Building a knowledge graph to enable precision medicine

Payal Chandak[1,⋆], Kexin Huang[2,⋆], and Marinka Zitnik[3,4,5,‡]
[1] Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139, USA
[2]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[3]Department of Biomedical Informatics, Harvard University, Boston, MA 02115, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[5]Harvard Data Science Initiative, Cambridge, MA 02138, USA

⋆ Equal contribution
‡ Corresponding author: marinka@hms.harvard.edu

Developing personalized diagnostic strategies and targeted treatments requires a deep understanding of disease biology and the ability to dissect the relationship between molecular and genetic factors and their phenotypic consequences. However, such knowledge is fragmented across publications, non-standardized research repositories, and evolving ontologies describing various scales of biological organization between genotypes and clinical phenotypes. Here, we present PrimeKG, a precision medicine-oriented knowledge graph that provides a holistic view of diseases. PrimeKG integrates 20 high-quality resources to describe 17,080 diseases with 4,050,249 relationships representing ten major biological scales, including disease-associated protein perturbations, biological processes and pathways, anatomical and phenotypic scale, and the entire range of approved and experimental drugs with their therapeutic action, considerably expanding previous efforts in disease-rooted knowledge graphs. PrimeKG supports drug-disease prediction by including an abundance of 'indications', 'contradictions' and 'off-label use' edges, which are usually missing in other knowledge graphs. We accompany PrimeKG's graph structure with text descriptions of clinical guidelines for drugs and diseases to enable multimodal analyses.

Background & Summary

Precision medicine takes a personalized approach to disease diagnosis and treatment that accounts for the variability in genetics, environment, and lifestyle across individuals[1]. In order to be precise, medicine must revolve around data, especially in generating, linking, and learning from both biomedical knowledge and health records[2]. Nevertheless, many barriers to linking and efficiently exploiting medical information across healthcare organizations and biological scales slow down the research and development of individualized care[2]. While many have acknowledged the difficulties in linking existing biomedical knowledge to patient level health records[2–4], few realize that biomedical knowledge is itself fragmented. Biomedical knowledge about complex diseases comes from different organizational scales, including genomics, transcriptomics, proteomics, molecular functions, intra- and inter- cellular communications, phenotypes, therapeutics, and environmental exposures. For any given disease, information from various organizational scales is scattered across individual publications, non-standardized data repositories, evolving ontologies, and clinical guidelines. Developing networked relationships between these sources can support research in disease-rooted precision medicine.

A resource that comprehensively describes the relationships of diseases to biomedical entities would enable the large-scale, data-driven study of human disease. Understanding the connections between diseases, drugs, phenotypes, and other entities could open the doors for many types of research to leverage recent computational advances, including but not limited to the study of disease phenotyping[5–7], disease etiology[8], disease similarity[9], disease diagnosis[10–12], disease treatments[13], drug-disease relationships[14–16], mechanisms of drug action[17], drug repurposing[18–20], drug discovery[21,22], adverse drug events[23,24], combination drug therapies[25], and so forth. Many researchers have developed knowledge graphs for individual diseases that have helped advance computational precision medicine within their respective disease domains [26–41]. Nevertheless, the costs and extended timelines of these individual efforts demonstrate a need for a systematic data resource that could unify existing biomedical knowledge to enable the investigation of diseases at scale.

While there exist many primary data resources containing information about diseases, consolidating them into a comprehensive, disease-rich and functional knowledge graph present three challenges. Firstly, existing approaches to network analysis of diseases require expert review and curation of data in the knowledge graph [28,29,42]. While incredibly detailed, this approach requires substantial manual labour and expensive expert input, making it difficult to scale. Further, there

lacks a consistent representation of diseases across biomedical datasets and clinical guidelines. Rather than have a standardized disease ontology, database developers select the ontology that best suits their function from a multitude of biorepositories[43–53]. Because each set of disease vocabulary was tailored for some to serve an unique purpose, their disease encodings overlap unsystematically and are often in conflict. For instance, ICD codes[49] are optimized for medical billing whereas MedGen[52], PhenoDB[50], and Orphanet[47] focus on rare and genetic diseases. Moreover, expertly curated disease descriptions in medical knowledge repositories do not follow any naming conventions [47,54]. The lack of standardized disease representations and the multimodal nature of the datasets makes it challenging to harmonizing biomedical knowledge at scale. Finally, the definition of an 'unique' disease remains medically and scientifically ambiguous. For instance, while autism spectrum disorder is considered a medical diagnosis, the condition has many subtypes linked to clinically divergent manifestations[55,56]. Clinically studied subtypes of a disease often do not correlate clearly with those defined in disease ontologies. Although only three subtypes of autism have be clinically identified[56], the Unified Medical Language System (UMLS)[45] describes 192, the Monarch Disease Ontology (MONDO)[43] describes 37, and Orphanet[47] describes 6. The challenge in reconciling diseases entities is only exacerbated by the variety of synonyms and abbreviations available for any particular disease[57] and the difficulty in linking in structured diseases entities to unstructured names in text[58]. Meaningful disease entity resolution across multimodal, non-standardized datasets is critical for developing knowledge graphs that will be useful for downstream precision medicine tasks.

While drug repurposing remains the focus of knowledge graph development[32,36,38,41,59–61], considerable effort has been devoted to building knowledge graphs from biomedical literature [27,30,39] and clinical records [28,29,33,62]. For example, the SPOKE network is a seminal effort that linked 29 heterogeneous biomedical databases to build a disease-centric knowledge graph [37]. Although SPOKE is limited to about 200 diseases and lacks multimodal connections between textual clinical guidelines and tabular molecular data, it has nevertheless enabled many precision medicine efforts including this study [34] to overlay individual patient-level information onto SPOKE's biomedical knowledge. Most recently, an initiative from the White House led to the development of The COVID-19 Open Research Dataset (CORD-19)[63]. CORD-19 was able to empower data-driven medicine during the pandemic by facilitating the development of neural search engines for healthcare workers[64,65] and provided insights into drug repurposing targets [66]. Collectively, these knowledge graphs have lent themselves to a variety of scientific discoveries[67,68],

methodological innovations[69–71] and biomedical benchmarking[31,35,72]. Large-scale knowledge graphs have facilitated fruitful research across a variety of problems faced by the biomedical community. Nevertheless, due to the medical heterogeneity of diseases, the multimodal nature of disease information, and the incompatibility of existing disease biorepositories, knowledge graphs focused on diseases have not achieved their the scale or impact of many other efforts in this space.

Here, we present the Precision Medicine Knowledge Graph (PrimeKG), a knowledge graph that provides a holistic view of diseases. We integrate 20 high-quality resources, biorepositories and ontologies to curate this knowledge graph. PrimeKG systematically captures information about 17,080 diseases with 4,050,249 relationships representing ten major biological scales, including disease-associated perturbations in the proteome, biological processes and pathways, anatomical and phenotypic scale, and the entire range of approved and experimental drugs together with their therapeutic action, thus considerably expanding previous efforts in disease-rooted knowledge graph creation. We demonstrate that disease nodes in our multi-relational knowledge graph are densely connected to every other node type, including phenotypes, exposures and seven others. We tune PrimeKG specifically to support drug-disease prediction by including an abundance of 'indications', 'contradictions' and 'off-label use' edges, which are usually missing or sparse in other knowledge graphs. We supplement PrimeKG's rich graph structure with textual descriptions of clinical guidelines for drug and disease nodes to enable multi-modal analyses. Finally, we address the disease entity resolution challenge by improving the correspondence between diseases in PrimeKG and disease subtypes found in the clinic, making PrimeKG-based analyses medically meaningful.

Methods

The Precision Medicine Knowledge Graph (PrimeKG) is heterogeneous, with 10 types of nodes and 30 types of undirected edges. To develop PrimeKG, we retrieved and collated the 20 primary data resources as visualized in Figure 2a, identified relations across these resources as shown in Figures 2b and 2c, harmonized them into a rich, heterogeneous network as illustrated in Figure 2c, and augmented the drug and disease nodes in this network with textual descriptions as depicted in Figure 2d.

A. Curating primary data resources

To develop a comprehensive knowledge graph to study diseases, we considered 20 primary resources and a number of additional repositories of biological and clinical information. Figure 2a provides an overview of all 20 resources. We selected datasets that provided widespread coverage of biomedical entities, including proteins, genes, drugs, diseases, anatomy, biological processes, cellular components, molecular functions, exposures, disease phenotypes and drug side effects. These were high-quality datasets, either expertly curated annotations such as DisGeNet and Mayo Clinic, widely-used standardized ontologies such as Mondo Disease Ontology, or direct readouts of experimental measurements such as Bgee and DrugBank. A complete list of primary resources along with their processing steps is listed in the Data Records section. All our data curation and processing approaches are transparent, fully reproducible, and can be continually adapted as individual data resources evolve and new data become available.

B. Standardizing and harmonizing data resources

To harmonize these primary data resources into the Precision Medicine Knowledge Graph, we selected ontologies for each node type, harmonized datasets into a standardized format, and resolved overlap across ontologies.

Defining node types and selecting common ontologies. Our knowledge graph consists of 10 types of nodes. The node types ‘drug’, ‘disease’, ‘anatomy’ and ‘pathway’ are respectively encoded as terms in DrugBank, Mondo, UBERON, and Reactome. Genes and proteins are treated as a single node type, ‘gene/protein’, and identified by Entrez Gene IDs. The node types ‘biological process’, ‘molecular function’, and ‘cellular component’ are defined using Gene Ontology terms. Disease phenotypes extracted from HPO and drug side effects extracted from SIDER are collapsed into a single node type, ‘effect/phenotype’, that is encoded using HPO IDs. Finally, ‘exposure’ nodes are defined using the ExposureStressorID field, which contains MeSH identifiers provided

by the Comparative Toxicogenomics Database. This is illustrated in Figure 2a where each database is coloured by the node type it defines. Moving forward, we interchangeably refer to 'gene/protein' nodes as proteins and 'effect/phenotype' nodes as phenotypes.

Harmonizing external data resources. We mapped the aforementioned processed datasets to ensure that all nodes were defined in their respective common ontologies. Next, we identified sources of information across different primary resources for each node type to maximise the number of relationships in PrimeKG (see Figure 2b).We then restructured the datasets to follow the following format. For each node in the knowledge graph, we provide 'node_index' which is a unique index to identify the node in our KG; 'node_id' which indicates the identifier of the node from it's ontology; 'node_type' which indicates the node type as defined in our knowledge graph; 'node_name' which indicates the name of the node as provided by the ontology; and 'node_source' which indicates the ontology from which 'node_id' and 'node_name' fields were extracted. For each edge in the knowledge graph, we provide 'relation' which is the name of the edge type that connects the two nodes; 'x_index' which links to the 'node_index' field; and 'y_index' which also links to 'node_index'. Finally, we renamed columns for consistency, dropped rows with NaN values, dropped duplicated edges and removed self loop edges from each individual dataset.

Resolving overlap between phenotype and disease nodes. Since both the Mondo Disease Ontology and Human Phenotype Ontology were developed by the Monarch Initiative, there was considerable overlap between phenotype nodes and disease nodes across the various datasets. Overlapping nodes were defined as effect/phenotype nodes in HPO that (i). had the same ID number as disease nodes in Mondo and (ii). could be mapped from HPO to Mondo using cross-references found in the Mondo ontology. To avoid duplicate nodes, these overlapping phenotype nodes were converted to disease nodes by manipulating edges in various datasets as follows. Let us define the set of overlapping phenotype nodes as P. Phenotype-phenotype edges extracted from the HPO ontology were converted to phenotype-disease edges if one phenotype node was in P and to disease-disease edges if both phenotype nodes were in P. These converted edges were then dropped from the original phenotype-phenotype resource. Protein-phenotype edges extracted from DisGeNet were converted to protein-disease relations if the phenotype node was in P and removed from the group of protein-phenotype edges. Finally, for disease-phenotype and drug-phenotype relations, we dropped any edges where the phenotype was in P. Adding these edges to drug-disease relations would only introduce unnecessary noise to the indication, contraindication, and off-label use edges.

C. Building Precision Medicine Knowledge Graph (PrimeKG)

To construct the network structure of PrimeKG, we merged the harmonized primary data resources into a graph and extracted it's largest connected component as shown in Figure 2c. We integrated the various processed, curated datasets and cleaned the graph by dropping NaN and duplicate edges, adding reverse edges, dropping duplicates again and removing self loops. This version of the knowledge graph is available on our Harvard Dataverse as 'kg_raw.csv'. To ensure that our knowledge graph was well-connected and did not have any isolated pockets, we extracted it's largest connected component. This giant component retained 0.99998% of edges that were present in the original graph. The largest connected component of the knowledge graph is available on our Harvard Dataverse as 'kg_giant.csv'.

D. Supplementing drug nodes with clinical information

As shown in Figure 2d, we extracted both textual and numerical features for drug nodes in the knowledge graph from DrugBank and Drug Central. Features from DrugBank mapped directly to the knowledge graph since drugs were coded using DrugBank identifiers. Some features had unique attributes for each drug, such as 'state', 'indication' and 'mechanism of action', and others had numerous attributes for each drug, such as 'group' and 'ATC level'. The latter set of features were converted to single text descriptions by joining features using conjunctions such as ';' and 'and'. Features in Drug Central were mapped to DrugBank IDs using their CAS identifiers, from the vocabulary that was retrieved from DrugBank. Once all features were mapped, text processing involved removing all tokens that are references in DrugBank (for example "[L64839]") with the help of regular expressions. For the half-life feature, we nullified locations where the text mentioned that no data was available. Finally, we converted numerical features into textual descriptions in order to standardize the feature set.

As an example, let's explore the features available for Prednisolone. Prednisolone is a glucocorticoid similar to cortisol used for its anti-inflammatory, immunosuppressive, anti-neoplastic, and vasoconstrictive effects. Prednisolone has a plasma half life of 2.1-3.5 hours. Prednisolone is indicated to treat endocrine, rheumatic, and hematologic disorders; [...] and other conditions like tuberculous meningitis. Corticosteroids binding to the glucocorticoid receptor mediates changes in gene expression that lead to [...]. Prednisolone's protein binding is highly variable, [...]. Corticosteroids bind to the glucocorticoid receptor, inhibiting pro-inflammatory signals, and promoting [...]. Prednisolone is a solid. Prednisolone is part of Adrenal Cortex Hormones ; Adrenals ; [...]

Prednisolone is approved and vet approved. Prednisolone uses Prednisone Action Pathway [...] The molecular weight is 360.45. Prednisolone has a topological polar surface area of 94.83. The log p value of is 1.42.

E. Supplementing disease nodes with clinical information

As shown in Figure 2d, we extracted textual features for diseases nodes in the knowledge graph from the Mondo Disease Ontology, Orphanet, Mayo Clinic, and UMLS. Features from all these sources were mapped to the 'node_id' field of disease nodes, that was defined using the Mondo Disease Ontology. Since disease nodes were grouped as previously described, many diseases defined in the Mondo Disease Ontology (ie. many 'node_id' values) we collapsed into a single node (ie. unique 'node_index' values). Since disease features are mapped to Mondo identifiers, or the 'node_id' field, it is possible for a single disease node in the knowledge graph, defined by a unique 'node_index', to have multiple feature values for a given feature. We provide the available features in their entirety since we did not have the medical authority to summarize them into single descriptors.

Disease definitions from the Mondo Disease Ontology were directly extracted from the ontology file and unique for each node_id. Disease descriptions extracted from UMLS were mapped from CUI terms to Mondo and as a result, numerous for each node_id. We removed tokens that were references and URLs from UMLS disease descriptions using regular expressions. From Orphanet, we extracted definitions, prevalence, epidemiology, clinical description, and management and treatment. We mapped the features from Orphanet IDs to Mondo, and as a result, there were multiple for each node_id. We used regular expressions to fix formatting errors in the prevalence and epidemiology features.

We extracted the following disease features from Mayo Clinic's website: symptoms, causes, risk factors, complications, and prevention. Since the Mayo Clinic web-scrapping did not provide a unique identifier in any ontology, we mapped disease names in Mayo Clinic to those in the Mondo Disease Ontology. To develop this mapping, we used a strategy for grouping disease names that is described in detail in the technical validation. Briefly, we conducted automated string matching followed by manual approval of disease name mappings based on their BERT embedding similarity. Automated string matching involved approving exact matches and encapsulated matches, where the name in Mayo was completely present in the name in Mondo. During processing of the symptoms feature, we used regular expressions to extract the end of the text description that explained when to see the doctor as a new and separate feature. Finally, we cleaned the text for

formatting errors.

As an example of the depth and breadth of information covered by the disease features, let's explore Hepatic Porphyria. Per the Mondo Disease Ontology, Hepatic Porphyria is a group of metabolic diseases due to deficiency of one of a number of liver enzymes in the biosynthetic pathway of heme. They are characterized by [...]. Clinical features include [...]. The UMLS has a very similar disease description. According to Orphanet, it's a rare sub-group of porphyrias characterized by the occurrence of neuro-visceral attacks with [...]. In the majority of European countries, the prevalence of acute hepatic porphyrias is around 1/75000. In 80% of cases the patients are female. All acute hepatic porphyrias can be accompanied by neuro-visceral attacks that appear as [...]. The attacks are most commonly triggered by [...]. When an acute attack is confirmed, urgent treatment with an injection of [...]. According to Mayo Clinic, signs and symptoms of acute porphyria may include: Severe abdominal pain, [...], Seizures. All types of porphyria involve a problem in the production of heme [...] and a shortage of a specific enzyme determines the type of porphyria. In addition to genetic risks, environmental factors may trigger development of [...]. Examples of triggers include: Exposure to sunlight, [...]. Possible complications depend on [...] During an attack, you may experience [...] Although there's no way to prevent porphyria, if you have the disease, avoid [...]. When to see a doctor, [...].

Data Records

We proceed with a detailed description of the 20 primary data resources used to build PrimeKG.

Bgee gene expression knowledge base in animals. Bgee[73] contains gene expression patterns across multiple animal species. We retrieved gene expression data for humans from ftp://ftp.bgee.org/current/download/calls/expr_calls/Homo_sapiens_expr_advanced.tsv.gz on 31 May 2021. Processing involved keeping only gold quality calls and ensuring that the anatomical entities were coded using the UBERON ontology. To extract only highly expressed genes in the anatomical entity, we empirically filtered the data to keep data with expression rank less than 25,000. After processing, we had 1,786,311 anatomy-protein associations where gene expression was found to be present or absent.

Comparative Toxicogenomics Database. The Comparative Toxicogenomics Database (CTD)[74] is focused on the impact of environmental exposures on human health. We retrieved information about exposures (05/21 version) from http://ctdbase.org/reports/CTD_exposure_events.csv.gz on 9 Jun 2021. Processing involved removing header comments from the csv file. After processing, our data contained 180,976 associations of exposures with proteins, diseases, other exposures, biological processes, molecular functions, and cellular components.

DisGeNET knowledgebase of gene-disease associations. DisGeNET[75] is a resource about the relationships between genes and human disease that has been curated by experts. We retrieved curated disease-gene associations (version 7.0) from https://www.disgenet.org/static/disgenet_ap1/files/downloads/curated_gene_disease_associations.tsv.gz on 31 May 2021. The raw data file, 'curated_gene_disease_associations.tsv' was not processed further and contains 84,038 associations of genes with diseases, and phenotypes.

Disease Ontology. Disease Ontology[46] groups diseases in many meaningful clusters by using clinically relevant characteristics. For instance, diseases are grouped by anatomical entity. We retrieved the ontology from https://raw.githubusercontent.com/DiseaseOntology/HumanDiseaseOntology/main/src/ontology/HumanDO.obo on 29 Jun 2021. The raw data 'HumanDO.obo' is mapped to disease nodes in our knowledge graph. Since the Mondo Disease Ontology is not grouped anatomically or by clinical speciality, this will allow users of PrimeKG to explore disease nodes in a medically meaningful format.

DrugBank. DrugBank[76] is a resource that contains pharmaceutical knowledge. We retrieved the complete database (version 5.1.8) from https://go.drugbank.com/releases/5-1-8/downloads

/all-full-database on 31 May 2021. Processing involved using the beautiful soup package to extract synergistic drug interactions. The processed data contains 2,682,157 associations. We also extracted drug features from the raw data. For over 14,000 drugs, we construct 12 drug features, including group, state, description, mechanism of action, ATC code, pharmacodynamics, half life, protein binding, and pathway.

We also retrieved information about drug targets from https://go.drugbank.com/releases/5-1-8/downloads/target-all-polypeptide-ids, about drug enzymes from https://go.drugbank.com/releases/5-1-8/downloads/enzyme-all-polypeptide-ids, about drug carriers from https://go.drugbank.com/releases/5-1-8/downloads/carrier-all-polypeptide-ids, about drug transporters from https://go.drugbank.com/releases/5-1-8/downloads/transporter-all-polypeptide-ids all on 31 May 2021. Processing involved combining all four resources and mapping gene names from UniProt IDs to NCBI Gene IDs using vocabulary retrieved from HNCG gene names https://www.genenames.org. The processed data contains 26,118 drug-protein interactions.

Drug Central. Drug Central[77] is a resource that curates information about drug-disease interactions. We retrieved the Drug Central SQL dump from https://drugcentral.org/ActiveDownload on 1 Jun 2021. The database was loaded into Postgres SQL and drug-disease relationships were extracted. The processed data contains 26,698 indication edges, 8,642 contraindication edges, and 1,917 off-label use edges. We also extracted drug features from the Drug Central SQL dump from the 'structures' and 'structure_type' tables. We extracted features for over 4500 drugs, representing each drug with features including topological polar surface area (TPSA), molecular weight and cLogP. As an example, the features for *Atorvastatin* are: organic structure, molecular weight of 558.65, TPSA of 111.79 and a ClogP value of 4.46.

Entrez Gene. Entrez Gene[78] is a resource maintained by the NCBI that contains vast amounts of gene-specific information. We retrieved data about relations between genes and Gene Ontology terms from https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz on 31 May 2021. Processing involved using the goatools package[79] to extract relations between genes and Gene Ontology terms. The processed data contains 297,917 associations of genes with biological processes, molecular functions, and cellular components.

Gene Ontology. The Gene Ontology[80] network describes molecular functions, cellular components, and biological processes. We retrieved the ontology from http://purl.obolibrary.org/obo/go/go-basic.obo on 31 May 2021. Processing involved using the goatools package[79] to extract

information for gene ontology terms and relations between go terms. The processed data contains 71,305 hierarchical associations between biological processes, molecular functions, and cellular components.

Human Phenotype Ontology. The Human Phenotype Ontology[?] (version hpo-obo@2021-04-13) provides information on phenotypic abnormalities found in diseases. We retrieved the ontology from http://purl.obolibrary.org/obo/hp.obo on 31 May 2021. Processing involved parsing the ontology file to extract phenotype terms in the ontology, parent-child relationships and cross references to other ontologies. The processed data contains disease-phenotype, protein-phenotype, and phenotype-phenotype edges. We also retrieved expertly curated annotations from http://purl.obolibrary.org/obo/hp/hpoa/phenotype.hpoa on 31 May 2021. Additionally, we extracted 218,128 curated positive and negative associations between diseases and phenotypes.

Mayo Clinic. Mayo Clinic is a nonprofit academic medical center and biomedical research institution focused on integrated health care[54]. On it's website https://www.mayoclinic.org/diseases-conditions, Mayo Clinic has curated information about symptoms, causes, risk factors, complications and prevention of 2,227 diseases and conditions. We web-scraped this data and extracted descriptions for these diseases and conditions using the *mayo.py* and *diseases.py* scripts on 28 March 2021. The raw data is available at 'mayo.csv'.

For example, we extracted features of 'Atrial fibrillation' from Mayo Clinic. 'Some people with atrial fibrillation have no symptoms [...] others may experience signs and symptoms such as: Palpitations, Weakness, [...] and Chest Pain. The disease occurs when 'the two upper chambers of your heart experience chaotic electrical signals [...] As a result, they quiver. The AV node is bombarded with impulses trying to get through to the ventricles'. Certain factors may increase your risk of developing atrial fibrillation including age, heart disease, [...] and obesity. Complications include: 'the chaotic rhythm causing blood to pool in your atria and form clots [...] leading to a stroke. [...] Atrial fibrillation, especially if not controlled, may weaken the heart and lead to heart failure'. To prevent atrial fibrillation, it's important to live a heart-healthy lifestyle [...] which may include increasing your physical activity, [...]. These snippets represent only an overview of over three pages of descriptive features available on Atrial Fibrillation.

Mondo Disease Ontology. Since the Mondo Disease Ontology[43] harmonizes diseases from a wide range of ontologies, including OMIM, SNOMED CT, ICD, and MedDRA, it was our preferred ontology for defining diseases. We retrieved the ontology from http://purl.obolibrary.org/obo/mo

ndo.obo on 31 May 2021. Processing involved parsing the ontology file to extract disease terms in the ontology, parent-child relationships, subsets of diseases, cross references to other ontologies, and definitions of disease terms. The processed data contains 64,388 disease-disease edges.

Orphanet. Orphanet[47] is a database that focuses on gathering knowledge about rare diseases. The Orphanet website https://www.orpha.net/consor/cgi-bin/Disease_Search_List.php?lng=EN has curated information about definitions, prevalence, management and treatment, epidemiology, and clinical description for 9348 rare diseases. We web-scraped this data and extracted disease features using code available at *orpha.py* on 10 May 2021.

For instance, the rare disease *Hurler syndrome* with Orphanet ID 93473 has the following features. Hurler syndrome is the most severe form of mucopolysaccharidosis type 1, a rare lysosomal storage disease, characterized by skeletal abnormalities, cognitive impairment, heart disease, [...] and reduced life expectancy. The prevalence of the Hurler subtype of MPS1 is estimated at 1/200,000 in Europe and one in a million in general. The clinical manifestation of the disease includes ' musculoskeletal alterations, cardiomyopathy, [...] and neurosensorial hearing loss within the first year of life'. Management of the disease is multidisciplinary: 'Hematopoietic stem cell transplantation is the treatment of choice as it can prolong survival. [...] Enzyme replacement therapy (ERT) with laronidase [...] is a lifelong therapy which alleviates non neurological symptoms.'. These descriptions only represent a brief snapshot of the expertly curated knowledge available in Orphanet.

Four integrated resources of physical protein-protein interactions. Protein-protein interactions are composed of experimentally-verified interactions between proteins. The interactions we consider are diverse in nature and include signalling, regulatory, metabolic-pathway, kinase-substrate and protein complex interactions, which are considered as unweighted and undirected. We use the human PPI network compiled by Menche *et al.*[81] as the starting resource. This resource integrates several protein-protein interaction databases, including TRANSFAC for regulatory interactions[82], MINT and IntAct for yeast to hybrid binary interactions[83,84], and CORUM for protein complex interactions[85]. Additionally, we retrieve protein-protein interaction information from from BioGRID[86] and STRING[87] databases. We also consider the human reference interactome (HuRI) generated by Luck *et al.*[88], specifically, we use the HI-union, a combination of HuRI and several related efforts to systematically screen for protein-protein interactions. The processed data contains 642,150 edges.

Reactome pathway database. Reactome[89] is an open-source, curated database for pathways. We retrieved information about pathways from https://reactome.org/download/current/ReactomePathways.txt, relationships between pathways from https://reactome.org/download/current/ReactomePathwaysRelation.txt and pathway-protein relations from https://reactome.org/download/current/NCBI2Reactome.txt on 31 May 2021. Processing involved extracting ontology information such as hierarchical relationships and extracting pathway-protein interactions. The processed data contains 5,070 pathway-pathway and 85,292 protein-pathway edges.

Side effect knowledgebases. The Side Effect Resource (SIDER)[90] contains data about adverse drug reactions. We retrieved side-effect data (SIDER 4.1 version) from http://sideeffects.embl.de/media/download/meddra_all_se.tsv.gz and SIDER's drug to Anatomical Therapeutic Chemical (ATC) classification mapping from http://sideeffects.embl.de/media/download/drug_atc.tsv on 31 May 2021. Processing involved extracting all side effects where the MedDRA term was coded at the "PT" or preferred term level, and then mapping drugs from STITCH ID to ATC ID. The processed data 202,736 contains drug-phenotype associations.

Uberon multi-species anatomy ontology. Uberon[91] is an ontology that contains information about the human anatomy. We retrieved the ontology from http://purl.obolibrary.org/obo/uberon/ext.obo on 31 May 2021. Processing involved extracting information about anatomy nodes and the relationships between them. The processed data 28,064 hierarchical relationships between anatomy nodes.

UMLS knowledgebase. The Unified Medical Language System (UMLS) Knowledge Source[45] contains information about biomedical and health related concepts. We retrieved the complete UMLS Metathesauras from https://download.nlm.nih.gov/umls/kss/2021AA/umls-2021AA-metathesaurus.zip on 31 May 2021 in '.RRF' format. To map UMLS CUI terms to the Mondo Disease Ontology, we used the 'MRCONSO.RRF' to extract UMLS Concept Unique Identifier (CUI) terms in English. We mapped UMLS CUI terms to Mondo terms in two ways. Firstly, we directly extracted cross references between the two from the Mondo ontology. Secondly, we indirectly mapped UMLS to Mondo using OMIM, NCIT, MESH, MedDRA, ICD 10 and SNOMED CT as intermediate ontologies.

Further, we used 'MRSTY.RRF' and 'MRDEF.RRF' files to extract definitions for UMLS terms. Of the 127 semantic types present in the 'MRSTY.RRF' file, we selected 11 that belonged to the Disorder semantic group in a manner that was consistent with prior work. [92] These semantic

types were Congenital Abnormality, Acquired Abnormality, Injury or Poisoning, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Cell or Molecular Dysfunction, Experimental Model of Disease, Signs and Symptoms, Anatomical Abnormality, and Neoplastic Process. We then used the 'MRDEF.RRF' file to extract definitions for CUI terms from sources that were in English.

Additional vocabularies. We retrieved gene names and mappings between NCBI Entrez IDs and UniProt IDs from https://www.genenames.org/download/custom/ on 31 May 2021. We retrieved the DrugBank drug vocabulary from https://go.drugbank.com/releases/5-1-8/downloads/all-drugbank-vocabulary on 31 May 2021. These were used to map nodes in the knowledge graph to consistent ontologies.

Technical Validation

As part of the technical validation, we explore the structure and connectivity of Precision Medicine Knowledge Graph.

Characterizing Precision Medicine Knowledge Graph

PrimeKG contains 129,375 nodes and 8,100,498 edges. Figure 1a shows a schematic overview of the graph structure, containing 10 types of nodes and 30 types of edges. We provide a breakdown of the number of nodes by node type and the number of edges by edge type in Tables 1 and 2, respectively. Figure 1b demonstrates that disease nodes are densely connected to other node types in the knowledge graph. Tables 3 and 4 show statistics on the number of features available for drug and disease nodes. Disease features include information on disease prevalence, symptoms, causes, risk factors, epidemiology, clinical description, management and treatment, complications, prevention and when to see a doctor. Drug features include information on molecular weight of chemical compounds, indications, mechanisms of action, pharmacodynamics, protein binding events, and pathway information, among others. This extensive clinical information describing the entire range of drugs and diseases is a unique characteristic of PrimeKG that makes PrimeKG stand out among its peer knowledge graphs. Figure 1c provides an example of the supporting information that is available across these features.

A case study in autism to evaluate relevance of PrimeKG to clinical presentation of autism

For downstream inferences made using PrimeKG to be conducive to studying human disease, disease nodes in PrimeKG would need to be medically relevant. To this end, we next analyze if PrimeKG's representation of diseases strongly relates to their clinical presentation by carrying out a case study in autism spectrum disorder. We were motivated to investigate autism because it not only has incredible clinical heterogeneity[93–95] but this heterogeneity has also been studied to identify clinically meaningfully subtypes[55,56]. We gauged the relevance of disease nodes related to autism in PrimeKG in two steps: first, by performing the entity resolution for autism concepts across all relevant primary data resources (see Methods), and second, by examining the relationship between these autism concepts and clinical subtypes of autism.

We start by exploring whether autism disease nodes in PrimeKG reconciled the variation in autism concepts across databases and ontologies. As demonstrated in Figure 3a, MONDO disease ontology has 37 disease concepts that relate to autism, whereas the UMLS has 192 autism-associated concepts and Orphanet has 6 autism-associated concepts. Although it is not immediately

clear how these concepts relate to each other, we cannot develop a coherent knowledge graph without establishing connections between these concepts. To this end, we overcome this challenge by defining all nodes using the MONDO disease ontology and mapping all other vocabularies to diseases in MONDO as outlined in Figure 3a.

Finally, before using MONDO disease concepts as disease nodes in PrimeKG, we need to assess whether autism disease concepts in MONDO correlate with clinical subtypes of autism. Autism has been shown to manifest as three clinical subgroups characterized primarily by seizures, multisystem and gastrointestinal disorders, and psychiatric disorders[56]. It was unclear how the 37 autism disease concepts in MONDO (see Figure 3a) relate to the three clinically defined subtypes. Rather, there were many disease concepts in autism such as 'Autism, susceptibility to, 1', 'Autism, susceptibility to, 2', 'Autism, susceptibility to, x-linked', etc, with no clear clinical meaning, suggesting that disease nodes in MONDO do not correspond one-to-one to clinical manifestation of autism. For this reason, we developed a strategy to group diseases from MONDO into medically relevant and coherent nodes in PrimeKG. We proceed with describing and evaluating that strategy.

Computational approaches to grouping disease nodes

As demonstrated in our case study of autism, disease concepts in MONDO may not correlate well with medical subtypes. In fact, MONDO contains many repetitive disease entities with no clear clinical correlation. For this reason, we were motivated to group diseases in MONDO into medically relevant entities. Ideally, we would have preferred to leverage expertise across a wide variety of disease areas when grouping these concepts. However, this approach was time-consuming, expensive and challenging to execute at scale. Further, disease sub-phenotyping is a relatively new paradigm, and so we anticipated low consensus among medical experts on what constitutes a unique disease.

Since manually grouping diseases with expert supervision was not feasible, we took a semi-automated unsupervised approach to grouping disease concepts in PrimeKG. Advances in natural language processing, specifically the Bidirectional Encoder Representations from Transformers (BERT) model[96], allowed us to study similarity between disease concept names. We grouped disease concepts with nearly identical names into a single node with string matching and BERT embedding similarity[96–100].

We identified disease groups using a string matching strategy across disease names [101]. In this strategy, we selected a disease that ended with a number, or a roman numeral, or any alphanumeric phrase with a length of less than 2, or 'type' as the second-last word. Once such a disease was

selected, we extracted the main disease phrase by dropping the ending and used this phrase to find matches. Matches included diseases with the same initial phrase and those containing all phrase words with no additional words regardless of word ordering. For the latter matching criteria, the words 'type' and '(disease)' were ignored. In this manner, we grouped disease concepts in MONDO with string matching.

We further tightened groupings identified using string matching by exploring word embedding similarities between disease names, which is visualized in Figure 3b. In natural language processing, word embeddings have been widely and successfully used to resolve conflicting and redundant entities in an unsupervised manner[101–103] and deep language models like BERT[96] are used to obtain semantically meaningful word embeddings. Specifically, ClinicalBERT[104] is a BERT language model that encodes medical notions of semantics because it has been pre-trained on biomedical knowledge from PubMed[105] and discharge summaries from MIMIC-III[106]. Here, we used ClinicalBERT to extract word embeddings for disease group-names identified during string matching and defined similarity between two disease names as the cosine distance between their ClinicalBERT embeddings. After applying an empirically chosen cutoff of similarity ≥ 0.98, we manually approved the suggested disease matches and assigned names to the new groups. Finally, these groupings were applied to the knowledge graph.

After this process, 22,205 disease concepts in MONDO were collapsed into 17,080 grouped diseases, which has resulted a higher average edge density across diseases and in more clinically relevant disease nodes. We anticipate that PrimeKG is a more powerful dataset with this grouping because disease representations are concentrated and robust, which, in turn, can enable biological insights gleaned from PrimeKG to be medically relevant.

Conclusion

The potential uses of PrimeKG are vast. Since PrimeKG describes drug features on a deeper biological level and disease features on a deeper clinical level, they can be used to explain why a genotype-phenotype association is occurring in terms of genes, pathways, or any other nodes in a large knowledge graph, like PrimeKG. Consequently, PrimeKG can be paired with machine learning[107] to discover new disease biomarkers, characterize disease processes, hone disease classification, identify phenotypic traits, predict biological mechanisms, and repurpose drugs. With implementation of machine learning functionality, we anticipate that PrimeKG and similar knowledge graphs will constitute a critical tool in advancing precision medicine.

Data availability. PrimeKG is hosted on Harvard Dataverse with the following persistent identifier https://doi.org/10.7910/DVN/IXA7BM. We have deposited the knowledge graph along with all relevant intermediate files at this repository.

Code availability. The PrimeKG's project website is at https://zitniklab.hms.harvard.edu/projects/PrimeKG. The code to reproduce results, together with documentation and tutorials, is available in the PrimeKG's Github repository: https://github.com/mims-harvard/PrimeKG.

Acknowledgements. We would like to thank Jingyi Liu for help with retrieving and processing primary data resources and to Michelle M. Li and Emily Alsentzer for helpful discussion on our precision medicine knowledge graph. M.Z. is supported, in part, by NSF under Nos. IIS-2030459 and IIS-2033384, US Air Force Contract No. FA8702-15-D-0001, Harvard Data Science Initiative, Amazon Research Award, Bayer Early Excellence in Science Award, AstraZeneca Research, and Roche Alliance with Distinguished Scientists Award. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Author contributions. P.C., K.H., and M.Z. contributed new analytic tools and wrote the manuscript. P.C. retrieved, processed, harmonized and datasets. P.C. analyzed the resulting knowledge graph. M.Z. designed the study.

Competing interests. The authors declare no competing interests.

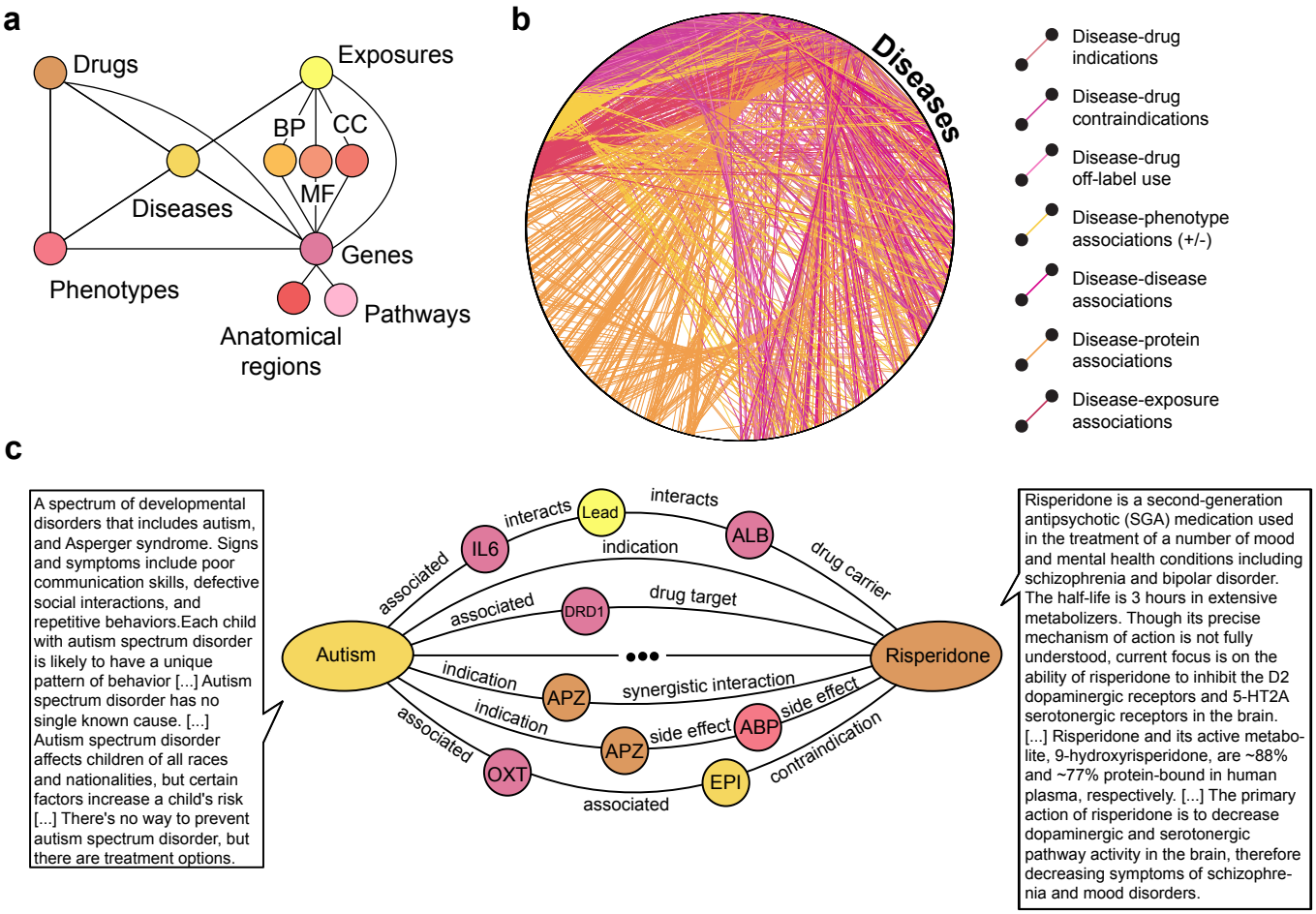

Figure 1: Overview of PrimeKG. a) Shown is a schematic overview of the various types of nodes in PrimeKG and the relationships they have with other nodes in the graph. **b)** All disease nodes in PrimeKG shown in a circular layout together with disease-associated information. All relationships between disease nodes and any other node type are depicted here. Disease nodes are densely connected to four other node types in PrimeKG through seven types of relations. **c)** Shown is an example of paths in PrimeKG between the disease node 'Autism' and the drug node 'Risperidone'. Intermediate nodes are colored by their node type from panel a. We also display snippets of text features for both nodes to demonstrate the multimodality of PrimeKG. Abbreviations - MF: molecular function, BP: biological process, CC: cellular component, APZ: Apiprazole, EPI: epilepsy, ABP: abdominal pain, + / - associations: positive and negative associations.

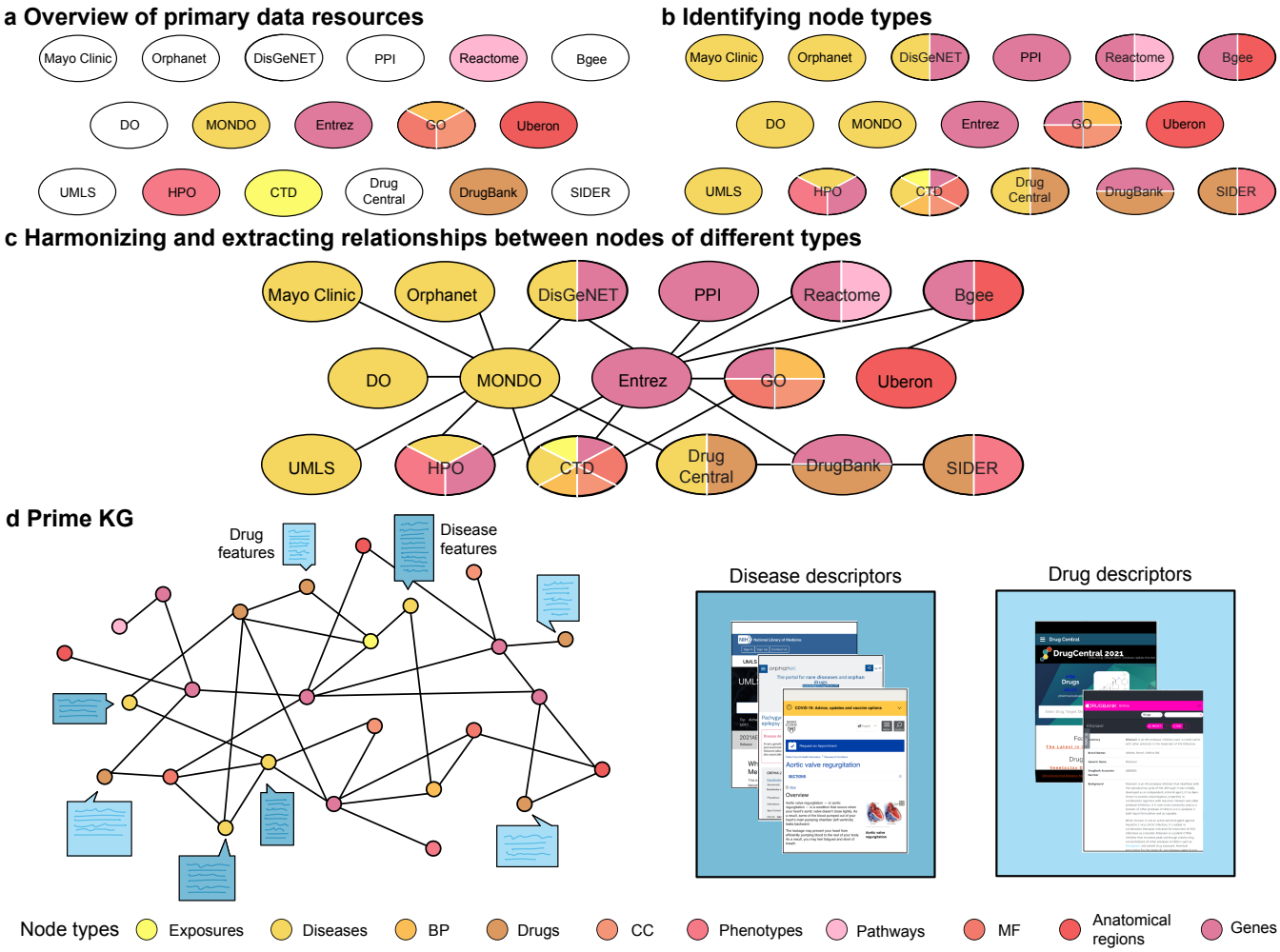

Figure 2: Building PrimeKG. The panels sequentially illustrate the process of developing the Precision Medicine Knowledge Graph. **a)** Shown are 20 primary data resources curated to develop PrimeKG. The colors indicate which database was used to define each node type. **b)** Primary resources are colored by each node type for which they possess information. **c)** Illustrated is the process of harmonizing these primary resources to extract relationships between node types. **d)** The left side illustrates PrimeKG and the right side shows all the textual sources of clinical information on drugs and diseases. The node type legend is consistent across the figure. Abbreviations - MF: molecular function, BP: biological process, CC: cellular component, PPI: protein protein interactions, DO: disease ontology, MONDO: mondo disease ontology, Entrez: entrez gene, GO: gene ontology, UMLS: unified medical language system, HPO: human phenotype ontology, CTD: comparative toxicogenomics database, SIDER: side effect resource.

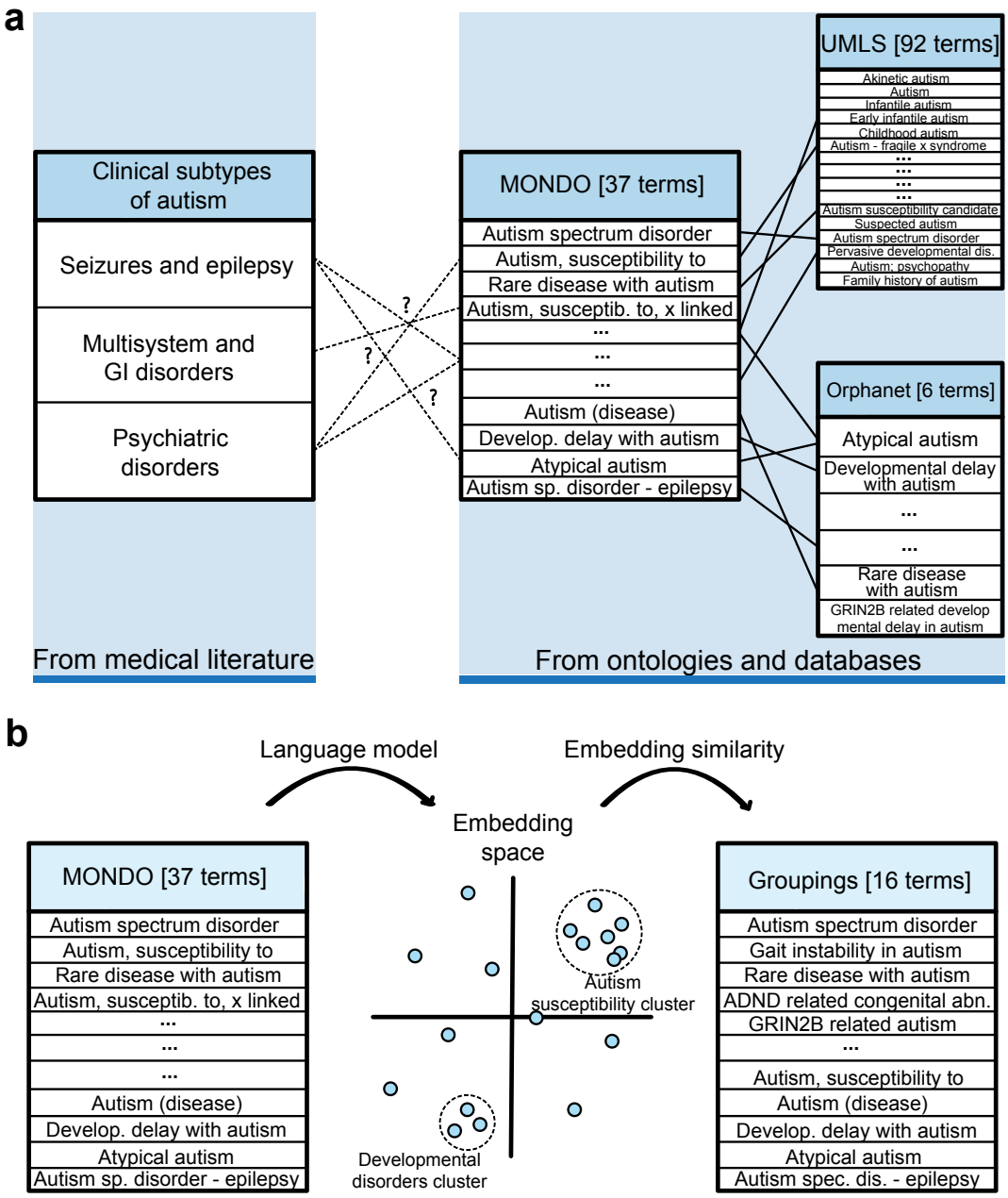

Figure 3: Reconciling autism disease nodes into more medically relevant entities. a) The left side shows three clinically determined subtypes of autism. The right side shows autism-related disease terms across three ontologies: MONDO, UMLS, and Orphanet. While we can identify mappings across the ontologies, it is unclear how the terms in any ontology connect to clinical subtypes. **b)** Illustration on how we use a language model, ClinicalBERT, to map terms from MONDO into a latent embedding space. Because the language model can group synonyms in the embedding space, we can cluster MONDO terms with similar semantic and medical meaning by calculating cosine similarity between embeddings of disease concepts. These clusters are created to develop disease groupings as shown on the right in panel b. Abbreviations - MONDO: MONDO disease ontology, UMLS: unified medical language system.

Node Type	Count	Percent (%)
Biological process	28,642	22.1
Protein	27,671	21.4
Disease	17,080	13.2
Phenotype	15,311	11.8
Anatomy	14,035	10.8
Molecular function	11,169	8.6
Drug	7,957	6.2
Cellular component	4,176	3.2
Pathway	2,516	1.9
Exposure	818	0.6
Total	129,375	100.0

Table 1: Statistics on nodes in the knowledge graph. Add average node degree.

Relation type	Count	Percent (%)
Anatomy - Protein (present)	3,036,406	37.5
Drug - Drug	2,672,628	33.0
Protein - Protein	642,150	7.9
Disease - Phenotype (positive)	300,634	3.7
Biological process - Protein	289,610	3.6
Cellular component - Protein	166,804	2.1
Disease - Protein	160,822	2.0
Molecular function - Protein	139,060	1.7
Drug - Phenotype	129,568	1.6
Biological process - Biological process	105,772	1.3
Pathway - Protein	85,292	1.1
Disease - Disease	64,388	0.8
Drug - Disease (contraindication)	61,350	0.8
Drug - Protein	51,306	0.6
Anatomy - Protein (absent)	39,774	0.5
Phenotype - Phenotype	37,472	0.5
Anatomy - Anatomy	28,064	0.3
Molecular function - Molecular function	27,148	0.3
Drug - Disease (indication)	18,776	0.2
Cellular component - Cellular component	9,690	0.1
Phenotype - Protein	6,660	0.1
Drug - Disease (off-label use)	5,136	0.1
Pathway - Pathway	5,070	0.1
Exposure - Disease	4,608	0.1
Exposure - Exposure	4,140	0.1
Exposure - Biological process	3,250	< 0.1
Exposure - Protein	2,424	< 0.1
Disease - Phenotype (negative)	2,386	< 0.1
Exposure - Molecular function	90	< 0.1
Exposure - Cellular component	20	< 0.1
Total	8,100,498	100.0

Table 2: Statistics on edges in the knowledge graph.

Source	Feature	Count	Unique	Percent (%)
Drug Central [77]	Molecular weight	2,797	2,308	35.2
	TPSA	2,718	2,718	34.2
	cLogP	2,574	980	32.3
DrugBank [76]	Group	7,957	7,903	100.0
	State	6,517	6,463	81.9
	Category	5,431	5,431	68.3
	Description	4,591	4,565	57.7
	Indication	3,393	3,076	42.6
	Mechanism of action	3,242	3,161	40.7
	ATC 4	2,818	1,040	35.4
	ATC 3	2,818	2,818	35.4
	ATC 2	2,818	2,818	35.4
	ATC 1	2,818	2,818	35.4
	Pharmacodynamics	2,659	2,617	33.4
	Half life	2,063	1,893	25.9
	Protein binding	1,669	1,487	21.0
	Pathway	598	598	7.5

Table 3: Statistics on drug features in the knowledge graph. The count column refers to the number of features including duplicates and the unique column refers to the number of unique features.

Source	Feature	Unprocessed KG		Processed KG	
		Count	Unique	Count	Unique
Any	Any	40,068	18,152	39,800	14,252
Mondo Disease Ontology [43]	Definition	15,238	15,238	15,238	12,001
UMLS [45]	Description	28,468	8,689	25,374	6,964
Orphanet [47]	Definition	6,564	6,548	6,562	5,645
	Prevalence	3,989	3,989	3,500	3,430
	Epidemiology	2,350	2,348	2,335	2,026
	Clinical description	2,294	2,292	2,293	1,972
	Management and treatment	1,732	1,731	1,722	1,553
Mayo Clinic [54]	Symptoms	6,642	5,789	5,140	4,470
	Causes	6,629	5,776	5,128	4,459
	Risk factors	6,284	5,501	4,898	4,299
	Complications	5,011	4,455	3,792	3,396
	Prevention	2,529	2,273	1,907	1,776
	When to see a doctor	5,862	5,234	4,531	4,058

Table 4: Statistics on disease features in the knowledge graph. Unprocessed KG refers to the initial knowledge graph assembled from the datasets. Processed KG refers to the fully processed knowledge graph, and includes disease groupings. The count column refers to the number of features including duplicates and the unique column refers to the number of unique features.

References

1. Adams, S. A. & Petersen, C. Precision medicine: opportunities, possibilities, and challenges for patients and providers. *Journal of the American Medical Informatics Association: JAMIA* **23**, 787–790 (2016).

2. Prosperi, M., Min, J. S., Bian, J. & Modave, F. Big data hurdles in precision medicine and precision public health. *BMC Medical Informatics and Decision Making* **18**, 139 (2018).

3. Hulsen, T. *et al.* From Big Data to Precision Medicine. *Frontiers in Medicine* **6** (2019).

4. Individualized Knowledge Graph.

5. Lussier, Y. A. & Liu, Y. Computational Approaches to Phenotyping: High-Throughput Phenomics. *Proceedings of the American Thoracic Society* **4**, 18–25 (2007).

6. Che, Z. & Liu, Y. Deep Learning Solutions to Computational Phenotyping in Health Care. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1100–1109 (IEEE, New Orleans, LA, 2017).

7. Che, Z., Kale, D., Li, W., Bahadori, M. T. & Liu, Y. Deep Computational Phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 507–516 (ACM, Sydney NSW Australia, 2015).

8. Kann, M. G. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Briefings in Bioinformatics* **8**, 333–346 (2007).

9. Cheng, L. *et al.* Computational Methods for Identifying Similar Diseases. *Molecular Therapy - Nucleic Acids* **18**, 590–604 (2019).

10. Jabbar, M. A., Deekshatulu, B. L. & Chandra, P. Computational intelligence technique for early diagnosis of heart disease. In *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, 1–6 (IEEE, Coimbatore, India, 2015).

11. Nahar, J., Imam, T., Tickle, K. S. & Chen, Y.-P. P. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications* **40**, 96–104 (2013).

12. Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science Translational Medicine* **6**, 252ra123–252ra123 (2014).

13. Mac Gabhann, F., Ji, J. W. & Popel, A. S. Multi-scale Computational Models of Pro-angiogenic Treatments in Peripheral Arterial Disease. *Annals of Biomedical Engineering* **35**, 982–994 (2007).

14. Lu, L. & Yu, H. DR2DI: a powerful computational tool for predicting novel drug-disease associations. *Journal of Computer-Aided Molecular Design* **32**, 633–642 (2018).

15. Martínez, V., Navarro, C., Cano, C., Fajardo, W. & Blanco, A. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. *Artificial Intelligence in Medicine* **63**, 41–49 (2015).

16. Zhou, R. *et al.* NEDD: a network embedding based method for predicting drug-disease associations. *BMC Bioinformatics* **21**, 387 (2020).

17. Roberts, P. D., Spiros, A. & Geerts, H. Simulations of symptomatic treatments for Alzheimer's disease: computational analysis of pathology and mechanisms of drug action. *Alzheimer's Research & Therapy* **4**, 50 (2012).

18. Wu, C., Gudivada, R. C., Aronow, B. J. & Jegga, A. G. Computational drug repositioning through heterogeneous network clustering. *BMC Systems Biology* **7**, S6 (2013).

19. Dudley, J. T., Deshpande, T. & Butte, A. J. Exploiting drug-disease relationships for computational drug repositioning. *Briefings in Bioinformatics* **12**, 303–311 (2011).

20. Xu, R. & Wang, Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC Bioinformatics* **14**, 181 (2013).

21. Lin, X., Li, X. & Lin, X. A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules* **25**, 1375 (2020).

22. Dai, Y.-F. & Zhao, X.-M. A Survey on the Computational Approaches to Identify Drug Targets in the Postgenomic Era. *BioMed Research International* **2015**, 1–9 (2015).

23. Tatonetti, N. P., Ye, P. P., Daneshjou, R. & Altman, R. B. Data-Driven Prediction of Drug Effects and Interactions. *Science Translational Medicine* **4**, 125ra31–125ra31 (2012).

24. Chandak, P. & Tatonetti, N. P. Using Machine Learning to Identify Adverse Drug Effects Posing Increased Risk to Women. *Patterns* **1**, 100108 (2020). Publisher: Elsevier.

25. Gayvert, K. M. *et al.* A Computational Approach for Identifying Synergistic Drug Combinations. *PLOS Computational Biology* **13**, e1005308 (2017).

26. Shenoi, S. J., Ly, V., Soni, S. & Roberts, K. Developing a Search Engine for Precision Medicine. *AMIA Summits on Translational Science Proceedings* **2020**, 579–588 (2020).

27. Xu, J. *et al.* Building a PubMed knowledge graph. *Scientific Data* **7**, 205 (2020).

28. Hasan, S. *et al.* Knowledge Graph-Enabled Cancer Data Analytics. *IEEE Journal of Biomedical and Health Informatics* **24**, 1952–1967 (2020).

29. Wang, L. *et al.* Construction of a knowledge graph for diabetes complications from expert-reviewed clinical evidences. *Computer Assisted Surgery* **25**, 29–35 (2020).

30. Rossanez, A., dos Reis, J. C., Torres, R. d. S. & de Ribaupierre, H. KGen: a knowledge graph generator from biomedical scientific literature. *BMC Medical Informatics and Decision Making* **20**, 314 (2020).

31. Zheng, S. *et al.* PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in Bioinformatics* **22**, bbaa344 (2021).

32. Zhu, Y. *et al.* Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics Journal* **26**, 2737–2750 (2020).

33. Zhu, Q. *et al.* An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). *Journal of Biomedical Semantics* **11**, 13 (2020).

34. Nelson, C. A., Butte, A. J. & Baranzini, S. E. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nature Communications* **10**, 3045 (2019).

35. Huang, K. *et al.* Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks* (2021).

36. Zhou, Y., Wang, F., Tang, J., Nussinov, R. & Cheng, F. Artificial intelligence in COVID-19 drug repurposing. *The Lancet Digital Health* **2**, e667–e676 (2020).

37. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**, e26726 (2017).

38. Morselli Gysi, D. *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences* **118**, e2025581118 (2021).

39. Percha, B. & Altman, R. B. A global network of biomedical relationships derived from text. *Bioinformatics* **34**, 2614–2624 (2018).

40. Nadkarni, R. *et al.* Scientific Language Models for Biomedical Knowledge Base Completion: An Empirical Study. *arXiv:2106.09700 [cs]* (2021). ArXiv: 2106.09700.

41. Hu, W. *et al.* Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv:2005.00687 [cs, stat]* (2021). ArXiv: 2005.00687.

42. Li, N. *et al.* KGHC: a knowledge graph for hepatocellular carcinoma. *BMC Medical Informatics and Decision Making* **20**, 135 (2020).

43. Shefchek, K. A. *et al.* The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research* **48**, D704–D715 (2020).

44. Köhler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Research* **45**, D865–D876 (2017).

45. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* **32**, 267D–270 (2004).

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

47. Weinreich, S., Mangon, R., Sikkens, J., Teeuw, M. E. e. & Cornel, M. [orphanet: a european database for rare diseases]. *Nederlands tijdschrift voor geneeskunde* **152**, 518—519 (2008).

48. Hamosh, A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **33**, D514–D517 (2004).

49. Organization, W. H. (ed.) *International statistical classification of diseases and related health problems* (World Health Organization, Geneva, 2004), 10th revision, 2nd edition edn.

50. Cheung, K.-H. *et al.* PhenoDB: An Integrated Client/Server Database for Linkage and Population Genetics. *Computers and Biomedical Research* **29**, 327–337 (1996).

51. Jaasu, N. M., Kamaraj, R. & Seetharaman, R. Meddra (medical dictionary for regulatory activities). *Research journal of pharmacy and technology* **11**, 4751–4754 (2018).

52. Louden, D. N. MedGen: NCBI's Portal to Information on Medical Conditions with a Genetic Component. *Medical Reference Services Quarterly* **39**, 183–191 (2020).

53. Vasant, D. *et al.* ORDO: An Ontology Connecting Rare Disease, Epidemiology and Genetic Data 4.

54. for Medical Education and Research (MFMER), M. F. Mayo Clinic Mission and Values (2020).

55. Luo, Y. *et al.* A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia. *Nature Medicine* **26**, 1375–1379 (2020).

56. Doshi-Velez, F., Ge, Y. & Kohane, I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. *Pediatrics* **133**, e54–e63 (2014).

57. Davis, A. P., Wiegers, T. C., Rosenstein, M. C. & Mattingly, C. J. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database* **2012**, bar065–bar065 (2012).

58. Karadeniz, & Özgür, A. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC Bioinformatics* **20**, 156 (2019).

59. Ioannidis, V. N. *et al.* Drkg - drug repurposing knowledge graph for covid-19. https://github.com/gnn4dr/DRKG/ (2020).

60. Zhang, R. *et al.* Drug Repurposing for COVID-19 via Knowledge Graph Completion. *Journal of Biomedical Informatics* **115**, 103696 (2021). ArXiv: 2010.09600.

61. Richardson, P. *et al.* Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *The Lancet* **395**, e30–e31 (2020).

62. Hong, C. *et al.* Clinical knowledge extraction via sparse embedding regression (keser) with multi-center large scale electronic health record data. *NPJ Digital Medicine* **4**, 151 (2021).

63. Wang, L. L. *et al.* CORD-19: The COVID-19 Open Research Dataset. *arXiv:2004.10706 [cs]* (2020). ArXiv: 2004.10706.

64. Bhatia, P. *et al.* AWS CORD-19 Search: A Neural Search Engine for COVID-19 Literature. *arXiv:2007.09186 [cs]* (2020). ArXiv: 2007.09186.

65. Zhang, E. *et al.* Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset. *arXiv:2007.07846 [cs]* (2020). ArXiv: 2007.07846.

66. Li, X. *et al.* Network bioinformatics analysis provides insight into drug repurposing for COVID-19. *Medicine in Drug Discovery* **10**, 100090 (2021).

67. Mohamed, S. K., Nounu, A. & Nováček, V. Drug target discovery using knowledge graph embeddings. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 11–18 (ACM, Limassol Cyprus, 2019).

68. Mohamed, S. K., Nováček, V. & Nounu, A. Discovering Protein Drug Targets Using Knowledge Graph Embeddings. *Bioinformatics* btz600 (2019).

69. Sosa, D. N. *et al.* A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases 18 (2020).

70. Crichton, G., Guo, Y., Pyysalo, S. & Korhonen, A. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics* **19**, 176 (2018).

71. Long, Y. *et al.* Pre-training Graph Neural Networks for Link Prediction in Biomedical Networks 8.

72. Breit, A., Ott, S., Agibetov, A. & Samwald, M. OpenBioLink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics* **36**, 4097–4098 (2020).

73. Bastian, F. B. *et al.* The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research* **49**, D831–D847 (2021).

74. Davis, A. P. *et al.* Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research* **49**, D1138–D1143 (2021).

75. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* gkz1021 (2019).

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

77. Avram, S. *et al.* DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Research* **49**, D1160–D1169 (2021).

78. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **39**, D52–D57 (2011).

79. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports* **8**, 10872 (2018).

80. The Gene Ontology Consortium *et al.* The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* **49**, D325–D334 (2021).

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

82. Matys, V. *et al.* TRANSFAC ® : transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31**, 374–378 (2003).

83. Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research* **38**, D532–D539 (2010).

84. Aranda, B. *et al.* The IntAct molecular interaction database in 2010. *Nucleic Acids Research* **38**, D525–D531 (2010).

85. Giurgiu, M. *et al.* Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research* **47**, D559–D563 (2019).

86. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **30**, 187–200 (2021).

87. Szklarczyk, D. *et al.* The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605–D612 (2021).

88. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).

89. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Research* gkz1031 (2019).

90. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Research* **44**, D1075–D1079 (2016).

91. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biology* **13**, R5 (2012).

92. Leaman, R., Khare, R. & Lu, Z. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics* **57**, 28–37 (2015).

93. Georgiades, S., Szatmari, P. & Boyle, M. Importance of studying heterogeneity in autism. *Neuropsychiatry* **3**, 123 (2013).

94. Jeste, S. S. & Geschwind, D. H. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nature Reviews Neurology* **10**, 74–81 (2014).

95. Lenroot, R. K. & Yeung, P. K. Heterogeneity within Autism Spectrum Disorders: What have We Learned from Neuroimaging Studies? *Frontiers in Human Neuroscience* **7** (2013).

96. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (2019). ArXiv: 1810.04805.

97. Bosselut, A. *et al.* COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. *arXiv:1906.05317 [cs]* (2019). ArXiv: 1906.05317.

98. Celikyilmaz, A., Bosselut, A., He, X. & Choi, Y. Deep Communicating Agents for Abstractive Summarization. *arXiv:1803.10357 [cs]* (2018). ArXiv: 1803.10357.

99. Malaviya, C., Bhagavatula, C., Bosselut, A. & Choi, Y. Commonsense Knowledge Base Completion with Structural and Semantic Context. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 2925–2933 (2020).

100. Bosselut, A. *et al.* Discourse-Aware Neural Rewards for Coherent Text Generation. *arXiv:1805.03766 [cs]* (2018). ArXiv: 1805.03766.

101. Passos, A., Kumar, V. & McCallum, A. Lexicon Infused Phrase Embeddings for Named Entity Resolution. *arXiv:1404.5367 [cs]* (2014). ArXiv: 1404.5367.

102. Souza, L. & Ferreira, A. An Entity Resolution Approach Based on Word Embeddings and Knowledge Bases for Microblog Texts. In *XVII Brazilian Symposium on Information Systems*, 1–8 (ACM, Uberlândia Brazil, 2021).

103. Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M. & Tang, N. DeepER – Deep Entity Resolution. *Proceedings of the VLDB Endowment* **11**, 1454–1467 (2018). ArXiv: 1710.00597.

104. Alsentzer, E. *et al.* Publicly available clinical BERT embeddings. *arXiv:1904.03323* (2019).

105. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* btz682 (2019).

106. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**, 160035 (2016).

107. Li, M. M., Huang, K. & Zitnik, M. Representation learning for networks in biology and medicine: Advancements, challenges, and opportunities. *arXiv:2104.04883* (2021).