Multimodal learning with graphs

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, & Marinka Zitnik

Artificial intelligence for graphs has achieved remarkable success in modelling complex systems, ranging from dynamic networks in biology to interacting particle systems in physics. However, the increasingly heterogeneous graph datasets call for multimodal methods that can combine different inductive biases — assumptions that algorithms use to make predictions for inputs they have not encountered during training. Learning on multimodal datasets is challenging because the inductive biases can vary by data modality and graphs might not be explicitly given in the input. To address these challenges, graph artificial intelligence methods combine different modalities while leveraging cross-modal dependencies through geometric relationships. Diverse datasets are combined using graphs and fed into sophisticated multimodal architectures, specified as image-intensive, knowledge-grounded and language-intensive models. Using this categorization, we introduce a blueprint for multimodal graph learning, use it to study existing methods and provide guidelines to design new models.

Deep learning on graphs has contributed to breakthroughs in biology, chemistry, physics and the social sciences. The predominant use of graph neural networks is to learn representations of various graph components — such as nodes, edges, subgraphs and entire graphs — based on neural message-passing strategies. The learned representations are used for downstream tasks, including label prediction via semi-supervised learning, self-supervised learning, and graph design and generation. In most existing applications, datasets explicitly describe graphs in the form of nodes, edges and additional information representing contextual knowledge, such as node, edge and graph attributes.

Modelling complex systems requires measurements that describe the same objects from different perspectives, at different scales, or through multiple modalities, such as images, sensor readings, language sequences and compact mathematical statements. Multimodal learning studies how such heterogeneous, complex descriptors can be optimized to create learning systems that are broadly generalizable, robust to changes in the underlying data distributions and can train more with fewer labelled data. While multimodal learning has been successfully used in settings where unimodal methods fail, it presents several challenges that must be overcome to enable its broad use in artificial intelligence. These challenges include finding representations optimized for machine learning analyses and fusing combined information from various modalities to create predictive models. These challenges have proven difficult. For example, multimodal methods tend to focus on only a subset of modalities that are most helpful during model training while ignoring modalities that might be informative for model implementation — a pitfall known as modality collapse. Moreover, in contrast to the frequent assumption that every object must exist in all modalities, the complete set of modalities is rarely available due to limitations of data collection and measurement technologies — a challenge known as missing modalities. Because different modalities can lead to intricate relational dependencies, simple modality fusion cannot fully leverage multimodal datasets. Graph learning models such data systems by connecting data points in different modalities as edges in optimally defined graphs and building learning systems for a wide range of tasks.

We introduce a blueprint for multimodal graph learning (MGL). The MGL blueprint provides a framework that can express existing algorithms and help develop new methods for multimodal learning.

1Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA, USA. 2Department of Biomedical Informatics, Harvard University, Boston, MA, USA. 3Harvard Data Science Initiative, Cambridge, MA, USA. 4Harvard College, Cambridge, MA, USA. 5Division of Pulmonary and Critical Care, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. 6Broad Institute of MIT and Harvard, Cambridge, MA, USA. 7These authors contributed equally: Yasha Ektefaie, George Dasoulas. e-mail: marinka@hms.harvard.edu
Leveraging graphs. This framework allows for learning fused graph representations and studying the aforementioned challenges of modality collapse and missing modalities. We apply this formulation across a broad spectrum of domains, ranging from computer vision and language processing to the natural sciences. We consider image-intensive graphs (IIGs) for image and video reasoning (‘MGL for images’), language-intensive graphs (LIGs) for processing natural and biological sequences (‘MGL for language datasets’) and knowledge-intensive graphs (KIGs) used to aid in scientific discovery (‘MGL in natural sciences’).

Graph neural networks for multimodal learning
Deep learning has created a wide range of fusion approaches for multimodal learning. For example, recurrent neural network and convolutional neural network (CNN) architectures have successfully been combined to fuse sound and image representations in video description problems. More recently, generative models have also proven very accurate for both language-dependent and physics-based multimodal data. Such models are based on an encoder–decoder framework, where in the encoder, the combined architectures are trained simultaneously (each one specialized for a modality), while the decoder aggregates information from individual architectures. When complex relations between modalities produce a network structure, graph neural networks (GNNs; Supplementary Note 1) provide an expressive and flexible strategy to leverage interdependencies in multimodal datasets.

Blueprint for graph-centric multimodal learning
The use of GNNs for multimodal learning is attractive because of their flexibility to model interactions both within and across different data types. However, data fusion through graph learning requires the construction of network topology and the application of inference algorithms over graphs. We present a methodology that given a collection of multimodal input data, yields output representations that are used in downstream tasks. We refer to this methodology as MGL. MGL can be seen as a blueprint consisting of four learning components that are connected in an end-to-end fashion. In Fig. 2a,b, we highlight the difference between a conventional combination of unimodal architectures for treating multimodal data and the suggested all-in-one multimodal architecture.

The first two components of MGL, identifying entities and uncovering topology, can be grouped as the structure-learning phase (Fig. 2c).

Identifying entities (MGL component 1). The first component identifies relevant entities in various data modalities and projects them into a shared namespace. For example, in precision medicine, the state of a patient might be described by matched pathology slides and clinical notes, giving rise to patient nodes with the combined image and language information. In another example from computer vision (Fig. 3), entity identification entails defining superpixels in an image.

Uncovering topology (MGL component 2). With the entities of our problem defined, the second component discovers the interactions and interaction types among the nodes across the modalities. Interactions are often explicitly provided, so the graph is given, and this component is responsible for the combination of the already existing graph structure with the rest of modalities (for example, in Fig. 3c, the uncovering topology component corresponds to combining protein surface information with the protein structure itself). When the data do not have an a priori network structure, the uncovering topology component explores possible adjacency matrices based on explicit (for example, spatial and visual characteristics) or implicit (for example, similarities in representations) features. For the latter case, examples from the natural language processing field consider the construction of graphs from text input that express relations among words (Fig. 4b).

After graphs are specified or adaptively optimized (structure-learning phase in MGL; Fig. 2c), various strategies can be used to learn on the graphs. The last two MGL components, known together as the learning-on-structure phase (Fig. 2c), capture these strategies.
Perspective

Propagating information (MGL component 3). The third component employs convolutional or message-passing steps to learn node representations based on graph adjacencies (see Supplementary Note 1 for more details on graph convolutions and message passing). In the case of multiple adjacency matrices, methods use independent propagation models or assume a hypergraph formulation that fuses adjacency matrices with a single propagation model.

Mixing representations (MGL component 4). The last component transforms learned node-level representations depending on downstream tasks. The propagation models output representations over the nodes that can be combined and mixed depending on the final representation level (for example, a graph-level or a subgraph-level label). Popular mixing strategies include simple aggregation operators (for example, summation or averaging) or more sophisticated functions that incorporate neural network architectures.

Visual comprehension

Visual comprehension remains a cornerstone of visual analyses, where MGL has proven helpful in classifying, segmenting and enhancing images. Image classification identifies the set of object categories present in an image. By contrast, image segmentation divides an image into segments and assigns each segment into a category. Finally, image restoration and denoising transform low-quality images into high-resolution counterparts. The information required for these tasks lies in objects, segments and image patches, as well as in the long-range context surrounding them.

MGL for images

IIGs are multimodal graphs where nodes represent visual features and edges represent spatial connections between image features. Structure image learning entails creating IIGs to encode geometric priors relevant to images, such as translational invariance and scale separation. Translational invariance describes how the output of a CNN must not change depending on shifts in the input image and is achieved by convolutional filters with shared weights. By contrast, scale separation specifies how to decompose long-range interactions between features across scales, focusing on localized interactions that can be propagated to coarser scales. For example, pooling layers follow convolution layers in CNNs to achieve scale separation. In addition, GNNs can model long-range dependencies of arbitrary shape that are important for image-related tasks such as image segmentation, image restoration or human–object interaction.

Figure 2c shows all MGL components, going from multimodal input data to optimized representations used for downstream tasks. Mathematical formulations are in Box 1 and summaries of MGL methods are in Supplementary Note 2.

MGL for images

IIGs are multimodal graphs where nodes represent visual features and edges represent spatial connections between image features. Structure image learning entails creating IIGs to encode geometric priors relevant to images, such as translational invariance and scale separation. Translational invariance describes how the output of a CNN must not change depending on shifts in the input image and is achieved by convolutional filters with shared weights. By contrast, scale separation specifies how to decompose long-range interactions between features across scales, focusing on localized interactions that can be propagated to coarser scales. For example, pooling layers follow convolution layers in CNNs to achieve scale separation. In addition, GNNs can model long-range dependencies of arbitrary shape that are important for image-related tasks such as image segmentation, image restoration or human–object interaction.

Visual comprehension

Visual comprehension remains a cornerstone of visual analyses, where MGL has proven helpful in classifying, segmenting and enhancing images. Image classification identifies the set of object categories present in an image. By contrast, image segmentation divides an image into segments and assigns each segment into a category. Finally, image restoration and denoising transform low-quality images into high-resolution counterparts. The information required for these tasks lies in objects, segments and image patches, as well as in the long-range context surrounding them.

MGL for images

IIGs are multimodal graphs where nodes represent visual features and edges represent spatial connections between image features. Structure image learning entails creating IIGs to encode geometric priors relevant to images, such as translational invariance and scale separation. Translational invariance describes how the output of a CNN must not change depending on shifts in the input image and is achieved by convolutional filters with shared weights. By contrast, scale separation specifies how to decompose long-range interactions between features across scales, focusing on localized interactions that can be propagated to coarser scales. For example, pooling layers follow convolution layers in CNNs to achieve scale separation. In addition, GNNs can model long-range dependencies of arbitrary shape that are important for image-related tasks such as image segmentation, image restoration or human–object interaction.

Visual comprehension

Visual comprehension remains a cornerstone of visual analyses, where MGL has proven helpful in classifying, segmenting and enhancing images. Image classification identifies the set of object categories present in an image. By contrast, image segmentation divides an image into segments and assigns each segment into a category. Finally, image restoration and denoising transform low-quality images into high-resolution counterparts. The information required for these tasks lies in objects, segments and image patches, as well as in the long-range context surrounding them.

MGL for images

IIGs are multimodal graphs where nodes represent visual features and edges represent spatial connections between image features. Structure image learning entails creating IIGs to encode geometric priors relevant to images, such as translational invariance and scale separation. Translational invariance describes how the output of a CNN must not change depending on shifts in the input image and is achieved by convolutional filters with shared weights. By contrast, scale separation specifies how to decompose long-range interactions between features across scales, focusing on localized interactions that can be propagated to coarser scales. For example, pooling layers follow convolution layers in CNNs to achieve scale separation. In addition, GNNs can model long-range dependencies of arbitrary shape that are important for image-related tasks such as image segmentation, image restoration or human–object interaction.

Visual comprehension

Visual comprehension remains a cornerstone of visual analyses, where MGL has proven helpful in classifying, segmenting and enhancing images. Image classification identifies the set of object categories present in an image. By contrast, image segmentation divides an image into segments and assigns each segment into a category. Finally, image restoration and denoising transform low-quality images into high-resolution counterparts. The information required for these tasks lies in objects, segments and image patches, as well as in the long-range context surrounding them.
regions define nodes used to extract feature maps and summary visual features for each region \(^2\text{4,45}\), whose attributes are initialized from CNNs such as FCN-1653 or VGG1954. Moreover, the nodes are connected to their \(k\)-nearest neighbours in the CNN learned feature space \(^2\text{4,44,51}\) (Fig. 3b), to spatially adjacent regions \(^2\text{4,43,52,53}\), or to an arbitrary number of neighbours based on a previously defined similarity threshold between nodes \(^4\text{4,52}\).

Once the structure-learning phase of MGL is completed, propagation models (MGL component 3) based on graph convolutions \(^4\text{2,46,52,55}\) and graph attention \(^4\) are used to weigh node neighbours in the graph based on learned attention scores \(^2\text{4,47}\). In addition, methods such as graph denoiser networks \(^4\), internal graph neural networks \(^4\) and residual graph convolutional networks \(^4\) consider edge similarities to indicate the relative distance between image regions.

**Visual reasoning**

Visual reasoning goes beyond recognizing visual elements by asking questions about the relationships between entities in images. These relationships can involve humans and objects as in human–object interaction (HOI) \(^2\text{6,40}\) or, more broadly, visual, semantic and numeric entities as in visual question answering (VQA) \(^4\text{0–43}\).

In HOI, the MGL methods identify two entities, human body parts (for example, hands, face and so on) and objects (for example, surfboard, bike and so on) \(^4\text{0,42,46}\), that interact in fully connected \(^4\text{0,45}\), bipartite \(^4\text{2,46}\) or partially connected topologies \(^4\text{4,46}\). MGL methods for VQA construct a new topology \(^4\) that spans interconnected visual, semantic and numeric graphs. Entities represent visual objects identified by an extractor, such as Faster R-CNN \(^4\), scene text identified by optical character recognition and number-type texts. Interactions between these entities are defined based on spatial localization: entities occurring near each other are connected by edges.

To learn about these structures (MGL component 3), methods distinguish between propagating information between entities of the same type and entities of different types. In HOI, knowledge about entities of the same kind (that is, intra-class neural messages) is exchanged by following edges and applying transformations defined by a graph attention \(^4\), which weighs neural messages by the similarity of latent vectors of nodes. By contrast, information between different entities (that is, inter-class neural messages) is propagated using a graph parsing neural network \(^4\) where the weights are adaptively learned \(^4\). Models can have multiple channels that reason over entities of the same class and share information across classes. For example, in HOI, relation parsing neural networks \(^4\) use a two-channel model where human and object-centric message passing is performed before mixing these representations for the final prediction (Fig. 3c). The same occurs in VQA, where visual, semantic and numeric channels perform independent message passing before sharing information via visual–semantic aggregation and semantic–numeric aggregation \(^4\text{4,45}\). Other neural architectures can serve as drop-in replacements to graph-based channels \(^4\text{4,45}\).

**MGL for language datasets**

With the ability to generate contextual language embeddings, language models have broadly reshaped analyses of natural language \(^7\). However, beyond words, structure in language exists at the level of sentences (syntax trees, dependency parsing), paragraphs (sentence-to-sentence...
for other language tasks, such as question answering, summarization to identify relations between words in a text, a capability important of text based on the usage and meaning of words (tokens). Graph most language analyses.

A LIG to represent this context is task specific. We describe these steps of language datasets.

Creating LIGs

At the highest level, a language dataset can be seen as a corpus of documents, then a single document, a group of sentences, a group of mentions, a group of entities and, finally, single words (Fig. 4a). MGL can consider these different levels of contextual information by constructing LIGs. The choice of context and how to include and how to create a LIG to represent this context is task specific. We describe these steps for text classification and relation extraction as these tasks underlie most language analyses.

In text classification, the model is asked to assign a label to a span of text based on the usage and meaning of words (tokens). Graph structure involving words is given by the relative position of words in a document or document cooccurrence. Relation extraction seeks to identify relations between words in a text, a capability important for other language tasks, such as question answering, summarization and knowledge graph reasoning. To capture sentence meaning, the structure among word entities is based on the underlying dependency tree. Beyond words, other entities are included to capture cross-sentence topology (Fig. 4a,b).

Learning on LIGs

Once a LIG is constructed, a model must be designed to learn on the LIG while incorporating inductive biases relevant to the language task. We illustrate strategies for learning on LIGs using aspect-based sentiment analysis (ABSA) as a downstream language task. ABSA assigns a sentiment (positive, negative) to a text or word/words or an aspect. Models must reason over syntactic structure and long-range relations between aspects and other words in the text to perform ABSA. To propagate information between distant words, aspect-specific GNNs mask non-aspect words in LIGs for long-range message passing. They also gate or perform element-wise multiplication between latent representations of query and aspect words. To include information about the syntactic structure, GNNs distinguish between the different types of relation in the dependency tree via type-specific message passing (Fig. 4c).

The sentiment of neighboring or similar sentences is essential to determine the aspect-based sentiment of the document. Cooperative
graph attention networks incorporate this via the cooperation between two graph-based modelling blocks: the inter- and intra-aspect modelling blocks (Fig. 4d). These blocks capture the relation of sentences to other sentences with the same aspect (intra-aspect) and to neighbouring sentences in the document that contain different aspects (inter-aspect). The outputs of the intra- and inter-aspect blocks are mixed in an interaction block, passing through a series of hidden layers. Finally, the intermediate representations between each hidden layer are fused via learned attention weights to create a final sentence representation (MGL component 4).

**MGL in natural sciences**
In addition to computer vision and language modelling, graphs are increasingly employed in the natural sciences. We call these graphs KIGs as they incorporate inductive biases relevant to a specific task or encode scientific knowledge in their structure.

**MGL in physics**
In particle physics, GNNs have been used to identify progenitor particles causing particle jets, sprays of particles that fly out from high-energy particle collisions. In these graphs, nodes are particles connected to their k-nearest neighbours. After rounds of message passing, aggregated node representations are used to identify progenitor particles.

Physics-informed GNNs have emerged as a promising approach for simulating physical systems governed by multiscale processes for which conventional methods fail. A typical goal is to discover hidden physics from available experimental data. GNNs are trained from available experimental data and information obtained by employing the physical laws and are then evaluated at points in the space–time domain. Such physics-informed architectures integrate multimodal data with mathematical models. For example, GNNs can express differential operators of the underlying dynamics as functions on nodes and edges. GNNs can also represent physical interactions between objects, such as particles in a fluid, joints in a robot and points in a power grid. Initial node representations describe the initial state of these particles and global constants such as gravity with edges indicating relative particle velocity. Message passing updates edge representations first to calculate the effect of relative forces in the system. It then uses the updated edge representation to update node representations and calculate the new state of particles as a result of the forces (Fig. 5a). This message-passing strategy advances the MGL’s third component (MGL for images) and has also been employed to solve combinatorial algorithms (Bellman–Ford and Prim’s algorithms) and chip floorplanning to design the physical layout of computer chips.

**MGL in chemistry**
In chemistry, MGL methods can predict intra- and intermolecular properties from the primary molecular structure by performing message passing on molecular graphs of atoms linked by bonds. Present efforts incorporate 3D spatial molecular information in addition to two-dimensional molecular details. When this information is unavailable, the MGL methods consider stereochemistry to aggregate neural messages and model molecules as sets of chemical substructures in addition to granular atom representations.

Stereoisomers are molecules with the same graph connectivity but different spatial arrangements. Aggregation functions in molecular graphs aggregate the same regardless of the orientation of atoms in 3D space. This can lead to poor performance, as stereoisomers can have different properties. To mitigate this issue, permutation (PERM) and permutation-concatenation (PERM-CAT) aggregation update every atom in a chiral group via a weighted sum of every permutation of its respective chiral group. Although the identity of the neighbours is the same in every permutation, the spatial arrangement varies. By weighing each permutation, PERM and PERM-CAT encode this inductive bias by modifying how information is propagated in the underlying graph (MGL component 3).

Moreover, MGL can help identify chemical products produced by molecules through reactions. For example, to predict whether two molecules react, QM-GNN, a quantum chemistry-augmented GNN, represents each reactant by its molecular graph with chemistry-informed initial representations for every atom and bond. After rounds of message passing, the atom representations are updated through a global attention mechanism. The attention mechanism uncovers a novel topology where atoms can interact with atoms on other molecules. It incorporates a principle from chemistry that

---

**Box 1**

**The blueprint for MGL**

The blueprint for graph-centric multimodal learning has four components.

1. Identifying entities. Information from different sources is combined and projected into a shared namespace. Nodes are identified independently as set elements, and no interactions are given yet. Let there be k modalities \( \mathcal{C} = \{ C_1, \ldots, C_k \} \), where \( C_k \) is an information matrix of the \( k \)th modality that describes every entity by an information vector. We define Identify, module for every modality \( i \) as:

\[
X_i \leftarrow \text{Identify}(C_i),
\]

which maps information of all modalities into the same namespace. If \( k=1 \), we get a reduced unimodal variant of MGL.

2. Uncovering topology. Let there be data modalities \( \mathcal{X} = \{ X_1, \ldots, X_m \} \). We define Connect, modules specify predefined neighbourhoods.

3. Propagating information. Neural messages are exchanged along edges in the adjacency matrices \( \mathcal{A} = \{ A_1, \ldots, A_m \} \) to produce node representations:

\[
H \leftarrow \text{Propagate}(\mathcal{X}, \mathcal{A}).
\]

When multiple adjacency matrices are given, the Propagate module can specify multiple independent propagation models (Supplementary Note 1) or operate on a combined adjacency matrix.

4. Mixing representations. Representations are mixed and transformed into latent representations optimized for a downstream task:

\[
Z \leftarrow \text{Mix}(H, \mathcal{A}).
\]

The mixing module Mix transforms node representations into final representations of entities \( Z \) on which downstream tasks are defined on. Established strategies to mix representations include aggregation operators, such as summation, averaging, multi-hop aggregation and methods using adjacency information \( \mathcal{A} \).
intermolecular interactions between particles inform reactivity. The final representations are combined with descriptors, such as atomic charges and bond lengths, and used for prediction. Such an approach integrates structural knowledge about molecules in a GNN with relevant chemistry knowledge, allowing for accurate prediction on small training datasets\cite{henderson2020structure}. The inclusion of domain knowledge by fusing GNN outputs illustrates the Mix module in MGL (Box 1). Graph learning on molecules created new opportunities for virtual drug screening\cite{cho2020molecular, deng2021molecular}, molecule generation and design\cite{li2020molecular, li2021molecular}, and drug target identification\cite{park2020drug, park2021drug}.

**MGL in biology**

Beyond individual molecules, MGL can help understand the properties of complex structures across multiple scales, the most pertinent of these structures being proteins. At the primary amino acid sequence scale, the hallmark task predicts the 3D structure from the amino acid sequence. AlphaFold constructs a KIG where nodes are amino acids with representations derived from sequence homology\cite{norville2022alphafold}. To propagate information in this KIG, AlphaFold introduces a triangle multiplicative update and triangle self-attention update. These triangle modifications integrate the inductive bias that learned representations must abide by the triangle inequality on distances to represent 3D structures. MGL, among other innovations, enabled AlphaFold to predict 3D protein structure from the amino acid sequence\cite{norville2022alphafold}.

Beyond 3D structure, molecular protein surfaces mediate critical roles in cellular function and disease, and thus modelling geometric and physical protein properties is essential\cite{zhang2022alphafold, norville2022alphafold}. For example, MaSIF\cite{norville2022alphafold} trains a GNN on molecular surfaces described as multimodal graphs to predict protein–protein interactions\cite{norville2022alphafold}. This robust topology provides a better prediction on tasks such as protein–ligand binding-affinity prediction.

**Outlook**

MGL is an emerging field with applications across natural sciences, vision and language domains. We anticipate the growth in MGL will be driven by fully multimodal graph architectures and new uses in the natural sciences and medicine. We also outline applications to understand when MGL is valuable or unhelpful and needs improvements to resolve challenges represented by multimodal inductive biases or a lack of explicit graphs.

**Fully multimodal graph architectures**

Prevailing approaches use domain-specialized architectures tailored to each data modality. However, advances in general-purpose architectures provide an expressive strategy to consider dependencies between modalities irrespective of whether they are given as images, language sequences, graphs or tabular datasets. Moreover, the MGL blueprint supports more complex graph structures, such as hypergraphs\cite{zhang2022alphafold, norville2022alphafold} and heterogeneous graphs\cite{zhang2022alphafold, norville2022alphafold}.

The blueprint can also pave the way for novel uses of graph-centric multimodal learning. For example, knowledge distillation aims to transfer knowledge from a teacher model to a smaller student model in a way that preserves performance while using...
fewer resources. KIGs\textsuperscript{119–121} can be used to design more efficient knowledge-distillation loss functions\textsuperscript{122,123}. In another example, visible neural networks specify the architecture such that nodes correspond to concepts (for example, molecules, pathways) at different scales of the cellular system, ranging from small complexes to extensive signal pathways\textsuperscript{2,324}, connected based on biological relationships, used in forward- and back-propagation. By incorporating such inductive biases, models can be trained in a data-efficient manner as they do not have to invent relevant fundamental principles but can know these from the start and thus need fewer data for training. Harmonizing algorithm design with domain knowledge can also improve model interpretability.

Algorithmic improvements to resolve multimodal challenges
Existing methods are limited in areas without prior knowledge or relational structure. For example, in tasks such as chemical reaction prediction\textsuperscript{104}, progenitor particle classification\textsuperscript{105}, physical interaction simulation\textsuperscript{106} and protein–ligand modelling\textsuperscript{107}, interactions relevant for the task are not a priori given, meaning that the methods must automatically capture novel, unspecified and relevant interactions. Some applications use node feature similarity to dynamically construct local adjacencies after each layer to discover new interactions\textsuperscript{108}. However, this cannot capture novel interactions among distant nodes as information is only passed among closely connected nodes in message passing. Methods address this limitation by incorporating attention layers with induced sparsity to discover these interactions\textsuperscript{109}. In applications without strong relational structure, such as molecular property prediction\textsuperscript{109–111}, particle classification\textsuperscript{112} and text classification\textsuperscript{113}, node features often have more predictive value than any encoded structure. As a result, other methods have been shown to lead to better performance than graph-based methods\textsuperscript{114,115}.

Ground-breaking applications in natural sciences and medicine
Using deep learning in natural sciences revealed the power of graph representations for modelling small and large molecular structures. Combining different types of data can create bridges between the molecular and organism levels for modelling physical, chemical or biological phenomena on a large scale. Recent knowledge graph applications have been introduced to enable precise medicine and make predictions across genomic, pharmaceutical and clinical applications\textsuperscript{116,117}. Multiscale learning systems are becoming valuable tools for protein structure prediction\textsuperscript{118–120}, protein property prediction\textsuperscript{120} and biomolecular interaction modelling\textsuperscript{121}. These methods can incorporate mathematical statements of physical relationships, knowledge graphs, prior distributions and constraints by modelling predefined graph structures or modifying message-passing algorithms. When such information exists, multimodal learning can enhance image denoising\textsuperscript{122}, image restoration\textsuperscript{123} and human–object interaction\textsuperscript{124} in vision systems.

Data availability
We summarize MGL methods and provide a continually updated summary at https://yashaektefaie.github.io/mlg. We host a live table where MGL methods are added to provide an evolving resource for the community.

References
1. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. 23, 40–55 (2022).
2. Yu, M. K. et al. Visible machine learning for biomedicine. Cell 173, 1562–1565 (2018).
3. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. Chem. Sci. 9, 513–530 (2017).
4. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In Proc. 34th International Conference on Machine Learning: Proc. Machine Learning Research Vol. 70 (eds Precup, D. & Teh, Y. W.) 1263–1272 (PMLR, 2017).
5. Sanchez-Gonzalez, A. et al. Graph networks as learnable physics engines for inference and control. In Proc. 35th International Conference on Machine Learning: Proc. Machine Learning Research Vol. 80 (eds Dy, J. & Krause, A.) 4470–4479 (PMLR, 2018).
6. Sanchez-Gonzalez, A. et al. Learning to simulate complex physics with graph networks. In Proc. 37th International Conference on Machine Learning: Proc. Machine Learning Research Vol. 119 (eds Daumé, H. III & Singh, A.) 8459–8468 (PMLR, 2020).
7. Liu, Q., Kusner, M. J. & Blunsom, P. A survey on contextual embeddings. Preprint at https://arxiv.org/abs/2003.07278 (2020).
8. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. IEEE Trans. Neural Netw. 20, 61–80 (2009).
9. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In Proc. 5th International Conference on Learning Representations (2017).
10. Kipf, T. N. & Welling, M. Variational graph auto-encoders. In NIPS Workshop on Bayesian Deep Learning (2016).
11. Grover, A., Zweig, A. & Emron, S. Graphite: iterative generative modeling of graphs. In Proc. 36th International Conference on Machine Learning: Proc. Machine Learning Research Vol. 972 (eds Chaudhuri, K. & Salakhutdinov, R.) 434–2444 (PMLR, 2019).
12. Guo, X. & Zhao, L. A systematic survey on deep generative models for graph generation. Preprint at https://arxiv.org/abs/2007.06686 (2020).
13. Baltrusaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41, 423–443 (2019).
14. Hong, C., Yu, J., Wan, J., Tao, D. & Wang, M. Multimodal deep autoencoder for human pose recovery. IEEE Trans. Image Process. 24, 5659–5670 (2015).
15. Khattar, D., Goud, J. S., Gupta, M. & Varma, V. MVAE: multimodal variational autoencoder for fake news detection. In The World Wide Web Conference 2915–2921 (Association for Computing Machinery, 2019).
16. Mao, J., Xu, J., Jing, Y. & Yuille, A. Training and evaluating multimodal word embeddings with large-scale web annotated images. In Proc. 30th International Conference on Neural Information Processing Systems 442–450 (Curran Associates, 2016).
17. Huang, Y., Lin, J., Zhou, C., Yang, H. & Huang, L. Modality competition: what makes joint training of multi-modal network fail in deep learning? (Provably). In Proc. 39th International Conference on Machine Learning: Proc. Machine Learning Research Vol. 162 (eds Chaudhuri, K. et al.) 9226–9259 (PMLR, 2022).
18. Xu, P., Zhu, X. & Clifton, D. A. Multimodal learning with transformers: a survey. Preprint at https://arxiv.org/abs/2206.06488 (2022).
19. Bayoudh, K., Knani, R., Hamdaoui, F. & Mtibaa, A. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. Vis.Comput. 38, 2939–2970 (2020).
20. Zhang, C., Yang, Z., He, X. & Deng, L. Multimodal intelligence: representation learning, information fusion, and applications. IEEE J. Sel. Top. Signal Process. 14, 478–493 (2020).
21. Javaloy, A., Meghdadi, M. & Valera, I. Mitigating modality collapse in multimodal VAEs via impartial optimization. In Proc. 39th International Conference on Machine Learning: Proc. Machine Learning Research Vol. 162 (eds Chaudhuri, K. et al.) 9938–9964 (PMLR, 2022).
22. Ma, M. et al. SMIL: multimodal learning with severely missing modality. Proc. AAAI Conf. Artif. Intell. 35, 2302–2310 (2021).
23. Pokluwar, P. et al. Geometric multimodal contrastive representation learning. In Proc. Mach. Learn. Res. 162, 17782–17800 (2022).
24. Zitnik, M. et al. Machine learning for integrating data in biology: principles, practice, and opportunities. Inf. Fusion 50, 71–91 (2019).
25. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
26. Somnath, V. R., Bunne, C. & Krause, A. Multi-scale representation learning on proteins. Adv. Neural Inf. Process. Syst. 34 (2021).
27. Walters, W. P. & Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. Acc. Chem. Res. 54, 263–270 (2021).
28. Wang, J., Hu, J., Qian, S., Fang, Q. & Xu, C. Multimodal graph convolutional networks for high quality content recognition. Neurocomputing 412, 42–51 (2020).
29. Mai, S., Hu, H. & Xing, S. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion. Proc. AAAI Conf. Artif. Intell. 34, 164–172 (2020).
30. Zhang, X., Zeman, M., Tsiligkaridis, T. & Zitnik, M. Graph-guided network for irregularly sampled multivariate time series. In International Conference on Learning Representations (2022).
31. Zheng, S. et al. Multi-modal graph learning for disease prediction. In IEEE Trans. Med. Imaging 41, 2207–2216 (2022).
32. Ramachandram, D. & Taylor, G. W. Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process. Mag. 34, 96–108 (2017).
33. Ngiam, J. et al. Multimodal deep learning. In Proc. 28th International Conference on International Conference on Machine Learning 689–696 (Omnipress, 2011).
34. Aafaq, N., Akhtar, N., Liu, W., Gilani, S. Z. & Mian, A. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 12487–12496 (IEEE, 2019).
35. Fang, Z., Gokhale, T., Banerjee, P., Baral, C. & Yang, Y. Video2Commonsense: generating commonsense descriptions to enrich video captioning. In Proc. 2020 Conference on Empirical Methods in Natural Language Processing 840–860 (Association for Computational Linguistics, 2020).
36. Kiros, R., Salakhutdinov, R. & Zemel, R. Multimodal neural language models. In Proc. 31st International Conference on Machine Learning: Proc. Machine Learning Research Vol. 32 (eds Xing, E. P. & Jebara, T.) 595–603 (PMLR, 2014).
37. Rezaei-Shoshtari, S., Hogan, F. R., Jenkin, M., Meger, D. & Dudek, G. Learning intuitive physics with multimodal generative models. Proc. AAAI Conf. Artif. Intell. 35, 610–6118 (2021).
38. Bronstein, M. M., Bruna, J., Cohen, T. & Velicković, P. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. Preprint at https://arxiv.org/abs/2104.13478 (2021).
39. Chen, Y. et al. Graph-based global reasoning networks. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 433–442 (IEEE, 2019).
40. Varga, V. & Lorincz, A. Fast interactive video object segmentation with graph neural networks. In International Joint Conference on Neural Networks 1–10 (IEEE, 2021).
41. Liu, Q., Kampffmeyer, M., Jenssen, R. & Salberg, A.-B. Self-constructing graph neural networks to model long-range pixel dependencies for semantic segmentation of remote sensing images. Int. J. Remote Sensing 42, 6184–6208 (2021).
42. Zhou, S., Zhang, J., Zhuo, W. & Loy, C. C. Cross-scale internal graph neural network for image super-resolution. Adv. Neural Inf. Process. Syst. 33, 3499–3509 (2020).
43. Mou, C. & Zhang, J. Graph attention neural network for image restoration. In 2021 IEEE International Conference on Multimedia and Expo 1–6 (IEEE, 2021).
44. Qi, S., Wang, W., Jia, B., Shen, J. & Zhu, S. C. Learning human-object interactions by graph parsing neural networks. In Computer Vision – ECCV 2018 407–423 (Springer, 2018).
45. Wang, H., Zheng, W.-s. & Yingbiao, L. Contextual heterogeneous graph network for human–object interaction detection. In Computer Vision—ECCV 2020: Proc. 16th European Conference Part XVII 248–264 (Springer-Verlag, 2020).
46. Avelar, P. C., Tavares, A. R., da Silveira, T. T., Jung, C. R. & Lamb, L. C. Superpixel image classification with graph attention networks. In 33rd SIGGRAPH Conference on Graphics, Patterns and Images 203–209 (IEEE Computer Society, 2020).
47. Lu, Y., Chen, Y., Zhao, D. & Chen, J. In Advances in Neural Networks: Lecture Notes in Computer Science Vol. 11554 (eds Lu, H. et al.) 97–105 (Springer, 2019).
48. Kim, J., Lee, J. K. & Lee, K. M. Deeply-recursive convolutional network for image super-resolution. In Proc. IEEE Conference on Computer Vision and Pattern Recognition 1637–1645 (IEEE, 2016).
49. Achanta, R. et al. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2274–2282 (2012).
50. Zeng, H., Liu, Q., Zhang, M., Han, X. & Wang, Y. Semi-supervised hyperspectral image classification with graph clustering convolutional networks. Preprint at https://arxiv.org/abs/2012.10932 (2020).
51. Wang, S. et al. Multiscale dynamic graph convolutional network for hyperspectral image classification. IEEE Trans. Geosc. Remote Sensing 58, 3162–3177 (2019).
52. Long, J., Shkelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition 3431–3440 (IEEE, 2015).
53. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proc. 3rd International Conference on Learning Representations (eds Bengio, Y. & LeCun, Y.) (2015).
54. Knyazev, B., Lin, X., Amer, M. R. & Taylor, G. W. Image classification with hierarchical multigraph networks. In British Machine Vision Conference (2019).
55. Velickovic, P. et al. Graph attention networks. In International Conference on Learning Representations (2018).
56. Balles, K., Nüsken, N., Schlögl, A. & Windisch, C. Hierarchical variational graph network for image denoising. In IEEE Trans. Image Process. 2226–2237 (2020).
57. Bresson, X. & Laurent, T. Residual gated graph ConvNets. Preprint at https://arxiv.org/abs/1711.07533 (2017).
58. Biten, A. F. et al. Scene text visual question answering. In Proc. IEEE/CVF International Conference on Computer Vision 4291–4301 (2019).
59. Singh, A. et al. Towards VQA models that can read. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 8317–8326 (IEEE, 2019).
60. Liu, C. et al. Graph structured network for image-text matching. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 10921–10930 (IEEE, 2020).
61. Zhang, F. Z., Campbell, D. & Gould, S. Spatially conditioned graphs for detecting human–object interactions. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 13319–13327 (IEEE, 2021).
63. Ulutan, O., Iftekhar, A. S. M. & Manjunath, B. S. VSGNet: spatial attention network for detecting human object interactions using graph convolutions. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 13617–13626 (IEEE, 2020).

64. Gao, C., Xu, J., Zou, Y. & Huang, J.-B. DRG: Dual relation graph for human–object interaction detection. In Computer Vision—ECCV 2020 (eds Vedaldi, A. et al.) 696–712 (Springer, 2020).

65. Zhou, P. & Chi, M. Relation passing neural network for human–object interaction detection. In Proc. IEEE/CVF International Conference on Computer Vision 843–851 (IEEE, 2019).

66. Gao, D., Li, K., Wang, R., Shan, S. & Chen, X. Multi-modal graph neural network for joint reasoning on vision and scene text. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 12746–12756 (IEEE, 2020).

67. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149 (2016).

68. Wu, Y. & Zhang, P. GINet: graph interaction network for scene parsing. In Computer Vision—ECCV 2020 (eds Vedaldi, A. et al.) 34–51 (Springer, 2020).

69. Wu, L. et al. Graph neural networks for natural language processing: a survey. Preprint at https://arxiv.org/abs/2106.06090 (2021).

70. Vaswani, A. et al. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, 5998–6008 (2017).

71. Li, I., Li, T., Li, Y., Dong, R. & Suzumura, T. Heterogeneous graph neural networks for multi-label text classification. Preprint at https://arxiv.org/abs/2103.14620 (2021).

72. Huang, L., Ma, D., Li, S., Zhang, X. & Wang, H. Text level graph neural network for text classification. In Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing 3444–3450 (Association for Computational Linguistics, 2019).

73. Zhang, Y. et al. Every document owns its structure: inductive text classification via graph neural networks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 334–339 (Association for Computational Linguistics, 2020).

74. Pan, J., Peng, M. & Zhang, Y. Mention-centered graph neural network for document-level relation extraction. Preprint at https://arxiv.org/abs/2103.08200 (2021).

75. Zhu, H. et al. Graph neural networks with generated parameters for relation extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics 1331–1339 (Association for Computational Linguistics, 2019).

76. Guo, Z., Zhang, Y. & Lu, W. Attention guided graph convolutional networks for relation extraction. In Proc. 57th Annual Meeting of the Association for Computational Linguistics 241–251 (Association for Computational Linguistics, 2019).

77. Zeng, S., Xu, R., Chang, B. & Li, L. Double graph based reasoning for document-level relation extraction. In Proc. 2020 Conference on Empirical Methods in Natural Language Processing 1630–1640 (Association for Computational Linguistics, 2020).

78. Chen, X. et al. Aspect sentiment classification with document-level sentiment preference modeling. In Proc. 58th Annual Meeting of the Association for Computational Linguistics 3667–3677 (Association for Computational Linguistics, 2020).

79. Zhang, C., Li, Q. & Song, D. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing 4568–4578 (Association for Computational Linguistics, 2019).

80. Zhang, M. & Qian, T. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In Proc. 2020 Conference on Empirical Methods in Natural Language Processing 3540–3549 (Association for Computational Linguistics, 2020).

81. Pouran Ben Veyseh, A. et al. Improving aspect-based sentiment analysis with gated graph convolutional networks and syntax-based regulation. In Findings of the Association for Computational Linguistics: EMNLP 2020 4543–4548 (Association for Computational Linguistics, 2020).

82. Shlomi, J., Battaglia, P. & Vlimant, J.-R. Graph neural networks in particle physics. Mach. Learn. Sci. Technol. 2, 021001 (2021).

83. Henrion, I. et al. Neural message passing for jet physics. In Deep Learning for Physical Sciences Workshop at the 31st Conference on Neural Information Processing Systems (2017).

84. Qasim, S. R., Kieseler, J., Iiyama, Y. & Pierini, M. Learning representations of irregular particle-detector geometry with distance-weighted graph networks. Eur. Phys. J. C 79, 608 (2019).

85. Mikuni, V. & Canelli, F. ABCNet: an attention-based method for particle tagging. Eur. Phys. J. Plus 135, 463 (2020).

86. Ju, X. et al. Graph neural networks for particle reconstruction in high energy physics detectors. Preprint at https://arxiv.org/abs/2003.11603 (2020).

87. Shukla, K., Xu, M., Trask, N. & Kamiadakis, G. E. Scalable algorithms for physics-informed neural and graph networks. Data Centric Eng. 3, e24 (2022).

88. Seo, S. & Liu, Y. Differentiable physics-informed graph networks. Preprint at https://arxiv.org/abs/1902.02950 (2019).

89. Li, W. & Deka, D. Physics based GNNs for locating faults in power grids. Preprint at https://arxiv.org/abs/2107.02275 (2021).

90. Battaglia, P. W. et al. Relational inductive biases, deep learning, and graph networks. Preprint at https://arxiv.org/abs/1806.01261 (2018).

91. Velickovic, P., Ying, R., Pavolov, M., Hasdell, R. & Blundell, C. Neural execution of graph algorithms. In International Conference on Learning Representations (2020).

92. Schuetz, M. J. A., Brubaker, J. K. & Katzgraber, H. G. Combinatorial optimization with physics-inspired graph neural networks. Nat. Mach. Intell. 4, 367–377 (2022).

93. Mirhoseini, A. et al. A graph placement methodology for fast chip design. Nature 594, 207–212 (2021).

94. Gasteiger, J., Gross, J. & Günnemann, S. Directional message passing for molecular graphs. In International Conference on Learning Representations (2020).

95. Jørgensen, P. B., Jacobsen, K. W. & Schmidt, M. N. Neural message passing with edge updates for predicting properties of molecules and materials. Preprint at https://arxiv.org/abs/1806.03146 (2018).

96. Gasteiger, J., Yeshwanth, C. & Günnemann, S. Directional message passing on molecular graphs via synthetic coordinates. Adv. Neural Inf. Process. Syst. 34, 15421–15433 (2021).

97. Liu, M. et al. Fast quantum property prediction via deeper 2D and 3D graph networks. Adv. for Science Workshop (NeurIPS, 2021).

98. St. John, P. C., Guan, Y., Kim, Y., Kim, S. & Paton, R. S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. Nat. Commun. 11, 2328 (2020).

99. Pattanaik, L. et al. Message passing networks for molecules with tetrahedral chirality. Preprint at https://arxiv.org/abs/2012.00094 (2020).

100. Foy, M., Yuen, J.-G. & Weichert, F. Hierarchical inter-message passing for learning on molecular graphs. Preprint at https://arxiv.org/abs/2006.12179 (2020).

101. Ariëns, E. Chirality in bioactive agents and its pitfalls. Trends Pharmacol. Sci. 7, 200–205 (1986).
102. Guan, Y. et al. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. Chem. Sci. 12, 2198–2208 (2021).
103. Coley, C. W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem. Sci. 10, 370–377 (2019).
104. Struble, T. J., Coley, C. W. & Jensen, K. F. Multitask prediction of site selectivity in aromatic C–H functionalization reactions. React. Chem. Eng. 5, 896–902 (2020).
105. Stuyver, T. & Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: performance, generalizability, and explainability. J. Chem. Phys. 156, 084104 (2022).
106. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. Cell 180, 689–702.e13 (2020).
107. Fu, T. et al. Differentiable scaffolding tree for molecule optimization. In International Conference on Learning Representations (2022).
108. Mercado, R. et al. Graph networks for molecular design. Mach. Learn. Sci. Technol. 2, 025023 (2021).
109. Torng, W. & Altman, R. B. Graph convolutional neural networks for predicting drug–target interactions. J. Chem. Inf. Model. 59, 4131–4149 (2019).
110. Moon, S., Zhung, W., Yang, S., Lim, J. & Kim, W. Y. PIGMnet: a physics-informed deep learning model toward generalized drug-target interaction predictions. Chem. Sci. 13, 3661–3673 (2022).
111. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat. Methods 17, 184–192 (2020).
112. Sanner, M. F., Olson, A. J. & Spehner, J.-C. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38, 305–320 (1996).
113. Sverrisson, F., Feydy, J., Correia, B. E. & Bronstein, M. M. Fast end-to-end learning on protein surfaces. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 15272–15281 (IEEE, 2021).
114. Feng, Y., You, H., Zhang, Z., Ji, R. & Gao, Y. Hypergraph neural networks. Proc. AAAI Conf. Artif. Intell. 33, 3558–3565 (2019).
115. Srinivasan, B., Zheng, D. & Karypis, G. Learning over Families of Sets—Hypergraph Representation Learning for Higher Order Tasks 756–764 (SIAM Activity Group on Data Science, 2021).
116. Jo, J. et al. Edge representation learning with hypergraphs. Adv. Neural Inf. Process. Syst. 34, 7534–7546 (2021).
117. Zhang, C., Song, D., Huang, C., Swami, A. & Chawla, N. V. Heterogeneous graph neural network. In Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 793–803 (Association for Computing Machinery, 2019).
118. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. Sci. Data 10, 67 (2023).
119. Lee, S. & Song, B. C. Graph-based knowledge distillation by multi-head attention network. In Proc. British Machine Vision Conference (eds Sidorov, K. & Hicks, Y.) 162.1–162.12 (BMVA, 2019).
120. Zhou, S. et al. Distilling holistic knowledge with graph neural networks. In Proc. IEEE/CVF International Conference on Computer Vision 10387–10396 (IEEE, 2021).
121. Sun, L., Gou, J., Yu, B., Du, L. & Tao, D. Collaborative teacher–student learning via multiple knowledge transfer. Preprint at https://arxiv.org/abs/2101.08471 (2021).
122. Park, W., Kim, D., Lu, Y. & Cho, M. Relational knowledge distillation. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 3967–3976 (IEEE, 2019).
123. Liu, Y. et al. Knowledge distillation via instance relationship graph. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 7096–7104 (IEEE, 2019).
124. Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods 15, 290–298 (2018).
125. Borisov, V. et al. Deep neural networks and tabular data: a survey. in IEEE Transactions on Neural Networks and Learning Systems https://doi.org/10.1109/TNNLS.2022.3229161 (2022).
126. Jiang, D. et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J. Cheminform. 13, 12 (2021).
127. Nicholson, D. N. & Greene, C. S. Constructing knowledge graphs and their biomedical applications. Comput. Struct. Biotechnol. J. 18, 1414–1428 (2020).
128. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In International Conference on Learning Representations (2019).
129. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. 30, 1024–1034 (2017).
130. Xu, K. et al. Representation learning on graphs with jumping knowledge networks. In Proc. 35th International Conference on Machine Learning: Proc. Machine Learning Research Vol. 80 (eds Dy, J. & Krause, A.) 5453–5462 (PMLR, 2018).

Acknowledgements
Y.E., G.D. and M.Z. gratefully acknowledge the support of US Air Force Contract No. FA8702-15-D-0001, and awards from Harvard Data Science Initiative, Amazon Research, Bayer Early Excellence in Science, AstraZeneca Research and Roche Alliance with Distinguished Scientists. Y.E. is supported by grant T32 HG002295 from the National Human Genome Research Institute and the NSDEG fellowship. G.D. is supported by the Harvard Data Science Initiative Postdoctoral Fellowship. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00624-6.

Correspondence should be addressed to Marinka Zitnik.

Peer review information Nature Machine Intelligence thanks Jianzhu Ma, Ying Ding and Shuiwang Ji for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.
© Springer Nature Limited 2023