Dimensional reduction of phenotypes from 53,000 mouse models reveals a diverse landscape of gene function

Tomasz Konopka[1,*], Letizia Vestito[1,2,3], Damian Smedley[1]

[1] William Harvey Research Institute, Queen Mary University of London, UK

[2] UCL Ear Institute, University College London, UK

[3] UCL Great Ormond Street Institute of Child Health, University College London, UK

* correspondence: t.konopka@qmul.ac.uk

Abstract

Animal models have long been used to study gene function and the impact of genetic mutations on phenotype. Through the research efforts of thousands of research groups, systematic curation of published literature, and high-throughput phenotyping screens, the collective body of knowledge for the mouse now covers the majority of protein-coding genes. We here collected data for over 53,000 mouse models with mutations in over 15,000 genomic markers and characterized by more than 254,000 annotations using more than 9,000 distinct ontology terms. We investigated dimensional reduction and embedding techniques as means to facilitate access to this diverse and high-dimensional information. Our analyses provide the first visual maps of the landscape of mouse phenotypic diversity. We also summarize some of the difficulties in producing and interpreting embeddings of sparse phenotypic data. In particular, we show that data preprocessing, filtering, and encoding have as much impact on the final embeddings as the process of dimensional reduction. Nonetheless, techniques developed in the context of dimensional reduction create opportunities for explorative analysis of this large pool of public data, including for searching for mouse models suited to study human diseases.

Introduction

Measuring the consequences of genetic mutations on organism-level phenotype is instrumental for describing gene function. It is a laborious process that requires breeding animals with controlled genotypes, performing a variety of assays, and describing phenotypes in a systematic fashion. For the mouse as a model organism, the collective knowledge of genotype-phenotype associations now covers around 15,000 genes (Blake et al., 2021). At the current pace of research, it may approach genome-wide coverage within a few years. A comprehensive phenomics dataset would impact many applications, for example, supporting efforts to identify the causes of rare genetic diseases (Justice & Dhillon, 2016; Meehan et al., 2017). However, there are currently few established methods to analyze phenomic data at scale, both for interactive exploration and for machine learning. Given recent advances in dimensional reduction, this promising approach may bring insight to mouse phenotype data and facilitate its integration with other omic datasets.

Phenotype data consist of links between the genetic characteristics of an animal model and sets of observations. For the mouse, the latter are usually tracked using the mammalian phenotype (MP) ontology (Smith & Eppig, 2015). The ontology is a collection of more than 13,000 concepts, also called MP terms, that are related through a hierarchy. For example, a phenotype describing ‘increased heart weight’ is a more precise annotation for the phenotype of ‘abnormal heart weight’, which in turn is a specific type of ‘cardiovascular system phenotype’. Individual terms in the ontology are thus not independent. However, despite the hierarchical connections, the space of possible phenotypic abnormalities is of high dimension. It covers all organ systems, fertility, and other factors, and animal models can be described by any combination of ontology terms.

High-dimensional data poses challenges both for explorative analysis and for machine learning (ML). Explorative analysis intends to place preliminary findings in context and to direct in-depth studies. It is often a manual process that relies on visualizations. Projection of large datasets into two dimensions is thus a common technique for this purpose. Machine-learning approaches are, in principle, not hindered by unintuitive data. However, training ML models requires fixing values for free parameters. When data is of high-dimension, ML models must estimate many more free parameters than with low-dimensional inputs. Using dimensional reduction may improve training efficiency, especially when there is limited data available.

A canonical approach to dimensional reduction of tabular data uses principal component analysis. It consists of a rotation - a linear transformation - followed by truncation to a fixed number of principal components, or dimensions, that incorporate the most variability. However, nonlinear embedding techniques can capture more relationships among the original data into the same number of dimensions. Examples include neural network auto-encoders (Hinton & Salakhutdinov, 2006), t-SNE (van der Maaten & Hinton, 2008), UMAP (Becht et al., 2019), EmbedSOM (Kratochvíl et al., 2019), PHATE (Moon et al. 2019), and Poincare maps (Klimovskaia et al. 2020). Approaches have also been implemented for non-tabular data. For example, node2vec provides embeddings of graphs (Grover & Leskovec, 2016). Importantly, recent implementations are computationally efficient and can process many thousands of data items. These methods have been instrumental in exploring atlases of transcriptomic data (The Tabula Muris Consortium, 2018; Han, et al., 2018; Rosenberg et al., 2018; Saunders et al., 2018; Cao et al., 2019; Kalucka et al., 2020). They also open possibilities to analyze the phenotypic landscape of all animal models.

Because ontology terms carry hierarchical relationships, dedicated methods have been proposed to embed the terms into low-dimensional spaces: Onto2vec (Smaili, et al., 2018), Opa2vec (Smaili et al., 2019), HiG2Vec (Kim et al., 2020), and Owl2vec (Chen et al., 2020). These approaches have also explored describing sets of ontology terms, particularly from the gene ontology. A common approach to embedding sets of ontology terms is through composition, or averaging, of the coordinates of the individual terms. This is motivated by observations in the context of word-based embeddings (Mikolov et al., 2013). It presents encouraging results also in the biological context, for example for the prediction of gene-gene interactions (Duong et al., 2020). However, linear composition conflicts with the nonlinearity inherent in dimensional reduction and is bound to lose effectiveness for sets with many terms. Moreover, ontology-based dimensional reduction techniques have not been explored in the context of animal phenotyping.

In this work, we examine a public dataset of mouse phenotypes and visualize the holistic landscape of mouse phenotypic variation. We contrast dimensional reduction approaches that are promising to some that are sub-optimal. In particular, we show that animal models can appear distorted when seen through embeddings of ontologies. Direct embeddings of biological entities such as mouse strains or human diseases provide less biased insights. Yet, because mouse phenotype data is sparse, results can be sensitive to incomplete phenotyping and pre-processing steps.

Results

Embeddings of ontologies offer insight for small annotation sets

Ontology structures are relatively stable and evolve only slowly in time. Furthermore, the mammalian phenotype (MP) ontology that catalogs possible phenotypic aberrations is agnostic to animal species. It is thus natural to consider dimensional reduction of the space spanned by MP terms as a basis for phenotype data. We produced embeddings of MP terms in two dimensions using two distinct approaches (Methods). In the first approach, we computed semantic similarities between text descriptions of MP terms and then created an embedding using uniform manifold approximation and projection (UMAP) (Becht et al., 2019). For comparison, we also constructed a graph of the hierarchical relations between MP terms and then created an embedding using node2vec (Grover & Leskovec, 2016). Of the two, the text-based approach placed the ontology root near the center, grouped ontology terms into small clusters, and produced a visual pattern that conveyed the multi-dimensionality of the ontology (Figure 1A). In contrast, the graph-based approach produced a monotonous layout that hid distinctions between phenotype domains (Figure S1).

We continued to investigate animal models and diseases via the text-based embedding of MP terms. We assembled annotations about animal models from the Mouse Genome Database (MGD) (Blake et al., 2021) and the International Mouse Phenotyping Consortium (IMPC) (Dickinson et al., 2016). This resulted in a collection of 53,629 models that describe mouse strains with mutations in 15,729 distinct genomic markers (Methods). Most markers correspond to protein-coding genes or non-coding genes, but also include other genomic constructs; we treated all on an equal level. The models carried 254,623 annotations to 9,907 different MP terms. We computed the position of these models in the embedding space by averaging the coordinates associated with their phenotypes. These projections appeared throughout the embedding space (Figure 1B), conveying that the animal models have diverse phenotypic features. Next, we investigated the distribution of the number of phenotypes per model. The distribution also had a strong skew. Although some models were associated with more than a hundred MP terms, 69% were associated with fewer than five, and 86% with fewer than ten (Figure 1C). The number of annotations created a bias in the position of the models within the embedding: the distance of the model from the center was anti-correlated to the number of

phenotypes (Figure 1D). Thus, out of the two coordinates in the visualization, one conveyed in part the number of annotations rather than the biological meaning of those annotations.

We repeated the projection calculation for phenotype profiles of human diseases (Methods). We translated disease phenotypes recorded using the HP ontology into the MP ontology, and then projected the translations into the embedding space. The profiles accumulated near the center (Figure 1E). Diseases were, on average, linked with more phenotypes than mouse models (Figure 1F). As a result, the correlation between the number of phenotypes and the distance from the center was more marked than for animal models (Figure 1G). This bias creates an impression that diseases with rich annotations are more similar to one another than diseases with few annotations, which is not justified *a-priori* from a phenotypic perspective.

While projecting phenotype sets into the ontology embedding through coordinate averaging may produce insight for small phenotype sets, our results demonstrate that this approach creates bias that becomes stronger with the number of annotations. As the bias is related to averaging, it is bound to appear with other embeddings of MP terms as well, for example those generated based on the ontology hierarchy graph. Furthermore, because phenotypic annotations are expected to become more detailed with time, such bias should be expected to grow as well. Thus, embeddings of ontology terms in low dimensions are problematic as tools for the visual data exploration of phenotype profiles.

Embeddings of full phenotype profiles capture the diversity of animal phenotypes

As an alternative to treating mouse models as sets of phenotypes and averaging coordinates for ontology terms, we produced embeddings for the mouse phenotype profiles directly. There are several possible approaches to encode phenotype profiles into a numerical form that can be processed with dimensional reduction algorithms (Figure S2).

In a first attempt, we constructed vector representations for individual mouse models using a previously described procedure (Konopka & Smedley, 2020). This procedure used prior probabilities for all phenotypes and applied phenotype inference based on the ontology hierarchy, producing vectors with non-binary values. We applied UMAP for dimensional reduction and the resultant embedding summarized the 53,629 mouse models from MGD and IMPC into a complex layout with some large clusters and many small clusters (Figure 2A). The

embedding was not dominated by technical variables such as data source, number of phenotypes, mouse genetic background, or zygosity of genetic mutations (Figure S3). It thus provides a concise visualization of the phenotypic diversity among mouse models.

For comparison, we produced embeddings with alternative procedures: using binary phenotype vectors, text-based semantic similarities, and a graph-based approach (Methods). These approaches all used more coarse-grained representations of phenotypes than our real-valued vectors. They resulted in embeddings that were qualitatively different (Figure S2). An embedding obtained using text-base semantic similarities also showed clusters of various sizes (Figure 2B). However, relative distances between mouse models differed: models that appeared nearby in the first embedding were far-apart in the second, and vice versa (Figure 2B). Such disparities are not unexpected given the ambiguities in turning phenotype annotations into a numerical representation. The text-based embedding also showed a more prominent separation of models according to MGD or IMPC data source. This may arise because IMPC data, which originate from a systematic screen rather than from bespoke experiments, have a more limited range of MP terms than MGD models.

Each embedding is rich in interpretable information. At a fine-grained level, each data point corresponds to a specific phenotype profile (Figure 2C) (Lyon et al., 1996; Hayes et al., 1998; Schreyer et al., 2001; Grigelioniene et al., 2019). Furthermore, each embedding is interpretable at a regional level. To demonstrate regional patterns, we set a “gating” in our first embedding and zoomed in on that region (Figure 2D). Enrichment analysis revealed an over-representation in certain phenotypes among the models within the gate (Figure 2E). Such enrichment is not surprising as the embedding was constructed so that similar phenotypes appear together. Indeed, other examples also exhibited enrichment (Figure S4), validating the embedding as well as providing a mechanism to assign regions with phenotypic interpretations.

These experiments demonstrate that our calculations produced more than one reasonable low-dimensional embedding of mouse models. Although it is unclear how to choose a single embedding as a reference map for mouse phenomics, these candidates can be used for data exploration.

Neighborhoods provide insight for phenotype prediction

Having demonstrated that embeddings summarize the diversity among mouse models and reveal qualitative patterns, we next investigated whether they can provide quantitative insight for

individual models. To this end, we computed predictions of phenotype profiles based on nearest neighbors (Figure 3A). For each model, we extracted a ranked list of nearest neighbors in the high-dimensional representations and in low-dimensional embeddings. We averaged the vector representations of the neighbors to create a prediction, and evaluated the error between the prediction and the model's true representation (Methods). We then investigated the mean prediction error as a function of the number of neighbors, denoted as k (Figure 3B). This approach is used in self-supervised learning to calibrate free parameters without a ground-truth (Batson & Royer, 2019). For predictions based on neighbors evaluated from the original data, i.e. from high-dimensional vectors, the optimal k was k=2. For approaches based on embeddings, the optimal number of neighbors varied with the embedding dimension, but all were below k=10. As expected, prediction errors were lowest when the neighbors were computed from the high-dimensional data, and highest when the neighbors originated from two-dimensional embeddings used for visualization.

To further investigate the factors that can affect prediction of phenotypes, we repeated these calculations with a series of approaches. We computed neighbors using high-dimensional representations based on non-binary phenotype vectors, binary phenotype vectors, and text descriptions. We considered dimensional reduction via UMAP, principal component analysis (PCA), and coordinate averaging of MP terms. For each combination of data encoding and dimensional reduction technique, we calibrated the optimal number of neighbors and reported the mean prediction error (Figure 3C). All errors were computed with respect to the non-binary vector representations. Given this choice, it is unsurprising that the smallest prediction error was achieved by the approach that used neighbors from high-dimensional non-binary vector data. However, this choice is useful because it ensures that the scales for all errors are comparable.

Among approaches that used dimensional reduction, we observed an improvement (reduction) in prediction error with increasing embedding dimension, d, for all dimensional reduction techniques (UMAP, PCA, MP coordinates). Interestingly, the improvement plateaued quickly with UMAP; there was little change between d=8 and d=10 dimensions. In contrast, prediction errors from PCA showed a steadier improvement with d. However, errors from PCA embeddings were higher than for UMAP, and remained higher for PCA with d=10 than for UMAP with d=2.

The encoding scheme (non-binary phenotype vectors, binary phenotype vectors, text descriptions) had a substantial effect on prediction errors. Variability between encodings was

large compared to differences between UMAP in various dimensions. This indicates that studies of mouse models are bound to be more influenced by how phenotype data are curated and encoded than by any loss of precision due to dimensional reduction. At the same time, there was relatively low discrepancy in prediction errors between the binary vector-based encoding and text-based approaches.

Average prediction errors hide variation within the cohort of mouse models. To investigate heterogeneity within the cohort, we stratified models according to the number of associated phenotypes (Figure 3D). Errors for models with few phenotypes were generally low. This is because when a model has a limited number of phenotype annotations, there may exist other models with the same characteristics. As an example, our dataset had 14 models annotated with the single phenotype of “deafness”. Averaging a small number of nearest neighbors that may have equivalent phenotypes produces predictions with zero error. For models with more annotations, the probability that other models have the same set of phenotypes by chance decreases.

Predictions that deviate from the annotated model representation highlight which of a model’s phenotypes are unusual. They can also suggest phenotypes that may be missing in the annotations. To illustrate such reasoning, we visualized the discrepancies between two models and their predictions (Figure 3E) (Luoh et al., 1997; Dickinson et al., 2016). In each case, part of the discrepancies originate from different weights assigned to measured MP terms or their ancestors, i.e. MP terms that are upstream in the ontology hierarchy. Other discrepancies arise from phenotypes that are not related to the models’ annotations. If nearest-neighbors were to be used as a denoising scheme, these discrepancies would result in imputed phenotypes. Even without formal imputation, if embedding coordinates were used as inputs for a machine-learning classifier, these phenotypes would influence how the model would be utilized by the classifier. Thus, the neighbor prediction can be informative for interpreting, or explaining, outcomes of downstream ML models.

Neighborhoods highlight consistency as well as heterogeneity of genotype-phenotype annotations

Models in our dataset are characterized by the mouse background strain, genetic allele or mutation, and mutation zygosity. While combinations of these features appear uniquely in the dataset, several models can be linked to the same genetic marker. We refer to all genetic

markers as genes for simplicity. A summary of the number of models per gene revealed a skewed distribution (Figure 4A). Thousands of genes were represented by a single mouse model in the dataset. Thousands of others have been studied in more than one model. The most studied genetic markers appeared in more than 200 models each (Figure 4A, inset).

The multiplicity of models linked with the same gene provides opportunities to study consistency and heterogeneity among genotype-phenotype associations. For illustration, we picked two of the most studied genes and highlighted their models in embeddings (Figure 4B). Some of the models appeared close together. This suggests consistency in the phenotypic annotations linked to the gene and robustness with respect, for example, to mutation construct or strain. At the same time, models linked to one gene also formed several close-knit clusters in distinct parts of the embedding. This suggests an opposite effect, i.e. heterogeneity in annotations. Heterogeneity may be due to incompatible curation, incomplete phenotyping measurements on some of the models, or genuine difference in phenotype due to the genotypes. The embedding cannot deconvolute these effects, but the visualization provides qualitative insight on the scale of the phenomenon in the mouse data.

To quantify the extent of consistency and heterogeneity, we studied the nearest neighbors of all models and assessed the proportion of times a model was near another model with the same gene (Figure 4C). This fraction increased with the multiplicity of mouse models. Among genes represented by more than 20 models, the proportion of models linked to another model with the same gene was above a third. This was substantially higher than expected if neighbors were assigned by chance. Nonetheless, since the level was below 0.5 for most genes, it is more likely for a randomly selected model to have neighbors with different genetic composition than to link to at least one model with the same mutated gene. As above, this can be due to difficulties in encoding the data, incompleteness in the dataset, or due to overlapping phenotypes associated with different genes.

Embedding diseases alongside animal models provides grounds for exploring disease-causing genes

Finally, we returned to the dataset of rare human diseases. For those diseases with annotations in the form of HP terms, we translated the phenotypes into MP terms and constructed vectors-based and text-based representations. We then projected the diseases into our previously-generated embeddings (Methods). With both vector-based (Figure 5A) and text-based (Figure

5B) approaches, diseases covered large areas of the embedding space and formed several disjoint clusters. Patterns were robust to how the disease phenotypes were translated from human to mammalian phenotype ontologies (Figure S5). Interestingly, disease projections avoided certain areas of the embeddings. As an example, one of the areas that was omitted by the diseases was enriched in phenotypes related to mortality and ageing (Figure S4).

Next, we asked to what extent similarities between diseases and mouse models can link diseases to their associated genes. For simplicity, we called all disease-associated genes as causative genes. We compared approaches based on numeric vectors, based on text descriptions, and two previously described methods for scoring disease-model associations (Smedley et al., 2013; Konopka & Smedley, 2020). The proportion of diseases that had a mouse model with the causative gene within 15 nearest neighbors was low: below ~4% (Figure 5C). Interestingly, the algorithm used to translate between human and mammalian phenotype ontologies had a smaller effect than the data encoding or the algorithm for computing neighbors. The best performer was a scheme specifically designed to score disease-model associations (Smedley et al., 2013). Strikingly, text-based approaches performed almost on par, better than vector-based approaches.

Given that text-based descriptions performed well in matching diseases to causative genes, we reasoned this approach could be a gateway to studying diseases without formal phenotype annotations. Among the dataset of human diseases, 61% did not have any HP annotations (Figure 5D). These diseases could not be included in calculations that rely on ontologies. They were also excluded from the performance assessments to ensure like-for-like comparisons (Figures 5A, 5B, 5C). However, these diseases have text descriptions, so they can be used in text-based calculations. Projections into the embedding of mouse models revealed they clustered into the same regions as before (Figure 5E). This indicates that un-annotated diseases span the whole phenotypic space, and also that our treatment of text descriptions did not introduce excessive biases that would place these cases apart from well-annotated diseases.

To illustrate explorative analyses based on text descriptions, we performed searches for two diseases without formal phenotype annotations (Figure 5F, 5G). An initial search for a glycogen deficiency (ORPHA:137625) correctly linked this disease to mouse models with abnormalities in glycogen homeostasis (Figure 5F). Because of the naive treatment of text in our calculations, some of the top-ranked models were characterized by opposite directional effects to the disease

description. Among the top hits were models with mutations in *Gys1* (Bouskila et al., 2010), the known causative gene for the disease. Other hits with similar phenotype profiles had mutations in *Ppp1r3a* and *Gyg*, both genes that participate in glycogen homeostasis. This confirms that text search can link human diseases to relevant mouse data. In this case, the top search hits all had consistent phenotype profiles, so a projection of the diseases into low-dimensional embeddings can also be expected to link the disease with a neighborhood of relevant mouse models.

Separately, we searched for an otorhinolaryngological disorder (ORPHA:141219) characterized by cysts around the nose and extending into the cranium (Figure 5G). The causative gene for the disorder is not known. The first two hits in text-based search matched the disease to mouse models with quite distinct phenotypes. The first was characterized by an intracranial phenotype (Hart et al., 2000); the second by phenotypes of the epidermis (Mill et al., 2009). Such hits can appear in distinct portions of an embedding. When several neighbors with drastically different profiles are utilized to project a data point into an embedding, the result may be blurring (Figure S5) or placement near one neighbor at the expense of the other. Both outcomes produce discrepancies between sets of neighbors computed in the original high-dimensional space and the low-dimensional embedding. This is a documented effect inherent to dimensional reduction (Cooley et al., 2020). In the disease context, it highlights the value in scrutinizing raw search results in addition to a low-dimensional visualization.

Discussion

Phenotyping animal models is a direct route to characterizing the impact of mutations in specific genomic elements. Studying genotype-phenotype relationships is often performed gene-by-gene. However, careful curation of the literature (Blake et al., 2021) as well as systematic phenotyping of hitherto-unstudied genes (Dickinson et al., 2016) mean that the collective data for the mouse will approach whole-genome coverage in the near future. This opens possibilities to utilize mouse phenotypes as a reference dataset in genomic analyses. As such, it is important to characterize the potential such a dataset offers for data exploration, machine learning, and downstream applications. In this work, we explored dimensional reduction for this data. The results visualize the heterogeneous landscape of mouse phenotypes. Our calculations also provide qualitative and quantitative observations about the strengths and limitations of this pool of data.

A challenge in dealing with large-scale phenotype data is that there are several plausible ways to encode sets of phenotypes so that they can be used for calculations. We explored vector-based, text-based, and graph-based approaches. There also exist many algorithms that can perform dimensional reduction. We focused on approaches that can be used for visualization and thus focused on dimensional reduction into two dimensions. Such embeddings enable interactive, human-led data exploration. However, we also investigated embeddings into higher dimensions, which can be beneficial for machine learning.

Strikingly, certain strategies provide sub-optimal visualizations in two dimensions. A strategy that first creates an embedding of an ontology and then projects phenotype sets into that space via coordinate averaging is prone to construct visualizations that are dominated by technical features, notably the number of phenotypes within a phenotype set (model or disease). This result has a mathematical justification: averaging is bound to summarize heterogeneous elements, in this case phenotypes, to a central value, with the variability of the outcome decreasing with the number of elements. It is also worth comparing this strategy with analysis pipelines in transcriptomics, which do not create embeddings based on genes and project expression profiles for samples into that space, but rather create embeddings for samples directly. Another suboptimal strategy is one that uses the MP ontology as a graph and performs a joint embedding of the ontology and a set of mouse models. This approach hides much of the rich phenotypic similarities and differences among models (Figure S2).

Strategies based on phenotype vectors as well as based on text descriptions produce visualizations that are interpretable at the level of single mouse models as well as the level of gated regions. The visualizations are not dominated by technical features of the dataset. Thus, the embeddings can be said to capture the true phenotypic diversity among mouse models. However, these embeddings may reflect, even amplify, peculiarities and limitations of the underlying data. Embeddings show many small, isolated groups with peculiar combinations of MP terms. Such combinations may arise due to targeted phenotyping efforts on those mouse models rather than true phenomenological specificity compared to other mutants. Moreover, the majority of mouse models are associated with a limited number of MP terms. Small annotation sets may be due to biased or incomplete phenotyping, for example if performed as part of a high-throughput screen (Dickinson et al., 2016). Such annotation sets may grow in time as further observations are catalogued and curated from literature (Blake et al., 2021). Analysis methods can be designed specifically to track changes of annotations due to steady growth (Konopka & Smedley, 2020). However, in the context of dimensional reduction, changes in the

annotation set should be expected to disrupt the positioning of individual models within an embedding. Thus, the embeddings should be expected to be unstable for models with low annotation counts. The overall structure of the embeddings may change too, albeit to a lesser extent.

Besides providing visualizations of the heterogeneous phenotype data, we investigated schemes for phenotype prediction based on nearest neighbors. These predictions are informative from at least three perspectives. First, they suggest new experimental assays on individual mouse models that may complete their phenotype profiles. Second, the predictions can be used for interpreting outputs from machine learning models trained using embedding coordinates. Third, we used prediction errors to quantify the information loss produced by various data-encoding and data-embedding approaches. The results confirm expected properties, namely that embedding data into low-dimensional spaces loses some information, that increasing the target dimension increases the fidelity of the embeddings, and that non-linear methods like UMAP preserve more information than linear methods like PCA. Interestingly, the results also show that discrepancies between predictions from original data and from embeddings can be comparable to discrepancies between different encodings of the original data (e.g. binary or real-valued vectors). This suggests that any analyses based on phenotype data are likely to be sensitive to how the raw data is prepared. Indeed, they may be more sensitive to data preparation than to dimensional reduction.

We corroborated this sensitivity in calculations projecting human diseases into embeddings of mouse models. We used nearest neighbors to link human diseases to mouse models and thereby to genes. Recall of established disease-gene associations varied depending on the encoding strategy (vector, text, etc.). Interestingly, approaches based on text similarities were among the most performant. Considering that these approaches are tunable (Konopka et al., unpublished), can integrate datasets other than phenotypic annotations, and that they execute two orders of magnitude faster than a dedicated scoring scheme for scoring disease-gene associations, they represent promising avenues for subsequent analyses.

Methods

Phenotype data

Definitions of the human phenotype (HP) ontology (Köhler et al., 2018) and the mammalian phenotype (MP) ontology (Smith & Eppig, 2015) were obtained through the OBO Foundry (HP version 2021-02-28, MP version 20201-01-12). Terms from the HP ontology were mapped onto terms in the MP ontology using owlsim (Washington et al., 2009), which is an ontology-aware algorithm, and using crossmap (Konopka et al., unpublished), which performs searches based on text similarity. Mappings with crossmap were performed using diffusion driven by the MP dataset and by a set of manual annotations (Konopka et al., unpublished). Both owlsim and crossmap map queries to multiple hits. Translations between the HP and MP ontologies were established using only the best-ranked mapping.

Mouse model definitions and associated phenotypes were obtained through the data portal of the International Mouse Phenotyping Consortium (data release 14.0) (Dickinson et al., 2016). Data downloaded from the IMPC included definitions of mouse models curated by the Mouse Genomics Database (Blake et al., 2021).

Disease definitions and associated phenotypes were obtained from Orphadata (http://www.orphadata.org; data version 2021-04-01). Downloaded data were then parsed using custom scripts (https://github.com/tkonopka/crossmap).

Ontology representations and embeddings

Embeddings for ontology terms were generated using two approaches (Figure S1). In a first approach, text strings were constructed for each term in the MP ontology by concatenating phenotype name, definition, synonyms, and comments. These strings were loaded into a crossmap knowledge-base (Konopka et al, unpublished; https://github.com/tkonopka/crossmap). The crossmap instance splits text into bags of k-mers, weights the k-mers according to their information content, and then builds a nearest-neighbors index. The crossmap instance was used to perform searches and compute sets of nearest neighbors for each ontology term. The nearest neighbors were provided to the UMAP algorithm (Becht et al., 2018) implemented in R to produce an embedding in two dimensions. Settings

were left at the default values, except for knn_repeats=3 to increase the quality of nearest-neighbor search, and min_dist=0.2 to increase space between adjacent points (for visualization).

In a second approach, MP ontology terms were treated as nodes in a graph. Edges between nodes were set if two MP terms were linked by 'is a' relationships in the ontology hierarchy (which is the only relationship type defined in the MP ontology). The resulting graph was processed using node2vec (Grover & Leskovec, 2016) to produce an embedding in two dimensions. Embeddings were produced using the snap implementation (https://github.com/snap-stanford/snap) with default parameters and the python implementation (https://pypi.org/project/node2vec/) with default settings. Since the results were noticeably different, the python implementation was also executed with non-default settings with num_walks=5 and walk_length=5.

Both UMAP and node2vec are stochastic algorithms and embeddings may differ slightly when repeated.

Mouse model representations

Mouse models were defined as sets of phenotypes associated with a specific mouse strain, gene knock-out genotype and zygosity. For IMPC data, models were further subcategorized by life stage (embryonic, early-adult, or late-adult).

Raw data for each mouse model were encoded in several ways: based on numeric vectors, on text, and using graphs (Figure S2). The numeric representations consisted of vectors of length equal to the size of the MP ontology. In a binary approach, values were set to zero by default and changed to one if an MP term was linked with a mouse model, or was an ancestor of such an MP term. Non-binary vector representations were constructed following a published protocol (Konopka & Smedley, 2020). Briefly, values within the vectors were initially set to prior probabilities for each MP phenotype, which were estimated from the ensemble of non-IMPC models. Values were updated through a Bayesian procedure with phenotype annotations and then propagated using the ontology hierarchy.

Two types of representations were constructed starting from text. A 'concise' representation was defined by concatenating the names of all MP terms associated with a model. A second 'complete' representation was constructed in the same way, but also including the names of all

ancestors of MP terms associated with a model. Text strings defined in these ways were loaded into an instance of a crossmap knowledge-base (Konopka et al., 2021). This software uses k-merization and domain-specific feature weighting, and provides an interface to search the data using nearest-neighbor algorithms.

For graph-based approaches, MP terms and mouse models were used as graph nodes. Edges were defined between MP nodes if the corresponding MP terms were linked by 'is a' relationships in the ontology. Edges were defined between mouse nodes and MP nodes from mouse model associations.

Mouse model embeddings

Embeddings of mouse models based on vector and text representations were performed with the UMAP algorithm (Becht et al., 2019) through an R implementation. For encodings based on non-binary and binary vectors, embeddings were produced starting from matrices of raw data, with vectors normalized to unit norm. Settings were left at default, except for knn_repeats=3 to increase the quality of nearest-neighbors and min_dist=0.2 to spread points for visualization purposes. In calculations exploring the impact of the embedding dimension, UMAP runs were provided with the same set of nearest neighbors, thus eliminating stochasticity effects due to approximate calculations of neighbors.

For encodings based on text, raw data were managed with a crossmap instance (https://github.com/tkonopka/crossmap). The crossmap instance was used to search for sets of nearest neighbors. Nearest neighbors were provided to the UMAP algorithm in R, which was run as described for the vector approaches. For graph-based approaches, embeddings were generated using node2vec (Grover & Leskovec, 2016) through the snap implementation (https://github.com/snap-stanford/snap) and python implementations (https://pypi.org/project/node2vec/) with default settings. The python implementation was also run using non-default settings walk_length=5 and num_walks=5.

Calculations of nearest neighbors with crossmap and embeddings with UMAP and node2vec rely on stochastic algorithms. Results may differ slightly when repeated.

While all embeddings were computed based on all the available data, some visualizations were truncated to enhance the presentation. Models that were not associated with any phenotype (MP:0002169, 'no abnormal phenotype detected') were excluded from visualizations, as were

models with only one phenotype. Such models can form isolated groups that are well separated from the others. Visualizations excluded these models to focus attention on the portion of the dataset with rich annotations. Visualizations also set quantile-based limits for axes to focus attention on the central parts of embeddings with. Quantile intervals were at least as wide as 2%-98% for each axis.

Projections of human diseases into embeddings of mouse models

Human diseases with phenotype annotations from the HP ontology were translated to the MP ontology by replacing their HP terms with best-matching MP terms. Sets of MP terms derived from diseases were encoded into non-binary vectors using the same procedures as for mouse models. Vectors with disease profiles were compared with all mouse models to identify nearest neighbors, and the position of each disease in an embedding was computed using UMAP. This calculation initially places a disease at the averaged location of its nearest mouse models, and subsequently adjusts this position using the UMAP optimization algorithm.

For encodings based on text, human diseases were taken to consist of a clinical summary paragraph, the names of associated phenotypes (MP translations), and the names of curated genes. These text documents were compared with phenotype descriptions of mouse models using crossmap to produce sets of nearest mouse models. Projections of the disease descriptions in embeddings of mouse models were computed by averaging the coordinates of the nearest mouse models. Optimization of this initial placement was not possible due to a technical limitation of the R UMAP package. Therefore, visualizations were produced using coordinate averaging of various numbers of mouse models (Figure S5). Given that coordinate averaging produces blurring and bias, the primary visualization was set as by placing each disease at the location of the single most-similar mouse model.

Availability

Source code for the scripts used in this work are available on github (https://github.com/tkonopka/mouse-embeddings). A snapshot of the raw and processed data files, including all embeddings, are available in a zenodo repository (doi 10.5281/zenodo.4916172).

Funding

This work was supported by a grant from the National Institutes of Health [Grant 5-UM1-HG006370 to D.S.].

References

Batson, J., & Royer, L. (2019). Noise2Self: Blind Denoising by Self-Supervision. *Proceedings of Machine Learning Research*, *97*.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38-44.

Blake, J. A., Baldarelli,, R., Kadin,, J. A., Richardson,, J. E., Smith, C. L., Bult, C. J., & the Mouse Genome Database Group. (2021). Mouse Genome Database (MGD): Knowledgebase for mouse–human comparative biology. *Nucleic Acids Research*, *49*, D981-987.

Bouskila, M., Hunter, R. W., Ibrahim, A. F., Delattre, L., Peggie, M., van Diepen, J. A., Voshol, P. J., Jensen, J., & Sakamoto, K. (2010). Allosteric regulation of glycogen synthase controls glycogen synthesis in muscle. *Cell metabolism*, *12*(5), 456-466.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496-502.

Chen, J., Hu, P., Jimenez-Ruiz, E., Holter, O. M., Antonyrajah, D., & Horrocks, I. (2020). OWL2Vec : Embedding of OWL Ontologies. arXiv:2009.14654.

Cooley, S. M., Hamilton, T., Ray, C. J., & Deeds, E. J. (2020). A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. bioRxiv:689851.

Dickinson, M. E., Flenniken, A. M., & Ji, X. (2016). High-throughput discovery of novel developmental phenotypes. *Nature*, *537*(7621), 508-514.

Duong, D., Uppunda, A., Ju, C., Zhang, J., Chen, M., Eskin, E., Li, J. J., & Chang, K.-W. (2020). Evaluating Representations for Gene Ontology Terms. biorxiv:765644.

Grigelioniene, G., Suzuki, H. I., & Taylan, F. (2019). Gain-of-function mutation of microRNA-140 in human skeletal dysplasia. *Nature Medicine*, *25*(4), 583-590.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD*.

Han, X., Wang, R., & Zhou, Y. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, *172*(5), 1091-1107.

Hart, A., Melet, F., Grossfeld, P., Chien, K., Jones, C., Tunnacliffe, A., Favier, R., & Bernstein, A. (2000). Fli-1 is required for murine vascular and megakaryocytic development and is hemizygously deleted in patients with thrombocytopenia. *Immunity*, *13*(2), 167-177.

Hayes, C., Lyon, M. F., & Morriss-Kay, G. M. (1998). Morphogenesis of Doublefoot (Dbf), a mouse mutant with polydactyly and craniofacial defects. *Journal of Anatomy*, *193*(1), 81-91.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*, 504-507.

Justice, M. J., & Dhillon, P. (2016). Using the mouse to model human disease: increasing validity and reproducibility. *Disease Models & Mechanisms*, *9*, 101-103.

Kalucka, J., de Rooij, L. P., & Goveia, J. (2020). Single-Cell Transcriptome Atlas of Murine Endothelial Cells. *Cell*, *180*(4), 764-779.

Kim, J., Kim, D., & Sohn, K.-A. (2021). HiG2Vec: Hierarchical Representations of Gene Ontology and Genes in the Poincaré Ball. *Bioinformatics*, btab193.

Klimovskaia, A., Lopez-Paz, D., Bottou, L., & Nickel, M. (2020). Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature Communications*, *11*(1), 1-8.

Köhler, S., Carmody, L., & Vasilevsky, N. (2018). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, *47*(D1), D1018-D1027.

Konopka, T., Ng, S., & Smedley, D. (unpublished). Diffusion enables integration of heterogeneous data and user-driven learning in a desktop knowledge-base. https://github.com/tkonopka/crossmap

Konopka, T., & Smedley, D. (2020). Incremental data integration for tracking genotype-disease associations. *PLoS Comp Bio*, *16*(1), e1007586.

Kratochvíl, M., Koladiya, A., Balounova, J., Novosadova, V., Sedlacek, R., Fišer, K., Vondrášek, J., & Drbal, K. (2019). SOM-based embedding improves efficiency of high-dimensional cytometry data analysis. *biorxiv*. https://doi.org/10.1101/496869

Luoh, S. W., Bain, P. A., Polakiewicz, R. D., Goodheart, M. L., Gardner, H., Jaenisch, R., & Page, D. C. (1997). Zfx mutation results in small animal size and reduced germ cell number in male and female mice. *Development*, *124*(11), 2275-2284.

Lyon, M. F., Quinney, R., Glenister, P. H., Kerscher, S., Guillot, P., & Boyd, Y. (1996). Doublefoot: a new mouse mutant affecting development of limbs and head. *Genetics Research*, *68*(3), 221-231.

Meehan, T., Conte, N., & West, D. (2017). Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nature Genetics*, *49*(8), 1231-1238.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Mill, P., Lee, A. W., Fukata, Y., Tsutsumi, R., Fukata, M., Keighren, M., Porter, R. M., McKie, L., Smyth, I., & Jackson, I. J. (2009). Palmitoylation regulates epidermal homeostasis and hair follicle differentiation. *PLoS Genetics*, *5*(11), e1000748.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S.

(2019). Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, *37*, 1482-1492.

Rosenberg, A. B., Roco, C. M., & Muscat, R. A. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, *360*(6385), 176-182.

Saunders, A., Macosko, E. Z., & Wysoker, A. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell*, *174*(4), 1015-1030.

Schreyer, S. A., Cummings, D. E., McKnight, G. S., & LeBoeuf, R. C. (2001). Mutation of the RIIbeta subunit of protein kinase A prevents diet-induced insulin resistance and dyslipidemia in mice. *Diabetes*, *50*(11), 2555-2562.

Smaili,, F. Z., Gao, X., & Hoehndorf*, R. (2018). Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, *34*(13), i52-i60.

Smaili, F. Z., Gao, X., & Hoehndorf, R. (2019). OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics*, *35*(12), 2133-2140.

Smedley, D., Oellrich, A., Köhler, S., Ruef, B., Sanger Mouse Genetics Project, Westerfield, M., Robinson, P., Lewis, S., & Mungall, C. (2013). PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database*.

Smith, C. L., & Eppig, J. T. (2015). Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. *Journal of Biomedical Semantics*, *6*(1), 1-7.

The Tabula Muris Consortium. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, *562*(7727), 367-372.

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., & Lewis, S. E. (2009). Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation. *PLoS Biology*, *7*(11), e1000247.

Figures

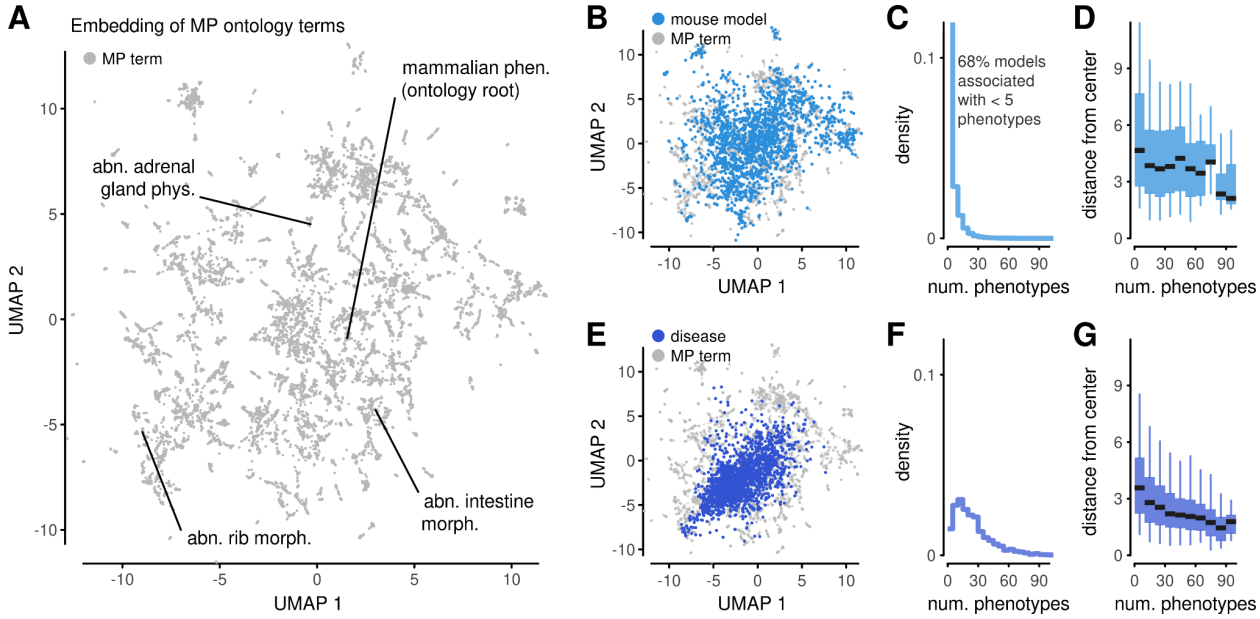

Figure 1. Embeddings of mammalian phenotypes. (A) Embedding of mammalian phenotype (MP) ontology terms based on text similarity. Labels point to selected ontology terms. (phen.: phenotype, abn.: abnormal, dev.: development) (B) Projection of mouse models into an embedding of ontology terms via averaging of coordinates of their annotated phenotypes. (C) Histogram of the number of phenotypes for all mouse models. (D) A summary of the position of mouse models in the projections in (B), stratified by the number of annotated phenotypes. Boxes represent 25%-75% intervals, whiskers represent 5%-95% intervals. (E-G) Analogous to (B-D) using phenotype profiles of human diseases translated into the mammalian phenotype ontology.

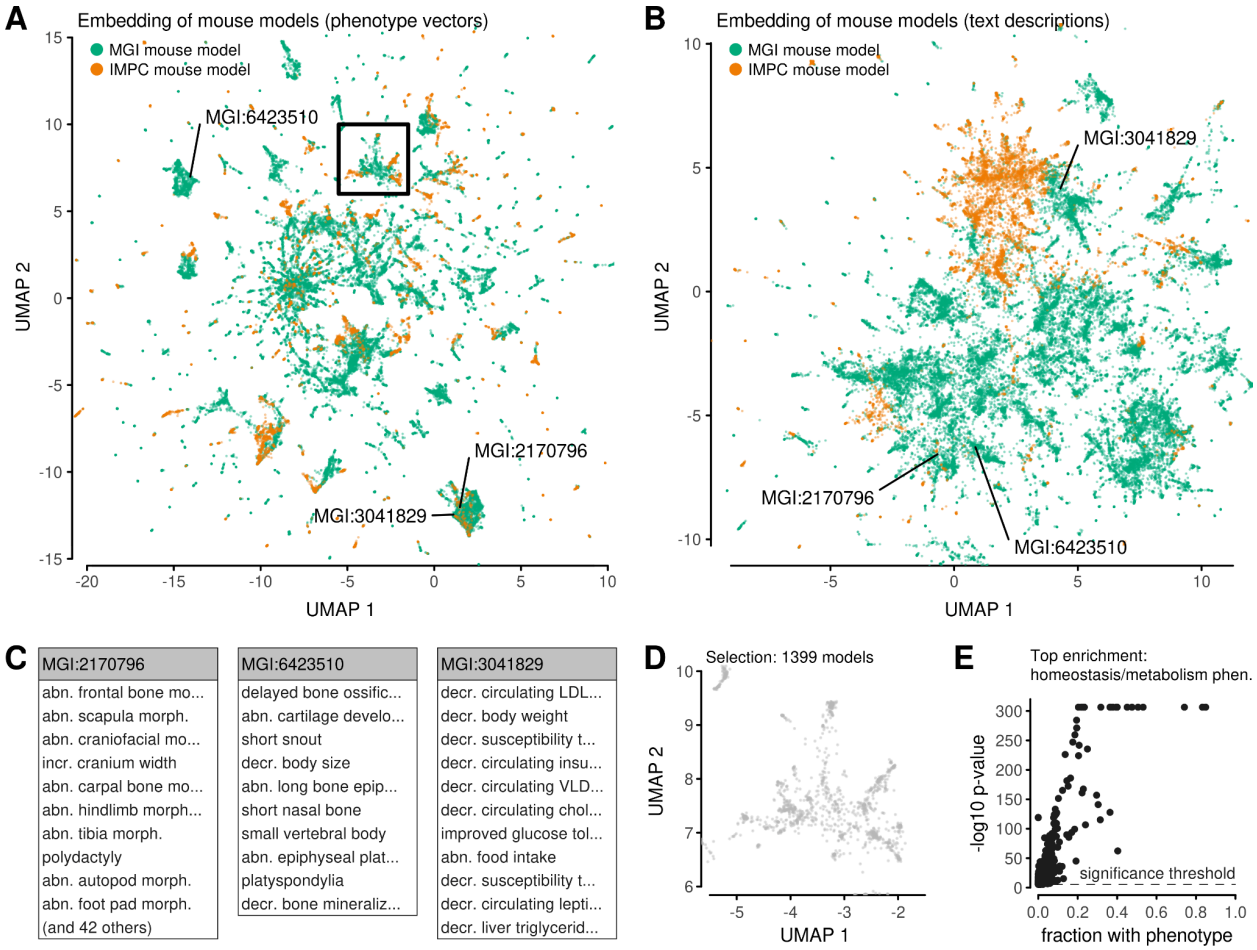

Figure 2. Embeddings of mouse models. (A) Embedding of mouse models based on vector representations of their phenotypes. Models are colored by the source of curated data. Labels and gate point to selected models. (B) Analogous to (A), but with the layout of mouse models produced using semantic similarities of text descriptions. (C) Lists of phenotypes associated to individual mouse models highlighted in (A). Some lists are truncated for this visualization. All phenotype names match definitions from the ontology (abn.: abnormal, morph.: morphology, incr.: increased, decr.: decreased) (D) A small region of the embedding in (A). (E) Enrichment analysis comparing the phenotypes associated with mouse models in (D) against all other models outside the gated region. The significance threshold is Bonferroni-corrected p=0.05. The most significant phenotype is labeled.

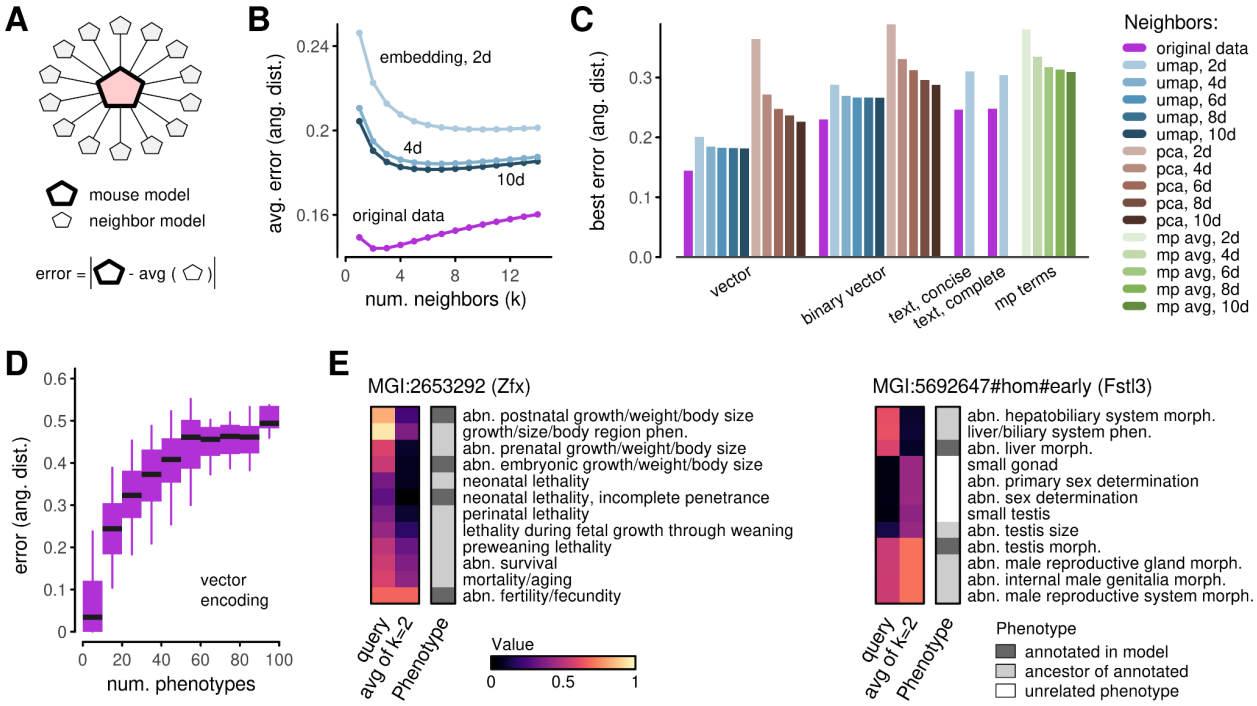

Figure 3. Phenotype prediction. (A) Schematic explaining phenotype prediction using neighbors. Given a mouse model, its predicted phenotype profile is defined as a simple average over its neighbors. An error is defined as the L2 norm between the model profile and the prediction. (B) Exploration of mean prediction error as a function of the number of neighbors used in the calculation. Lines correspond to distinct ways of identifying neighbors: from original vector representations, or from embeddings in various dimensions. (C) Summary of best-achieved errors for prediction approaches using original vector data, original binary vector data, embeddings in various dimensions, and using text-based similarity measures. (D) Stratification of mouse models by the number of model phenotypes. Boxes represent 25%-75% intervals, whiskers represent 5%-95% intervals. (E) Examples of mouse model phenotype vectors and predictions based on two nearest neighbors. Heatmaps only show a small number of phenotypes that contribute the most to prediction errors. Categorical phenotype annotations indicate whether a listed phenotype is one of the models' annotated phenotypes, an ancestor of an annotated phenotype, or a phenotype unrelated to model annotations.

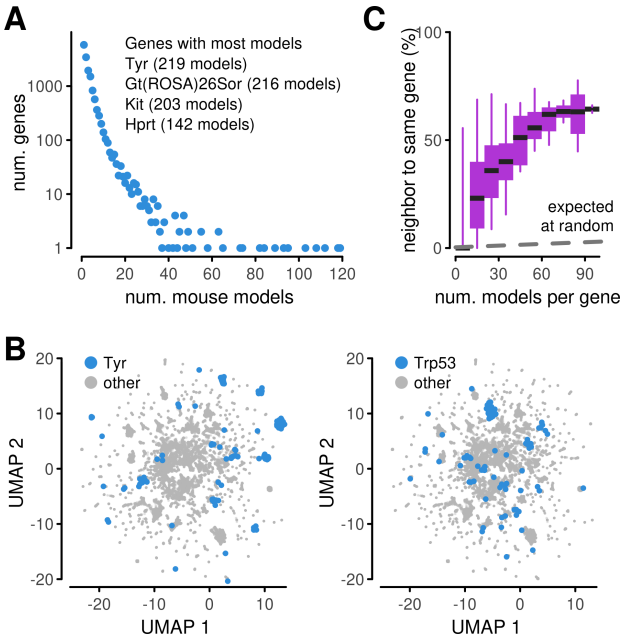

Figure 4. Phenotype heterogeneity. (A) Summary of the multiplicity of models available for individual genes. The most represented genes are listed in the inset. (B) Embeddings of mouse models highlighting the location of models with selected genes knocked-out. Highlighted models are jittered to better display the number of models. (C) Summary of the proportion of genes for which the nearest-neighbors of a mouse model contain another model with the same gene knocked-out. The summary is stratified by the number of models available for a gene. Boxes represent 25%-75% intervals, whiskers represent 5%-95% intervals. Dashed line indicates an expected level under a null hypothesis that neighbors are selected at random.

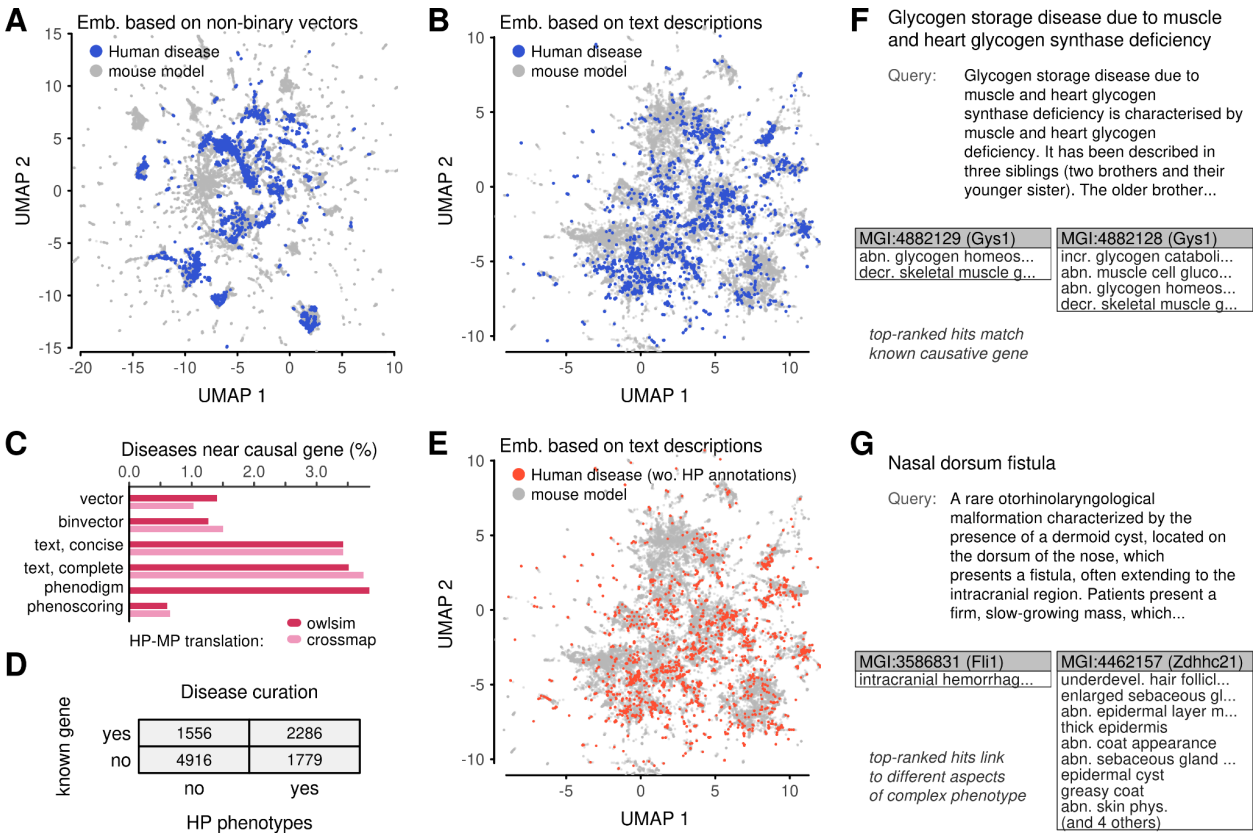

Figure 5. Embedding of human diseases in the mouse phenotypic space. (A) Projection of human diseases into an embedding of mouse models created using phenotype vectors. (B) Analogous to (A), but with the underlying embedding produced using semantic similarities of text-based descriptions. (C) Summary of causal-gene extraction from 15 nearest neighbors. (D) Summary of ORPHANET disease annotations in terms of phenotype ontology terms and causative genes. (E) Projection of human diseases without HP annotation into an embedding of mouse models based on text similarity. (F,G) Examples of text-based disease descriptions along with two mouse models, selected manually from among the top five search hits.

Supplementary Figures

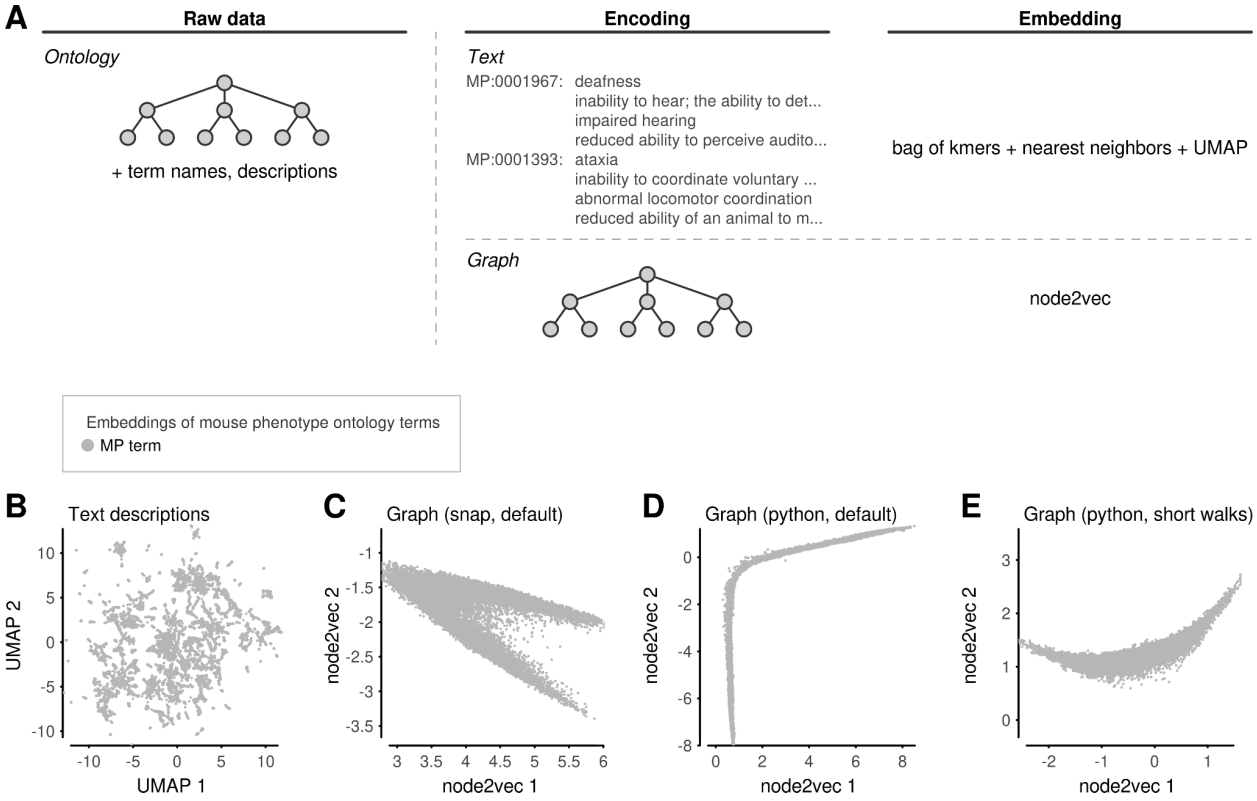

Figure S1. Embeddings of mammalian phenotype ontology terms. (A) Schematic of ontology data,possibilities for numerical encoding, and algorithms for creating embeddings. (B) Embedding based on text descriptions (name, definition, synonyms, comments, and name of parent term). Similarities computed using crossmap and layout generated with UMAP. (C) Embedding based on the hierarchy relations between ontology terms, generated using snap implementation of node2vec with default settings. (D) Embedding based on the ontology hierarchy graph, generated using python implementation of node2vec with default settings. (E) Similar to E, generated with non-default settings with walk-length and num-walks set to 5.

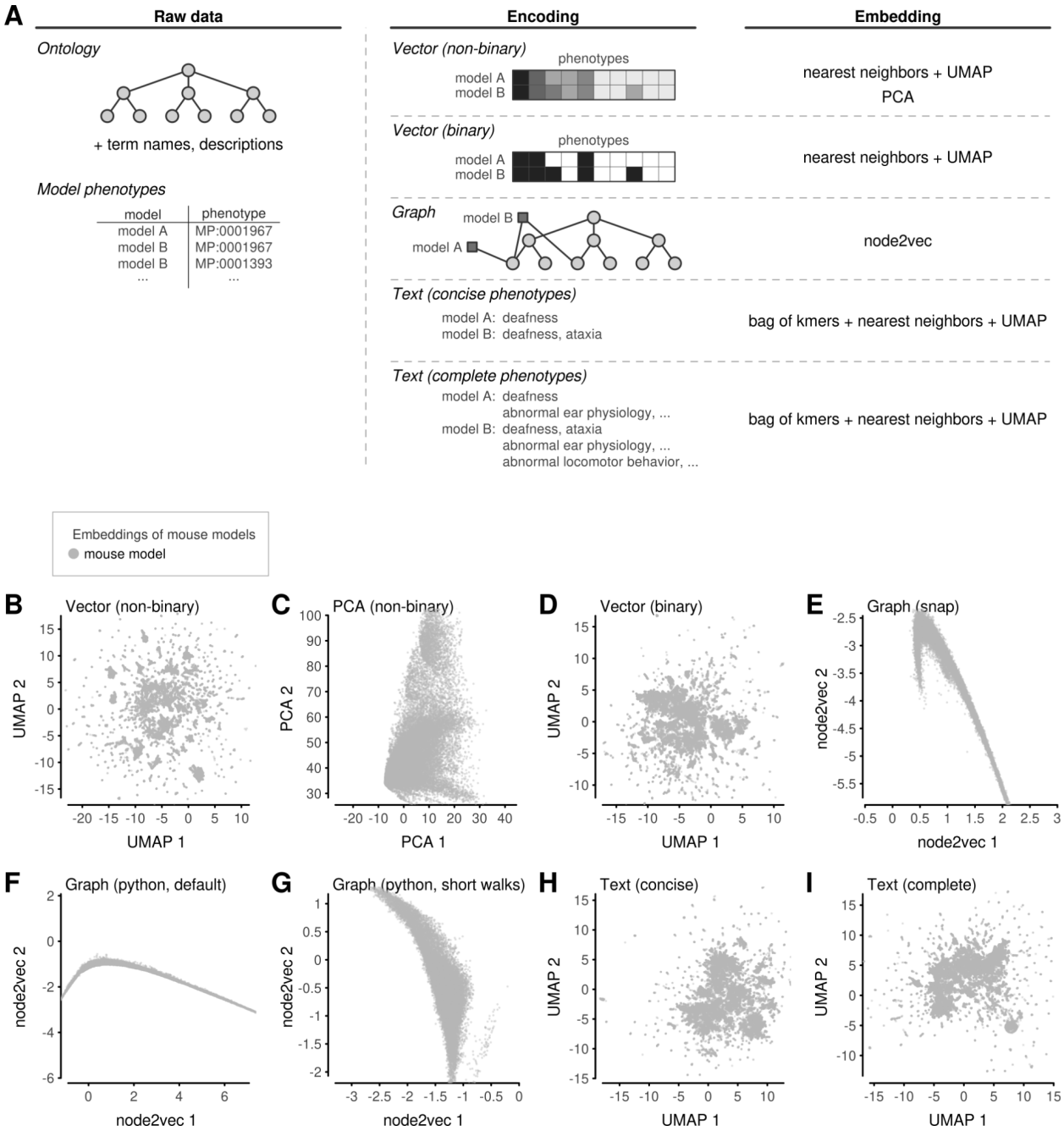

Figure S2. Embeddings of mouse models. (A) Schematic of phenotype data for mouse models, data encoding possibilities, and algorithms for creating embeddings. (B) UMAP embedding based on non-binary vector representations. (C) PCA based on non-binary vector representations. (D) UMAP embedding based on binary vector representations. (E) Node2vec embedding based on a graph connecting mouse models to their ontology phenotypes, generated using snap implementation of node2vec with default settings. (F) Similar to (E), generated using python implementation of node2vec with default settings. (G) Similar to (E),

generated with python implementation of node2vec with non-default settings using walk-length and num-walks set to 5. (H) Embedding based on text descriptions of mouse phenotypes. (I) Similar to (H), but using text descriptions of complete phenotypes.

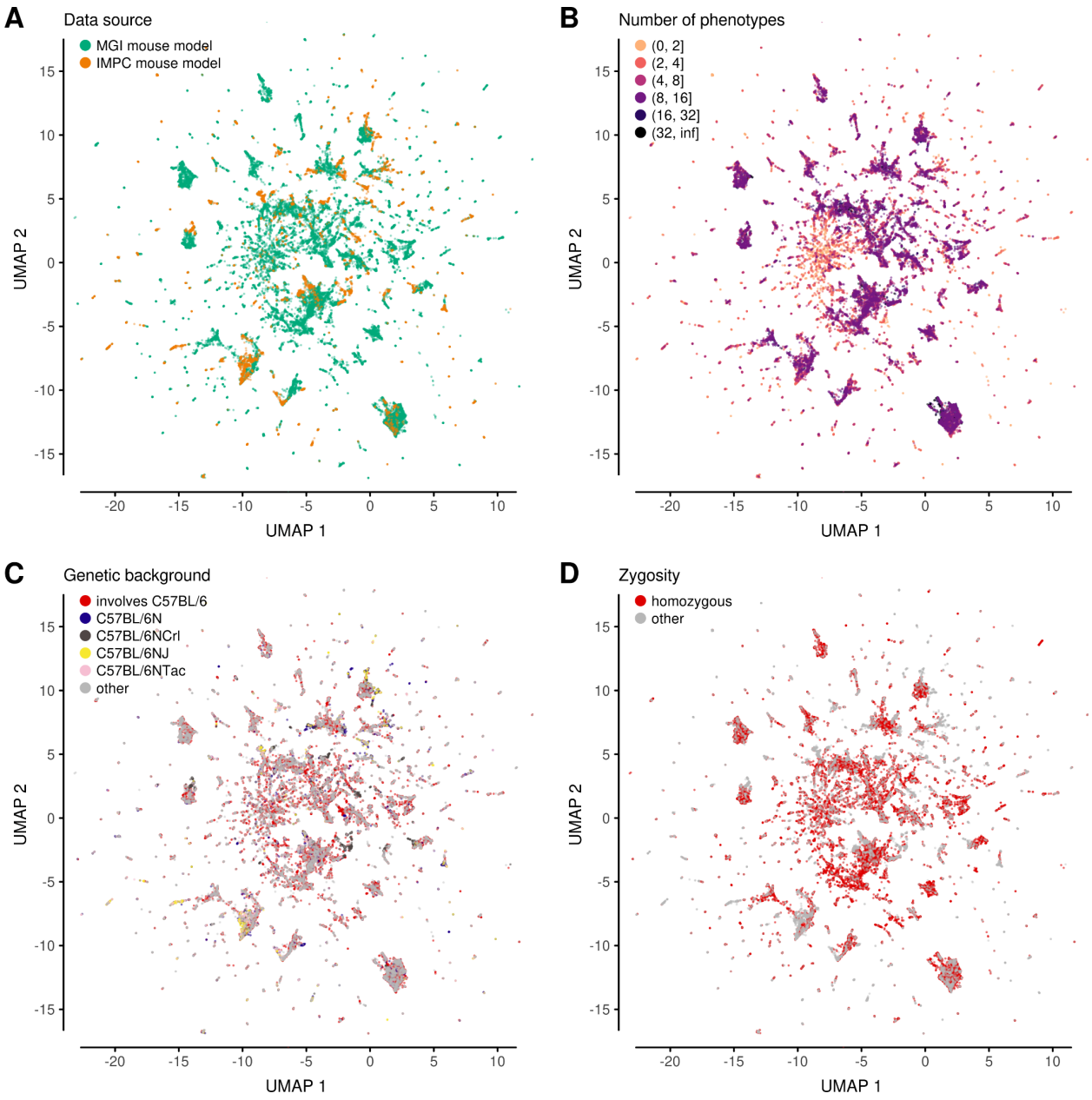

Figure S3. Mouse model covariates. All panels show embeddings of mouse models based on non-binary vectors of phenotypes. Panels differ by stratification strategy: (A) phenotyping source; (B) number of registered phenotypes; (C) animal genetic background; (D) zygosity of gene knock-out.

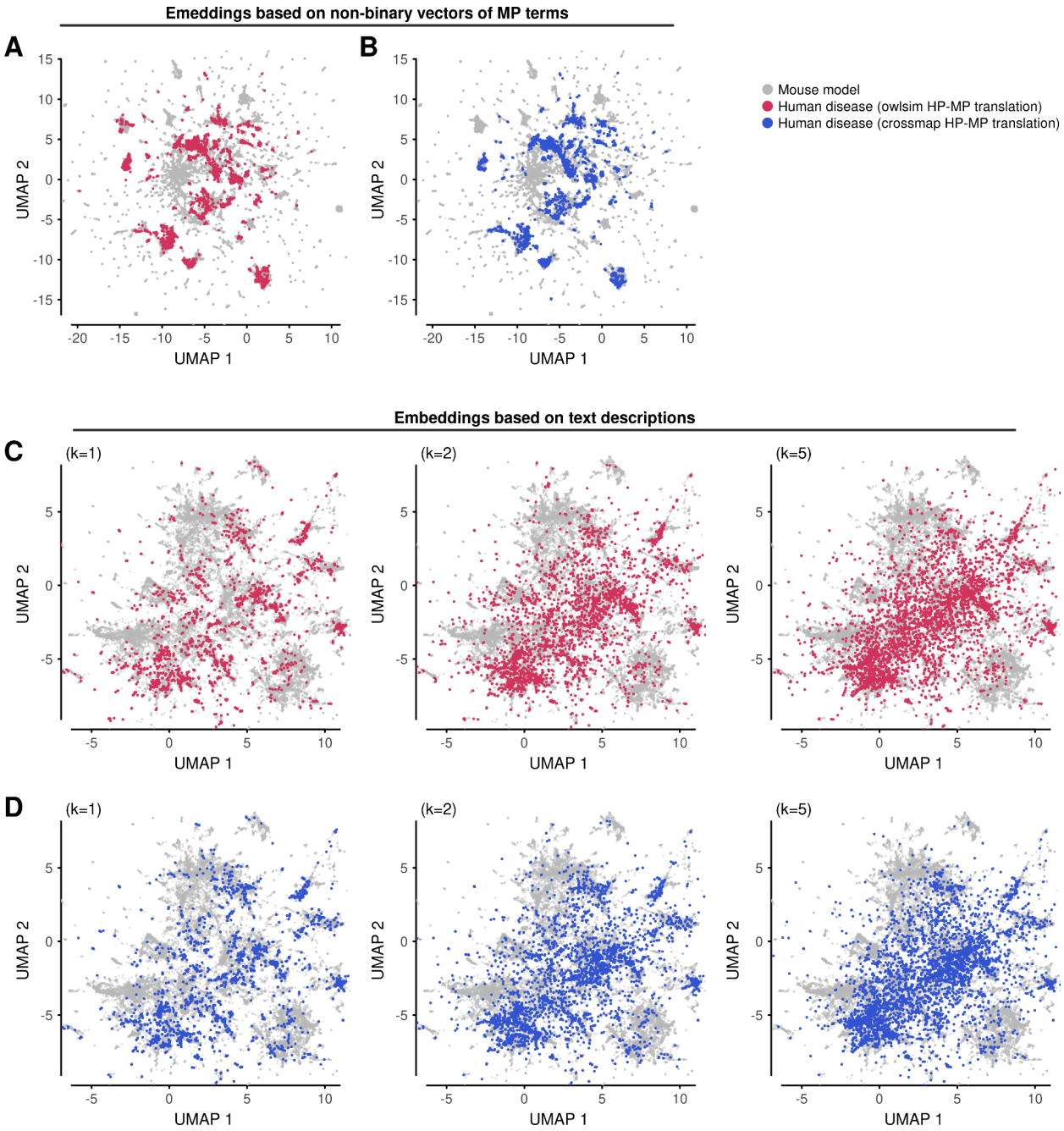

Figure S4. Embeddings of diseases. (A) Projection of human disease phenotype profiles (colored points) into an embedding of mouse models (gray dots). The embedding was created based on non-binary vectors with mouse model phenotypes, without information about diseases. Disease phenotype profiles were translated into mammalian phenotype (MP) ontology terms using owlsim, encoded into vectors, and projected into the embedding using UMAP. (B) Similar to (A), but with the translations between human and mouse phenotypes carried out using crossmap. (C) Series of three embeddings where human diseases are projected into an

embedding based on text descriptions of mouse models. Descriptions of human diseases consisted of a clinical summary, and translations of phenotypes into mammalian phenotype ontology terms using owlsim. Disease descriptions were projected into the embedding by searching for k=1, k=2, k=5 mouse models and then averaging their coordinates. (D) Similar to (C), but with translations between human and mouse phenotype carried out using crossmap.