotic cell cycle), tissues (i.e., muscle atrophy), organ systems (i.e., activation of the innate immune response), and the whole organism (i.e., anatomical structure development) (34,777 edges between proteins and biological functions, 22,545 edges between biological functions; Gene Ontology)[40,41]. By modeling the effect of drugs and diseases on both proteins and biological functions, our multiscale interactome can model the range of drug treatments that rely on both[20–24].

Overall, our multiscale interactome provides a large, 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[31,42,43], exceeding the largest prior network-based study by 10X[13] (Anatomical Therapeutic Classification; Supplementary Fig. 4).

Propagation of the effects of drugs and diseases through the multiscale interactome. To learn how the effects of drugs and diseases propagate through proteins and biological functions, we harnessed network diffusion profiles (Fig. 1c). A network diffusion profile propagates the effects of a drug or disease across the multiscale interactome, revealing the most affected proteins and biological functions. The diffusion profile is computed by biased

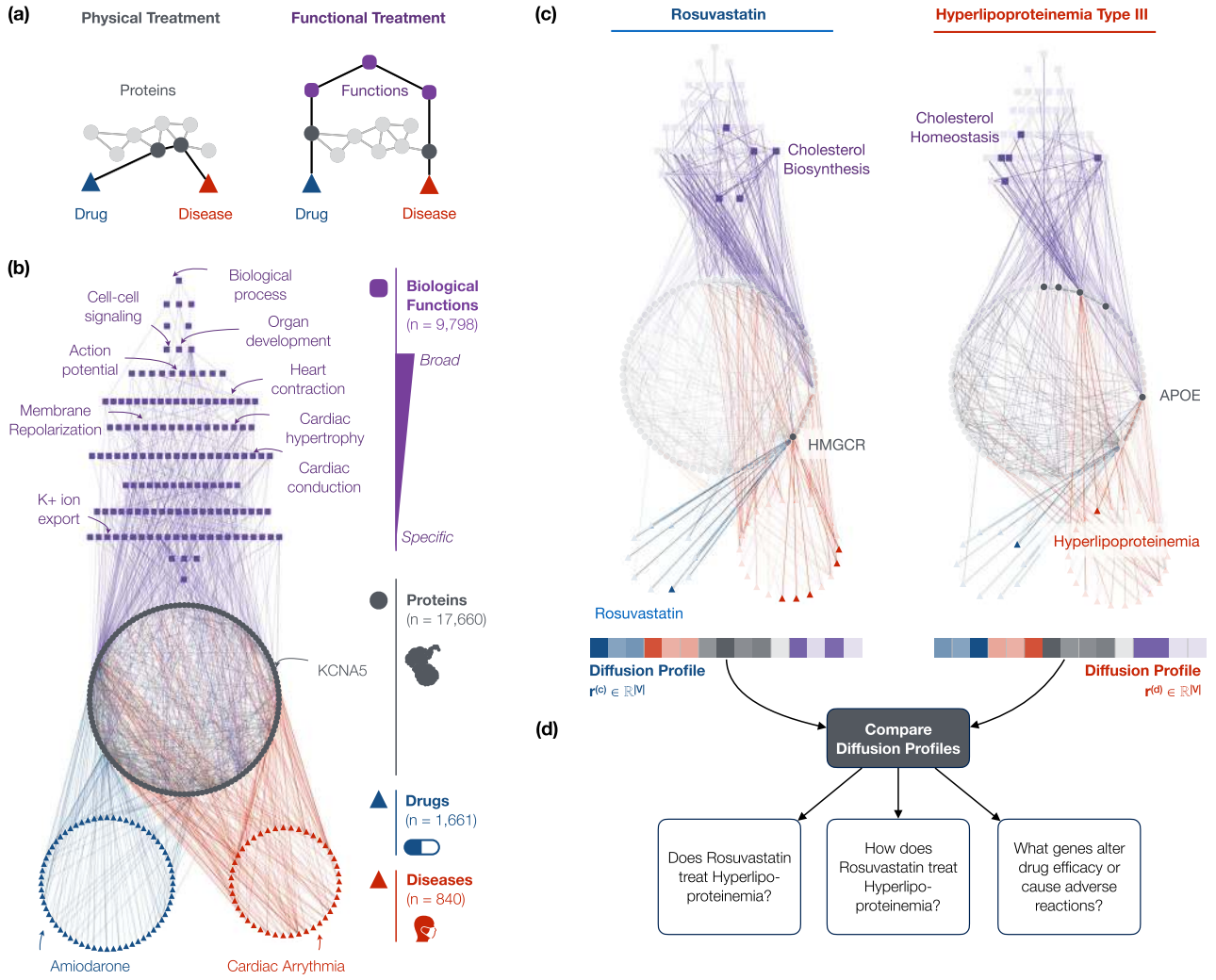

Fig. 1 The multiscale interactome models drug treatment through both proteins and biological functions. a Existing systematic network approaches assume that drugs treat diseases by targeting proteins that are proximal to disease proteins in a network of physical interactions[10–14]. However, drugs can also treat diseases by targeting distant proteins that affect the same biological functions (Supplementary Fig. 3)[20–25]. **b** The multiscale interactome models drug-disease treatment by integrating both proteins and a hierarchy of biological functions (Supplementary Fig. 1). **c** The diffusion profile of a drug or disease captures its effect on every protein and biological function. The diffusion profile propagates the effect of the drug or disease via biased random walks which adaptively explore proteins and biological functions based on optimized edge weights. Ultimately, the visitation frequency of a node corresponds to the drug or disease's propagated effect on that node (see the "Methods" section). **d** By comparing the diffusion profiles of a drug and disease, we compare their effects on both proteins and biological functions. Thereby, we predict whether the drug treats the disease (Fig. 2a–c), identify proteins and biological functions related to treatment (Fig. 2d–h), and identify which genes alter drug efficacy or cause dangerous adverse reactions (Fig. 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[50–54]. 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[55,56]. By comparing these diffusion profiles, we thus predict that Rosuvastatin treats Hyperlipoproteinemia Type III, identify the HMGCR and APOE-driven cholesterol metabolic functions relevant to treatment, and predict that mutations in APOE and HMGCR may interfere with treatment and thus alter drug efficacy or cause dangerous adverse reactions.

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 biological function.

Diffusion profiles contribute three methodological advances. First, diffusion profiles provide a general framework to adaptively integrate physical interactions between proteins and a hierarchy of biological functions. When continuing its walk, the random walker jumps between proteins and biological functions at different hierarchical levels based on optimized edge weights. These edge weights encode the relative importance of different types of nodes: w_{drug}, w_{disease}, w_{protein}, $w_{\text{biological function}}$, $w_{\text{higher-level biological function}}$, $w_{\text{lower-level biological function}}$. These weights are hyperparameters which we optimize when predicting the drugs that treat a given disease (see the "Methods" section). For drug and disease treatments, these optimized edge weights encode the knowledge that proteins and biological functions at different hierarchical levels have different importance in the effects of drugs and diseases[20,21]. By adaptively integrating both proteins and biological functions in a hierarchy, therefore, diffusion profiles model 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 effects by disrupting or binding to proteins which recursively affect other proteins and biological functions. The effect propagates via two principles[8,9]. First, proteins and biological functions closer to the drug or disease are affected more strongly. Similarly in diffusion profiles, proteins and biological functions 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 effects on proteins and biological functions, a richer comparison. Our approach is thus consistent with recent machine learning advances which harness diffusion profiles to represent nodes[44,45].

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%) (Fig. 2a, b, see the "Methods" section). Note that drug–disease treatment relationships are never directly encoded into our network. Instead, the multiscale interactome learns to effectively predict drug–disease treatment relationships it has never previously seen.

Moreover, the multiscale interactome accurately models classes of drugs that rely on biological functions and which molecular-scale interactome approaches thus cannot model effectively. Indeed, the top overall performing drug classes (i.e., sex hormones, modulators of the genital system; Supplementary Fig. 6) and the top drug classes for which the multiscale interactome outperforms the molecular-scale interactome (i.e., pituitary, hypothalamic hormones, and analogs; Fig. 2c, Supplementary Fig. 7) harness biological functions that describe processes across the body. 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[30,46,47]. Treatment by Vasopressin, and by pituitary and hypothalamic hormones more broadly, relies on biological functions that describe processes across the body and that are modeled by the multiscale interactome.

The multiscale interactome identifies proteins and biological functions relevant in complex treatments. Existing interactome approaches to systematically study treatment are black-boxes: they predict what drug treats a disease but cannot explain how the drug treats the disease through specific proteins and biological functions[10–15] (Fig. 2d). By contrast, drug and disease diffusion profiles identify proteins and biological functions relevant to treatment (Fig. 2e, Supplementary Note 3). For a given drug and disease, we identify proteins and biological functions relevant to treatment by inducing a subgraph on the k most frequently visited nodes in the drug and disease diffusion profiles which correspond to the proteins and biological functions most affected by the drug and disease.

Gene expression signatures validate the biological relevance of diffusion profiles (Fig. 2f). We find that drugs with more similar diffusion profiles have more similar gene expression signatures (Spearman $\rho = 0.392$, $p = 5.8 \times 10^{-7}$, $n = 152$)[48,49], indicating that diffusion profiles reflect the effects of drugs on proteins and biological functions.

Furthermore, case studies validate the proteins and biological functions that diffusion profiles identify as relevant to treatment. Consider the treatment of Hyperlipoproteinemia Type III by Rosuvastatin (i.e., Crestor). In Hyperlipoproteinemia Type III, defects in apolipoprotein E (APOE)[50–52] and apolipoprotein A-V (APOA5)[53,54] lead to excess blood cholesterol, eventually leading to the onset of severe arteriosclerosis[51]. Rosuvastatin is known to treat Hyperlipoproteinemia Type III by inhibiting HMG-CoA reductase (HMGCR) and thereby diminishing cholesterol production[55,56]. Crucially, diffusion profiles identify proteins and biological functions that recapitulate these key steps (Fig. 2g). Notably, there is no direct path of proteins between Hyperlipoproteinemia Type III and Rosuvastatin. Instead, treatment operates through biological functions (i.e., cholesterol biosynthesis and its regulation). Consistently, the multiscale interactome identifies Rosuvastatin as a treatment for Hyperlipoproteinemia Type III far more effectively than a molecular-scale interactome approach, ranking Rosuvastatin in the top 4.33% of all drugs rather than the top 72.7%. The multiscale interactome explains treatments that rely on biological functions, a feat which molecular-scale interactome approaches cannot accomplish.

Similarly, consider the treatment of cryopyrin-associated periodic syndromes (CAPS) by Anakinra. In CAPS, mutations in NLRP3 and MME lead to immune-mediated inflammation through the Interleukin-1 beta signaling pathway[57]. Anakinra treats CAPS by binding IL1R1, a receptor which mediates regulation of the Interleukin-1 beta signaling pathway and thus prevents excessive inflammation[30,58]. Again, diffusion profiles identify proteins and biological functions that recapitulate these key steps (Fig. 2h). Crucially, diffusion profiles identify the regulation of inflammation and immune system signaling, complex biological functions which are not modeled 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 in the top 10.9% of all drugs rather than the top 71.8%.

The multiscale interactome identifies genes that alter patient-specific drug efficacy and cause adverse reactions. A key goal of precision medicine is to understand how changes in genes alter patient-specific drug efficacy and cause adverse reactions[59] (Fig. 3a). For particular treatments, detailed mechanistic models have been developed which can predict and explain drug resistance among genes already identified as relevant to treatment[26–29]. More systematically, however, current tools of precision medicine struggle to predict the genes that interfere with patient-specific treatment[60] and explain how such genes interfere with treatment[61].

We find that genetic variants that alter drug efficacy and cause serious adverse reactions occur in genes that are highly visited in the corresponding drug and disease diffusion profiles (Fig. 3b). We define the treatment importance of a gene according to the visitation frequency of the corresponding protein in the drug and disease diffusion profiles (see the "Methods" section). Genes that alter drug efficacy and cause adverse reactions exhibit substantially higher treatment importance scores than other genes (median network importance = 0.912 vs. 0.513; $p = 2.95 \times 10^{-107}$, Mood's median test), indicating that these treatment altering genes occur at highly visited nodes. We thus provide evidence that the topological position of a gene influences its ability to alter drug efficacy or cause serious adverse reactions.

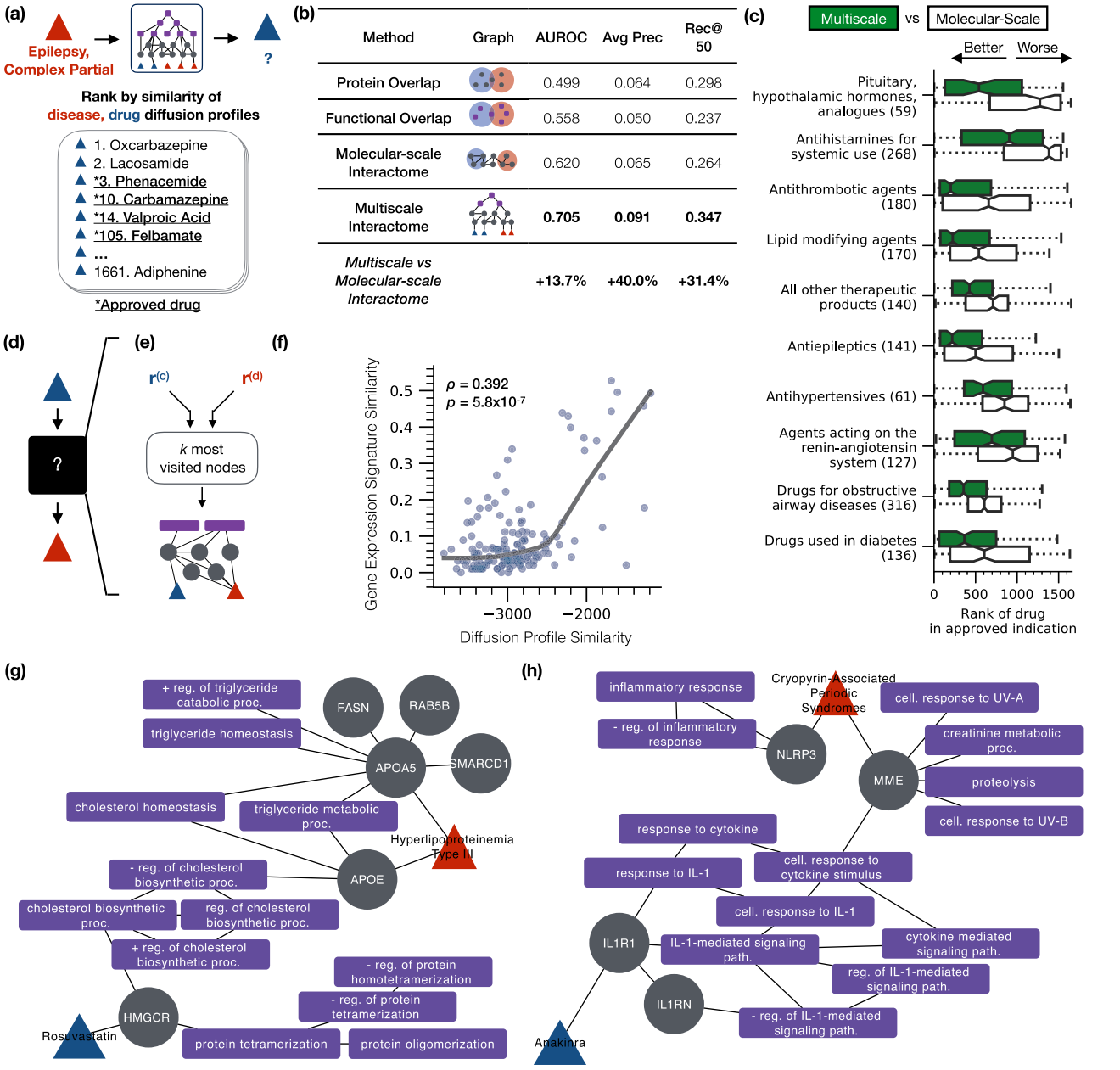

Fig. 2 The multiscale interactome accurately predicts what drugs treat a disease and systematically identifies proteins and biological functions related to treatment. a To predict whether a drug treats a disease, we compare the drug and disease diffusion profiles according to a correlation distance. **b** By incorporating both proteins and biological functions, 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 (see the "Methods" section); multiscale interactome values are in bold. **c** The multiscale interactome outperforms the molecular-scale interactome most greatly on drug classes known to harness biological functions that describe processes across the body (i.e., pituitary, hypothalamic hormones and analogs). **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 functions[10–15]. **e** By contrast, the drug and disease diffusion profiles ($\mathbf{r}^{(c)}$ and $\mathbf{r}^{(d)}$) reveal the proteins and biological functions relevant to treatment. For each drug and disease pair, we induce a subgraph on the k most frequently visited nodes in the drug and disease diffusion profiles to explain treatment. **f** Drugs with more similar diffusion profiles have more similar gene expression signatures (Spearman $\rho = 0.392$, $p = 5.8 \times 10^{-7}$, $n = 152$, two-sided), suggesting that drug diffusion profiles capture their biological effects. **g** The multiscale interactome explains treatments 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 functions affected by Hyperlipoproteinemia Type III[50–56]. **h** Anakinra treats cryopyrin-associated periodic syndromes (CAPS) by binding to IL1R1 which regulates immune-mediated inflammation through the Interleukin-1 beta signaling pathway[30,58]. Inflammation is a hallmark of CAPS[57]. Abbreviations: reg. regulation, path. pathway, proc. process, cell. cellular, + positive, − negative. Boxplots: median (line); 95% CI (notches); 1st, 3rd quartiles (boxes); data within 1.5 × the inter-quartile range from the 1st, 3rd quartiles (whiskers). Sample sizes in parentheses.

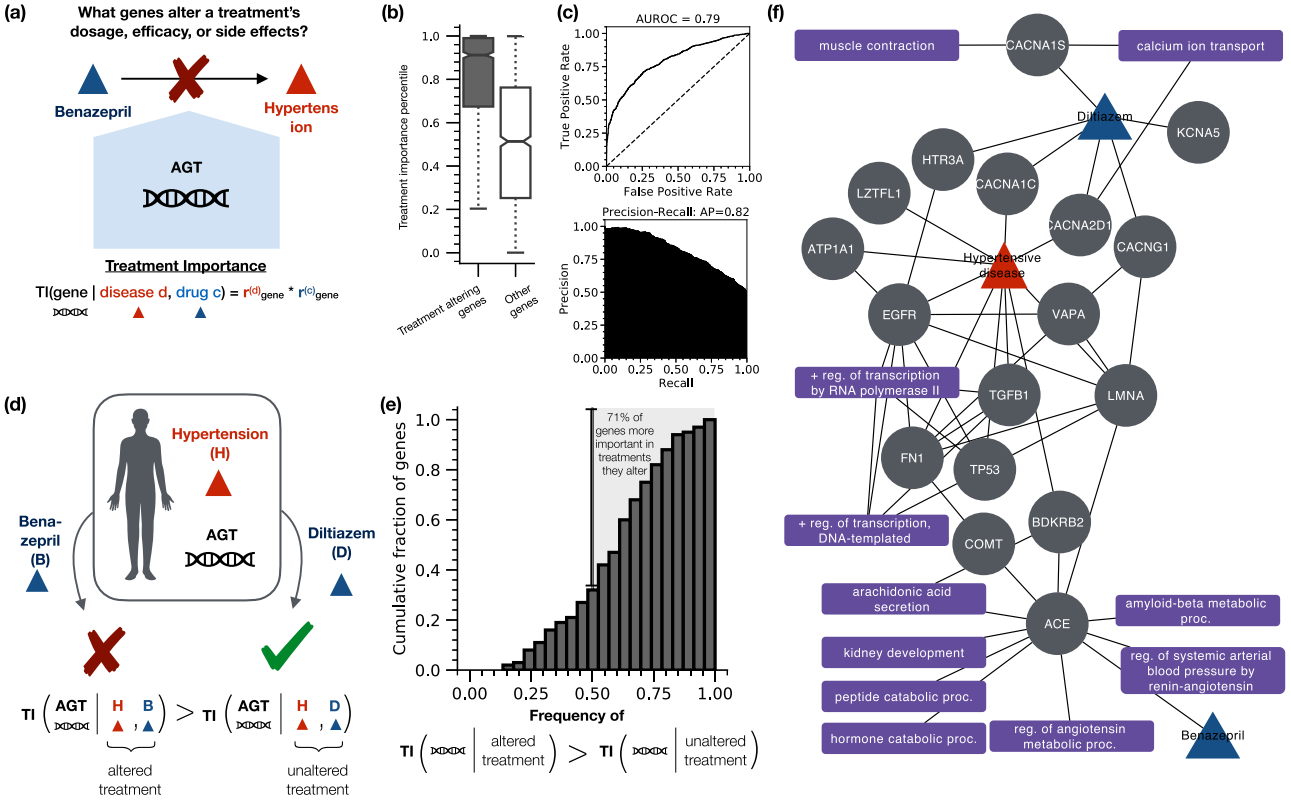

Fig. 3 Diffusion profiles identify which genes alter drug efficacy and cause serious adverse reactions and identify biological functions that help explain the alteration in treatment. a Genes alter drug efficacy and cause serious adverse reactions in a range of treatments[62]. A pressing need exists to systematically identify genes that alter drug efficacy and cause serious adverse reactions for a given treatment and explain how these genes interfere with treatment[60]. **b** Genetic variants alter drug efficacy and cause serious adverse reactions by targeting genes of high network importance in treatment (median network importance of treatment altering genes = 0.912 vs. 0.513; $p = 2.95 \times 10^{-107}$, Mood's median test, two-sided; $n = 1{,}223$ vs. 1,223). We define the network treatment importance of a gene according to its visitation frequency in the drug and disease diffusion profiles (see the "Methods" section). **c** The treatment importance of a gene in the drug and disease diffusion profiles predicts whether that gene alters drug efficacy and causes serious adverse reactions for that particular treatment (AUROC = 0.79, average precision = 0.82). **d** Genes uniquely alter efficacy in one indicated drug but not another by primarily targeting the genes and biological functions used in treatment by the affected drug. In patients with Hypertensive Disease, a mutation in AGT alters the efficacy of Benazepril but not Diltiazem. Indeed, AGT exhibits a higher network importance in Benazepril treatment than in Diltiazem treatment, ranked as the 45th most important gene rather than the 418th most important gene. **e** Overall, 71.0% of genes known to alter efficacy in one indicated drug but not another exhibit higher network importance in treatment by the affected drug. **f** Diffusion profiles can identify biological functions that may help explain alterations in treatment. Shown are the proteins and biological functions identified as relevant to the treatment of Hypertensive Disease by Benazepril and Diltiazem. AGT, which uniquely alters the efficacy of Benazepril, is a key regulator of the renin–angiotensin system, a biological function harnessed by Benazepril in treatment but not by Diltiazem[70–72]. Abbreviations: reg. regulation, proc. process, + positive, − negative. Boxplots: median (line); 95% CI (notches); 1st, 3rd quartiles (boxes); data within 1.5 × the inter-quartile range from the 1st, 3rd quartiles (whiskers).

We find that the network importance of a gene in the drug and disease diffusion profiles predicts whether that gene alters drug efficacy and causes adverse reactions for that particular treatment (AUROC = 0.79, average precision = 0.82) (Fig. 3c). Importantly, the knowledge that a gene alters a given treatment is never directly encoded into our network. Instead, diffusion profiles predict treatment altering relationships that the multiscale interactome has never previously seen. Our diffusion profiles thereby provide a systematic approach to identify genes with the potential to alter treatment. Our finding is complementary to high-resolution, temporal approaches such as discrete dynamic models which model drug resistance and adverse reactions by first curating genes and pathways deemed relevant to a particular treatment[26–29]. Diffusion profiles may help provide candidate genes and pathways for inclusion in these detailed approaches, including genes not previously expected to be relevant. New treatment altering genes, if validated experimentally and clinically, could ultimately affect patient stratification in clinical trials and personalized therapeutic selection[62].

Finally, we find that when a gene in a diseased patient alters the efficacy of one indicated drug but not another, that gene primarily targets the genes important to treatment for the resistant drug (Fig. 3d, e). Overall, 71.0% of the genes known to alter the efficacy of 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 changes to genes can alter efficacy and cause adverse reactions in only some drugs indicated to treat a disease.

Consider Benazepril and Diltiazem, two drugs indicated to treat hypertensive disease (Fig. 3f). A mutation in the *AGT* gene alters the efficacy of Benazepril but not Diltiazem[63–65]. Indeed, our approach gives higher treatment importance to *AGT* in treatment by Benazepril than in treatment by Diltiazem, ranking *AGT* as the 45th most important gene for Benazepril treatment but only the 418th most important gene for Diltiazem treatment. Moreover, our approach explains why *AGT* alters the efficacy of Benazepril but not Diltiazem (Fig. 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 lower blood pressure[30,66–68]. By contrast, Benazepril operates at a systems-scale: Benazepril binds to ACE which affects the renin–angiotensin system, a systems-level biological function that controls blood pressure through hormones[30,69,70]. Crucially, *AGT* or Angiotensinogen, is a key component of the renin–angiotensin system[70–72]. Therefore, *AGT* affects the key biological function used by Benazepril to treat hypertensive disease. By contrast, *AGT* plays no direct role in the calcium receptor-driven pathways used by Diltiazem. Thus when a gene alters the efficacy of a drug, the multiscale interactome can identify biological functions that may help explain the alteration in treatment.

Discussion

The multiscale interactome provides a general approach to systematically understand how drugs treat diseases. By integrating physical interactions and biological functions, 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 identifies proteins and biological functions relevant to treatment. By contrast, existing systematic network approaches are black-boxes which make predictions without providing mechanistic insight. Finally, the multiscale interactome predicts what genes alter drug efficacy or cause severe adverse reactions for drug treatments and identifies biological functions that may explain how these genes interfere with treatment.

The multiscale interactome demonstrates that integrating biological functions into the interactome improves the systematic modeling of drug–disease treatment. Historically, systematic approaches to study treatment via the interactome have primarily focused on physical interactions between proteins[8–10,13]. Here, we find that integrating biological functions into a physical interactome improves the systematic modeling of nearly 6,000 treatments. We find drugs and drug categories which depend on biological functions for treatment. More broadly, incorporating biological functions may improve systematic approaches that currently use physical interactions to study disease pathogenesis[73–76], disease comorbidities[6], and drug combinations[22–24]. Harnessing the multiscale interactome in these settings may thus help answer key pharmacological questions. Moreover, the multiscale interactome can be readily expanded to add additional node types relevant to the problem at hand (i.e., microRNAs to study cancer initiation and progression[77]). Our finding is consistent with systematic studies which demonstrate, in other contexts, that networks involving functional information can strengthen prediction of cellular growth[25,78], identification of gene function[79–81], inference of drug targets[82], and general discovery of relationships between biological entities[83,84].

Moreover, we find that diffusion profiles incorporating both proteins and biological functions provide predictive power and interpretability in modeling drug–disease treatments. Diffusion profiles predict what drugs treat a given disease and identify proteins and biological functions relevant to treatment. In other pharmacological contexts, diffusion profiles incorporating proteins and biological functions may thus improve systematic approaches which currently employ proximity or other non-interpretable methods[6,16,17,33]. In studying the efficacy of drug combinations[17], diffusion profiles may identify synergistic effects on key biological functions. In studying the adverse reactions of drug combinations[16], diffusion profiles may identify biological functions which help explain polypharmacy side effects. In disease comorbidities[6,33], diffusion profiles may predict new comorbidities and identify biological functions which help explain the development of the comorbidity.

Finally, our study shows that both physical interactions and biological functions can propagate the effects of drugs and diseases. We find that many drugs neither directly target the proteins associated with the disease they treat nor target proximal proteins. Instead, these drugs affect the same biological functions disrupted by the disease. This view expands upon the current view of indirect effects embraced in other biological phenomena. In the omnigenic model of complex disease[85,86], 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, incorporating both physical interactions and biological functions, may help propagate indirect effects in complex disease. Altogether, the multiscale interactome provides a general computational paradigm for network medicine.

Methods

The multiscale interactome. The multiscale interactome captures how drugs use both a network of physical interactions and a rich hierarchy of biological functions to treat diseases. In the multiscale interactome, 1,661 drugs connect to the proteins they target (8,568 edges)[30,31]. 840 diseases connect to the proteins they disrupt through effects like genomic alterations, altered expression, or post-translational modification (25,212 edges)[32]. 17,660 proteins connect to other proteins based on physical interactions such as regulatory, metabolic, kinase-substrate, signaling, or binding relationships (387,626 edges)[33–39]. Proteins connect to the 9,798 biological functions they affect (34,777 edges)[40,41]. Finally, biological functions connect to each other in a rich hierarchy ranging from specific processes (i.e., embryonic heart tube elongation) to broad processes (i.e., heart development) (22,545 edges)[40,41]. Biological functions can describe processes involving molecules (i.e., DNA demethylation), cells (i.e., the mitotic cell cycle), tissues (i.e., muscle atrophy), organ systems (i.e., activation of the innate immune response), and the whole organism (i.e., anatomical structure development).

We visualize a representative subset of the multiscale interactome using Cytoscape[87] (Fig. 1b).

Drug–protein interactions. We map drugs to their protein targets using DrugBank[30] and the Drug Repurposing Hub[31]. For DrugBank, we map the Uniprot Protein IDs to Entrez IDs using HUGO[88]. 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[88]. We filter drug–target relationships to only include proteins that are represented in the network of physical interactions between proteins (see the "Methods" subsection "Protein–protein interactions"). All drug–target interactions are provided in Supplementary Data 1.

Disease–protein interactions. We map diseases to genes they affect through effects like genomic alterations, altered expression, or post-translational modification by using DisGeNet[32]. 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. 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 the "Methods" subsection "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[34] (BioGRID; 309,187 interactions between 16,352 proteins). BioGRID manually curates both physical and genetic interactions between proteins from 71,713 high-

throughput and low-throughput publications. We map BioGRID proteins to Entrez IDs by using HUGO[88]. We only include protein–protein interactions from BioGRID that result from experiments indicating a physical interaction between the proteins, as described by BioGRID[34], 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[36] (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[88]. We allow all experimental methods from DIP since they all capture physical interactions[36]. We use the "Hsapi20170205.txt" file.
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[39]: 2,611 interactions between 1,522 proteins; HI-II-14[35] 13,426 interactions between 4,228 proteins; Venkatesan-09[37]: 233 interactions between 229 proteins; Yu-11[38] 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[33] (138,425 interactions between 13,393 proteins). Finally, we integrate the physical protein–protein interaction network compiled by Menche et al. [33]. 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 Vinayagam et al. [89]. All proteins are provided in Entrez format and thus do not require further mapping.

Protein–biological function interactions. We map proteins to the biological functions they affect by using the human version of the Gene Ontology[40,41] (7,993 proteins; 6,387 biological functions; 34,777 edges). We only allow experimentally verified associations between genes and biological functions 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–biological function relationships that are inferred from physical interactions to avoid redundancy with the physical network of interacting proteins. We also exclude protein–biological function relationships inferred from gene expression patterns since the Gene Ontology states that such interactions are challenging to map to specific proteins[40,41]. 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 biological functions in the Gene Ontology have multiple synonymous IDs. For each biological function, we use the "master IDs" provided by GOATOOLS 0.8.4[90]. All protein–biological function interactions are provided in Supplementary Data 4.

Biological function–biological function interactions. We construct a hierarchy of biological functions by using the Gene Ontology's Biological Processes[40,41]. The Gene Ontology represents a curated hierarchy of biological functions, where highly specific biological functions are children of more general biological functions according to numerous relationship types. For example, "negative regulation of response to interferon-gamma" $\overset{\text{is a}}{\rightarrow}$ "negative regulation of innate immune response" $\overset{\text{is a}}{\rightarrow}$ "negative regulation of immune response" $\overset{\text{negativelyregulates}}{\rightarrow}$ "immune response." We allow relationships between biological functions of the following types: regulates, positively regulates, negatively regulates, part of, and is a. In order to allow the model to focus on the biological functions most relevant to treatment, we only consider biological functions which are associated with at least one drug target or one disease protein, either directly or implicitly through their children. All biological function–biological function 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[42] ($n_{\text{pairs}} = 2{,}538$; $n_{\text{drugs}} = 996$, $n_{\text{diseases}} = 463$), the Drug Repurposing Hub[31] ($n_{\text{pairs}} = 1{,}449$; $n_{\text{drugs}} = 908$, $n_{\text{diseases}} = 265$), and the Drug Indication Database[43] ($n_{\text{pairs}} = 3{,}304$; $n_{\text{drugs}} = 1{,}147$, $n_{\text{diseases}} = 615$). 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[30] and diseases are mapped to unique identifiers from the National Library of Medicine[91] (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[42]. 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[31]. We map drugs to DrugBank IDs by comparing their provided names and PubChem IDs to DrugBank's external links mapping[30]. We map diseases to UMLS CUIDs by using the UMLS Metathesaurus's REST API[91]. 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)[43]. 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[30]. 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[32].
   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[92]. 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 Diseases" 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 biological functions. 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[93]. 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[30] or the Drug Repurposing Hub[31] and

diseases with at least one associated gene in the curated version of DisGeNet[32] as these are the only drugs and diseases that the multiscale interactome represents (see the "Methods" subsection 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 biological functions 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 integrate proteins and biological functions. Diffusion profiles rely on a set of scalar weights which encode the relative importance of different types of nodes: $W = \{w_{\text{drug}}, w_{\text{disease}}, w_{\text{protein}}, w_{\text{biological function}}, w_{\text{higher-level biological function}}, w_{\text{lower-level biological function}}\}$. These weights are hyperparameters which we optimize when predicting the drugs that treat a given disease (see the "Methods" subsection "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 biological function neighbors, it is $\frac{w_{\text{protein}}}{w_{\text{biologicalfunction}}}$ times more likely to jump to the protein neighbors than the biological function neighbors. Notice that proteins connect to drugs, diseases, proteins, and biological functions, making $\{w_{\text{drug}}, w_{\text{disease}}, w_{\text{protein}}, w_{\text{biological function}}\}$ the relevant weights for a random walker currently at a protein. By contrast, biological functions connect to proteins, higher-level biological functions, and lower-level biological functions, making $\{w_{\text{protein}}, w_{\text{higher-level biological function}}, w_{\text{lower-level biological function}}\}$ the relevant weights for a random walker at a biological function. By providing separate weights for higher-level and lower-level biological functions, the random walker learns to explore different levels of the hierarchy of biological functions 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 biological function 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[94–96]. The diffusion profile computation takes as input:

1. $G = (V, E)$ the unweighted, undirected multiscale interactome with V nodes and E edges.
2. $W = \{w_{\text{drug}}, w_{\text{disease}}, w_{\text{protein}}, w_{\text{biological function}}, w_{\text{higher-level biological function}}, w_{\text{lower-level biological function}}\}$ 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-diffusion 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 biological functions. 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 (i.e., edges now include drug ↔ protein, disease ↔ protein, protein ↔ protein, protein ↔ biological function, and lower-level biological function ↔ higher-level biological function). 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 biological functions. 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:
$$p_t = \frac{w_t}{\sum_{t' \in T} w_{t'}}. \quad (1)$$
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:
$$\mathbf{M}_{ij} = p_{i \to j} = \frac{p_t}{n_t}. \quad (2)$$

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

$$\mathbf{r}^{(k+1)} = \overbrace{(1-\alpha)\mathbf{s}}^{\text{Restart walk}} + \alpha \overbrace{\left(\underbrace{\mathbf{r}^{(k)}\mathbf{M}}_{\text{from node with out-edges}} + \underbrace{\mathbf{s} \sum_{j \in J} \mathbf{r}_j^{(k)}}_{\text{from node without out-edges}} \right)}^{\text{Continue walk...}}. \quad (3)$$

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 $\epsilon = 1 \times 10^{-6}$. Pseudocode to compute diffusion profiles through power iteration is presented below.

```
% Algorithm: Diffusion profiles through power iteration
% Initialize diffusion profile
r_i^(0) = 1/|V| ∀i
% While not converged
while ||r^(k+1) − r^(k)||_1 > ε do
    % Start new walk at drug or disease node or continue walk.
    r^(k+1) = (1 − α)s + α(r^(k)M + s Σ_{j∈J} 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 biological functions 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 biological functions. 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 or distance. We compute the negative of the distance metrics.

1. L2 norm:
$$\sqrt{\sum_i |\mathbf{r}_i^{(c)} - \mathbf{r}_i^{(d)}|^2}, \quad (4)$$
2. L1 norm:
$$\sum_i |\mathbf{r}_i^{(c)} - \mathbf{r}_i^{(d)}|, \quad (5)$$
3. Canberra distance:
$$\sum_i \frac{|\mathbf{r}_i^{(c)} - \mathbf{r}_i^{(d)}|}{|\mathbf{r}_i^{(c)}| + |\mathbf{r}_i^{(d)}|}, \quad (6)$$
4. Cosine similarity:
$$\frac{\mathbf{r}^{(c)} \cdot \mathbf{r}^{(d)}}{||\mathbf{r}^{(c)}||_2 ||\mathbf{r}^{(d)}||_2}, \quad (7)$$

5. Correlation distance:

$$1 - \frac{(\mathbf{r}^{(c)} - \bar{\mathbf{r}}^{(c)}) \cdot (\mathbf{r}^{(d)} - \bar{\mathbf{r}}^{(d)})}{||(\mathbf{r}^{(c)} - \bar{\mathbf{r}}^{(c)})||_2 ||(\mathbf{r}^{(d)} - \bar{\mathbf{r}}^{(d)})||_2}. \quad (8)$$

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

Baseline metrics to predict what drugs will treat a disease. To predict what drugs will treat a given disease, we consider baselines that measure (1) the overlap between drug targets and disease proteins, (2) the overlap between the functions of drug targets and disease proteins, and (3) the state-of-the-art proximity metric on a molecular-scale interactome (Fig. 2b). First, we compute the "protein overlap" baseline which we define as the Jaccard Similarity between the set of drug targets T and the set of disease proteins S:

$$\frac{|T \cap S|}{|T \cup S|}. \quad (9)$$

Second, we compute the "functional overlap" baseline which we define as SimIC which measures the semantic similarity between the GO terms U associated with the drug targets and the GO terms V associated with the disease proteins[97]. We tested 17 functional overlap baselines, of which this was the best performing (see the "Methods" subsection "Baseline metrics of functional overlap between drug targets and disease proteins"; Supplementary Fig. 5). Third, we compute the state-of-the-art proximity metric on a molecular-scale interactome which is the closest distance metric in[10,13]. Let T be the set of drug targets, S be the set of disease proteins, and $l(s, t)$ be the shortest path length between nodes s and t. The state-of-the-art proximity metric first computes the "closest" distance

$$d(S, T) = \frac{1}{|T|} \sum_{t \in T} \min_{s \in S} l(s, t) \quad (10)$$

between S and T. Next, this distance is compared to a reference distance distribution which measures $d(S, T)$ when S and T are randomly permuted to 1000 sets of proteins that match the size and degrees of the original disease proteins and drug targets in the network. Finally, the state-of-the-art proximity metric is computed by taking a z-score of $d(S, T)$ with respect to the reference distribution:

$$z(S, T) = \frac{d(S, T) - \mu_{d(S,T)}}{\sigma_{d(S,T)}}. \quad (11)$$

Baseline metrics of functional overlap between drug targets and disease proteins. We tested 17 baseline methods that predict what drugs treat a disease by considering the biological functions affected by drug targets and disease proteins (Supplementary Fig. 5).

First, we tested baseline methods that compare the functional overlap between drug targets and disease proteins. Let U and V be the sets of Gene Ontology (GO) terms associated with drug targets and disease proteins respectively, either directly or through their descendant terms. Let U' and V' be the multisets of GO terms associated with drug targets and disease proteins respectively. Let U'' and V'' be the sets of GO terms enriched among drug targets and disease proteins according to Gene Set Enrichment Analysis (GSEA), respectively[98] (computed using GOATOOLS 0.8.4[90]). Note that in the multisets U' and V', U'_i and V'_i correspond to the number of occurrences of the ith element in the multiset.

We measure the following baselines:

- The Jaccard Similarity or Intersection between the set of GO terms associated with the drug targets and the set of GO terms associated with the disease proteins:

$$\frac{|U \cap V|}{|U \cup V|} \text{ or } |\mathrm{U} \cap \mathrm{V}|, \quad (12)$$

- The Jaccard Similarity or Intersection between the multiset of GO terms associated with the drug targets and the multiset of GO terms associated with the disease proteins:

$$\frac{\sum_i \min(U'_i, V'_i)}{\sum_i \max(U'_i, V'_i)} \text{ or } \sum_{\mathrm{i}} \min(\mathrm{U}'_{\mathrm{i}}, \mathrm{V}'_{\mathrm{i}}), \quad (13)$$

- The Jaccard Similarity or Intersection between the set of GO terms enriched among drug targets and the set of GO terms enriched among disease proteins according to Gene Set Enrichment Analysis[90,98]:

$$\frac{|U'' \cap V''|}{|U'' \cup V''|} \text{ or } |\mathrm{U}'' \cap \mathrm{V}''|, \quad (14)$$

- The z-scored Jaccard Similarity or Intersection between the set of GO terms associated with the drug targets and the set of GO terms associated with the disease proteins:

$$z\left(\frac{|U \cap V|}{|U \cup V|}\right) \text{ or } z(|\mathrm{U} \cap \mathrm{V}|), \quad (15)$$

- The z-scored Jaccard Similarity or Intersection between the multisets of GO terms associated with the drug targets and the set of GO terms associated with the disease proteins:

$$z\left(\frac{\sum_i \min(U'_i, V'_i)}{\sum_i \max(U'_i, V'_i)}\right) \text{ or } z\left(\sum_{\mathrm{i}} \min(\mathrm{U}'_{\mathrm{i}}, \mathrm{V}'_{\mathrm{i}})\right). \quad (16)$$

We compute reference distributions for z-scored metrics by following the approach in refs. [10,13]. Specifically, we randomly permute the set of disease proteins S and the set of drug targets T to 1000 sets of proteins that match the size and degrees of the original disease proteins and drug targets in the network. We then generate the GO sets and multisets that correspond to the permuted S and T, compute the relevant baseline metric, and repeat this for random permutations of S and T to generate a reference distribution. Finally, we compute a z-score by comparing the baseline metric for the true S and T to the reference distribution.

Second, we tested baseline methods that calculate the semantic similarity between the GO terms associated with the drug targets and those associated with the disease proteins[99]. Consider U and V, now defined as the sets of GO terms directly associated with drug targets and disease proteins, respectively. Semantic similarity methods first define a similarity $\mathrm{sim}(u, v)$ between a GO term directly associated with drug targets u and a GO term directly associated with disease proteins v. The similarity of the sets U and V are subsequently calculated by aggregating across the similarities of pairwise GO terms u and v.

We used the following semantic similarity metrics as as they are among the most common and best-performing metrics in a variety of settings[99].

- The Resnik Similarity[100,101] between u and v measures the information content of the most informative common ancestor between u and v:

$$\mathrm{sim}(\mathrm{u}, \mathrm{v}) = \mathrm{Resnik}(\mathrm{u}, \mathrm{v}) = \mathrm{IC}[\mathrm{MICA}(\mathrm{u}, \mathrm{v})]. \quad (17)$$

Let $p(u)$ be the fraction of proteins in the multiscale interactome that are associated with a GO term u or its descendants. The information content IC of term u is defined as

$$\mathrm{IC}(\mathrm{u}) = -\log[\mathrm{p}(\mathrm{u})]. \quad (18)$$

The maximum informative common ancestor (MICA) between two GO terms u and v is defined as

$$\mathrm{MICA}(u, v) = \underset{x \in \text{ ancestors } (u,v)}{\mathrm{argmax}} \mathrm{IC}(x). \quad (19)$$

- simIC[97] integrates both the information content of GO terms and the structural information of the GO hierarchy to determine the similarity between GO terms u and v:

$$\mathrm{sim}(\mathrm{u}, \mathrm{v}) = \mathrm{simIC}(\mathrm{u}, \mathrm{v}) = \frac{2\log[\mathrm{p}(\mathrm{MICA}(\mathrm{u}, \mathrm{v})]}{\log[\mathrm{p}(\mathrm{u})] + \log[\mathrm{p}(\mathrm{v})]} \times \left(1 - \frac{1}{1 + \mathrm{IC}[\mathrm{MICA}(\mathrm{u}, \mathrm{v})]}\right). \quad (20)$$

- simGIC[102] which considers the information content of all common ancestors of the GO terms directly associated with the drug targets U and the GO terms directly associated with the disease proteins V:

$$\mathrm{sim}(\mathrm{u}, \mathrm{v}) = \mathrm{simGIC}(\mathrm{U}, \mathrm{V}) = \frac{\sum_{\mathrm{x} \in \mathrm{A}(\mathrm{U}) \cap \mathrm{A}(\mathrm{V})} \mathrm{IC}(\mathrm{x})}{\sum_{\mathrm{x} \in \mathrm{A}(\mathrm{U}) \cup \mathrm{A}(\mathrm{V})} \mathrm{IC}(\mathrm{x})}. \quad (21)$$

Here, $A(X)$ is the set of terms within X and all their ancestors in the GO hierarchy.

We aggregated the Resnik Similarity and simIC across U and V by using the average, maximum, and best match average approaches:

- Average:

$$\frac{1}{|U||V|} \sum_{u \in U} \sum_{v \in V} \mathrm{sim}(\mathrm{u}, \mathrm{v}), \quad (22)$$

- Max:

$$\max_{u,v \in U \times V} \mathrm{sim}(\mathrm{u}, \mathrm{v}), \quad (23)$$

- Best match average[103]:

$$\frac{1}{|U|+|V|}\left[\sum_{u\in U}\max_{v\in V}\mathrm{sim}(u,v)+\sum_{v\in V}\max_{u\in U}\mathrm{sim}(u,v)\right]. \quad (24)$$

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 (Fig. 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 $W = \{w_{\text{drug}}, w_{\text{disease}}, w_{\text{protein}}, w_{\text{biological function}}, w_{\text{higher-level biological function}}, w_{\text{lower-level biological function}}\}$ 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 training set, averaged across folds.

After initial coarse explorations for each hyperparameter, we sweep across 486 combinations of hyperparameters sampled linearly within the following ranges $w_{\text{drug}} \in [3, 9]$, $w_{\text{disease}} \in [3, 9]$, $w_{\text{protein}} \in [3, 9]$, $w_{\text{higher-level biological function}} \in [1.5, 4.5]$, $w_{\text{lower-level biological function}} \in [1.5, 4.5]$, $\alpha \in [0.85, 0.9]$and set $w_{\text{biological function}} = w_{\text{higher-level biological function}} + w_{\text{lower-level biological function}}$. We also sweep across the seven comparison metrics described above. We repeat this procedure for 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 $\mathbf{r}^{(d)}$ (Fig. 2c, Supplementary Note 1, Supplementary Fig. 7). The optimal weights for the multiscale interactome are $w_{\text{drug}} = 3.21$, $w_{\text{disease}} = 3.54$, $w_{\text{protein}} = 4.40$, $w_{\text{higher-level biological function}} = 2.10$, $w_{\text{lower-level biological function}} = 4.49$, $w_{\text{biological function}} = 6.58$ with $\alpha = 0.860$ and use the correlation distance to compare $\mathbf{r}^{(c)}$ and $\mathbf{r}^{(d)}$ (Fig. 2b, c). We utilize these optimal weights for the multiscale interactome for all subsequent sections. Optimized diffusion profiles are provided in Supplementary Data 10. Additional information on selecting the edge weight ranges is provided as Supplementary Note 2.

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)[104]. We map all drugs to their ATC class by using DrugBank's XML database "full_database.xml"[30]. 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 Fig. 6) and second to understand the differential performance of the best multiscale interactome model compared to the best molecular-scale interactome model using diffusion profiles (Fig. 2c; Supplementary Fig. 7). The ATC classification for the drugs in our study is provided in Supplementary Data 7.

Diffusion profiles identify proteins and biological functions related to treatment. For a given drug–disease pair, diffusion profiles identify the proteins and biological functions related to treatment. For each drug–disease pair, we select the top k proteins and biological functions in the drug diffusion profile and in the disease diffusion profile. To explain the relevance of these proteins and biological functions to treatment, we induce a subgraph on these nodes and remove any isolated components. We set $k = 10$ for the case studies in Figs. 2g, h, and 3f. We focus on these nodes since the nodes ranked most highly in the diffusion profiles have the highest propagated effect and are thus considered the most relevant to treatment. Additionally, these top nodes also capture a substantial fraction of the overall visitation frequency in the diffusion profile (i.e., about 50% for Fig. 2g, h). We additionally include the rankings of the top 20 proteins and biological functions for each case study as Supplementary Figs. 16–18.

Validation of diffusion profiles through gene expression signatures. To validate drug diffusion profiles, we compare drug diffusion profiles to the drug gene expression signatures present in the Broad Connectivity Map[48,49] (Fig. 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[30].

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 quality control metrics and guidelines[48,49] 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 expression 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 three 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:

$$\frac{1}{2}\left[\frac{|U_{\text{drug1}} \cap U_{\text{drug2}}|}{|U_{\text{drug1}} \cup U_{\text{drug2}}|} + \frac{|D_{\text{drug1}} \cap D_{\text{drug2}}|}{|D_{\text{drug1}} \cup D_{\text{drug2}}|}\right]. \quad (25)$$

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 variants that alter treatment. We compile genetic variants that alter treatment by using the Pharmacogenomics Knowledgebase (PharmGKB)[65]. PharmGKB is a gold-standard database mapping the effect of genetic variants 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 variants at the gene level alter treatments. PharmGKB's "clinical_ann_metadata.tsv" file provides triplets of drugs, diseases, and genetic variants known to alter treatment. Treatment alteration occurs when a genetic variant 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 gene based on diffusion profiles. We define the treatment importance (TI) of 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 gene i is

$$\mathrm{TI}(\mathrm{i}|\mathrm{c},\mathrm{d}) = \mathbf{r}_{\mathrm{i}}^{(\mathrm{c})} * \mathbf{r}_{\mathrm{i}}^{(\mathrm{d})}. \quad (26)$$

We define the treatment importance percentile as the percentile rank of TI($i|c$, d) compared to all other genes for the same drug and disease. Intuitively, 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 genes known to alter a treatment with the treatment importance of other genes (Fig. 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[65]. 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 genes and a distribution of the treatment importance for other genes (Fig. 3b).

Predicting genes that alter a treatment based on treatment importance. We evaluate the ability of treatment importance to predict the genes that will alter a given treatment (Fig. 3c). For each (drug, disease, gene) triplet, we use the treatment importance of the gene 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 the "Methods" subsection "Comparing treatment importance of treatment altering genetic mutations vs. other genetic mutations"). We assess performance using AUROC and average precision (Fig. 3c).

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

Formally, let i be a gene. Consider a triplet (d, c_{altered}, $c_{\mathrm{unaltered}}$) 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 $c_{\mathrm{unaltered}}$ 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_{altered}, $c_{\mathrm{unaltered}}$) for which the treatment importance of i is higher in the (c_{altered}, d) treatment than in the ($c_{\mathrm{unaltered}}$, d) treatment, as shown below. We only consider genes for which $n_{\mathrm{triplets}} \geq 100$.

$$f[\,\mathrm{TI}(i|c_{\mathrm{altered}},d) > \mathrm{TI}(i|c_{\mathrm{unaltered}},d)] = \frac{\sum_{\forall(d,c_{\mathrm{altered}},c_{\mathrm{unaltered}})} \mathbb{1}\{\,\mathrm{TI}(i|c_{\mathrm{altered}},d) > \mathrm{TI}(i|c_{\mathrm{unaltered}},d)\}}{n_{\mathrm{triplets}}} \quad (27)$$

Analyzing whether distant proteins can have common biological functions. We analyzed whether two proteins can be more distant than expected by random chance in a physical protein–protein interaction (PPI) network yet affect the same function (Supplementary Fig. 2). To run this analysis, we first compute the set of all protein pairs that are both present in the protein–protein interaction network described previously (see the "Methods" subsection "Protein–protein interactions") and are also associated with a common biological function. We only consider direct associations of proteins to biological functions (i.e., we do not propagate associations up the GO hierarchy) in order to ensure that shared biological functions are specific and not generic (i.e., shared associations with the GO term 'Biological Process').

For each protein pair with a common biological function, we then:

1. Compute the shortest path distance in the PPI network between these two proteins.
2. Construct a reference distribution of shortest paths for these two protein pairs by following the approach in refs. [10,13]. Specifically, we repeatedly, randomly sample other proteins in the network with similar degree to the original proteins and measure the shortest path distance between them. These randomly sampled proteins do not necessarily share a common biological function.
3. Using the true shortest path distance between the proteins and the random reference distribution of shortest path distances, we compute a z-score. The z-score captures whether the proteins with a shared function are closer or further away than expected by random chance in the PPI network.

Construction of alternative multiscale interactomes that explicitly represent cells, tissues, and organs. We constructed three alternative multiscale interactomes which explicitly represent cells, tissues, and organs (Supplementary Note 4, Supplementary Fig. 8). In these alternative multiscale interactomes, the nodes and edges in the original multiscale interactome are all present. Additionally, (1) human cells, tissues, and organs are added as additional nodes; (2) edges between these cell, tissue, and organ nodes are added according to relationships defined in established anatomical ontologies; and (3) edges between GO biological function nodes and cell, tissue, and organ nodes are added according to relationships provided in Gene Ontology Plus (GO Plus)[105]. GO Plus maintains a curated set of relationships between the biological functions in GO and the cell, tissue, and organ nodes present in two key anatomical ontologies: Uberon and the Cell Ontology. We thus constructed three alternative multiscale interactomes incorporating human subsets of Uberon, the Cell Ontology, and both Uberon and the Cell Ontology.

1. Multiscale Interactome + Uberon: Uberon is an ontology covering anatomical structures in animals[106,107]. Uberon nodes include tissues (i.e., cardiac muscle tissue UBERON:0001133), organs (i.e., heart UBERON:0000948), and organ systems (i.e., cardiovascular system UBERON:0004535). We utilized GO Plus (i.e., "go-plus.owl") to link GO biological function nodes present in our original network to Uberon nodes present in a human-specific subset of Uberon (i.e., "subsets/human-view.obo"). Edges between Uberon nodes, which encode anatomical relationships, were also added according to "subsets/human-view.obo".
2. Multiscale Interactome + Cell Ontology: The Cell Ontology is an ontology for the representation of in vivo cell types[108,109]. Nodes consist primarily of cell types and their hierarchical relationships (i.e., epithelial cell CL:0000066, epithelial cell of pancreas CL:0000083, pancreatic A cell CL:0000171). We utilized a human-specific subset of the Cell Ontology previously prepared by the Human Cell Atlas Ontology[110]. We utilized GO Plus to link GO biological function nodes in our original network to Cell Ontology terms and the Cell Ontology (i.e., "cl-basic.obo") to link Cell Ontology terms with one another.
3. Multiscale Interactome + Uberon + Cell Ontology: The Multiscale Interactome + Uberon + Cell Ontology network contains all nodes and edges present in our original network as well as nodes and edges added via GO Plus, Uberon, and Cell Ontology as described above.

Prediction of what drugs treat a given disease in alternative multiscale interactomes. We evaluate the ability of diffusion profiles to predict what drugs treat a given disease in the alternative multiscale interactomes (see the "Methods" subsection "Construction of alternative multiscale interactomes that explicitly represent cells, tissues, and organs"; Supplementary Note 4, Supplementary Fig. 8). Given the presence of new node types, we modify the edge weight hyperparameters used in the calculation of diffusion profiles. We then sweep over the full set of edge weight hyperparameters according to the broad hyperparameter sweep described in Supplementary Note 2, in which we sample 560 combinations of hyperparameters sampled linearly in the range [1, 100]. The new sets of edge weight hyperparameters and their optimal values are present below:

1. Multiscale Interactome + Uberon: The optimal weights for Multiscale Interactome + Uberon are $w_{\mathrm{drug}} = 55.2$, $w_{\mathrm{disease}} = 27.3$, $w_{\mathrm{protein}} = 76.8$, $w_{\mathrm{biological\ function}} = 66.1$, $w_{\mathrm{uberon}} = 82.2$, $w_{\mathrm{higher\text{-}level\ biological\ function\ or\ uberon}} = 67.1$, $w_{\mathrm{lower\text{-}level\ biological\ function\ or\ uberon}} = 45.7$ with $\alpha = 0.76$ and use the correlation distance to compare $\mathbf{r}^{(c)}$ and $\mathbf{r}^{(d)}$.
2. Multiscale Interactome + Cell Ontology: The optimal weights for Multiscale Interactome + Cell Ontology are $w_{\mathrm{drug}} = 39.0$, $w_{\mathrm{disease}} = 17.1$, $w_{\mathrm{protein}} = 72.4$, $w_{\mathrm{biological\ function}} = 60.0$, $w_{\mathrm{cell\ ontology}} = 23.1$, $w_{\mathrm{higher\text{-}level\ biological\ function\ or\ cell\ ontology}} = 25.7$, $w_{\mathrm{lower\text{-}level\ biological\ function\ or\ cell\ ontology}} = 22.8$ with $\alpha = 0.83$ and use the correlation distance to compare $\mathbf{r}^{(c)}$ and $\mathbf{r}^{(d)}$.
3. Multiscale Interactome + Uberon + Cell Ontology: The optimal weights for Multiscale Interactome + Uberon + Cell Ontology are $w_{\mathrm{drug}} = 60.2$, $w_{\mathrm{disease}} = 12.8$, $w_{\mathrm{protein}} = 42.3$, $w_{\mathrm{biological\ function}} = 78.4$, $w_{\mathrm{uberon}} = 70.0$, $w_{\mathrm{cell\ ontology}} = 91.7$, $w_{\mathrm{higher\text{-}level\ biological\ function\ or\ uberon\ or\ cell\ ontology}} = 26.7$, $w_{\mathrm{lower\text{-}level\ biological\ function\ or\ uberon\ or\ cell\ ontology}} = 76.1$ with $\alpha = 0.82$ and use the correlation distance to compare $\mathbf{r}^{(c)}$ and $\mathbf{r}^{(d)}$.

Statistics and reproducibility. All boxplots depict the median (line), 95% CI (notches), and 1st and 3rd quartiles (boxes). Whiskers depict data within 1.5 × the inter-quartile range from the 1st and 3rd quartiles. Data beyond the whiskers are considered outliers.

No new experimental findings are reported in this manuscript. Reproducibility of the computational analyses in the manuscript are ensured through clear representation of the methods used and the public release of both code and data. The findings in this study are based on the random walk-based model described in the manuscript and the resulting analyses are based on this model. All attempts at replication were successful.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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 publicly available at github.com/snap-stanford/multiscale-interactome[111]. This manuscript uses and compiles data from numerous public data sources including: DrugBank (5.1.1, accessed July 2018; https://go.drugbank.com/)[30], the Drug Repurposing Hub (September 2018; https://clue.io/repurposing)[31], the Drug Repurposing Database (May 2018; http://apps.chiragjpgroup.org/repoDB/)[42], the Drug Indication Database[43], DisGeNet (March 2018; https://www.disgenet.org/)[32], Disease Ontology (July 5, 2018; https://disease-ontology.org/)[93], HUGO (October 2018; https://www.genenames.org/)[88], the Unified Medical Language System (https://www.nlm.nih.gov/research/umls/index.html)[91], the Biological General Repository for Interaction Datasets (3.5.178, November 2019; https://thebiogrid.org/)[34], the Database of Interacting Proteins (February 2017; https://dip.doe-mbi.ucla.edu/dip/Main.cgi)[36], the Human Reference Protein Interactome Mapping Project (http://www.interactome-atlas.org/)[35,37–39], Menche-2015[33], the Gene Ontology[40,41] and Gene Ontology Plus (February 2018; July 2020; http://geneontology.org/)[105,112], the Broad Connectivity Map (June 2019; https://clue.io/cmap)[48,49], the Pharmacogenomics Knowledgebase (September 2018; https://www.pharmgkb.org/)[65], Uberon (July 2020; http://uberon.github.io/)[106,107], the Cell Ontology (August 2020; http://www.obofoundry.org/ontology/cl.html)[108,109], and the Human Cell Atlas Ontology (August 2020; https://github.com/HumanCellAtlas/ontology)[110].

Code availability

Python implementation of our methodology is available at github.com/snap-stanford/multiscale-interactome[111]. All analyses were performed using Python 3.7, NetworkX 2.3, NumPy 1.16.2, Pandas 0.24.2, Scipy 1.3.0, GOATOOLS 0.8.4. Additional packages used are present in the requirements.txt file at the GitHub repository. Please read the README for information on downloading and running the code.

Received: 22 May 2020; Accepted: 4 February 2021;
Published online: 19 March 2021

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. *Nat. Methods* **12**, 615–621 (2015).
3. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **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. *Nat. Genet.* **47**, 106–114 (2015).
5. Nikolsky, Y., Nikolskaya, T. & Bugrim, A. Biological networks and analysis of experimental data in drug discovery. *Drug Discov. Today* **10**, 653–662 (2005).
6. Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.* **17**, 615–629 (2016).
7. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res.* **25**, 142–154 (2015).
8. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
9. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).
10. Cheng, F. et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9**, 2691 (2018).
11. Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **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. *Brief. Bioinform.* **19**, 878–892 (2018).
13. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **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. *Nat. Commun.* **8**, 573 (2017).
16. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).
17. Cheng, F., Kovacs, I. A. & Barabasi, A.-L. Network-based prediction of drug combinations. *Nat. Commun.* **10**, 1197 (2019).
18. Hu, Y. et al. Optimal control nodes in disease-perturbed networks as targets for combination therapy. *Nat. Commun.* **10**, 2180 (2019).
19. Firestone, A. J. & Settleman, J. A three-drug combination to treat BRAF-mutant cancers. *Nat. Med.* **23**, 913–914 (2017).
20. Zhao, S. & Iyengar, R. Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. *Annu. Rev. Pharmacol. Toxicol.* **52**, 505–521 (2012).
21. Walpole, J., Papin, J. A. & Peirce, S. M. Multiscale computational models of complex biological systems. *Annu. Rev. Biomed. Eng.* **15**, 137–154 (2013).
22. van Hasselt, J. C. & Iyengar, R. Systems pharmacology: defining the interactions of drug combinations. *Annu. Rev. Pharmacol. Toxicol.* **59**, 21–40 (2019).
23. Han, K. et al. Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat. Biotechnol.* **35**, 463–474 (2017).
24. Jia, J. et al. Mechanisms of drug combinations: interaction and network perspectives. *Nat. Rev. Drug Discov.* **8**, 111–128 (2009).
25. Yu, M. K. et al. Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell Syst.* **2**, 77–88 (2016).
26. Zañudo, J. G. T., Scaltriti, M. & Albert, R. A network modeling approach to elucidate drug resistance mechanisms and predict combinatorial drug treatments in breast cancer. *Cancer Converg.* **1**, 5 (2017).
27. Zañudo, J. G., Steinway, S. N. & Albert, R. Discrete dynamic network modeling of oncogenic signaling: Mechanistic insights for personalized treatment of cancer. *Curr. Opin. Syst. Biol.* **9**, 1–10 (2018).
28. Trachana, K. et al. Taking systems medicine to heart. *Circ. Res.* **122**, 1276–1289 (2018).
29. Montagud, A. et al. Conceptual and computational framework for logical modelling of biological networks deregulated in diseases. *Brief. Bioinform.* **20**, 1238–1249 (2019).
30. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2017).
31. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
32. Piñero, J. et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2016).
33. Menche, J. et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
34. Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2019).
35. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
36. Salwinski, L. et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
37. Venkatesan, K. et al. An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
38. Yu, H. et al. Next-generation sequencing to generate interactome datasets. *Nat. Methods* **8**, 478–480 (2011).
39. Rual, J.-F. et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).
40. Gene Ontology Consortium. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2018).
41. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
42. Brown, A. S. & Patel, C. J. A standard database for drug repositioning. *Sci. Data* **4**, 170029 (2017).
43. Sharp, M. E. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. *J. Biomed. Semant.* **8**, 2 (2017).
44. Donnat, C., Zitnik, M., Hallac, D. & Leskovec, J. Learning structural node embeddings via diffusion wavelets. In *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (eds Guo, Y. & Farooq F.) 1320–1329 (Assocation for Computing Machinery, 2018).
45. Cao, M. et al. Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLOS ONE* **8**, e76339 (2013).
46. Nielsen, S. et al. Vasopressin increases water permeability of kidney collecting duct by inducing translocation of aquaporin-CD water channels to plasma membrane. *Proc. Natl. Acad. Sci. USA* **92**, 1013–1017 (1995).
47. Holmes, C. L., Landry, D. W. & Granton, J. T. Science review: vasopressin and the cardiovascular system part 1–receptor physiology. *Crit. Care* **7**, 427–434 (2003).
48. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
49. Lamb, J. et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).

50. 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 Lett.* **56**, 352–355 (1975).
51. Utermann, G. et al. Polymorphism of apolipoprotein E: genetics of hyperlipoproteinemia type III. *Clin. Genet.* **15**, 37–62 (1979).
52. Ghiselli, G., Schaefer, E. J., Gascon, P. & Breser, H. Type III hyperlipoproteinemia associated with apolipoprotein E deficiency. *Science* **214**, 1239–1241 (1981).
53. Wang, J. et al. APOA5 genetic variants are markers for classic hyperlipoproteinemia phenotypes and hypertriglyceridemia. *Nat. Clin. Pract. Cardiovasc. Med.* **5**, 730–737 (2008).
54. Evans, D., Seedorf, U. & Beil, F. Polymorphisms in the apolipoprotein a5 (APOA5) gene and type III hyperlipidemia. *Clin. Genet.* **68**, 369–372 (2005).
55. Moghadasian, M. H. Clinical pharmacology of 3-hydroxy-3-methylglutaryl coenzyme a reductase inhibitors. *Life Sci.* **65**, 1329–1337 (1999).
56. Holdgate, G., Ward, W. & McTaggart, F. Molecular mechanism for inhibition of 3-hydroxy-3-methylglutaryl CoA (HMG-CoA) reductase by rosuvastatin. *Biochem. Soc. Trans.* **31**, 528–531 (2003).
57. Shinkai, K., McCalmont, T. & Leslie, K. Cryopyrin-associated periodic syndromes and autoinflammation. *Clin. Exp. Dermatol.* **33**, 1–9 (2008).
58. Kone-Paut, I. & Galeotti, C. Anakinra for cryopyrin-associated periodic syndrome. *Expert Rev. Clin. Immunol.* **10**, 7–18 (2014).
59. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).
60. Goldstein, D. B., Tate, S. K. & Sisodiya, S. M. Pharmacogenetics goes genomic. *Nat. Rev. Genet.* **4**, 937–947 (2003).
61. Hansen, N. T., Brunak, S. & Altman, R. Generating genome-scale candidate gene lists for pharmacogenomics. *Clin. Pharmacol. Ther.* **86**, 183–189 (2009).
62. Karczewski, K. J., Daneshjou, R. & Altman, R. B. Chapter 7: Pharmacogenomics. *PLoS Comput. Biol.* **8**, e1002817 (2012).
63. 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).
64. Yu, H. et al. A core promoter variant of angiotensinogen gene and interindividual variation in response to angiotensin-converting enzyme inhibitors. *J. Renin-Angiotensin-Aldosterone Syst.* **15**, 540–546 (2014).
65. Whirl-Carrillo, M. et al. Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).
66. Nayler, W. G. & Dillon, J. Calcium antagonists and their mode of action: an historical overview. *Br. J. Clin. Pharmacol.* **21**, 97S–107S (1986).
67. Sutton, M. S. J. & Morad, M. Mechanisms of action of diltiazem in isolated human atrial and ventricular myocardium. *J. Mol. Cell. Cardiol.* **19**, 497–508 (1987).
68. O'Connor, S. E., Grosset, A. & Janiak, P. The pharmacological basis and pathophysiological significance of the heart rate-lowering property of diltiazem. *Fundam. Clin. Pharmacol.* **13**, 145–153 (1999).
69. Balfour, J. A. & Goa, K. L. Benazepril. *Drugs* **42**, 511–539 (1991).
70. Lavoie, J. L. & Sigmund, C. D. Minireview: overview of the renin–angiotensin system—an endocrine and paracrine system. *Endocrinology* **144**, 2179–2183 (2003).
71. Caulfield, M. et al. Linkage of the angiotensinogen gene to essential hypertension. *New Engl. J. Med.* **330**, 1629–1633 (1994).
72. Jeunemaitre, X. et al. Molecular basis of human hypertension: role of angiotensinogen. *Cell* **71**, 169–180 (1992).
73. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
74. Jones, D. Pathways to cancer therapy. *Nat. Rev. Drug Discov.* **7**, 875–876 (2008).
75. Jones, S. et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
76. Parsons, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
77. Di Leva, G., Garofalo, M. & Croce, C. M. MicroRNAs in cancer. *Annu. Rev. Pathol.* **9**, 287–314 (2014).
78. Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 (2018).
79. Cho, H., Berger, B. & Peng, J. Compact integration of multi-network topology for functional analysis of genes. *Cell Syst.* **3**, 540–548 (2016).
80. Wang, S., Cho, H., Zhai, C., Berger, B. & Peng, J. Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* **31**, i357–i364 (2015).
81. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
82. Yamanishi, Y., Kotera, M., Kanehisa, M. & Goto, S. Drug–target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* **26**, i246–i254 (2010).
83. Balaji, S., Mcclendon, C., Chowdhary, R., Liu, J. S. & Zhang, J. IMID: integrated molecular interaction database. *Bioinformatics* **28**, 747–749 (2012).
84. Bell, L., Chowdhary, R., Liu, J. S., Niu, X. & Zhang, J. Integrated bio-entity network: a system for biological knowledge discovery. *PLoS ONE* **6**, e21474 (2011).
85. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
86. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034 (2019).
87. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
88. Braschi, B. et al. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.* **47**, D786–D792 (2019).
89. Vinayagam, A. et al. A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* **4**, rs8–rs8 (2011).
90. Klopfenstein, D. V. et al. GOATOOLS: a python library for gene ontology analyses. *Sci. Rep.* **8**, 1–17 (2018).
91. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
92. Davis, A. P. et al. The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Res.* **47**, D948–D954 (2019).
93. Schriml, L. M. et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).
94. Langville, A. N. & Meyer, C. D. A survey of eigenvector methods for web information retrieval. *SIAM Rev.* **47**, 135–161 (2005).
95. Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report (Stanford InfoLab., 1999).
96. Hagberg, A., Swart, P. & Schult, D. Exploring network structure, dynamics, and function using NetworkX. In *Proc. 7th Python in Science Conferences (SciPy)*, (eds Gael, V., Travis V. & Jarrod, M.) 11–16 (Los Alamos National Lab, 2008).
97. Li, B., Luo, F., Wang, J. Z., Feltus, F. A. & Zhou, J. Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins. In *International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, (eds Gael, V. et al.) 166–172 (CSREA Press, 2010).
98. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
99. Pesquita, C. Semantic similarity in the Gene Ontology. In *The Gene Ontology Handbook*, (eds Dessimoz, C. & Škunca, N.) 161–173 (Humana Press, 2017).
100. Lord, P. W., Stevens, R. D., Brass, A. & Goble, C. A. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**, 1275–1283 (2003).
101. Resnik, P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11**, 95–130 (1999).
102. Pesquita, C. et al. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinform.* **9**, S4 (2008).
103. Azuaje, F., Wang, H. & Bodenreider, O. Ontology-driven similarity approaches to supporting gene functional assessment. In *Proc. ISMB'2005 SIG Meeting on Bio-ontologies*, Vol. 2005, 9–10 (ISMB, 2005).
104. World Health Organization. *The Anatomical Therapeutic Chemical Classification System with Defined Daily doses-ATC/DDD* (World Health Organization, 2009).
105. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
106. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).
107. Haendel, M. A. et al. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J. Biomed. Semant.* **5**, 21 (2014).
108. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
109. Diehl, A. D. et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.* **7**, 1–10 (2016).
110. Welter, D., Jupp, S. & Osumi-Sutherland, D. Human Cell Atlas Ontology. In *Proc. 9th International Conference on Biological Ontology (ICBO)* (eds Jaiswal, P., Cooper, L., Haendel, M. A. & Mungall, C. J.) Vol. 2285 (CEUR-WS.org, 2018).
111. Ruiz, C., Zitnik, M. & Leskovec, J. *Identification of Disease Treatment Mechanisms Through the Multiscale Interactome, GitHub* https://doi.org/10.5281/zenodo.4435258 (2021).
112. Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* **38**, D331–D335 (2010).

Acknowledgements

We thank Dr. Emma Pierson and Dr. Maria Brbic for helpful discussions and feedback on our manuscript. C.R. is supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1656518 and a Stanford Enhancing Diversity in Graduate Education (EDGE) Fellowship. M.Z. is supported, in part, by NSF grant nos. IIS311 2030459 and IIS-2033384 and by the Harvard Data Science Initiative. We also gratefully acknowledge the support of DARPA under Nos. N660011924033 (MCS); ARO under Nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under Nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions), IIS-2030477 (RAPID); Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Chan Zuckerberg Biohub, Amazon, JPMorgan Chase, Docomo, Hitachi, JD.com, KDDI, NVIDIA, Dell, Toshiba, and UnitedHealth Group. J.L. is a Chan Zuckerberg Biohub investigator.

Author contributions

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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21770-8.

Correspondence and requests for materials should be addressed to J.L.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.