Integrating and formatting biomedical data in the Bioteque, a comprehensive repository of pre-calculated knowledge graph embeddings

Adrià Fernández-Torras[1], Miquel Duran-Frigola[1,2], Martino Bertoni[1], Martina Locatelli[1] and Patrick Aloy[1,3,*]

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
2. Ersilia Open Source Initiative, Cambridge, United Kingdom
3. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

* Corresponding author: patrick.aloy@irbbarcelona.org

Abstract

Biomedical data is accumulating at a fast pace and integrating it into a unified framework is a major challenge, so that multiple views of a given biological event can be considered simultaneously. Here we present the Bioteque, a resource of unprecedented size and scope that contains pre-calculated biomedical embeddings derived from a gigantic knowledge graph, displaying more than 450 thousand biological entities and 30 million relationships between them. The Bioteque integrates, harmonizes and formats data collected from over 150 data sources, including 12 biological entities (e.g., genes, diseases, drugs) linked by 67 types of associations (e.g., 'drug *treats* disease', 'gene *interacts with* gene'). We show how Bioteque embeddings facilitate the assessment of high-throughput interactome data, pharmacogenomics and drug repurposing exercises, and demonstrate that they can be used off-the-shelf in downstream machine learning tasks without loss of performance with respect to using original data. The Bioteque thus offers a thoroughly processed, tractable, and highly optimized assembly of the biomedical knowledge available in the public domain.

Introduction

Systematic measurements of biological samples through 'omics' technologies, together with efforts to distill the scientific literature into structured databases, are providing an ever-growing corpus of biomedical and biomolecular information[1]. Indeed, the data stored in the EMBL-EBI has increased sixfold in the last few years, from 40 petabytes in 2014 to over 250 in 2021[2]. Associated with this phenomenon, a variety of nomenclatures have been proposed, along with identifiers, levels of resolution (e.g., protein isoforms or gene splicing variants) and experimental conditions, making data integration and harmonization across platforms a challenging step[3]. As a result, even though as many as 1,641 resources were listed in the 2021 Online Molecular Biology Database Collection[4], only a small portion are broadly used, and hundreds remain isolated with their own particular formats[5,6]. Aware of the situation, several initiatives have emerged to standardize biological data by establishing common vocabularies and formats. For instance, the pioneering Harmonizome[7] was able to integrate knowledge from several gene-centric databases by representing data (e.g., gene expression, disease genetics, etc.) in a simple discretized format that was applicable to each type of data.

Nowadays, in an attempt to capture the complexity of biological systems, multiple omics profiles are often measured simultaneously (i.e., trans-omics analyses)[8,9] so that complementary views of a given phenotype or event can be considered in parallel and as a whole[10]. However, current methods mainly adapt and combine existing strategies developed to analyse individual omics data, and often the net result is that most conclusions are drawn from the most informative single data type, while the rest are used as support. It is thus fundamental to devise strategies able to capture the coordinated interplay of the many regulatory layers present in biological systems. Himmelstein *et al*. suggested the use of 'knowledge graphs' (KG) as a tool to integrate heterogeneous biomolecular data[11,12]. This concept has recently been expanded to include clinical entities[13]. In a KG, nodes represent biological or chemical entities (e.g., genes, cell lines, diseases, drugs, etc.), and edges capture the interactions or relationships between them.

However, large biomedical networks are intractable by conventional graph analytics techniques[14], thus prompting the development of graph embedding strategies to reduce the dimensionality of the data while preserving the topological information and the connectivity of the original network[15]. Moreover, the vectorial format resulting from network embedding approaches is better suited as an input for machine learning algorithms. For instance, Zitnik and Leskovek presented a set of protein embeddings that consider the protein interactions within each human tissue, as well as inter-tissue relationships, and showed their potential to predict tissue-specific protein functions[16]. Later on, the same authors embedded several networks (i.e., protein-protein, drug-target and disease-gene interactions) to explore the mechanisms of drug action[17]. Recently, Cantini *et al.* evaluated the capacity of several dimensionality reduction methods to integrate continuous multi-omics data (e.g., gene expression, copy number variation, miRNAs and methylation)[18], assessing their ability to preserve the structure of the original data and their prediction performance in different tasks. Overall, embedding-based descriptors provide a scalable and standard means to capture complex relationships between biological entities and they integrate the myriad of omics experiments associated with them[19,20].

To make biomedical knowledge embeddings available to the broad scientific community, we have developed the Bioteque, a resource of unprecedented size and scope that contains pre-calculated embeddings derived from a gigantic heterogeneous network (more than 450k nodes and 30M edges). The Bioteque harmonizes data extracted from over 150 data sources, including 12 distinct biological entities (e.g., genes, diseases, compounds) linked through 67 types of relationships (e.g., 'compound *treats* disease', 'gene *interacts with* gene'). We demonstrate that Bioteque embeddings retain the information contained in the large biological network and illustrate with examples how this concise representation of the data can be used to evaluate, characterize and predict a wide set of experimental observations. Finally, we offer an online resource to facilitate access and exploration of the pre-calculated embeddings (https://bioteque.irbbarcelona.org).

Results

A comprehensive biomedical knowledge graph (KG)

To build a KG that integrates biological and biomedical knowledge available in the public domain, we first defined the basic entities (nodes) of the network and the relationships between them (edges). As shown in Figure 1A, the resource is gene-centric. Thus, genes and gene products (GEN) are represented in the centre of the KG scheme and are involved in most associations. To better characterize genes and proteins, we collected their molecular function (MFN), cellular component localization (CMP), functional structure or domains (DOM), and biological processes or pathways (PWY). Additionally, we included information on cell lines (CLL), one of the most studied entities in biology, as well as their anatomical ensembles, namely the tissues (TIS). Analogously, chemical compounds (CPD) are depicted together with pharmacological classes (PHC) and chemical entities (CHE), two common vocabularies for medicinal compounds. Diseases (DIS) are abnormal conditions that have been widely studied in various fields, giving rise to a wide diversity of interactions between different nodes. Furthermore, although CPD and DIS are two of the major perturbational agents found in repositories like GEO[21] and LINCS[22], we also considered other biological entities such as miRNA, shRNA and overexpression vectors that can also act as perturbagens (PGN). To connect the entities in the Bioteque, we defined 67 types of associations reflecting biological relationships between them. An example of such an association would be a gene that is associated with a given pathway (GEN-ass-PWY) and might be downregulated in a certain cell (GEN-dwr-CLL) or tissue type (GEN-dwr-TIS), or a drug compound that is used to treat a disease (CPD-trt-DIS). A comprehensive list of all the biological and chemical entities included in the Bioteque, as well as the different associations, are summarized in Figure 1A, Table S1 and Table S2.

Having defined the biological entities and their interactions, we populated the Bioteque with data collected from representative datasets and resources. We first incorporated data from the Harmonizome[7], the most complete compendium of biological datasets to date, and added data from another 100 reference datasets. Each dataset was mapped to the KG scheme (or metagraph) depicted in Figure 1A. Inspired by the Harmonizome strategy, we processed each dataset separately following author guidelines, when possible (*Methods*). In brief, we binarized continuous data so that it could be represented in a network format, and we standardized identifiers from multiple sources.

The current version of the KG contains over 450k nodes, belonging to 12 types of biological entities (metanodes), and over 30M edges, representing 67 types of relationships (metaedges) (Figure 1B). Not surprisingly, genes and proteins account for most of the edges (25M) and metaedges (42) in the graph (Figure 1C-D). In terms of the number of reference datasets, protein interactions (GEN-ppi-GEN) and gene-disease associations (GEN-ass-DIS) are the most represented metaedges, supported by 17 and 15 datasets, respectively (Figure 1C). A comparison of data extracted from each dataset revealed that, although there is some overlap, most sets cover distinct associations, probably due to differences in the focus of the underlying experiments (i.e., physical[23] vs. functional[24] PPIs or drug-driven[25] vs. genomics-driven[26] gene associations) (Figure 1E).

Calculation of network embeddings across the KG

To integrate the biological knowledge gathered, we devised an approach to obtain, for a given node in the KG, a set of embeddings capturing different 'contexts' defined by one or more types of relationships between this node and other entities (Figure 2A). For example, the pharmacological context of a certain compound can be captured by 'compound *interacts with* protein' associations, while its clinical context may be captured by 'compound *treats* disease' links. The embedding procedure is as follows. We first define the types of biological entities (metanodes) to be connected and the sequence of relationships (metaedges) between them that we wish to explore. This sequence of relationships is called metapath. We then systematically examined all possible paths from the 'source' and 'target' nodes of the metapath, down-weighting highly connected nodes to ensure exhaustive exploration of the network[11]. This step yields a simplified homogeneous (when source and target metanodes belong to the same type) or bipartite (when source and target metanodes belong to different types) graph that can be explored with conventional network embedding techniques. We chose to use a 'random walk'-based method, where the trajectories of an agent that explores the network are retained and eventually fed into a text-embedding algorithm[27]. As a result, for each node in the network, a 128-dimensional vector (i.e. an embedding) is obtained, so that similar vectors are given to nodes that are proximal in the network. During this process, we mostly keep different datasets separately (i.e., without merging equivalent networks in different sources) to preserve the characteristic provenance information captured in them[28]. A more detailed description of the protocol is provided in the *Methods* section.

We have created a resource of pre-calculated biomedical embeddings, the Bioteque, where we have exhaustively considered most metapaths of length 1 and 2 extracted from the KG (i.e., direct connections between source and target nodes, or with one intermediate node between them). In addition, we have curated a collection of 135 metapaths of length ≥3 (Figure S1). Overall, the Bioteque currently holds a total of 83, 783, and 175 embeddings (unique metapaths) of length 1, 2, and ≥3, respectively (Figure 2C and Supplementary Table S3). Length 1 (L1) metapaths correspond to direct associations in the knowledge graph and provide the simplest domain knowledge representations of the entities. Larger metapaths (>L1), on the other hand, are either dedicated to connecting different entities through a third one (i.e. CPD-int-GEN-ass-DIS) or extend L1 associations to similar entities (i.e. CPD-int-GEN-ppi-GEN or CPD-trt-DIS-ass-GEN-ass-DIS), allowing the identification of more complex relationships between biological entities (i.e.,

two compounds may target different proteins yet affect the same pathway, or CPD-int-GEN-ass-PWY).

Given that the constructed KG is gene-centric, genes (GEN) are the most frequently embedded biological entity in the resource (516 unique metapaths from 44 different datasets) followed by compounds (CPD), cell lines (CLL), and diseases (DIS) (199, 169 and 150 unique metapaths, respectively) (Figure 2B). Furthermore, most of the metapaths used gene entities, such as those derived from omics experiments or literature curated annotations, as bridges to connect distinct entities (Figure S1). Compounds also play an important role, connecting pharmacological classes and chemical entities to the rest of the graph and being a major source of metapaths embedding cell lines, diseases and tissues.

Overall, the Bioteque provides a collection of 1,041 embeddings obtained from 748 unique metapaths, covering all entities defined in the biological KG (Figure 2D).

Embeddings retain the interactions in the original KG, and reveal relationships between biological entities depending on the scope and type of data

Having obtained embeddings for all nodes in the KG, we performed a set of analyses to, on the one hand, validate that the embeddings retained the connectivity observed in the KG and, on the other, to characterize each embedding space in the light of other (orthogonal) datasets in the Bioteque. As an illustrative example, Figure 3 shows the analysis of the metapath CPD-int-GEN-ass-DIS, corresponding to compounds that interact with genes, which are, in turn, associated with a disease.

To validate the embeddings, we calculated their cosine distances pairwise, and checked that proximal embeddings corresponded to edges in the KG (Figure 3B), measured with the Area Under the Receiver Operating Characteristic (AUROC) metric. Similarly, when we used the embedding distances to rank entity pairs, we found their known neighbours in the closest 10% of possible nodes (Figure 3D). Note that the goal of this study is not to benchmark the embedding method (which is already a well-accepted implementation in the field[27]), but to provide an assessment of the approach across a comprehensive set of cases.

Analogously, distances between embeddings can be used to measure whether the dimensional space preserves similarities among entities that share biological traits (i.e., cell lines sharing tissue of origin or genes sharing molecular functions). Following this rationale, we can characterize the type of biological signal captured by a given metapath by comparing its embeddings to a battery of reference biological traits, an approach already used to benchmark drug-drug similarities on the basis of shared chemical features[29]. The use of embeddings allows for straightforward comparison of entities of the same type (for example, similarity of cell lines according to their up-regulated genes can be measured by computing distances of CLL entities in the CLL-upr-GEN embedding). Likewise, it is easy to compare and uncover correlations between different types of associations. For instance, the correlation between copy number amplification and upregulation can be assessed by considering similarities in the CLL-cnu-GEN and CLL-upr-GEN embedding spaces. In the CPD-int-GEN-ass-DIS example, drug targets and gene-disease associations are among the biomedical traits that are better recapitulated by the

compound and disease embeddings (Figure 3E). Accordingly, we see how compounds and diseases associated with similar treatments are close in the embedding space. We also observe that compound-disease treatment similarity is achieved at the edge level (AUROC: 0.7), suggesting that not only compounds and diseases with similar treatments are close in the embedding space but compound-disease treatment pairs are often found in the same vicinity. Indeed, compound and disease-associated genes have proven useful in drug treatment prediction exercises[12,30].

A projection of the 128-dimensional embeddings onto a 2D space reveals clusters of drugs and treatments which, by the definition of the metapath, have identifiable targets (Figure 3A). We find, for instance, drug-disease groups associated with the treatment of leukemia (e.g., Etoposide and Daunorubicin), hormonal disorders (e.g., Somatostatin and Sermorelin), nervous system disorders (e.g., Carbidopa, Betahistine, and Protriptyline), and inflammatory conditions (e.g., Cortisone and Prednisolone). We observe that most of these drugs target a small subset of proteins or protein families directly related to the diseases, such as the growth hormone-releasing hormone receptor (GHRHR) for hypogonadism treatment, the somatostatin receptor (SSTR) for acromegaly treatment, and the DOPA decarboxylase to prevent dopamine formation in the treatment of Parkinson's disease. Additionally, the analysis reveals that drugs approved to treat either Leukemia or Kaposi's sarcoma cluster, all of them sharing the Topoisomerase II alpha (TOP2A) enzyme as target (Figure 3C). Indeed, comorbidity between these two diseases has been reported in several studies[31-33]. However, to the best of our knowledge, TOP2A has never been directly related to this relationship between diseases.

The repertoire of embeddings encoded in the Bioteque enables exploration of a given biomedical entity from multiple perspectives, often corresponding to different biological contexts, such as genes with the same biological role yet expressed in different tissues, or cell lines with similar transcriptional profiles but dissimilar at the proteome and drug response levels (Figure S2A). When performed systematically, this analysis quantifies the relationship of a certain metapath with the other metapaths in our collection, which in turn helps assess the types of biological traits that it captures. Figure S2B shows ten of the top metapaths recapitulating gene molecular function and compound pharmacological class. We see that genes targeted by the same compounds or having similar domains tend to share molecular function while, as expected, sets of interacting compounds, or those with similar binding profiles, tend to belong to the same pharmacological class.

Additionally, one can explore differences among datasets within a single metapath. In Figure S2C, we embedded three well-known protein-protein interaction (PPI) networks, representing functional interactions (STRING[24]), physical interactions (IntAct[23]), and protein-signalling interactions (OmniPath[34]), and quantified the capacity of these networks to capture a variety of biological features, from cellular localization to protein complexes. The diversity of functional interactions contained in STRING favours recapitulation of most of the features explored, especially those involving similar biological pathways (AUROC: 0.93), protein complexes (AUROC: 0.89) and protein domains (AUROC: 0.83). Not surprisingly, IntAct better preserves physical interactions (AUROC: 0.88) and shows good performance with protein complexes (AUROC: 0.86). Finally, OmniPath shows an enrichment in signalling processes such as kinase-substrate (AUROC: 0.9), phosphatase-substrate (AUROC: 0.96) and transcription factor

interactions (AUROC: 0.94), in good agreement with the type of resources used to build this network.

In general, the different considerations followed to populate these networks may favour some domains of knowledge, hence suiting different tasks, which can be efficiently and systematically revealed by transforming them into embeddings. We present three illustrative examples below on how these biological embeddings can be used off-the-shelf in a variety of tasks.

Gene expression embeddings as biological descriptors of cell lines

Gene Expression (GEx) experiments have been widely used to characterize cellular identity and state, as they broadly recapitulate tissues of origin[35] and they are notable genomic biomarkers for anticipating drug response[36]. However, these experiments typically measure the expression of 15-20k genes, yielding numerical profiles that are computationally demanding and prone to overfitting problems when used as input in machine learning approaches with limited data[37,38].

We thus explored whether our more succinct 128-dimensional vectors were able to retain the information contained within the full GEx profile. Taking the Genomics of Drug Sensitivity in Cancer (GDSC)[36] panel as a reference, we collected, for each cell line, the basal (raw) GEx (17.7K Genes) and the corresponding Bioteque metapath embedding CLL-dwr+upr-GEN-dwr+upr-CLL (hereafter CLL-gex-CLL), aimed at capturing gene expression similarities between cell lines.

We first examined the similarity landscape of the cell lines by performing a 2D projection of the raw and embedded GEx. By colouring the cell lines according to their tissue of origin, we visually verified the capacity of the CLL-gex-CLL embedding to resemble the raw GEx data (Figure 4A). Indeed, cosine similarities between CLL-gex-CLL vectors up-ranked CLLs sharing tissue of origin with a similar rate to the one obtained using correlations between raw GEx vectors (AUROC: 0.75 and 0.76, respectively) (Figure 4B).

Next, we assessed the capacity of our embeddings to predict the drug response of each cell line. To this end, we trained a standard machine learning model (a random forest classifier) for each of the 262 drugs in the panel and predicted sensitive/resistant responses using the raw GEx and our embeddings independently (*Methods*). Indeed, we found that the capacity of the CLL-gex-CLL embedding to recapitulate drug response is equivalent to that observed when the raw GEx data is used (average AUROC: 0.70 and 0.71, respectively). Moreover, the models based on embeddings had strong concordance with the raw GEx model (0.94 Pearson correlation) (Figure 4C). This level of agreement is remarkable and represents a clear advantage for the embeddings since they are smaller, easier to handle and do not require expert knowledge to pre-process the raw data. A disadvantage of the embedding approach is the less obvious interpretability of predictions.

After verifying that the Bioteque GEx embeddings retain the basal transcriptional information from the cell lines, we used them to compare profiles obtained from different cell line panels. Specifically, we compared the GDSC with the Cancer Cell Line Encyclopedia (CCLE)[39]. In agreement with previous reports, we observed a strong correspondence between the two panels, measured as CLL-gex-CLL similarities in the embedding space (AUROC: 0.89) (Figure 4D). To assess whether these similarities were driven by the up- or down-regulation of the same genes,

we repeated the analysis focusing on the CLL-upr-GEN and CLL-dwr-GEN embeddings and checked whether the CLL-GEN similarities in the GDSC panel were also preserved in the CCLE. In general, the recovery score of cell line-specific up-/down-regulated genes (i.e. CLL-GEN pairs) was lower (AUROC: 0.78) (Figure 4D). We obtained similar results when we reversed the exercise and used CCLE embeddings to recapitulate GDSC similarities (Figure S3). This finding suggests that, while cell line similarities between panels are robust (i.e., cell lines sharing similar transcriptional signatures in one panel also share similar ones in the other), the specific transcriptional changes of a given cell line may differ. The characterization of the CLL-CLL and GEN-GEN distances further confirmed the better recapitulation of cell line similarity in comparison to gene similarity between panels (AUROC: 0.9 and 0.8 for the CLL-CLL and GEN-GEN similarities, respectively) (Figure 4E). Furthermore, the CLL-CLL similarity characterization revealed a strong concordance between protein and transcript levels (AUROCs: 0.9 and 0.8 for protein abundance and deficiency, respectively), which was partially driven by the same CLL-GEN pairs (AUROC: 0.72 and 0.63 for the protein abundance and protein deficiency CLL-GEN pairs, respectively) (Figure S3). In addition to tissue of origin, we also observed resemblances between cell lines used to model a given disease (AUROC: 0.78), sharing fitness profiles (AUROC: 0.72 for negative and 0.69 for positive fitness profiles) and similar drug responses (AUROC: 0.7). Finally, the GEN-GEN similarities also revealed a mild recapitulation of known co-expressed gene pairs (AUROC: 0.64 and 0.69, for the up- and down-regulated gene similarities, respectively), thereby suggesting that some of the genes commonly up- or down-regulated by the same cell lines from different panels may share the same transcriptional regulatory programs.

On the whole, our approach retains meaningful information from the original data into a reduced number of dimensions (128 vs ~20k), even when the data comes from a much noisier source such as transcriptomic technologies. We believe that the standardized and dense format of our embeddings provides a by-default way to integrate and compare omics datasets.

Assessing the uniqueness of new omics datasets

Since the consolidation of high-throughput omics technologies, several long-term initiatives have been established to comprehensively characterize certain levels of biological systems (i.e., genetic interactions in yeast[40] or the transcriptomes of cell line panels and human tissues[39,41]). After several years running, all these efforts have had to balance a potential decrease in novelty and an increase in costs as the screens approach saturation. The Bioteque provides a corpus of biological data cast to a single format and, as such, it offers a means to quantify the degree of novelty of new data releases and omics experiments. As an illustrative example, we analyse the systematic charting of the Human Reference Interactome (HuRI) with the yeast two-hybrid methodology, which has already identified over 50,000 protein-protein interactions (PPIs) of high quality over the last 15 years[42-44].

To estimate the level of support from different experiments and assess the novelty of the latest HuRI release (HuRI-III[44]), we used the embedding space of relevant metapaths to determine the biological context of each pair of interacting proteins. In brief, for each gene-gene pair, we calculated an empirical P-value corresponding to the measured similarity in the embedding space, which allowed for commensurate comparison of distance/similarity measures performed in

different embedding spaces (see *Methods*). Note that, to have a fair representation of the known physical interactions, we embedded an older version of the protein-interaction network, without including any of the entries from HuRI-III. We then categorized each interaction in HuRI-III into four groups, depending on the level of support contained in the Bioteque embeddings. In this regard, we labelled them as (i) known and supported interactions (covered by GEN-ppi-GEN and at least another metapath), (ii) known interactions (only covered by GEN-ppi-GEN), (iii) supported interactions (covered by other metapaths but not GEN-ppi-GEN) and (iv) potentially novel interactions (with no apparent support in any of the metapaths screened) (Figure 5A). Remarkably, after three updated versions of HuRI, almost half of the interactions can be classified as potentially novel according to the selected metapaths. Moreover, although only 5,825 (11%) of the interactions were supported by GEN-ppi-GEN embeddings, mostly coming from previous versions of HuRI[42,43], our analysis suggests that a higher proportion can be recovered. In fact, at 0.05 FDR (*Methods*), the GEN-ppi-GEN embedding recovered 18% of HuRI-III, retrieving 5,456 (94%) of previously known interactions while finding 3,994 new pairs (Figure 5B). On the other hand, we observed a substantial number of physical interactions presumably involved in similar pathways (GEN-ass-PWY), cellular components (GEN-has-CMP), or protein domains (GEN-has-DOM). At 0.05 FDR, these metapaths alone recovered 6,905 unique interactions of which 4484 (65%) were not obvious from the physical interaction space (Figure 5C). To delve into the correlation and relative importance of the metapath for explaining PPIs, we used the P-values as features for a tree-based machine learning model trained to identify HuRI-III edges. We then assessed the importance of each metapath for the prediction using Shapley values[45]. As visually anticipated from the heatmap, the model achieved a reasonable performance (AUROC: 0.69), mostly relying on previously known physical interactions, cellular components, protein domains, and pathways, all of them showing a certain degree of agreement (Figure S4). Interestingly, we also identified successfully predicted cases with little to no evidence from physical PPIs. For instance, our metapath distance-based model predicted the interaction between the neuronal proteins HOMER1 and SHANK2, the tRNA-splicing endonuclease TSEN54 and the polyribonucleotide CLP1, and the adenosine deaminase ADARB1 and the protein kinase PRKRA, none of which had any reported evidence in protein interaction databases but showed strong positive support in the GEN-ass-PWY, GEN-has-CMP, and GEN-has-DOM metapaths, respectively (Figure 5D). Indeed, some of these associations have been related in other contexts[46-48], but with no indication of physical interactions before HuRI-III.

We have shown how the continuous and interpretable dimensional space of the Bioteque embeddings provides a powerful framework for characterizing individual observations, which can, in turn, be exploited to guide the interpretation of the entire dataset and, to some extent, assess the novelty of the data.

Discovery of drug repurposing opportunities using the multiple scopes offered by the embeddings

Drug repurposing is often regarded as an attractive opportunity to quickly develop new therapies[49]. However, perhaps with the exception of cancer, where abundant models and molecular data are available, it is difficult to generate data-driven predictors to suggest new uses for approved or investigational drugs, mainly due to the lack of disease descriptors and the small

number of known drug-disease indications. Indeed, according to the last update of repoDB (https://repodb.net), half of the drugs (1,097) have only one approved indication, and a third of the diseases (458) are treated with only one drug (Figure S5). Thus, training models with all the known drug-disease associations and later transfer of the insights gained to underexplored treatment areas would be highly desirable[50,51].

To explore whether the Bioteque could be useful in this scenario, we set out to predict new compound-disease indication pairs introduced in repoDB in 2020 (v2) training a model on the previous version (v1), launched in 2017 (*Methods*). We mapped all disease terms to the Disease Ontology, removed redundant indications (according to the ontology), and trained a conventional random forest classifier to predict whether a given CPD-DIS corresponds to a true therapeutic indication. We used two sets of metapath embeddings: one in which we used L1 metapaths (*Short*) based on the drug targets (CPD-int-GEN) and gene associations (DIS-ass-GEN), and another in which we used L3 metapath (*Long)* linking the pharmacological class and the treatment of known CPD and DIS to those sharing drug target (CPD-int-GEN-int-CPD-has-PHC) or gene associations (DIS-ass-GEN-ass-DIS-trt-CPD), respectively. We chose to use drug targets and gene associations because we observed that their embeddings broadly recapitulate the pharmacological class and the disease treatment for a sufficient number of nodes (Figure S5). Moreover, to assess the capacity of the gene-based similarities to correctly infer the treatment, we also tested a metapath (*Long-b*) in which we prevented the CPDs and DISs from being linked, thus making the association with PHC or treatment purely based on the gene-driven similarity to other CPD or DIS. To avoid trivial predictions, we removed associations with PHCs or treatments for drugs and disease unique to the repoDB v2 in all *Long* metapaths. As a basal model, we used chemical fingerprints (ECFP4, 2048 bits) for the CPDs and either one-hot identity vectors (*Basal1*) or binary gene annotations (*Basal2*) for the DISs.

We considered two use cases: a drug repurposing exercise, in which we ranked all the diseases predicted to be potentially treated with a given compound, and a prescription exercise, in which we ranked all compounds that might be useful to treat a given disease. In both scenarios, the three metapath embeddings showed remarkable predictive power compared to the basal models, with the model built from *Long* embeddings being the one with superior performance (Figure 6A-B). Specifically, for half the tested compounds, the *Long* embeddings model found a new validated therapeutic purpose within the top 2% of disease predictions (corresponding to the top 10 ranked diseases). Analogously, for roughly 50% of the diseases, the model found a correct treatment within the top 1% of compound predictions (corresponding to the top 8 ranked compounds). Furthermore, although with poorer performance, our biological embeddings were able to yield correct predictions for compounds and diseases with minimal evidence available (i.e., with only one known indication or treatment in repoDB v1) (Figure 6A-D, dotted lines). In contrast, the best performing basal model (*Basal2*) found correct predictions for 32% of the compounds and 41% of the diseases within the same ranking range. Moreover, the Bioteque-based models were better at consistently up-ranking indications (or treatments) of compounds (or diseases) with multiple new annotations in repoDB v2 (Figure 6C-E). Finally, we verified that these predictions covered a broad range of therapeutic areas and disease families. Indeed, we found that within the top 1% of predictions, the *Long* model successfully predicted one indication or treatment for 20% of all the compounds and diseases in each therapeutic area or disease family (Figure 6C-D and S6).

These results were reproduced with the *Long-b* model, showing that, as expected, the genes associated with drugs or diseases of known treatment can indeed be used to better infer the activity of drugs and diseases with unknown indication (Figure S6).

Overall, we showed how Bioteque embeddings can be directly plugged into machine learning models, and how, by combining different context associations into larger metapaths, they can increase the performance of drug-disease prediction models. Indeed, we used a preliminary version of Bioteque embeddings to successfully identify potential targets for a set of kinase inhibitors from perturbational profiles, including drug-induced transcriptional changes and cell sensitivity data, in several cell lines[52].

The Bioteque resource

We built an online resource to facilitate access to all the pre-calculated Bioteque embeddings (https://bioteque.irbbarcelona.org). The Bioteque web offers a visual way to explore over one thousand metapaths by selecting the nodes to connect, as well as the type of relationship between them. For a selected metapath, we provide an analytical card displaying a 2D representation of the embedding, as well as a validation ROC curve. Additionally, we display distance distributions of the embeddings generated for each entity and the biological associations that are best recapitulated by the metapath of interest. Finally, there is a downloadable file containing the embeddings for each node, a matrix with distances between them, and the analytical card displayed on the web.

The Bioteque also provides information on the specific sources used to construct each metapath, and some general statistics on the contents of the current version of this web resource. We also provide a link to our GitLab repository, which contain the full code necessary to pre-process the data to generate and analyse biological embeddings (http://gitlabsbnb.irbbarcelona.org/bioteque). The entire resource, including the underlying data and biological embeddings, will be updated once per year, or as soon as a major dataset is released.

Discussion

With the accumulation of large-scale molecular and cell biology datasets, coming from ever-growing literature, omics experiments and high-throughput screenings, new frameworks for integrative data analysis are necessary. For a given biological entity (e.g., a gene), we are now able to stack multiple layers of its biological complexity (e.g., its structure, function, regulation, or interactions), which offers an opportunity for a more complete, systemic view of biological phenomena, but brings along several challenges, including the handling of different data structures, nomenclatures, signal strengths, and variable dimensionalities.

To tackle these challenges, we have developed the Bioteque, a resource of pre-calculated, fixed-format vector embeddings built from a comprehensive biomedical knowledge graph (KG). The KG contains physical entities like genes, cell lines, and compounds, as well as concepts like pathways, molecular functions, and pharmacological classes. Embeddings capture the

connections between nodes in the KG according to a certain metapath, i.e., a sequence of semantic and/or mechanistic relationships between entities. We have shown how this approach is useful to (i) produce compact descriptors that broadly preserve the original data, (ii) systematically characterize biological datasets such as cancer cell line transcriptional signatures, (iii) assess the novelty of a given omics experiment, and (iv) mine for drug repurposing opportunities based on multiple associations between drugs and diseases.

In the Bioteque, we have incorporated datasets from over 150 distinct sources, keeping the integrity of the original data to a feasible extent and applying standard transformations when required. Note that the accuracy of the Bioteque is determined by the quality of the source data. As experimental technologies continue to evolve, new information will populate these databases and novel standards will emerge, opening the door for more comprehensive and higher quality embeddings. In addition, as a first attempt, we used a network embedding technique that purely relies on the graph topology built from the biomedical data, in contrast to other techniques that also leverage node and edge attributes, since incorporating other layers of information could compromise the controlled identity of the metapaths. Nevertheless, Bioteque descriptors can be easily recycled as node features for new task-specific networks, thus transferring the learning encoded from orthogonal biomedical datasets to more complex, attribute-aware models. Finally, we would like to point out that there are parts of the current biomedical knowledge that have not yet been included in the resource, such as antibody-target interactions and metabolomics. As a molecular/cell-centric resource, the Bioteque also lacks patient-derived data[13], including interactions with the microbiome[53]. Updated versions of the Bioteque will have to be complemented with the incorporation of other fields of biological knowledge, the re-accommodation of the datasets in the resource (based on updated standards), and the improvement of embedding strategies. Moreover, future developments will explore the adoption of biological descriptors as features for a variety of downstream-specific tasks, including a systematic screening of the biological support of wet lab experiments or the modelling of complex diseases to guide the generation of new chemical entities to tackle them[20].

Methods

Building the metagraph

All gathered data was stored in a graph database (KG) in which nodes represent biological or chemical entities and edges represent associations between them.

Nodes (entities)

The nodes in the graph can belong to one of 12 types (a.k.a. metanodes). For each entity type, we predefined a universe of nodes and chose a reference vocabulary based on standard terminologies. These 12 entity types are (in alphabetical order):

Chemical entities (CHE): Chemistry terminologies extracted from the Chemical Entities of Biological Interest (ChEBI) ontology[54].

Cells (CLL): Cell lines used in biomedical research and extracted from the Cellosaurus resource[55].

Cellular Components (CMP): Biomolecular structures and complexes as defined by the Gene Ontology[56] (extracted from the basic filtered ontology).

Compounds (CPD): Small molecules codified with the standard InChIKey. As we do not use any predefined library of compounds, the universe will be determined by the union of compounds included in other datasets (e.g., drug-target interactions).

Diseases (DIS): Abnormal conditions, drug side effects and symptoms. We used the Disease Ontology[57] as a reference vocabulary.

Domains (DOM): Functional and structural protein domains extracted from InterPro[58].

Genes and Proteins (GEN): Genes and proteins were unified and stored by Uniprot[59] accession code (UniProtAC). We worked on the reviewed Human proteome.

Molecular Functions (MFN): Biological function of the proteins defined by the basic Gene Ontology[56].

Perturbagens (PGN): CRISPR, overexpression, and shRNA perturbations. Note that PGNs are always mapped to the corresponding perturbed gene when constructing the metapath. Therefore, instead of providing PGN labels, we provide the UniProtACof the perturbed genes.

Pharmacologic Classes (PHC): Pharmacologic classes defined by the Anatomical Therapeutic Chemical (ATC) code (http://www.whocc.no).

Pathways (PWY): Biological pathways and processes. We used Reactome[60] as a reference vocabulary.

Tissues (TIS): Anatomical tissues and cell types defined by the BRENDA Tissue Ontology[61].

Please note that in the datasets containing ontological terms (CMP, DIS, MFN and PWY), we removed the least informative terms (i.e., those that are higher up in the ontology). These terms were identified by calculating the information content[62]. The node universe for each entity and the list of removed terms are available in Table S4.

Vocabulary mapping

To integrate terminologies, we extracted curated cross-references from the official terminology sources and associated ontologies. As the nomenclatures used to identify diseases and pathways were particularly diverse and rarely cross-referenced, we further increased the mapping of these terms by inferring similarities within concepts as detailed below.

Diseases were mapped by calculating disease term similarities through shared cross-references to the Unified Medical Language System (UMLS), obtained from the DisGeNET mapping resources (https://www.disgenet.org/downloads). Specifically, we encoded each disease term into a binary vector spanning the universe of UMLS terms of all nomenclatures. We then transformed the binary vectors with the corresponding term frequency-inverse document frequency (TF-IDF) values and computed pairwise cosine distances between the Disease Ontology and the rest of the vocabularies. Using the similarities obtained from curated cross-references as reference, we found a cosine similarity cutoff of 0.5 to correspond to an empirical P-value of $5x10^{-4}$.

Pathway cross-references were extracted from the ComPath resource[63] and extended following the PathCards[64] approach. This approach first clusters the pathways into SuperPaths based on overlapping genes and then uses Jaccard similarities between the SuperPaths genes to define pathway similarity. We used the same parameters described in the PathCards paper (0.9 for the overlap cutoff, 20 minimum genes in the pathways, and a Jaccard similarity of at least 0.7).

Edges (associations)

Edges in the graph are used to link biological and/or chemical entities. Since two entities may be connected by multiple edge types (i.e., 'compound treats disease' or 'compound causes disease'), we define the associations as triplets (metapaths) of entity-relationship-entity (CPD-trt-DIS, CPD-cau-DIS).

Homogeneous associations are those concerning entities (metanodes) of the same type (e.g., gene is coexpressed with GEN, GEN-cex-GEN), while heterogeneous associations are related to entities of different types (e.g., tissue has cell, TIS-has-CLL). Note that we annotated only one direction of the heterogeneous associations (in fact, we kept CLL-has-TIS instead of TIS-has-CLL), although both directions are valid when defining metapaths. On the other hand, edges were treated as directional whenever a homogeneous association had only one valid directionality, like in the case of kinase-substrate interactions (gene phosphorylates gene, GEN-pho-GEN) or transcription factor regulations (GEN-reg-GEN). Finally, edges corresponding to similarity measures required a predefined set of nodes for pairwise comparison, and they were computed only after the rest of the graph was populated.

Populating the knowledge graph with data

For each type of association or metaedge, we can have one or more datasets (Supplementary Table S2). Datasets are not merged but kept as individual sources so that they can be embedded individually or in combination within a given metapath. The dataset processing pipeline consisted of two steps. In the first step, nomenclatures were standardized and cutoffs were applied. In the second, applied only to ontological data, terminologies were mapped and the network was pruned.

Dataset standardization

We processed each dataset individually in order to handle the diversity of formats and data types. The guiding principles of data processing were those defined by the Harmonizome[7].

Datasets that already provided binary data were integrated naturally by converting them into the network format of the KG. If the database provided a measure of confidence (e.g., edge weights or P-values), we applied default cutoffs (if given) and/or followed author recommendations in order to remove spurious interactions. Occasionally, the same dataset can be further divided into different subsets on the basis of a given categorical variable (e.g., curated/inferred). We kept these subsets as independent datasets when applicable. For instance, there is a 'curated' version of DisGeNET and an 'inferred' version of it.

Continuous data requires the application of a cutoff before its integration in the KG. Below, we detail how these cut-offs were chosen depending on the nature of the data.

Transcriptomics and proteomics data: We first mapped the samples and genes to our reference vocabulary and collapsed the duplicates by their mean value. A log2 transformation was then applied followed by a quantile normalization of the genes (unless the dataset was already transformed by the data providers). Next, we subtracted the median and scaled the data according to the quantile range of each gene. Finally, the top 250 most positive and negative genes were selected for each sample and kept in the corresponding metaedges (e.g., CLL-upr-GEN and CLL-dwr-GEN).

Drug sensitivity: To binarize drug sensitivity data, we used the waterfall method described by Barretina *et al*.[65]. This method ranks cell lines on the basis of a drug response measure, for instance, the area under the growth inhibition curve (AUC), and uses the shape of the plot to define a sensitivity threshold. The waterfall method was applied for each compound in the dataset, keeping at least 1% but no more than 20% of sensitive cell lines and requiring an AUC sensitivity value lower than 0.9.

Perturbation experiments: Gene perturbation data required a preliminary step to differentiate the type of perturbation (e.g., 'CRISPR modification silences gene A') from its outcome (e.g., 'silencing gene A results in overexpression of gene B'). First, for each perturbation in the dataset, we created a perturbagen (PGN) node with a unique identifier. We then simplified the two-step relationship (e.g., 'perturbagen that silences gene A upregulates gene B') into a 'perturbagen upregulates gene B' association, in this case (PGN-upr-GEN).

Other datasets: For some datasets containing continuous data, we had to apply customized approaches to convert them into a network format. All Python scripts used to process each particular dataset are provided in: http://gitlabsbnb.irbbarcelona.org/bioteque.

Terminologies and pruning

Six terminologies (namely, CMP, DOM, MFN, PHC and PWY) had semantic relationships between them. In these cases, we propagated all the reported relationships with other terms (e.g., GEN) through the parents of their corresponding ontologies. To maximize coverage, propagation was done before cross-referencing.

Selection of metapaths

We chose a controlled set of metapaths for which we precomputed embeddings. These are the embeddings that are deposited in the Bioteque resource. The metapaths were selected as follows.

Length 1 (L1): All possible metapaths of length 1 are embedded except for those capturing cross-references (DIS-xrf-DIS), ontologies (PWY-hsp-PWY), compound-compound similarities (CPD-sim-CPD), and PGN associations. Note that PGN nodes are mapped to the corresponding perturbed genes through the PGN-pdw-GEN or PGN-pup-GEN metapaths (thus, >L1 metapaths).

Length 2 (L2): Only the mimicking (e.g. CLL-dwr+upr-GEN-dwr+upr-CLL) or reversion (CLL-upr+dwr-GEN-dwr+upr-CLL) of both directions (up/down) are used for metapaths connecting entities through transcriptomic, proteomic or transcription factor signatures. CLL and TIS are always connected through the CLL-has-TIS association. Finally, only the following associations

are allowed when linking cells and genes within a metapath: CLL-upr-GEN, CLL-dwr-GEN, CLL-mut-GEN.

Length 3 (L3): L3 metapaths are constructed by linking L1 metapaths with any of the following L2 metapaths: CLL-dwr+upr-GEN-dwr+upr-CLL; CLL-has-TIS-has-CLL; CMP-has-GEN-has-CMP; CPD-has-PHC-has-CPD; CPD-int-GEN-int-CPD; DIS-ass-GEN-ass-DIS; DOM-has-GEN-has-DOM; MFN-has-GEN-has-MFN; TIS-dwr+upr-GEN-dwr+upr-TIS; or PWY-ass-GEN-ass-PWY. GENs from the PGN-pup-GEN or PGN-pdw-GEN are linked through heterogeneous or directed homogeneous associations but not through undirected homogeneous associations.

Length >3 (>L3): Generated when mapping the source or target PGN to the perturbed genes in L3 metapaths.

In the case of directed homogeneous associations, we used the '_' mark next to the entity that acted as the source of the association. For instance, GEN-_pho-GEN-ass-PWY links the kinases to the pathways associated with their substrates while GEN-pho_-GEN-ass-PWY links the substrates with the pathways associated with their kinases.

Finally, metapaths whose embedding did not preserve the original network or that failed to keep most of the nodes in a single connected component were removed as described in the following section. The entire list of the embedded metapaths is provided in Table S3.

Obtaining Bioteque embeddings

To obtain the embeddings we used the node2vec algorithm[27], a well-accepted approach based on random walk trajectories[66], in which metapaths are used as single networks and fed to the node2vec algorithm.

Homogeneous and bipartite networks

L1 metapaths already correspond to homogenous or bipartite networks. For >L1 metapaths, the source and target nodes were connected by computing degree-weighted path counts (DWPC)[11] through the corresponding datasets and associations in the metapath. To this end, we sorted the datasets according to the associations of the metapath, represented them as adjacent matrices and kept the same source (rows) and target (columns) node universe as the target and source nodes of the previous and following datasets, respectively. Following the DWPC method, we first downweighted the degree of the nodes in each of the datasets by raising the degrees to the *-0.5* power. We then calculated the DWPC values by concatenating the matrix multiplication from the source to the target dataset. As a result, we obtained a new n x m matrix where n are the source nodes of the first dataset and m are the target nodes of the last dataset. The values of the matrix are the DWPC between the source and target nodes, which are used as weights during the random walker exploration. Finally, we limited the number of edges for each node to 5% of the total possible neighbours (with a minimum of 3 and maximum of 250 edges per node).

Occasionally, we used more than one dataset within the same association or we combined two metapaths into one. This is a common case for >L1 metapaths with transcriptomic signatures where the two directions (CLL-upr-GEN and CLL-dwr-GEN) are often combined (CLL-dwr+upr-GEN-dwr+upr-CPD). To handle these cases, we first obtained an individual network for each

metapath or dataset following the approach detailed above. We then merged all the networks by taking the union of the edges (L1 metapaths) or adding the DWPC values (>L1 metapaths).

At the end of the process, we removed network components that cover less than 5% of the entities from the network. And we also removed from the source metapaths that fail to retain 50% of the total nodes within their network components.

Node2vec parameters

The node2vec algorithm consists of a random walk-driven exploration of the network followed by a feature vector learning through a skip gram neural network architecture.

We implemented a custom random walker (with the node2vec parameters *p* and *q* set to 1) and ran 100 walks of length 100 for each node of the network. For >L1 metapaths, we scaled the DWPC values for each node to sum 1 and used them as probabilities to bias the random walker. We used the C++ skip-gram implementation provided by Dong *et al*.[67] with default parameters to obtain a 128-dimensional vector for each node.

Embedding evaluation

We used opt-SNE to generate the 2D representation of the embeddings[68]. To assess the quality of the embeddings, we reassembled the network obtained from the metapath using the embedding vectors. To this end, we first computed the cosine distance of each edge in the network using the embedding vectors of the nodes. Next, we generated 100 random permutations for each edge in the network and calculated the cosine distances between them. Finally, we sorted all the distances and computed the area under the ROC curve (AUROC) using the network edges and the random permutations as the positive and negative sets, respectively. When assessing >L1 metapaths, we repeated the same exercise using 3 extra network subsets obtained by keeping, for each node, the top 1%, 25% and 50% closest neighbours according to the DWPC weights of their edges. Embeddings with an AUROC below 0.8 were removed from the resource.

Embedding characterization

To characterize the embeddings, we first preselected a collection of reference networks representing commonly used biological associations. Then, given a set of embeddings corresponding to a certain metapath, we tested their capacity to recapitulate edges from other (orthogonal) datasets (i.e., the reference networks). Two measures were kept, the coverage (i.e., the number of overlapping nodes) and the AUROC, following the approach described above.

Aiming to extend this characterization, for each metapath we sought to characterize nodes separately, on the basis of their entity type. We first calculated the term frequency-inverse document frequency (TF-IDF) values of the nodes from each reference network in our collection. Next, within the same entity type and network, we used the TF-IDF-transformed vectors to compute pairwise cosine similarities between nodes. Finally, we built the entity similarity network by keeping the top 5 closest neighbours for each node. Note that from one heterogeneous (bipartite) network this process yields two homogeneous networks, one for each entity type.

Some of the networks in our collection required customized pre-processing. To represent perturbation associations, we directly linked the perturbed genes (PGN-pup-GEN or PGN-pdw-GEN) and the outcome of such perturbation (e.g., PGN-bfn-CLL or PGN-upr-GEN) through the corresponding associations and datasets. We computed the CHE-has-CPD similarity networks by directly linking each node with the top 3 partners that shared more neighbours. Additionally, some entity similarity networks were gathered from other sources, like the CPD-CPD mechanism of action similarity obtained from our Chemical Checker resource[69].

Embedding-based gene expression analysis of cancer cell lines

We downloaded the RMA-normalized gene expression (GEx) and the drug sensitivity data from the GDSC1000[36] web resource (https://www.cancerrxgene.org). We mapped the cell lines and genes to our reference vocabularies and took the mean value whenever duplicates occurred. We used the tissue of origin annotations from the CLUE cell app (https://clue.io/cell-app), which were already part of our graph (CLL-has-TIS, cl_tissue_clueio). Regarding CCLE data, we used the next-generation data[39] from the Broad Institute Portal (https://portals.broadinstitute.org/ccle/about). We processed the RNAseq data and produced three embeddings (CLL-upr-GEN, CLL-dwr-GEN and CLL-dwr+upr-GEN-dwr+upr-CLL) following the pipeline detailed in the “Dataset standardization” and “Obtaining the embeddings” sections.

In the drug sensitivity prediction exercise, we trained a random forest (RF) classifier for each drug and each GEx input data (i.e., the raw GEx or any of the GEx-derived embeddings). After removing drugs with less than 10 sensitive or resistant cell lines, we modelled 262 drugs. We used the SciKit Learn implementation of the RF algorithm, with a 10-fold stratified cross-validation scheme, and optimized RF hyperparameters over 20 iterations of Hyperopt[70].

Analysis of the HuRI-III protein-protein interaction network

We downloaded HuRI-III from the Human interactome atlas (http://www.interactome-atlas.org/). Next, we considered all L1 metapaths containing a GEN metanode, keeping the dataset with higher coverage for each metapath and discarding those covering less than 10% of the HuRI-III network. As a representative of PPI interactions (GEN-ppi-GEN), we used a version of IntAct dated December 2019 (before publication of the HuRI-III from which we removed all entries belonging to the HuRI-III screening (IMEX: IM-25472). Next, we calculated the cosine distance between each PPI in each of the metapath embedding spaces and ranked the distances according to the distance distribution of each of the proteins. Distances and rankings were obtained with FAISS[71]. To derive empirical P-values, we transformed the rankings into percentiles by normalizing them by the total number of covered genes in each metapath and kept the geometric mean of the normalized co-ranked pairs.

In parallel, we generated 1000 random permutations of HuRI-III by randomly swapping each of the HuRI-III edges 10 times using the BiRewire bioconductor package (https://doi.org/doi:10.18129/B9.bioc.BiRewire) and, likewise, calculated P-values for each metapath. For each permuted network, we calculated the recovery of the edges with a sliding P-value cutoff (between 1 and 0.001) and averaged the counts at each cutoff. After repeating this

process with the HuRI-III network, we were able to derive, for each metapath, the expected fold change (FC) across different P-value cutoffs (i.e., the number of covered HuRI-III edges at a given P-value cutoff divided by the average number of covered edges in the permuted networks). Moreover, the permuted networks were also used to estimate an empirical FDR for a given P-value. For instance, for each metapath, we found the P-value cutoff associated with a 0.05 FDR by calculating the minimum P-value needed to cover no more than 5% of the permuted network edges. Finally, to build the matrix shown in Figure 5A, we selected the top 20 metapaths with the highest FC (i.e., FC average in the P-value range between 0.1 and 0.001), and used their P-values to cluster the PPIs with the fastcluster package[72] and the ward distance update formula.

To obtain the Shapley values, we trained a XGBoost model to classify GEN-GEN edges as 'positive' (i.e., present in HuRI-III) or 'negative' using the P-values across metapaths as features. To sample negative pairs, we used the instance of the permuted networks hitting fewer HuRI-III edges (~3%) in order to avoid having the same edge as positive and negative instance at the same time. Furthermore, since the objective of this exercise was to study the interplay between the metapaths, we removed edges that were covered by less than 10 (50%) metapaths, resulting in a dataset of 60k positive and negative pairs. A simple mean imputation was applied to the missing P-values. At training time, we implemented a 20-fold stratified cross-validation split scheme and fine-tuned the hyperparameters using 20 iterations of Hyperopt[70]. Finally, we obtained the Shapley values from the test splits by implementing the TreeExplainer method[45]. All subsequent analyses and figures were obtained using the SHAP package (https://github.com/slundberg/shap).

Drug repurposing based on drug and disease embeddings

The first release of the repoDB (v1) data was downloaded from http://apps.chiragjpgroup.org/repoDB while the updated release (v2) was obtained from https://unmtid-shinyapps.net/shiny/repodb. Compounds were mapped to InChIKeys and diseases to the Disease Ontology (DO) forcing a 1:1 mapping. As features, we used the following metapaths (datasets) from the Bioteque resource: CPD-int-GEN (curated_targets); DIS-ass-GEN (disgenet_curated+disgenet_inferred); CPD-int-GEN-int-CPD-has-PHC (curated_targets-curated_targets-atc_drugs); and DIS-ass-GEN-ass-DIS-trt-CPD (disgenet_curated+disgenet_inferred-disgenet_curated+disgenet_inferred-repoDB).

Additionally, we obtained the 2,048-bit Morgan fingerprints (ECDF4) of the compounds using RDKIT (http://rdkit.org) and used the adjacency matrix of the disease-gene network from DisGeNET as binary descriptors of diseases. Having defined the features of the model, we filtered out those drugs and diseases from repoDB that fell outside the embedding universe and removed redundant pairs by de-propagating the associations to the most specific drug-disease terms according to the Disease Ontology. As a result, the train (repoDB v1) and test (repoDB v2) splits consisted of 2,522 and 1,187 unique drug-disease associations, respectively (Figure S5). Additionally, to prevent the model from focusing on the most frequently annotated drug and disease entities, we further processed the train data to balance the number of associations (degree of the nodes). More specifically, we capped the number of drug or disease associations to 5% of all possible associations (44 diseases and 26 drugs, respectively). Therefore, the

associations of those drugs or diseases exceeding this limit were subsampled by performing a K-means clustering (where K was set to the capping limit) using the CPD-int-GEN or DIS-ass-GEN embeddings as features, and by randomly selecting a representative association from each of the clusters (Figure S5). This step slightly decreased the number of training data to 2,326 drug-disease associations.

Next, we produced train negative pairs by aggregating 20 negative networks obtained by randomly swapping the edges of the training data (thus, forcing a ratio of 1:20 between the positive and negative instances), while preventing inconsistencies in the Disease Ontology (i.e., having a negative association that would be obtained by propagating a positive drug-disease association through the ontology). Note that, to comply with the time-split scenario, we did not remove any negative drug-disease pair reported to be positive in the repoDB v2 release.

Once the training data was ready, we ran an RF classifier for each of the explored models using 20 iterations of hyperopt[70] to fine-tune the hyperparameters. At prediction time, drug-disease associations in repoDB v2 were considered positive test pairs, whereas all the remaining drug-disease pairwise combinations were considered negative pairs. To avoid inconsistencies, we removed those negative pairs that were semantically related to positive pairs according to the Disease Ontology. As a result, we obtained between [460-500] diseases and [750-800] drug predictions for each drug and disease, respectively. As most of the drugs and diseases only had one or two positive instances, we assessed the performance of the models by ranking all the predictions individually for each entity (ranks were used as percentages). Additionally, we calculated ROC curves for those drugs and diseases that had at least 5 positive instances. Finally, we obtained the pharmacological action of the drugs by mapping them to the uppermost level of the Anatomical Therapeutic Chemical (ATC) classification, when available. Likewise, disease families were derived by propagating the disease terms to the first and second levels of the Disease Ontology.

Acknowledgements

P.A. acknowledges the support of the Generalitat de Catalunya (RIS3CAT Emergents CECH: 001-P-001682 and VEIS: 001-P-001647), the Spanish Ministerio de Ciencia, Innovación y Universidades (PID2020-119535RB-I00), the Instituto de Salud Carlos III (IMPaCT-Data), and the European Commission (RiPCoN: 101003633). AF-T is a recipient of an FPI fellowship (BES-2017-083053). We also acknowledge institutional funding from the Spanish Ministry of Science and Innovation through the Centres of Excellence Severo Ochoa Award, and from the CERCA Programme / Generalitat de Catalunya.

References

1 Baker, M. Big biology: The 'omes puzzle. *Nature* **494**, 416-419, doi:10.1038/494416a (2013).

2 Cantelli, G. *et al.* The European Bioinformatics Institute (EMBL-EBI) in 2021. *Nucleic Acids Res* **50**, D11-D19, doi:10.1093/nar/gkab1127 (2022).

3 Rouillard, A. D., Wang, Z. & Ma'ayan, A. Reprint of "Abstraction for data integration: Fusing mammalian molecular, cellular and phenotype big datasets for better knowledge extraction". *Computational Biology and Chemistry* **59**, 123-138, doi:10.1016/j.compbiolchem.2015.08.005 (2015).

4 Rigden, D. J. & Fernandez, X. M. The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res* **49**, D1-D9, doi:10.1093/nar/gkaa1216 (2021).

5 Ma'ayan, A. *et al.* Lean Big Data integration in systems biology and systems pharmacology. *Trends Pharmacol. Sci.* **35**, 450-460, doi:10.1016/j.tips.2014.07.001 (2014).

6 Hoehndorf, R., Schofield, P. N. & Gkoutos, G. V. The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.* **16**, 1069-1080, doi:10.1093/bib/bbv011 (2015).

7 Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)* **2016**, doi:10.1093/database/baw100 (2016).

8 Kawata, K. *et al.* Trans-omic Analysis Reveals Selective Responses to Induced and Basal Insulin across Signaling, Transcriptional, and Metabolic Networks. *iScience* **7**, 212-229, doi:10.1016/j.isci.2018.07.022 (2018).

9 Vitrinel, B. *et al.* Exploiting Interdata Relationships in Next-generation Proteomics Analysis. *Mol Cell Proteomics* **18**, S5-S14, doi:10.1074/mcp.MR118.001246 (2019).

10 Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, e8124, doi:10.15252/msb.20178124 (2018).

11 Himmelstein, D. S. & Baranzini, S. E. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Comput. Biol.* **11**, e1004259, doi:10.1371/journal.pcbi.1004259 (2015).

12 Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* **6**, doi:10.7554/eLife.26726 (2017).

13 Santos, A. *et al.* A knowledge graph to interpret clinical proteomics data. *Nat Biotechnol*, doi:10.1038/s41587-021-01145-6 (2022).

14 Cai, H., Zheng, V. W. & Chang, K. C.-C. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. *arXiv*, doi:arXiv:1709.07604 (2017).

15 Li, M., Huang, K. & Zitnik, M. Representation Learning for Networks in Biology and Medicine: Advancements, Challenges, and Opportunities. *arXiv* **arXiv:2104.04883** (2021).

16 Zitnik, M. & Leskovec, J. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* **33**, i190-i198, doi:10.1093/bioinformatics/btx252 (2017).

17 Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nat Commun* **12**, 1796, doi:10.1038/s41467-021-21770-8 (2021).

18 Cantini, L. *et al.* Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat Commun* **12**, 124, doi:10.1038/s41467-020-20430-7 (2021).

19 Duran-Frigola, M., Fernández-Torras, A., Bertoni, M. & Aloy, P. Formatting biological big data for modern machine learning in drug discovery. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **9**, doi:10.1002/wcms.1408 (2019).

20 Fernandez-Torras, A., Comajuncosa-Creus, A., Duran-Frigola, M. & Aloy, P. Connecting chemistry and biology through molecular descriptors. *Curr Opin Chem Biol* **66**, 102090, doi:10.1016/j.cbpa.2021.09.001 (2021).

21 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991-D995, doi:10.1093/nar/gks1193 (2012).

22 Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e1417, doi:10.1016/j.cell.2017.10.049 (2017).

23 Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358-363, doi:10.1093/nar/gkt1115 (2014).

24 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607-D613, doi:10.1093/nar/gky1131 (2019).

25 Carvalho-Silva, D. *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* **47**, D1056-D1065, doi:10.1093/nar/gky1133 (2019).

26 Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845-D855, doi:10.1093/nar/gkz1021 (2020).

27 Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *KDD* **2016**, 855-864, doi:10.1145/2939672.2939754 (2016).

28 Vlietstra, W. J., Vos, R., Sijbers, A. M., van Mulligen, E. M. & Kors, J. A. Using predicate and provenance information from a knowledge graph for drug efficacy screening. *J. Biomed. Semantics* **9**, 23, doi:10.1186/s13326-018-0189-6 (2018).

29 Iskar, M. *et al.* Drug-Induced Regulation of Target Expression. *PLoS Computational Biology* **6**, e1000925, doi:10.1371/journal.pcbi.1000925 (2010).

30 Wu, G., Liu, J. & Yue, X. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC Bioinformatics* **20**, 134, doi:10.1186/s12859-019-2644-5 (2019).

31 Kose, F., Kocer, N. E., Sumbul, A. T., Sezer, A. & Yilkan, O. Kaposi's Sarcoma following Chronic Lymphocytic Leukemia: A Rare Entity. *Case Rep. Oncol.* **5**, 271-274, doi:10.1159/000339228 (2012).

32 Belur, A. A., Raajasekar, A. K. A., Nannapaneni, S. & Chelliah, T. A Case of Kaposi's Sarcoma in a HIV Negative Patient with CLL Treated with Rituximab. *Blood* **124**, 4970-4970, doi:10.1182/blood.v124.21.4970.4970 (2014).

33 Vučinić, D. *et al.* Kaposi's sarcoma in an HIV-negative chronic lymphocytic leukemia patient without immunosuppressive therapy: A case report. *SAGE Open Med Case Rep* **6**, 2050313X18799239, doi:10.1177/2050313X18799239 (2018).

34 Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966-967, doi:10.1038/nmeth.4077 (2016).

35 Taskesen, E. & Reinders, M. J. T. 2D Representation of Transcriptomes by t-SNE Exposes Relatedness between Human Tissues. *PLoS One* **11**, e0149853, doi:10.1371/journal.pone.0149853 (2016).

36 Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754, doi:10.1016/j.cell.2016.06.017 (2016).

37 Hawkins, D. M. The problem of overfitting. *J Chem Inf Comput Sci* **44**, 1-12, doi:10.1021/ci0342472 (2004).

38 Fernandez-Torras, A., Duran-Frigola, M. & Aloy, P. Encircling the regions of the pharmacogenomic landscape that determine drug response. *Genome Med* **11**, 17, doi:10.1186/s13073-019-0626-x (2019).

39 Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508, doi:10.1038/s41586-019-1186-3 (2019).

40 Costanzo, M. *et al.* Environmental robustness of the global yeast genetic interaction network. *Science* **372**, doi:10.1126/science.abf8424 (2021).

41 Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, doi:10.1126/science.aaz8528 (2020).

42 Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178, doi:10.1038/nature04209 (2005).

43 Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212-1226, doi:10.1016/j.cell.2014.10.050 (2014).

44 Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402-408, doi:10.1038/s41586-020-2188-x (2020).

45 Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56-67, doi:10.1038/s42256-019-0138-9 (2020).

46 Yoon, S. *et al.* Homer1 promotes dendritic spine growth through ankyrin-G and its loss reshapes the synaptic proteome. *Mol Psychiatry* **26**, 1775-1789, doi:10.1038/s41380-020-00991-1 (2021).

47 Paushkin, S. V., Patel, M., Furia, B. S., Peltz, S. W. & Trotta, C. R. Identification of a Human Endonuclease Complex Reveals a Link between tRNA Splicing and Pre-mRNA 3′ End Formation. *Cell* **117**, 311-321, doi:10.1016/s0092-8674(04)00342-3 (2004).

48 Chung, H. *et al.* Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown. *Cell* **172**, 811-824 e814, doi:10.1016/j.cell.2017.12.038 (2018).

49 Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* **18**, 41-58, doi:10.1038/nrd.2018.168 (2019).

50 Cai, C. *et al.* Transfer Learning for Drug Discovery. *J Med Chem* **63**, 8683-8694, doi:10.1021/acs.jmedchem.9b02147 (2020).

51 Ma, J. *et al.* Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat Cancer* **2**, 233-244, doi:10.1038/s43018-020-00169-2 (2021).

52 Douglass, E. F. A community challenge for a pancancer drug mechanisms of action inference from perturbational profile data. *Cell Reports Medicine* **3**, 100492, doi:doi.org/10.1016/j.xcrm.2021.100492 (2022).

53 Forslund, S. K. *et al.* Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature* **600**, 500-505, doi:10.1038/s41586-021-04177-9 (2021).

54 Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214-1219, doi:10.1093/nar/gkv1031 (2016).

55 Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *J. Biomol. Tech.* **29**, 25-38, doi:10.7171/jbt.18-2902-002 (2018).

56 The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338, doi:10.1093/nar/gky1055 (2019).

57 Bello, S. M. *et al.* Disease Ontology: improving and unifying disease annotations across species. *Dis. Model. Mech.* **11**, doi:10.1242/dmm.032839 (2018).

58 Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa977 (2020).

59 The UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506-D515, doi:10.1093/nar/gky1049 (2018).

60 Rigden, D. J. & Fernández, X. M. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. *Nucleic Acids Res.* **48**, D1-D8, doi:10.1093/nar/gkz1161 (2020).

61 Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* **39**, D507-513, doi:10.1093/nar/gkq968 (2011).

62 Seco, N., Veale, T. & Hayes, J. in *Proceedings of the 16th European Conference on Artificial Intelligence* 1089-1090 (IOS Press, 2004).

63 Domingo-Fernández, D., Hoyt, C. T., Bobis-Álvarez, C., Marín-Llaó, J. & Hofmann-Apitius, M. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst Biol Appl* **5**, 3, doi:10.1038/s41540-018-0078-8 (2019).

64 Belinky, F. *et al.* PathCards: multi-source consolidation of human biological pathways. *Database* **2015**, doi:10.1093/database/bav006 (2015).

65 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

66 Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536-542, doi:10.1038/s41586-021-04115-9 (2021).

67 Dong, Y., Chawla, N. V. & Swami, A. in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 135-144 (Association for Computing Machinery, 2017).

68 Belkina, A. C. *et al.* Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat Commun* **10**, 5415, doi:10.1038/s41467-019-13055-y (2019).

69 Duran-Frigola, M. *et al.* Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nat Biotechnol* **38**, 1087-1096, doi:10.1038/s41587-020-0502-7 (2020).

70 Bergstra, J., Yamins, D. & Cox, D. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. *Proceedings of the 12th Python in Science Conference*, doi:10.25080/majora-8b375195-003 (2013).

71 Johnson, J., Douze, M. & Jegou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 1-1, doi:10.1109/tbdata.2019.2921572 (2019).

72 Müllner, D. fastcluster: Fast Hierarchical, Agglomerative Clustering Routines forRandPython. *Journal of Statistical Software* **53**, doi:10.18637/jss.v053.i09 (2013).

73 Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074-D1082, doi:10.1093/nar/gkx1037 (2017).

74 Corsello, S. M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med* **23**, 405-408, doi:10.1038/nm.4306 (2017).

75 Smirnov, P. *et al.* PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res.* **46**, D994-D1002, doi:10.1093/nar/gkx911 (2017).

Figures and legends

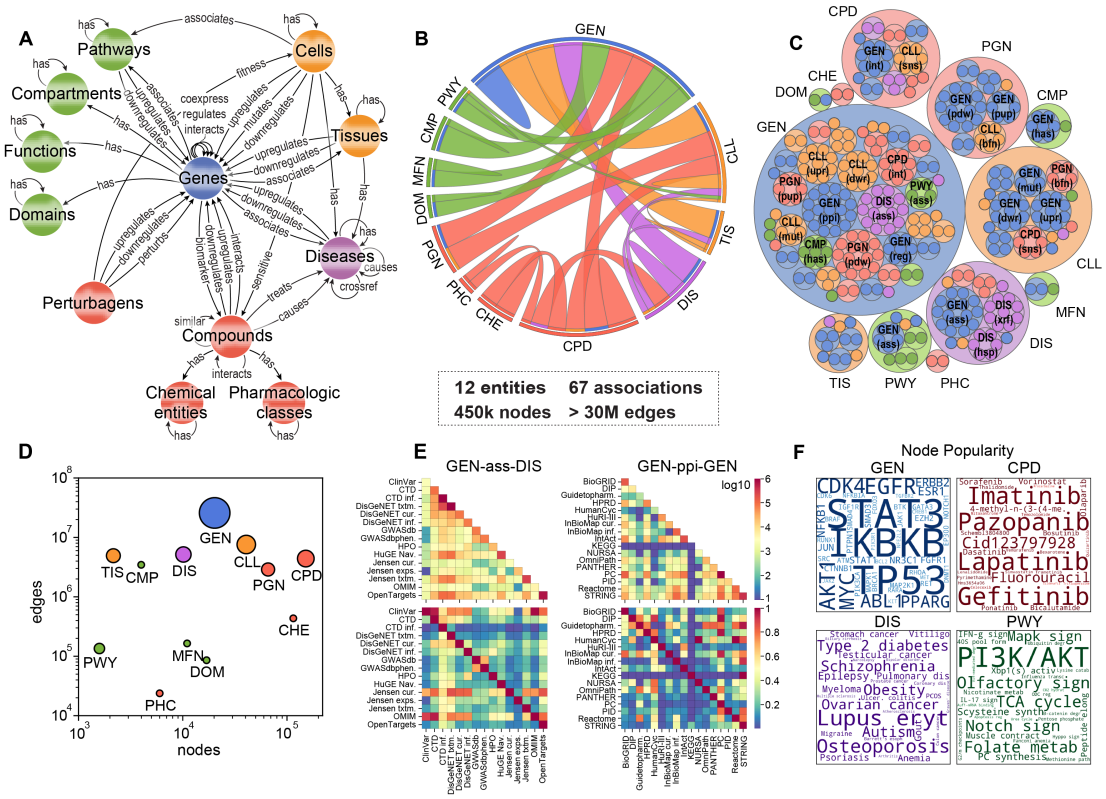

Figure 1. Building the Bioteque Knowledge Graph (KG). **A)** Metagraph of the Bioteque, showing all the entities and the most representative associations (metaedges) between them. **B)** Circos plot representation of the KG, showing the relationships between nodes. **C)** Treeplot showing the number of datasets used to construct each metaedge. **D)** Total number of nodes (x-axis) and edges (y-axis) available for each entity type. The size of the circles is proportional to the number of metaedges in which the entities participate. **E)** Number of edges (top row) and overlap (bottom row) between the datasets inside the *GEN-ass-DIS* (left) and *GEN-ppi-GEN* (right) associations. **F)** Most popular nodes in the KG within the gene (GEN, blue), compound (CPD, red), disease (DIS, purple) and pathway (PWY, green) universe.

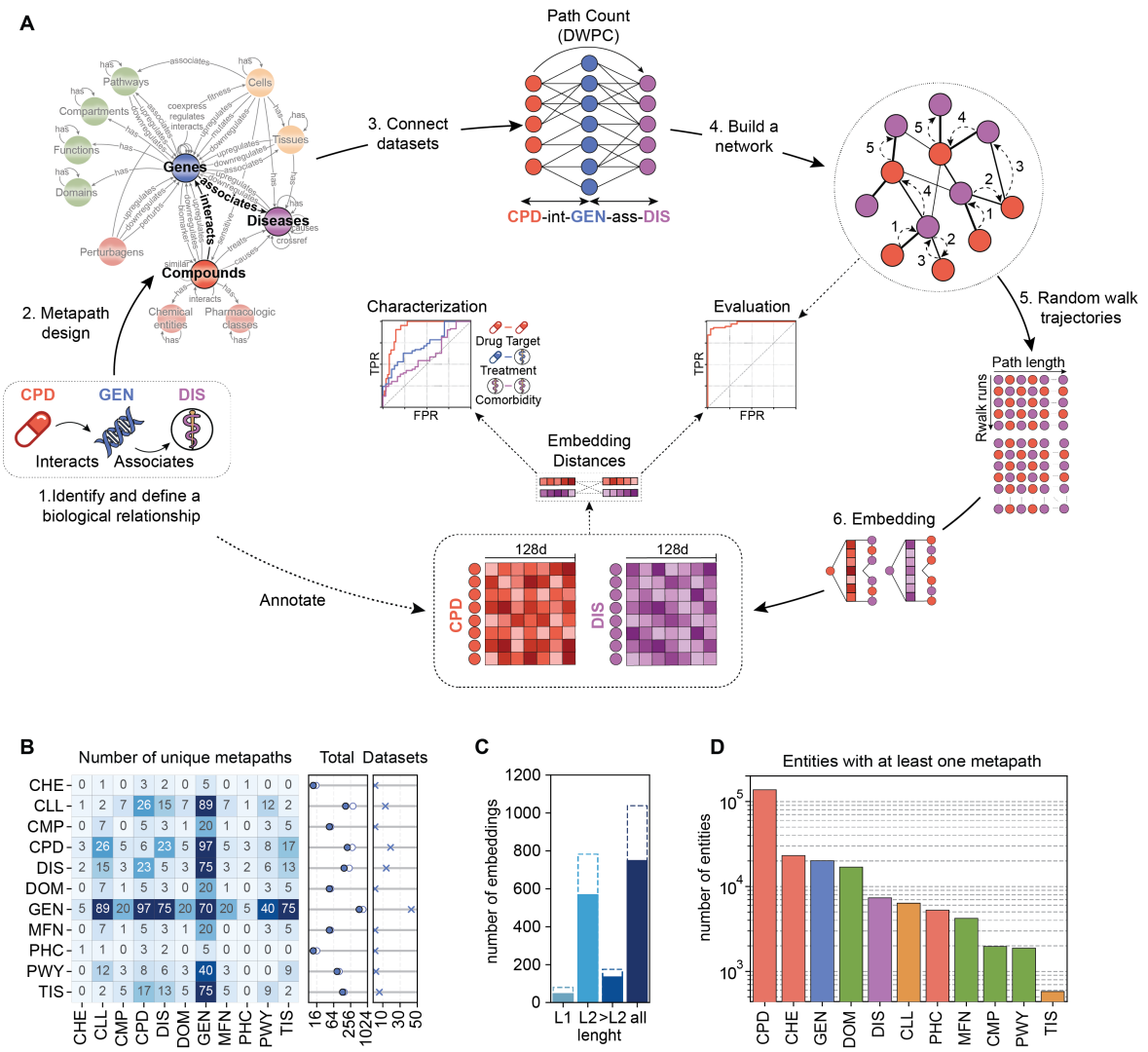

Figure 2. Generating the Bioteque embeddings. **A)** Scheme of the methodology. First, we define the biological entities to be connected and the specific context to be explored. Then a source-target network is derived by traversing all the paths available from the source to the target nodes of a given metapath. The vicinity of each node in the network is then explored by a random walker and embedded in a 128-dimensional space. Finally, embeddings are evaluated and characterized. **B)** Number of unique metapath embeddings linking each entity. In the middle plot, the filled dots indicate the total number of unique metapaths while the empty dots show the total number of metapath-dataset combinations. In the rightmost plot, we show the number of entity-specific datasets used in the metapaths. **C)** Number of metapath-dataset embedding combinations obtained at each metapath length. Solid bars highlight the number of unique metpaths. **D)** Number of nodes within each entity with at least one embedding in the Bioteque resource. Note that during metapath construction, perturbagen (PGN) entities are always mapped to the corresponding perturbed genes. Thus, although used to construct several metapaths, PGN nodes are not explicitly embedded, i.e., they are not the first or last nodes in the metapaths.

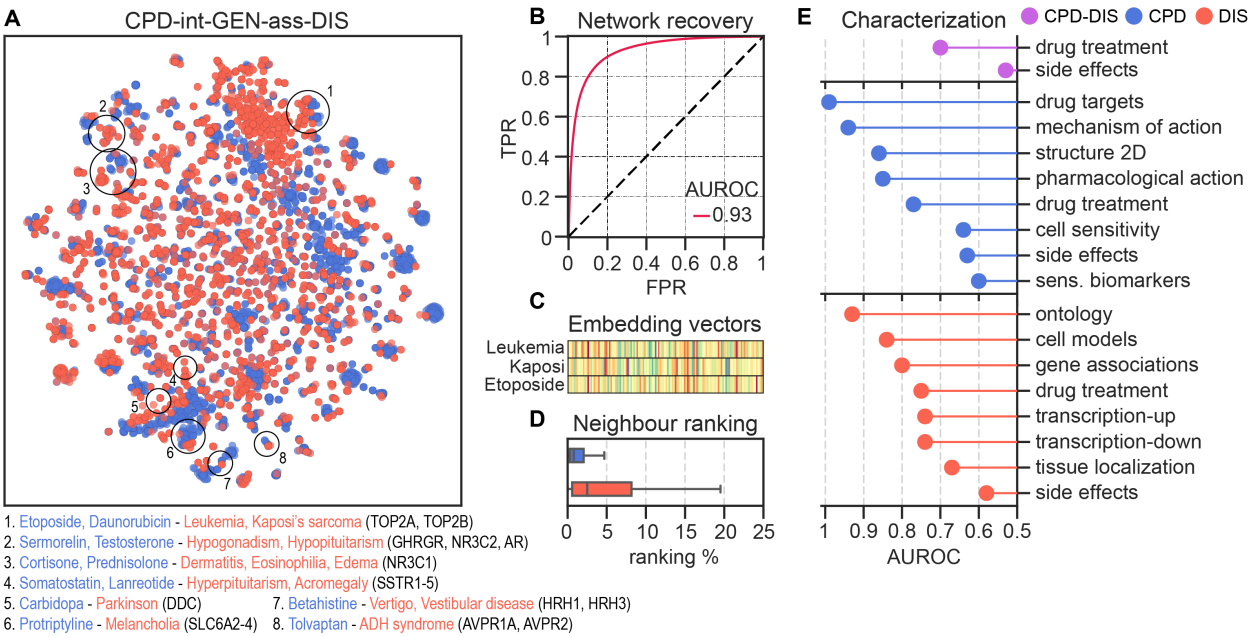

Figure 3. A Bioteque embedding summary card. **A)** 2D projection (opt-SNE) of the CPD (blue) and DIS (red) embeddings from the metapath CPD-int-GEN-ass-DIS. We highlight clusters of compounds and diseases sharing treatment evidence. We highlight some representative compounds and diseases found in these clusters, together with the drug targets associated with the diseases. **B)** ROC curve validation when reconstructing the original network with the corresponding embeddings. **C)** Visual representation of the embedding vectors of Leukemia (top) and Kaposi's sarcoma (middle), together with the drug Etoposide (bottom). **D)** Ranking proportion in which the putative CPD (blue) and DIS (red) neighbours are found. **E)** Recapitulation of orthogonal associations by using embedding distances. The AUROC (x-axis) summarizes the performance obtained when ranking the orthogonal associations. Drug targets are collected from Drugbank[73], the Drug Repurposing Hub[74] and PharmacoDB[75], and gene-disease associations are obtained from OpenTargets[25].

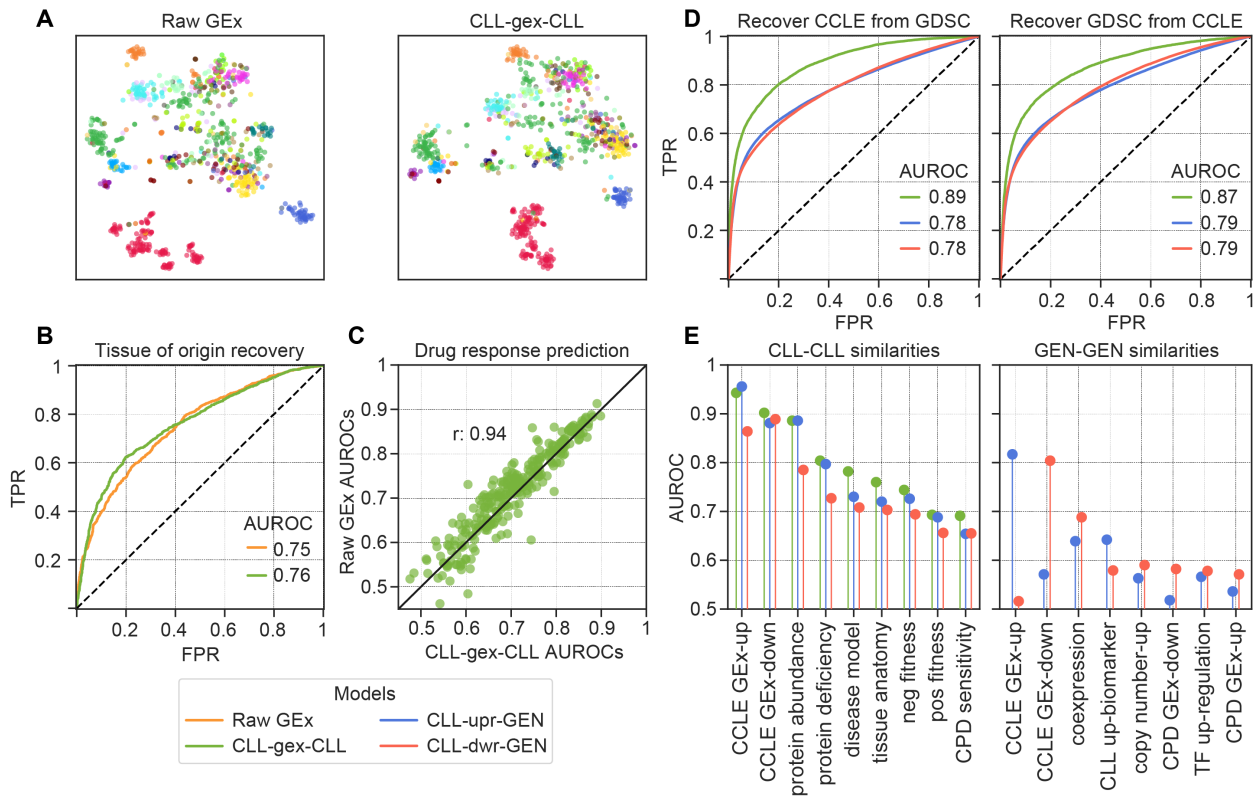

Figure 4. Analysis of gene expression (GEx) embeddings. **A)** 2D projection of the raw GEx (left) and the corresponding Bioteque CLL-gex-CLL embedding (right). Each dot corresponds to one cell line and is coloured by tissue of origin. **B)** Tissue recovery by the raw GEx and the CLL-gex-CLL embedding. **C)** Drug response prediction performance (AUROC) for each drug in the GDSC from models trained with either the raw GEx (y-axis) or the CLL-gex-CLL embeddings (x-axis). **D)** Recovering CCLE (left) and GDSC (right) CLL-CLL similarities (green), CLL-GEN upregulation similarities (blue) and CLL-GEN downregulation similarities (red) using embedding distances from the GDSC and the CCLE embedding spaces, respectively. **E)** Characterization of the CLL-CLL (left) and GEN-GEN (right) embedding similarities for three metapaths: CLL-gex-CLL (green), CLL-upr-GEN (blue) and CLL-dwr-GEN (red).

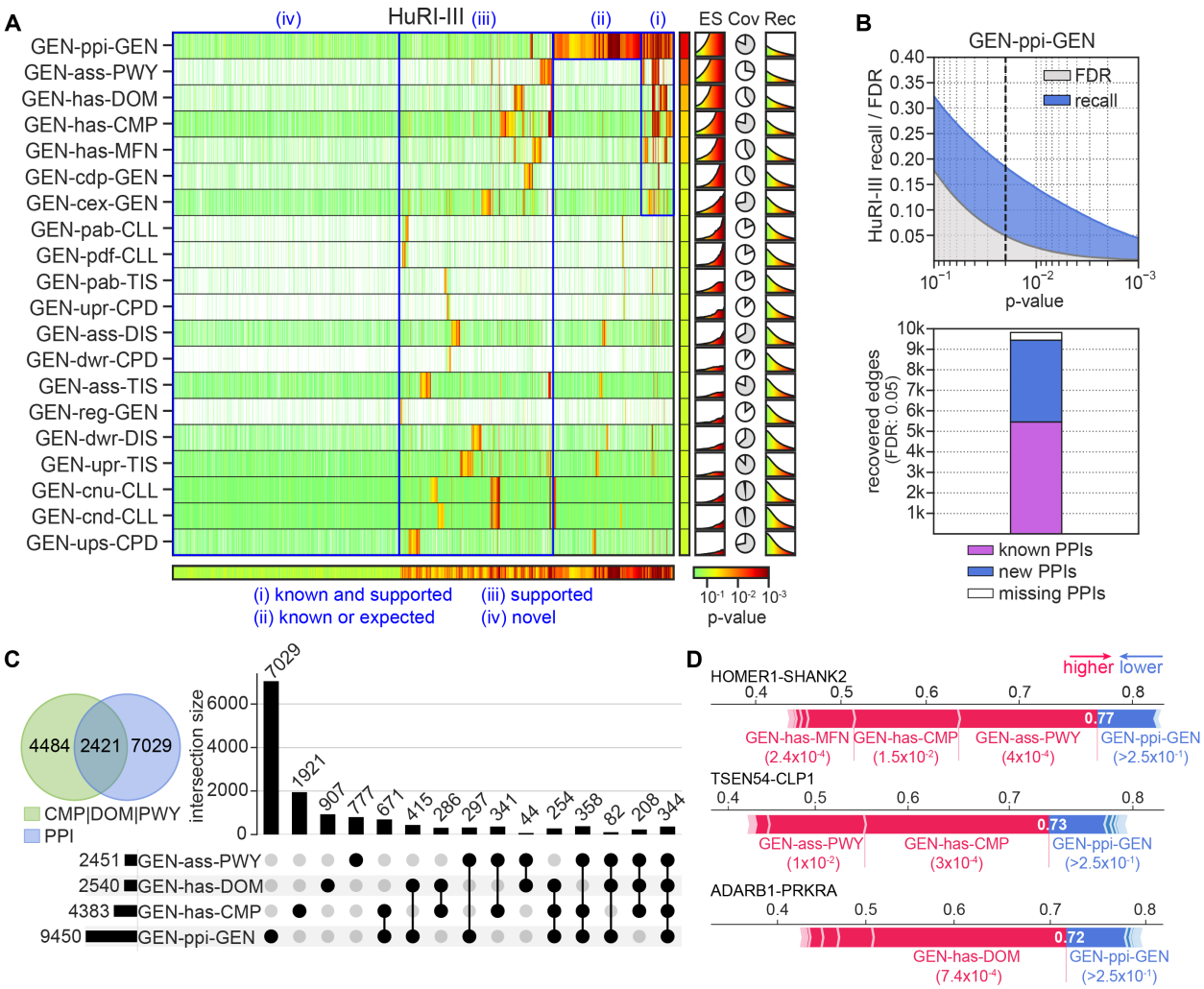

Figure 5. Assessing the novelty of the HuRI-III interactome. **A)** Embedding distance P-values (*Methods*) are calculated for each PPI in HuRI-III (x-axis) using the corresponding GEN-GEN embeddings from a subset of metapaths (y-axis). Red tones (corresponding to lower P-values) indicate similarity according to a given embedding space. The column and row next to the heatmap show the 10th percentile of the P-value distribution for each metapath and the lowest P-value for each edge, respectively. In blue, we grouped edges according to four levels of support. In the ES column, we derived an enrichment score based on the fold change (FC) between the recovered HuRI-III edges and the expected recovery from a set of random networks (*Methods*). The y-axis shows the FC (between 1 and 5) while the x-axis shows the range of measured P-values. In the Cov column, we provide the coverage by each of the metapaths. Finally, in the Rec column, we show the cumulative recall as the P-value decreases. **B)** (Top) Recall (blue) and FDR (grey) of the recovery of HuRI-III edges (y-axis) by GEN-ppi-GEN embeddings across the P-values (x-axis). The dashed line is placed at the 0.05 FDR (corresponding to a P-value of 0.02). (Bottom) Number of HuRI-III interactions recovered by the GEN-ppi-GEN embedding at 0.05 FDR. Colours show the known interactions (i.e., covered in the original network, coloured in purple), the new interactions (not available in the network, hence, predicted by the embeddings, coloured in blue) and the interactions present in the original network but not covered at the given

P-value (not coloured). **C)** Number of unique HuRI-III edges recovered at 0.05 FDR by the GEN-ppi-GEN and/or the three most supportive metapaths, including cellular components (CMP), protein domains (DOM) and pathway similarities (PWY). **D)** Shapley force plots corresponding to the prediction of three PPIs with no direct evidence of physical interaction before HuRI-III was released. Red segments are metapath-specific P-values that 'pushed' predictions toward a high probability of interactions, while blue segments 'pulled' predictions towards a low probability. Length of the segments is proportional to their impact on the prediction. The final output probability given by the model is found where both forces equalize (shown in white).

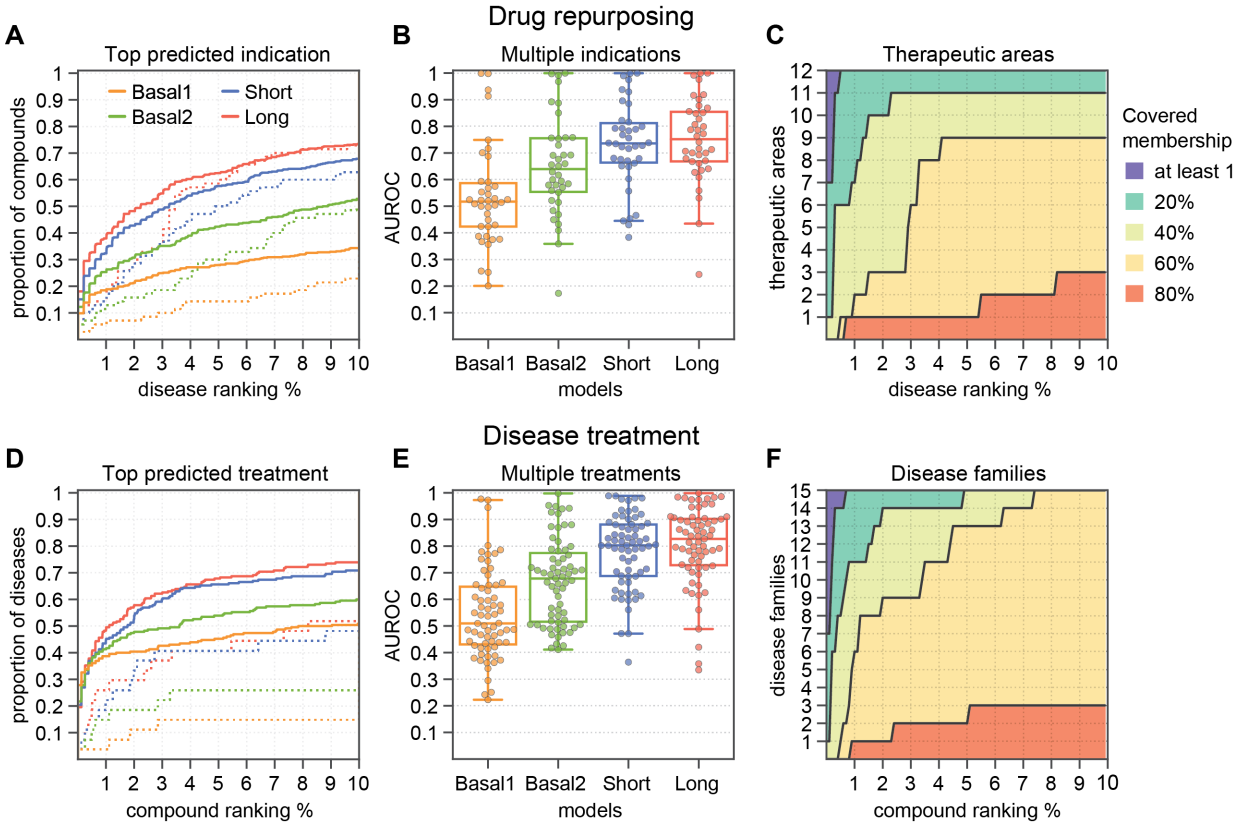

Figure 6. Prediction of drug indications and disease treatments from repoDB. **A)** Cumulative distribution (y-axis) of compounds (top) and diseases (bottom) according to the ranked position (x-axis) of the top predicted disease indication (top) or compound treatment (bottom) for the four tested models. The rankings are shown in percentages and only for the first 10% of compound/disease predictions (corresponding to the top 50 and 80 diseases and compounds, respectively). Dotted lines show the distribution for those compounds or diseases with only one positive indication in repoDB v1. **B)** Classification performance obtained for those compounds (top) and diseases (bottom) with multiple (≥5) new indications reported in repoDB v2. **C)** Number of different therapeutic areas (top) and disease families (bottom) covered by the predictions of the *Long* model. We considered a given therapeutic area or disease family to be covered when the model predicted one true indication or treatment (as in panel A) for at least 1%, 20%, 40%, 60%, or 80% of its instances.

Supplementary Figures and Tables

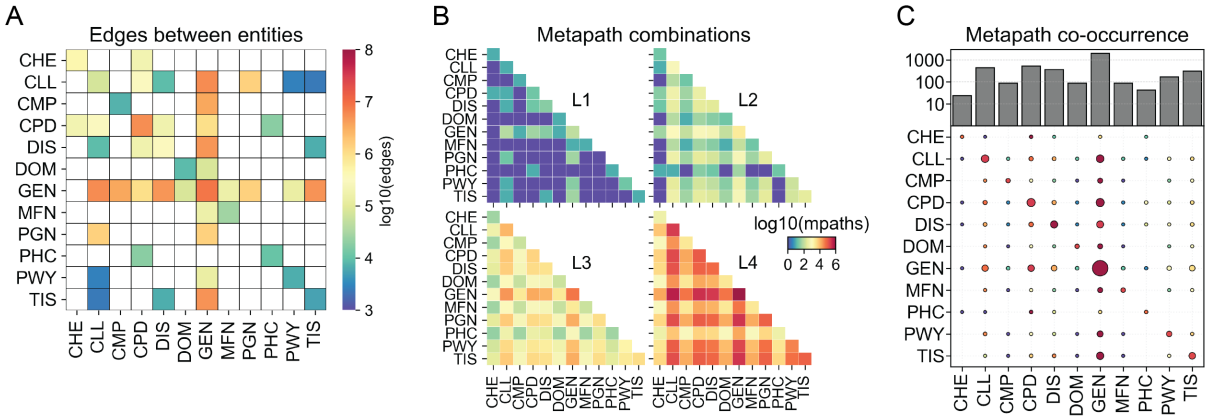

Figure S1. A) Number of possible edges between entities considering all associations and datasets available in the graph. **B)** Theoretical number of metapaths of length 1, 2, 3, and 4. **C)** Number of times that every entity (x-axis) co-occurs in a metapath with another entity (y-axis). The color scale illustrates the most used entities in each row, red and blue being the most and less used entities, respectively. The size is proportional to the number of times a given entity participates in a metapath.

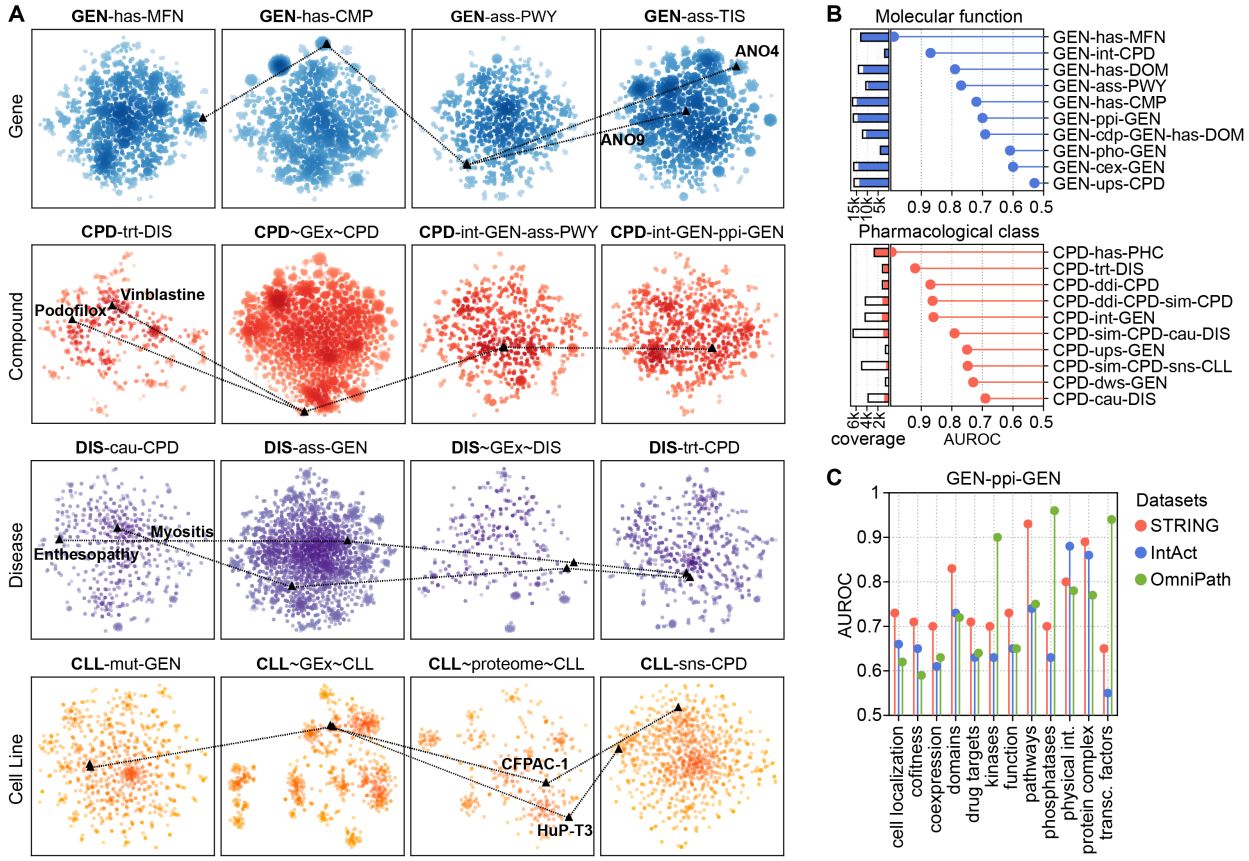

Figure S2. Comparison of embeddings built from different metapaths and datasets. **A)** Four illustrative examples showing pairs of genes, compounds, diseases and cell lines with similarities or differences depending on the metapaths. **B)** Top metapaths (y-axis) recapitulating (AUROC, x-axis) gene molecular function (blue) and compound pharmacological class (red). The coloured bars indicate the proportion of nodes in the metapath that could be assessed (i.e., with annotated molecular function or pharmacological classes). **C)** Gene embedding characterization of three reference PPI datasets, namely STRING, IntAct and OmniPath. We limited the analysis to the common gene universe (9,395 genes) between the three sources.

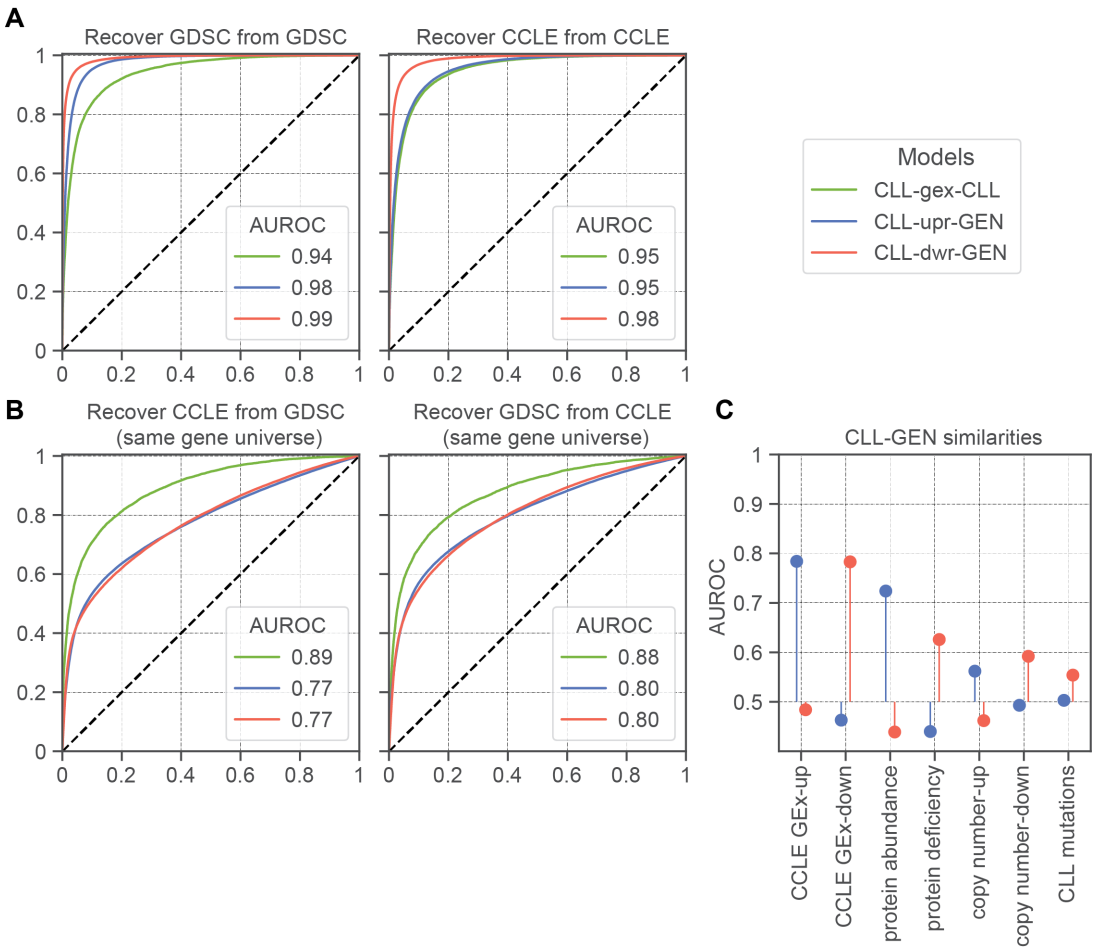

Figure S3. **A)** Recovery of the original GDSC (left) and CCLE (right) network by their corresponding Bioteque embeddings. **B)** Recovery of the CCLE (left) and GDSC (right) panels using the GDSC embeddings and CCLE embeddings, respectively. In contrast to Figure 5D, embeddings were obtained from a GDSC and CCLE version in which we kept only those genes in common between both panels. **C)** Characterization of the CLL-GEN similarities for the *CLL-upr-GEN* and *CLL-dwr-GEN* metapath embeddings.

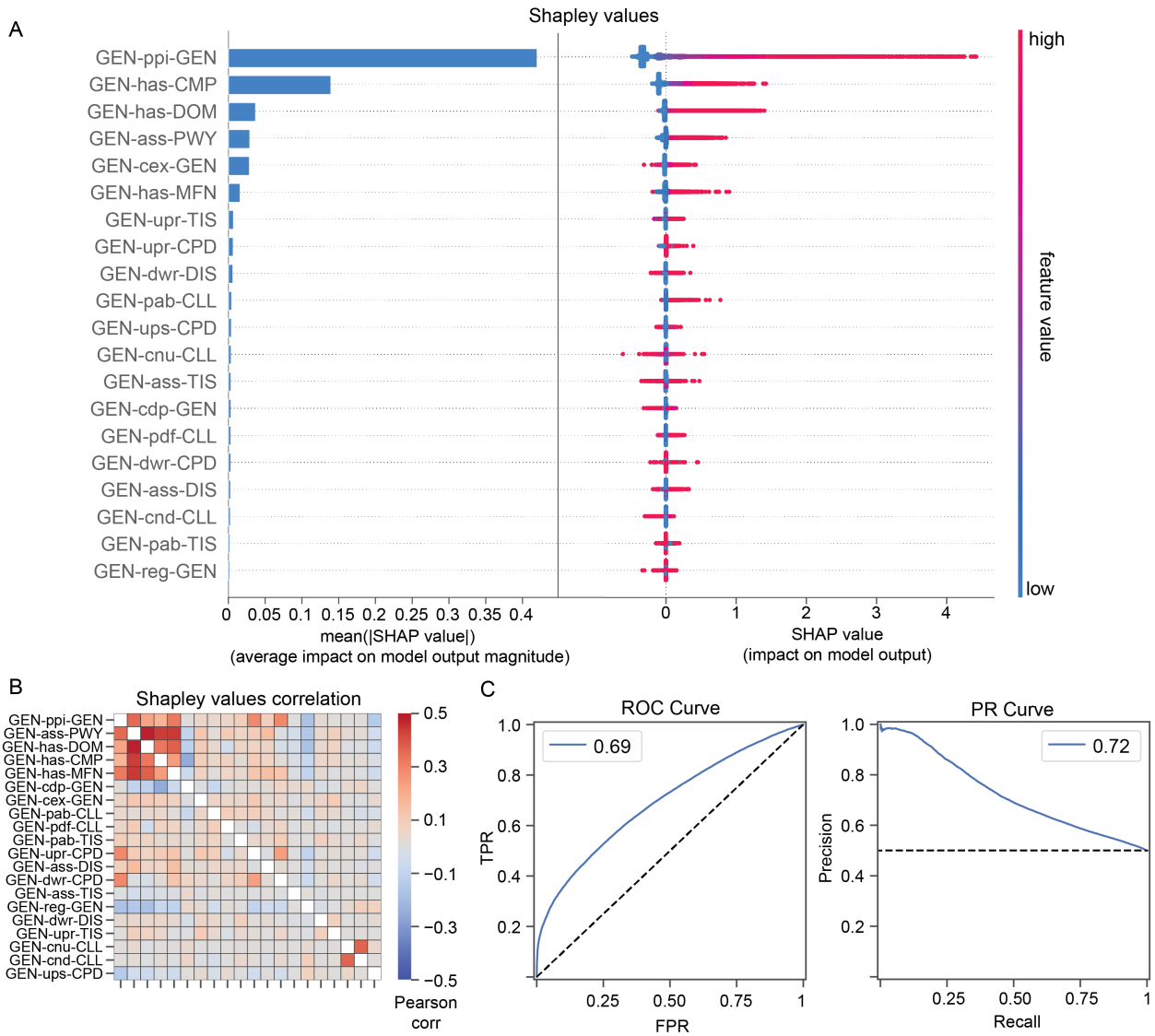

Figure S4. A) Feature importance measured as Shapley values (x-axis) for each metapath (y-axis) when predicting HuRI-III PPIs. The higher the Shapley value the higher the impact when predicting the correct class. (Left) Mean absolute Shapley value for each metapath. (Right) Individual Shapley values for each prediction (i.e., each dot corresponds to a predicted PPI). In the colour scale, red and blue indicate lower and higher P-values for the corresponding metapath, respectively. **B)** Pairwise Pearson correlation matrix of the Shapley value vectors for each metapath. The higher the correlation, the higher the agreement between metapaths when classifying a PPI. **C**) ROC (left) and Precision-Recall (right) curves obtained from the model. A precision of 0.5 is expected at recall 1 as negative pairs were down-sampled to achieve a 1:1 balance with positive pairs.

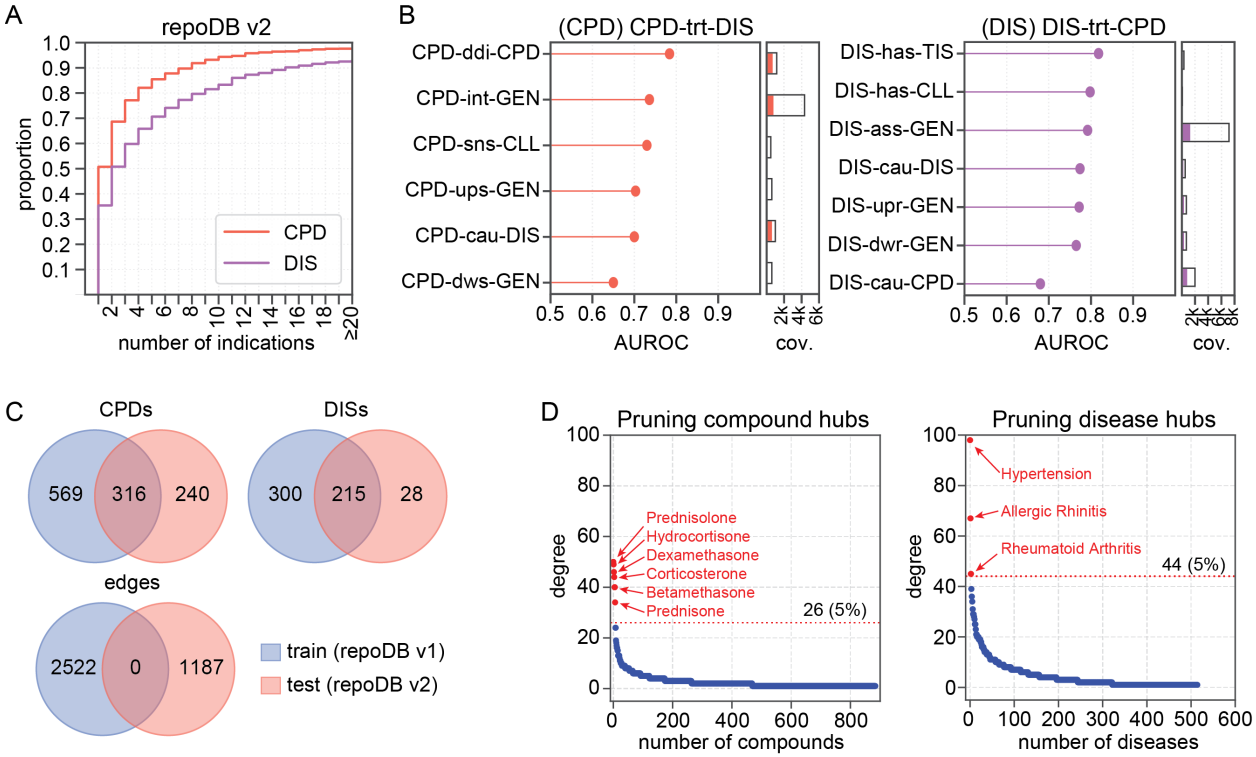

Figure S5. A) Cumulative distribution of the number of treatment indications reported by repoDB v2 for the compounds (CPD, red) and diseases (DIS, purple). **B)** Top L1 metapaths recapitulating CPD-CPD (left) and DIS-DIS (right) treatment similarities. This was measured by assessing (AUROC) how the different embeddings up-ranked CPD-CPD (or DIS-DIS) pairs associated with the same treatment. The bar plots show the number of nodes available by each metapath, colouring those that were covered by the CPD-trt-DIS network**. C)** Number of unique compounds (CPDs), diseases (DISs) and edges in the train (repoDB v1) and test (repoDB v2) splits after mapping the entities to the CPD-int-GEN and DIS-ass-GEN embedding universes. **D)** Compounds (left) and diseases (right) of the train split were sorted according to their node degree in the repoDB CPD-trt-DIS network. The red line shows the degree corresponding to 5% of total possible associations. We highlight those compounds and diseases whose degree exceeded this limit and were, therefore, pruned.

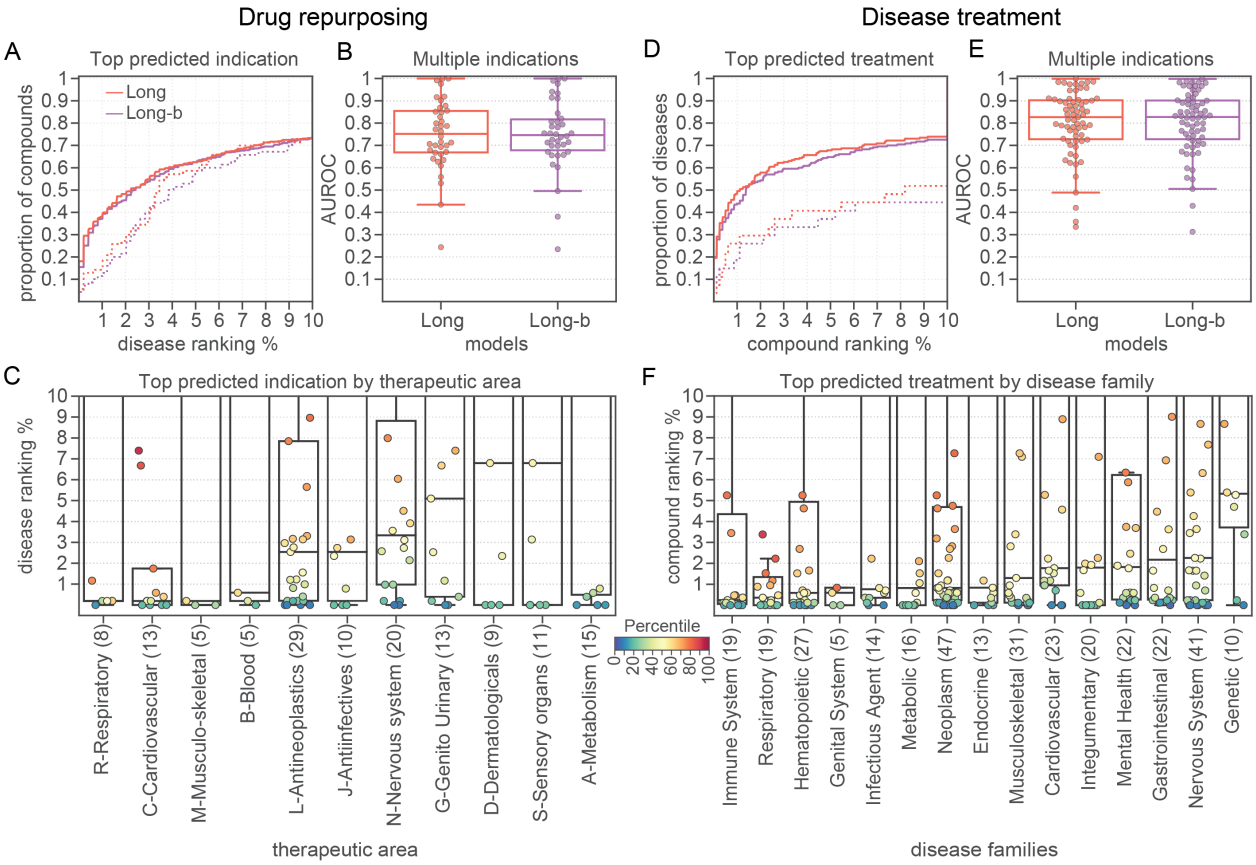

Figure S6. A, D) Cumulative distribution (y-axis) of compounds (A) and diseases (D) according to the ranked position (x-axis) of the best predicted disease indication (A) or compound treatment (A') for the *Long* and *Long-b* models. The rankings are shown in percentages and only for the first 10% of compounds or disease predictions. Dotted line shows the distribution for those compounds or diseases with only one positive indication in repodDB (v1). **B, E)** AUROC performance obtained for those compounds (B) and diseases (E) with multiple (>=5) new indications reported in repoDB (v2). **C, F)** We categorized the compounds (C) and diseases (F) according to their therapeutic area and disease family (x-axis) and showed the ranking of the best predicted indication and treatment (y-axis), respectively. Each dot corresponds to either a drug or a disease. Since the ranking was cut at the closest 10% of predictions, we coloured each drug or disease by the percentile it represents in the population of the corresponding group (i.e., note that yellow dots (~50th percentile) are placed at the median of the boxplots).

Table S1. Biological and Chemical entities in the Knowledge Graph (KG). We show the number of nodes, metaedges and edges contained in the KG for each metanode, as well as some examples of metaedges.

Table S2. Description of the associations between nodes (i.e., metaedges) and different datasets extracted from the knowledge graph (KG).

Table S3. Nodes universe considered in the resource for each of the embedded entities. The node name, together with some metadata, is provided, along with the identifier. The 'pruned general terms' tab contains those entities that were removed from ontologies.

Table S4. List of all the metapath embeddings available in the Bioteque resource.