Multimodal learning with graphs

Yasha Ektefaie[1,2,*], **George Dasoulas**[2,5,*], **Ayush Noori**[2,3], **Maha Farhat**[2,6], **Marinka Zitnik**[2,4,5,‡]

[1]Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA 02115, USA

[2]Department of Biomedical Informatics, Harvard University, Boston, MA 02115, USA

[3]Harvard College, Cambridge, MA 02138, USA

[4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[5]Harvard Data Science Initiative, Cambridge, MA 02138, USA

[6]Division of Pulmonary and Critical Care, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

Abstract

Artificial intelligence for graphs has achieved remarkable success in modeling 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—the set of assumptions that algorithms use to make predictions for inputs they have not encountered during training. Learning on multimodal datasets presents fundamental challenges because the inductive biases can vary by data modality and graphs might not be explicitly given in the input. To address these challenges, multimodal graph AI methods combine different modalities while leveraging cross-modal dependencies using graphs. 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.

1 Introduction

Deep learning on graphs has contributed to breakthroughs in biology [1, 2], chemistry [3, 4], physics [5, 6], and the social sciences [7]. The predominant use of graph neural networks [8] 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 [9], self-supervised learning [10], and graph design and generation [11, 12]. In most existing

[‡]Correspondence: marinka@hms.harvard.edu.
[*]Equal contribution

applications, datasets explicitly describe graphs in the form of nodes, edges, and additional information representing contextual knowledge, such as node, edge, and graph attributes.

Modeling 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 [13] 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 less labeled data. While multimodal learning has been successfully used in settings where unimodal methods fail [14, 15, 16], it presents several challenges that must be overcome to enable its broad use in AI [13, 17]. These challenges include finding representations optimized for machine learning analyses and fusing combined information from various modalities to create predictive models [18, 19, 20]. 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* [21]. 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* [22, 23]. Because different modalities can lead to intricate relational dependencies, simple modality fusion cannot fully leverage multimodal datasets [24]. Graph learning models such data systems [25, 26, 27] by connecting data points in different modalities as edges in optimally defined graphs [28, 29, 30] and building learning systems for a wide range of tasks [31, 32].

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 leveraging graphs. This framework allows for learning fused graph representations and studying the aforementioned challenges of modality collapse and missing modalities [18, 13]. We apply this formulation across a broad spectrum of domains, ranging from computer vision and language processing to the natural sciences (Figure 1). We consider image-intensive graphs (IIG) for image and video reasoning (Section 3), language-intensive graphs (LIG) for processing natural and biological sequences (Section 4), and knowledge-intensive graphs (KIG) used to aid in scientific discovery (Section 5).

2   Graph Neural Networks for Multimodal Learning

Deep learning has created a wide range of fusion approaches for multimodal learning [33, 34]. For example, recurrent neural network (RNN) and convolutional neural network (CNN) architectures have successfully been combined to fuse sound and image representations in video description problems [35, 36]. More recently, generative models have also proven very accurate for both language-dependent [37] and physics-based multimodal data [38]. 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.

2.1   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 *multimodal graph learning* (MGL). MGL can be seen as a blueprint consisting of four learning components that are connected in an end-to-end fashion. In Figure 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 (SL) phase (Figure 2c) :

Component 1: Identifying Entities.—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 (Figure 3), entity identification entails defining superpixels in an image.

Component 2: Uncovering Topology.—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 (*e.g.*, in Figure 5c, the Uncovering Topology component corresponds to combining protein surface information with the protein structure itself). When the data does not have an *a priori* network structure, the uncovering topology component explores possible adjacency matrices based on explicit (*e.g.*, spatial and visual characteristics) or implicit (*e.g.*, 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 (Figure 4b).

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

Component 3: Propagating Information.—The third component employs convolutional or message-passing steps to learn node representations based on graph adjacencies (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.

Component 4: Mixing Representations.—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 (*e.g.*, a graph-level or a subgraph-level label). Popular mixing strategies include simple aggregation operators (*e.g.*, summation or averaging) or more sophisticated functions that incorporate neural network architectures.

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 multimodal graph learning methods are in Supplementary Note 2.

3 Multimodal Graph Learning for Images

Image-intensive graphs (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 [42]. 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. In 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 [42]. In addition, GNNs can model long-range dependencies of arbitrary shape that are important for image-related tasks [43] such as image segmentation [44, 45], image restoration [46, 47], or human object interaction [48, 49].

Visual Comprehension

Visual comprehension remains a cornerstone of visual analyses, where multimodal graph learning has proven helpful in classifying, segmenting, and enhancing images. Image classification identifies the set of object categories present in an image [51]. In contrast, image segmentation divides an image into segments and assigns each segment into a category [52, 44, 45]. Finally, image restoration and denoising transform low-quality images into high-resolution counterparts [53]. The information required for these tasks lies in objects, segments, and image patches, as well as in the long-range context surrounding them [52].

IIG construction (corresponding to MGL Components 1 and 2) begins with a segmentation algorithm such as simple linear iterative clustering (SLIC) [54] to identify meaningful regions [55, 56, 44] (Figure 3a). These regions define nodes used to extract feature maps and summary visual features for each region [45, 52], whose attributes are initialized from CNNs like FCN-16 [57] or VGG19 [58]. Moreover, the nodes are connected to their k nearest neighbors in the CNN learned feature space [55, 45, 46, 47] (Figure 3b), to spatially adjacent

regions [51, 59, 44, 56], or to an arbitrary number of neighbors based on a previously defined similarity threshold between nodes [47, 56].

Once the SL phase of MGL is completed, propagation models (MGL Component 3) based on graph convolutions [52, 59, 56, 45] and graph attention [60] (GAT) are used to weigh node neighbors in the graph based on learned attention scores [51, 47]. In addition, methods such as graph denoiser networks (GCDNs) [61], internal graph neural networks (IGNNs) [46], and residualGCNs [62, 44] 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 [48] (HOI) or, more broadly, visual, semantic, and numeric entities as in visual question answering [63, 64, 65] (VQA).

In HOI, the MGL methods identify two entities, human body parts (*e.g.*, hands, face, etc.) and objects (*e.g.*, surfboard, bike, etc.) [48, 50], that interact in fully connected [48, 49], bipartite [50, 66], or partially connected topologies [67, 68]. MGL methods for VQA construct a new topology [69] that spans interconnected visual, semantic, and numeric graphs. Entities represent visual objects identified by an extractor, such as Faster R-CNN [70], 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 (*i.e.*, intra-class neural messages) is exchanged by following edges and applying transformations defined by a GAT [60], which weighs neural messages by the similarity of latent vectors of nodes. In contrast, information between different entities (*i.e.*, inter-class neural messages) is propagated using a GPNN [48] where the weights are adaptively learned [49]. 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 [68] use a two-channel model where human and object-centric message passing is performed before mixing these representations for the final prediction (Figure 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 [69, 71]. Other neural architectures can serve as drop-in replacements to graph-based channels [66, 67].

4   Multimodal Graph Learning for Language

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 relations), and documents (paragraph-to-paragraph links) [72]. Transformers, a prevailing class of language models [73], can capture such structure but have strict

computational and data requirements. MGL methods mitigate these issues by infusing language structure into models. Specifically, these methods rely on language-intensive graphs (LIGs), explicit or implicit graphs where nodes represent semantic features linked by language dependencies.

Creating Language-Intensive Graphs

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 (Figure 4a). Multimodal graph learning can consider these different levels of contextual information by constructing LIGs. The choice of context 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 [74] based on the usage and meaning of words (tokens). Graph structure involving words is given by the relative position of words in a document [75, 74] or document cooccurrence [76]. 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 [77, 78]. To capture sentence meaning, the structure among word entities is based on the underlying dependency tree [79]. Beyond words, other entities are included to capture cross-sentence topology [77, 80] (Figure 4a–b).

Learning on Language-Intensive Graphs

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 [81]. ABSA assigns a sentiment (positive, negative) of a text to a word/words or an aspect [81]. Models must reason over syntactic structure and long-range relations between aspects and other words in the text to perform ABSA [82, 83]. To propagate information between distant words, aspect-specific GNNs mask non-aspect words in LIGs for long-range message passing [82]. They also gate or perform element-wise multiplication between latent representations of query and aspect words [84]. To include information about the syntactic structure, GNNs distinguish between the different types of relations in the dependency tree via type-specific message passing [82, 83, 84] (Figure 4c).

The sentiment of neighboring or similar sentences is essential to determine the aspect-based sentiment of the document [81]. Cooperative graph attention networks (CoGAN) incorporate this via the cooperation between two graph-based modeling blocks: the inter- and intra-aspect modeling blocks (Figure 4d) [81]. These blocks capture the relation of sentences to other sentences with the same aspect (intra-aspect) and to neighboring 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).

5 Multimodal Graph Learning in Natural Sciences

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

Multimodal Graph Learning 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 [85]. In these graphs, nodes are particles connected to their k-nearest neighbors. After rounds of message passing, aggregated node representations are used to identify progenitor particles [86, 87, 88, 89].

Physics-informed GNNs have emerged as a promising approach for simulating physical systems governed by multiscale processes for which conventional methods fail [90]. 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 [91]. GNNs can also represent physical interactions between objects, such as particles in a fluid [6], joints in a robot [5], and points in a power grid [92]. Initial node representations describe the initial state of these particles and global constants like gravity [6] with edges indicating relative particle velocity [5]. 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 [93] (Figure 5a). This message-passing strategy advances the MGL's third component (Section 2) and has also been employed to solve combinatorial algorithms (Bellman-Ford and Prim's algorithms) [94, 95] and chip floorplanning to design the physical layout of computer chips [96].

Multimodal Graph Learning in Chemistry

In chemistry, MGL methods can predict intra- and inter-molecular properties from the primary molecular structure by performing message passing on molecular graphs of atoms linked by bonds [4, 97, 98, 99, 100, 101]. Present efforts incorporate 3D spatial molecular information in addition to 2D molecular details. When this information is unavailable, the MGL methods [97, 99, 100] consider stereo-chemistry to aggregate neural messages [102] and model molecules as sets of chemical substructures in addition to granular atom representations [103].

Stereoisomers are molecules with the same graph connectivity but different spatial arrangements [102]. Aggregation functions in molecular graphs aggregate the same regardless of the orientation of atoms in three-dimensional space. This can lead to poor performance, as stereoisomers can have different properties [104]. To mitigate this issue, permutation (PERM) and permutation-concatenation (PERM-CAT) aggregation [102]

update every atom in a chiral group via a weighted sum of every permutation of its respective chiral group. Though the identity of the neighbors 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 [105, 106, 107, 108]. For example, to predict whether two molecules react, QM-GNN [105], 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 (Figure 5b). The attention mechanism uncovers a novel topology where atoms can interact with atoms on other molecules. It incorporates a principle from chemistry that 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 [105]. The inclusion of domain knowledge by fusing GNN outputs illustrates the Mix module in MGL (Section 2, Box 1). Graph learning on molecules created new opportunities for virtual drug screening [109], molecule generation and design [110, 111, 27], and drug target identification [112, 113].

Multimodal Graph Learning 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 [25]. 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. Multimodal graph learning, among other innovations, enabled AlphaFold to predict 3D protein structure from amino acid sequence [25].

Beyond 3D structure, molecular protein surfaces mediate critical roles in cellular function and disease, and thus modeling geometric and physical protein properties is essential [1, 114, 115]. For example, MaSIF [114] trains a GNN on molecular surfaces described as multimodal graphs to predict protein interactions. The initial representation of the nodes is based on geometric and chemical features. Next, Gaussian kernels are defined on every node to propagate information, encoding complex geometric shapes of molecular surfaces and extending the notion of a convolution. The final representations can be used to predict protein-protein interactions [114], structural configurations of protein complexes [116], and protein-ligand binding [26].

6 Outlook

Multimodal graph learning is an emerging field with applications across natural sciences, vision and language domains. We anticipate the growth in MGL 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 [117, 118, 119] and heterogeneous graphs [120, 121].

The blueprint can also pave the way for novel uses of graph-centric multimodal learning. For example, knowledge distillation (KD) aims to transfer knowledge from a teacher model to a smaller student model in a way that preserves performance while using fewer resources. Knowledge-intensive graphs [122, 123, 124] can be used to design more efficient KD loss functions [125, 126]. In another example, visible neural networks specify the architecture such that nodes correspond to concepts (*e.g.*, molecules, pathways) at different scales of the cellular system, ranging from small complexes to extensive signal pathways [2, 127], 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 [105], progenitor particle classification [85], physical interaction simulation [6], and protein-ligand modeling [114], 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 [85]. However, this cannot capture novel interactions among distant nodes since 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 [105]. In applications without strong relational structure, such as molecular property prediction [99, 100, 101], particle classification [85], and text classification [74], 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 [129, 130].

Groundbreaking Applications in Natural Sciences and Medicine

Using deep learning in natural sciences revealed the power of graph representations for modeling small and large molecular structures. Combining different types of data can create bridges between the molecular and organism levels for modeling physical, chemical, or biological phenomena on a large scale. Recent knowledge graph applications have been introduced to enable precision medicine and make predictions across genomic, pharmaceutical, and clinical applications [121, 128]. Multi-scale learning systems are becoming valuable tools for protein structure prediction [25], protein property prediction [26], and biomolecular interaction modeling [77]. These methods can incorporate mathematical statements of physical relationships, knowledge graphs, prior distributions, and constraints by modeling predefined graph structures or modifying message-passing algorithms. When such information exists, multimodal learning can enhance image denoising [53], image restoration [53], and human-object interaction [48] in vision systems.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Data and code availability

We summarize multimodal graph learning (MGL) methods and provide a continually updated summary at https://yashaektefaie.github.io/mgl. We host a live table where future MGL methods will be added as a resource to the community.

References

[1]. Greener JG, Kandathil SM, Moffat L & Jones DT A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology 23, 40–55 (2022). [PubMed: 34518686]

[2]. Yu MK et al. Visible machine learning for biomedicine. Cell 173, 1562–1565 (2018). [PubMed: 29906441]

[3]. Wu Z et al. Moleculenet: a benchmark for molecular machine learning. Chemical science 9, 513–530 (2017). [PubMed: 29629118]

[4]. Gilmer J, Schoenholz SS, Riley PF, Vinyals O & Dahl GE Neural message passing for quantum chemistry. In Precup D & Teh YW (eds.) Proceedings of the 34th International Conference on Machine Learning, vol. 70 of Proceedings of Machine Learning Research, 1263–1272 (PMLR, 2017).

[5]. Sanchez-Gonzalez A et al. Graph networks as learnable physics engines for inference and control. In Dy J & Krause A (eds.) Proceedings of the 35th International Conference on Machine Learning, vol. 80 of Proceedings of Machine Learning Research, 4470–4479 (PMLR, 2018).

[6]. Sanchez-Gonzalez A et al. Learning to simulate complex physics with graph networks. In III HD & Singh A (eds.) Proceedings of the 37th International Conference on Machine Learning, vol. 119 of Proceedings of Machine Learning Research, 8459–8468 (PMLR, 2020).

[7]. Liu Q, Kusner MJ & Blunsom P A survey on contextual embeddings. In CoRR (2020). 2003.07278.

[8]. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M & Monfardini G The graph neural network model. IEEE Transactions on Neural Networks 20, 61–80 (2009). [PubMed: 19068426]

[9]. Kipf TN & Welling M Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations, ICLR '17 (2017).

[10]. Kipf TN & Welling M Variational graph auto-encoders. NIPS Workshop on Bayesian Deep Learning (2016).

[11]. Grover A, Zweig A & Ermon S Graphite: Iterative generative modeling of graphs. In Chaudhuri K & Salakhutdinov R (eds.) Proceedings of the 36th International Conference on Machine Learning, vol. 97 of Proceedings of Machine Learning Research, 2434–2444 (PMLR, 2019).

[12]. Guo X & Zhao L A systematic survey on deep generative models for graph generation. In CoRR (2020).

[13]. Baltrušaitis T, Ahuja C & Morency L-P Multimodal machine learning: A survey and taxonomy. CoRR (2017).

[14]. Hong C, Yu J, Wan J, Tao D & Wang M Multimodal deep autoencoder for human pose recovery. IEEE Transactions on Image Processing 24, 5659–5670 (2015). [PubMed: 26452284]

[15]. Khattar D, Goud JS, Gupta M & Varma V Mvae: Multimodal variational autoencoder for fake news detection. In The World Wide Web Conference, WWW '19, 2915–2921 (Association for Computing Machinery, New York, NY, USA, 2019).

[16]. Mao J, Xu J, Jing Y & Yuille A Training and evaluating multimodal word embeddings with large-scale web annotated images. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, 442–450 (Curran Associates Inc., Red Hook, NY, USA, 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 Chaudhuri K et al. (eds.) Proceedings of the 39th International Conference on Machine Learning, vol. 162 of Proceedings of Machine Learning Research, 9226–9259 (PMLR, 2022).

[18]. Xu P, Zhu X & Clifton DA Multimodal learning with transformers: A survey (2022).

[19]. Bayoudh K, Knani R, Hamdaoui F & Mtibaa A A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer 38, 2939–2970 (2022). [PubMed: 34131356]

[20]. Zhang C, Yang Z, He X & Deng L Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing 14, 478–493 (2020).

[21]. Javaloy A, Meghdadi M & Valera I Mitigating modality collapse in multimodal VAEs via impartial optimization. In Chaudhuri K et al. (eds.) Proceedings of the 39th International Conference on Machine Learning, vol. 162 of Proceedings of Machine Learning Research, 9938–9964 (PMLR, 2022).

[22]. Ma M et al. Smil: Multimodal learning with severely missing modality. Proceedings of the AAAI Conference on Artificial Intelligence 35, 2302–2310 (2021).

[23]. Poklukar P et al. Gmc – geometric multimodal contrastive representation learning. In Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research (arXiv, 2022).

[24]. Zitnik M et al. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion 50, 71–91 (2019). [PubMed: 30467459]

[25]. Jumper J et al. Highly accurate protein structure prediction with alphafold. Nature 596, 583–589 (2021). [PubMed: 34265844]

[26]. Somnath VR, Bunne C & Krause A Multi-scale representation learning on proteins. In Beygelzimer A, Dauphin Y, Liang P & Vaughan JW (eds.) Advances in Neural Information Processing Systems (2021).

[27]. Walters WP & Barzilay R Applications of deep learning in molecule generation and molecular property prediction. Accounts of Chemical Research 54, 263–270 (2021). [PubMed: 33370107]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

[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. Proceedings of the AAAI Conference on Artificial Intelligence 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, ICLR (2022).

[31]. Zhao F & Wang D Multimodal Graph Meta Contrastive Learning, 3657–3661 (Association for Computing Machinery, New York, NY, USA, 2021).

[32]. Zheng S et al. Multi-modal graph learning for disease prediction. CoRR 2107.00206 (2021). 2107.00206.

[33]. Ramachandram D & Taylor GW Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine 34, 96–108 (2017).

[34]. Ngiam J et al. Multimodal deep learning. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, 689–696 (Omnipress, Madison, WI, USA, 2011).

[35]. Aafaq N, Akhtar N, Liu W, Gilani SZ & Mian A Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[36]. Fang Z, Gokhale T, Banerjee P, Baral C & Yang Y Video2Commonsense: Generating commonsense descriptions to enrich video captioning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 840–860 (Association for Computational Linguistics, Online, 2020).

[37]. Kiros R, Salakhutdinov R & Zemel R Multimodal neural language models. In Xing EP & Jebara T (eds.) Proceedings of the 31st International Conference on Machine Learning, vol. 32 of Proceedings of Machine Learning Research, 595–603 (PMLR, Bejing, China, 2014).

[38]. Rezaei-Shoshtari S, Hogan FR, Jenkin M, Meger D & Dudek G Learning intuitive physics with multimodal generative models. Proceedings of the AAAI Conference on Artificial Intelligence 35, 6110–6118 (2021).

[39]. Xu K, Hu W, Leskovec J & Jegelka S How powerful are graph neural networks? In International Conference on Learning Representations (2019).

[40]. Hamilton W, Ying Z & Leskovec J Inductive representation learning on large graphs. In Guyon I et al. (eds.) Advances in Neural Information Processing Systems, vol. 30 (Curran Associates, Inc., 2017).

[41]. Xu K et al. Representation learning on graphs with jumping knowledge networks. In Dy J & Krause A (eds.) Proceedings of the 35th International Conference on Machine Learning, vol. 80 of Proceedings of Machine Learning Research, 5453–5462 (PMLR, 2018).

[42]. Bronstein MM, Bruna J, Cohen T & Veli kovi P Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. CoRR (2021). 2104.13478.

[43]. Chen Y et al. Graph-based global reasoning networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[44]. Varga V & Lorincz A Fast interactive video object segmentation with graph neural networks. In International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18–22, 2021, 1–10 (IEEE, 2021).

[45]. 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. International Journal of Remote Sensing 42, 6184–6208 (2021).

[46]. Zhou S, Zhang J, Zuo W & Loy CC Cross-scale internal graph neural network for image super-resolution. In Advances in Neural Information Processing Systems (2020).

[47]. Mou C & Zhang J Graph attention neural network for image restoration. In 2021 IEEE International Conference on Multimedia and Expo (ICME) (2021).

[48]. Qi S, Wang W, Jia B, Shen J & Zhu S-C Learning human-object interactions by graph parsing neural networks. In Proceedings of the European Conference on Computer Vision (ECCV) (2018).

[49]. Wang H, Zheng W. s. & Yingbiao L Contextual heterogeneous graph network for human-object interaction detection. In Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII, 248–264 (Springer-Verlag, Berlin, Heidelberg, 2020).

[50]. Zhang FZ, Campbell D & Gould S Spatially conditioned graphs for detecting human-object interactions. In CVPR, 13319–13327 (2021).

[51]. Avelar PC, Tavares AR, da Silveira TT, Jung CR & Lamb LC Superpixel image classification with graph attention networks. In 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 203–209 (IEEE Computer Society, Los Alamitos, CA, USA, 2020).

[52]. Lu Y, Chen Y, Zhao D & Chen J Graph-fcn for image semantic segmentation. In Lu H, Tang H & Wang Z (eds.) Advances in Neural Networks – ISNN 2019, 97–105 (Springer International Publishing, Cham, 2019).

[53]. Kim J, Lee JK & Lee KM Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).

[54]. Achanta R et al. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 2274–2282 (2012). [PubMed: 22641706]

[55]. Zeng H, Liu Q, Zhang M, Han X & Wang Y Semi-supervised hyperspectral image classification with graph clustering convolutional networks. arXiv preprint arXiv:2012.10932 (2020).

[56]. Wan S et al. Multiscale dynamic graph convolutional network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 58, 3162–3177 (2019).

[57]. Long J, Shelhamer E & Darrell T Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015).

[58]. Simonyan K & Zisserman A Very deep convolutional networks for large-scale image recognition. In Bengio Y & LeCun Y (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015).

[59]. Knyazev B, Lin X, Amer MR & Taylor GW Image classification with hierarchical multigraph networks. In British Machine Vision Conference (BMVC) (2019).

[60]. Veli kovi P et al. Graph attention networks. In International Conference on Learning Representations (2018).

[61]. Valsesia D, Fracastoro G & Magli E Deep graph-convolutional image denoising. In CoRR (2019). 1907.08448.

[62]. Bresson X & Laurent T Residual gated graph convnets. In Computing Research Repository, vol. 1711.07553 (2017). 1711.07553.

[63]. Biten AF et al. Scene text visual question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019).

[64]. Singh A et al. Towards vqa models that can read. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[65]. Liu C et al. Graph structured network for image-text matching. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020).

[66]. Ulutan O, Iftekhar ASM & Manjunath BS Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020).

[67]. Gao C, Xu J, Zou Y & Huang J-B Drg: Dual relation graph for human-object interaction detection. In Vedaldi A, Bischof H, Brox T & Frahm J-M (eds.) Computer Vision – ECCV 2020, 696–712 (Springer International Publishing, Cham, 2020).

[68]. Zhou P & Chi M Relation parsing neural network for human-object interaction detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019).

[69]. Gao D, Li K, Wang R, Shan S & Chen X Multi-modal graph neural network for joint reasoning on vision and scene text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020).

[70]. 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). [PubMed: 27295650]

[71]. Wu T et al. Ginet: Graph interaction network for scene parsing. In Vedaldi A, Bischof H, Brox T & Frahm J-M (eds.) Computer Vision – ECCV 2020, 34–51 (Springer International Publishing, Cham, 2020).

[72]. Wu L et al. Graph neural networks for natural language processing: A survey. In CoRR (2021). 2106.06090.

[73]. Vaswani A et al. Attention is all you need. In Guyon I et al. (eds.) Advances in Neural Information Processing Systems, vol. 30 (Curran Associates, Inc., 2017).

[74]. Li I, Li T, Li Y, Dong R & Suzumura T Heterogeneous Graph Neural Networks for Multi-label Text Classification. arXiv:2103.14620 [cs] (2021). ArXiv: 2103.14620.

[75]. Huang L, Ma D, Li S, Zhang X & Wang H Text level graph neural network for text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3444–3450 (Association for Computational Linguistics, Hong Kong, China, 2019).

[76]. Zhang Y et al. Every document owns its structure: Inductive text classification via graph neural networks. CoRR (2020). 2004.13826.

[77]. Pan J, Peng M & Zhang Y Mention-centered graph neural network for document-level relation extraction. In CoRR (2021). 2103.08200.

[78]. Zhu H et al. Graph Neural Networks with Generated Parameters for Relation Extraction. arXiv:1902.00756 [cs] (2019). ArXiv: 1902.00756.

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

[80]. Zeng S, Xu R, Chang B & Li L Double graph based reasoning for document-level relation extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1630–1640 (Association for Computational Linguistics, Online, 2020).

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

[82]. Zhang C, Li Q & Song D Aspect-based sentiment classification with aspect-specific graph convolutional networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4568–4578 (Association for Computational Linguistics, Hong Kong, China, 2019).

[83]. Zhang M & Qian T Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3540–3549 (Association for Computational Linguistics, Online, 2020).

[84]. 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, Online, 2020).

[85]. Shlomi J, Battaglia P & Vlimant J-R Graph neural networks in particle physics. Machine Learning: Science and Technology 2, 021001 (2021).

[86]. 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 (NeurIPS) (2017).

[87]. Qasim SR, Kieseler J, Iiyama Y & Pierini M Learning representations of irregular particle-detector geometry with distance-weighted graph networks. The European Physical Journal C 79 (2019).

[88]. Mikuni V & Canelli F Abcnet: an attention-based method for particle tagging. The European Physical Journal Plus 135 (2020).

[89]. Ju X et al. Graph neural networks for particle reconstruction in high energy physics detectors. In CoRR (2020). 2003.11603.

[90]. Shukla K, Xu M, Trask N & Karniadakis GE Scalable algorithms for physics-informed neural and graph networks. Data-Centric Engineering 3, e24 (2022).

[91]. Seo S & Liu Y Differentiable physics-informed graph networks. In CoRR (2019). 1902.02950.

[92]. Li W & Deka D Physics based gnns for locating faults in power grids. In CoRR (2021). 2107.02275.

[93]. Battaglia PW et al. Relational inductive biases, deep learning, and graph networks. In CoRR (2018). 1806.01261.

[94]. Veli kovi P, Ying R, Padovano M, Hadsell R & Blundell C Neural execution of graph algorithms. In International Conference on Learning Representations (2020).

[95]. Schuetz MJA, Brubaker JK & Katzgraber HG Combinatorial optimization with physics-inspired graph neural networks. Nature Machine Intelligence 4, 367–377 (2022).

[96]. Mirhoseini A et al. A graph placement methodology for fast chip design. Nature 594, 207–212 (2021). [PubMed: 34108699]

[97]. Gasteiger J, Groß J & Günnemann S Directional message passing for molecular graphs. In International Conference on Learning Representations (2020).

[98]. Jørgensen PB, Jacobsen KW & Schmidt MN Neural message passing with edge updates for predicting properties of molecules and materials. CoRR (2018). 1806.03146.

[99]. Gasteiger J, Yeshwanth C & Günnemann S Directional message passing on molecular graphs via synthetic coordinates. In Ranzato M, Beygelzimer A, Dauphin Y, Liang P & Vaughan JW (eds.) Advances in Neural Information Processing Systems, vol. 34, 15421–15433 (Curran Associates, Inc., 2021).

[100]. Liu M et al. Fast quantum property prediction via deeper 2d and 3d graph networks. In CoRR (2021). 2106.08551.

[101]. St. John PC, Guan Y, Kim Y, Kim S & Paton RS Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. Nature Communications 11, 2328 (2020).

[102]. Pattanaik L et al. Message passing networks for molecules with tetrahedral chirality. In CoRR (2020). 2012.00094.

[103]. Fey M, Yuen J-G & Weichert F Hierarchical inter-message passing for learning on molecular graphs. In CoRR (2020). 2006.12179.

[104]. Ariëns E Chirality in bioactive agents and its pitfalls. Trends in Pharmacological Sciences 7, 200–205 (1986).

[105]. 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).

[106]. Coley CW et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem. Sci 10, 370–377 (2019). [PubMed: 30746086]

[107]. Struble TJ, Coley CW & Jensen KF Multitask prediction of site selectivity in aromatic c–h functionalization reactions. React. Chem. Eng 5, 896–902 (2020).

[108]. Stuyver T & Coley CW Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. J Chem Phys 156, 084104 (2022). [PubMed: 35232175]

[109]. Stokes JM et al. A deep learning approach to antibiotic discovery. Cell 180, 688–702.e13 (2020). [PubMed: 32084340]

[110]. Fu T et al. Differentiable scaffolding tree for molecule optimization. In International Conference on Learning Representations (2022).

[111]. Mercado R et al. Graph networks for molecular design. Machine Learning: Science and Technology 2, 025023 (2021).

[112]. Torng W & Altman RB Graph convolutional neural networks for predicting drug-target interactions. Journal of Chemical Information and Modeling 59, 4131–4149 (2019). 10.1021/acs.jcim.9b00628. [PubMed: 31580672]

[113]. Moon S, Zhung W, Yang S, Lim J & Kim WY Pignet: a physics-informed deep learning model toward generalized drug–target interaction predictions. Chemical Science 13, 3661–3673 (2022). [PubMed: 35432900]

[114]. Gainza P et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods 17, 184–192 (2020). [PubMed: 31819266]

[115]. Sanner MF, Olson AJ & Spehner J-C Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38, 305–320 (1996). [PubMed: 8906967]

[116]. Sverrisson F, Feydy J, Correia BE & Bronstein MM Fast end-to-end learning on protein surfaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15272–15281 (2021).

[117]. Feng Y, You H, Zhang Z, Ji R & Gao Y Hypergraph neural networks. Proceedings of the AAAI Conference on Artificial Intelligence 33, 3558–3565 (2019).

[118]. 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).

[119]. Jo J et al. Edge representation learning with hypergraphs. In Ranzato M, Beygelzimer A, Dauphin Y, Liang P & Vaughan JW (eds.) Advances in Neural Information Processing Systems, vol. 34, 7534–7546 (Curran Associates, Inc., 2021).

[120]. Zhang C, Song D, Huang C, Swami A & Chawla NV Heterogeneous graph neural network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, 793–803 (Association for Computing Machinery, New York, NY, USA, 2019).

[121]. Chandak P, Huang K & Zitnik M Building a knowledge graph to enable precision medicine. Scientific Data (2023).

[122]. Lee S & Song BC Graph-based knowledge distillation by multi-head attention network. In Sidorov K & Hicks Y (eds.) Proceedings of the British Machine Vision Conference (BMVC), 162.1–162.12 (BMVA Press, 2019).

[123]. Zhou S et al. Distilling holistic knowledge with graph neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10387–10396 (2021).

[124]. Sun L, Gou J, Yu B, Du L & Tao D Collaborative teacher-student learning via multiple knowledge transfer. In CoRR (2021). 2101.08471.

[125]. Park W, Kim D, Lu Y & Cho M Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[126]. Liu Y et al. Knowledge distillation via instance relationship graph. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[127]. Ma J et al. Using deep learning to model the hierarchical structure and function of a cell. Nature Methods 15, 290–298 (2018). [PubMed: 29505029]

[128]. Nicholson DN & Greene CS Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal 18, 1414–1428 (2020). [PubMed: 32637040]

[129]. Borisov V et al. Deep neural networks and tabular data: A survey. ArXiv abs/2110.01889 (2021).

[130]. 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. Journal of Cheminformatics 13, 12 (2021). [PubMed: 33597034]

Box 1:

The blueprint for multimodal graph learning.

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 $\mathscr{C} = \{\mathbf{C}_1, …, \mathbf{C}_k\}$, where $\mathbf{C}_i$ is an information matrix of i-th modality that describes every entity by an information vector. We define Identify$_i$ module for every modality i as:

$$\mathbf{X}_i \leftarrow Identify_i(\mathbf{C}_i), \qquad (1)$$

    that 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 $\mathscr{X} = \{\mathbf{X}_1, …, \mathbf{X}_k\}$. We define *Connect$_j$* modules, $j = 1, …, m$, to specify connections between entities in $\mathscr{X}$ based on m distance measures as:

$$\mathbf{A}_j \leftarrow Connect_j(\mathscr{X}). \qquad (2)$$

    If $\mathbf{X}_i$ is already given as an adjacency matrix, the associated *Connect$_j$* modules specify predefined neighborhoods.

3. ***Propagating Information***: Neural messages are exchanged along edges in the adjacency matrices $\mathscr{A} = \{\mathbf{A}_1, …, \mathbf{A}_m\}$ to produce node representations:

$$\mathbf{H} \leftarrow Propagate(\mathscr{X}, \mathscr{A}). \qquad (3)$$

    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:

$$\mathbf{Z} \leftarrow Mix(\mathbf{H}, \mathscr{A}). \qquad (4)$$

    The mixing module Mix transforms node representations into final representations of entities $\mathbf{Z}$ on which downstream tasks are defined on. Established strategies to mix representations include aggregation operators, such as summation [39], averaging [40], multi-hop aggregation [41], and methods using adjacency information $\mathscr{A}$.
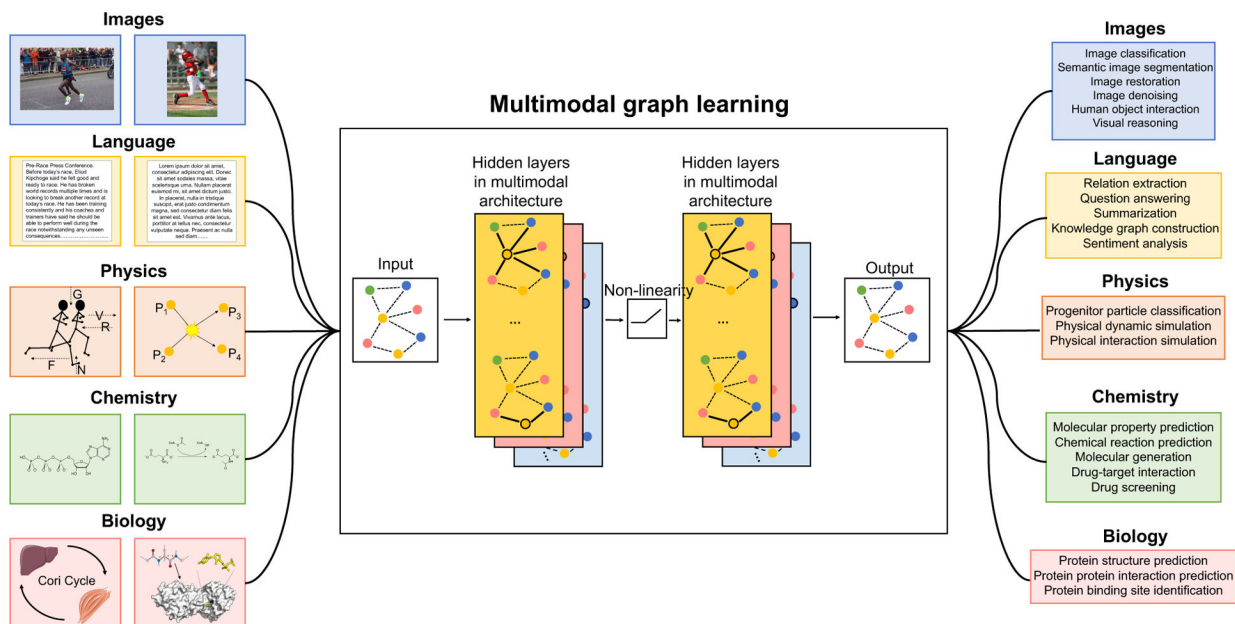
Fig. 1. Graph-centric multimodal learning.
Shown on the left are the different data modalities. Shown on the right are machine learning tasks for which multimodal graph learning has proved valuable. We introduce the multimodal graph learning (MGL) blueprint that serves as a unifying framework for multimodal graph neural architectures realized through learning systems in computer vision, natural language processing, and natural sciences.
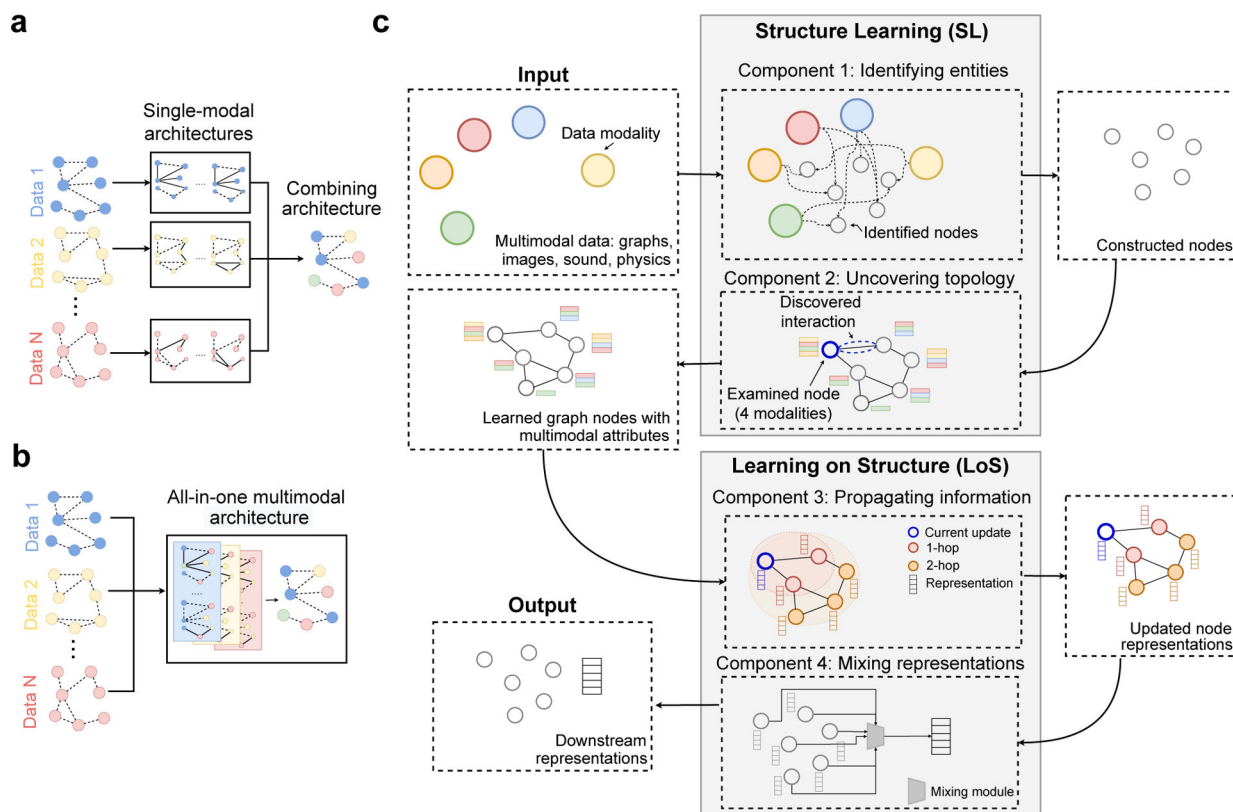
Fig. 2. Overview of multimodal graph learning (MGL) blueprint.
a, A standard approach to multimodal learning involves combining different unimodal architectures, each optimized for a distinct data modality. **b,** In contrast, an all-in-one multimodal architecture considers inductive biases specialized for each data modality and optimizes model parameters in an end-to-end manner, enabling expressive data fusion. **c,** The MGL blueprint comprises four components: identifying entities, uncovering topology, propagating information, and mixing representations. These components are grouped into two phases: structure learning and learning on the structure.
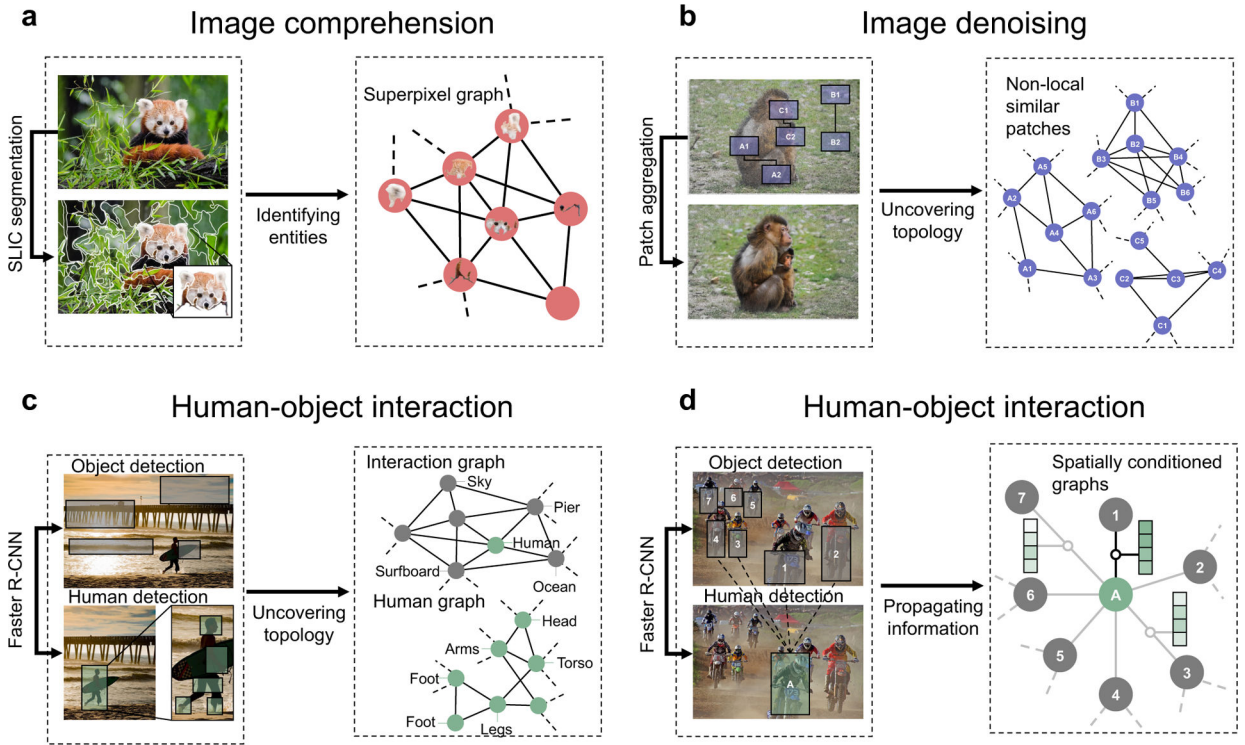
Fig. 3. Application of multimodal graph learning blueprint to images.
a, Modality identification for image comprehension where nodes represent aggregated regions of interest, or superpixels, generated by the SLIC segmentation algorithm. **b,** Topology uncovering for image denoising where image patches (nodes) are connected to other non-local similar patches. **c,** Topology uncovering in human-object interaction where two graphs are created. A human-centric graph maps body parts to their anatomical neighbors, and an interaction connects body parts relative to the distance to other objects in the image. **d,** Information propagation in human-object interaction where spatially conditioned graphs modify message passing to incorporate edge features that enforce the relative direction of objects in an image [50].
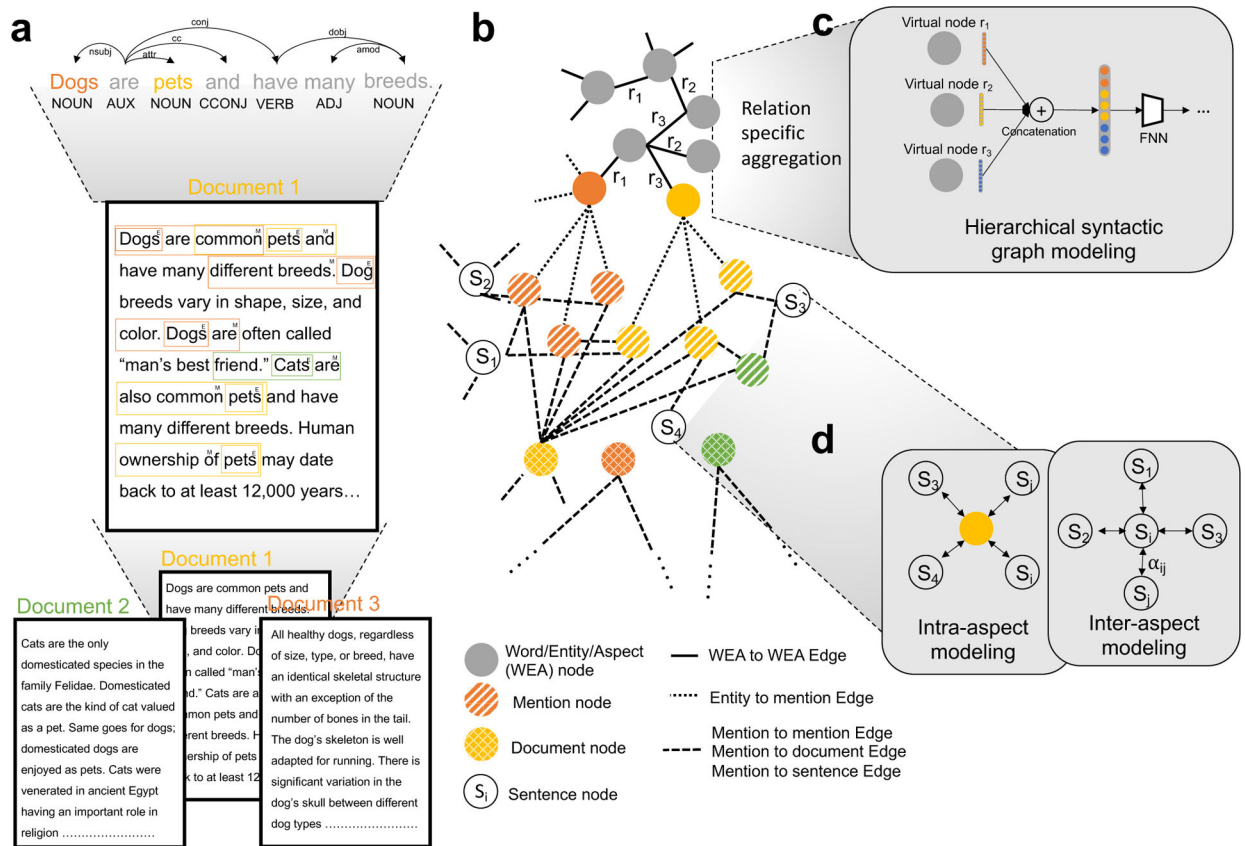
Fig. 4. Application of multimodal graph learning blueprint to language.
a, The different levels of context in text inputs from sentences to documents and the individual units identified at each context level. This is an example of modality identification's first component of the MGL blueprint. **b,** The simplified construction of a language-intensive graph from text input, an application of the topology uncovering component of the MGL blueprint. **c,** and **d,** visualize examples of learning on LIGs for aspect-based sentiment analysis (ABSA), which aims to assign a sentiment (positive, negative, or neutral) to a sentence with regards to a given aspect. By grouping by relation type from within a sentence (shown in **c**) or modeling relations between sentences and aspects (shown in **d**), these methods integrate inductive biases relevant to ABSA and innovate in MGL's third component, information propagation.
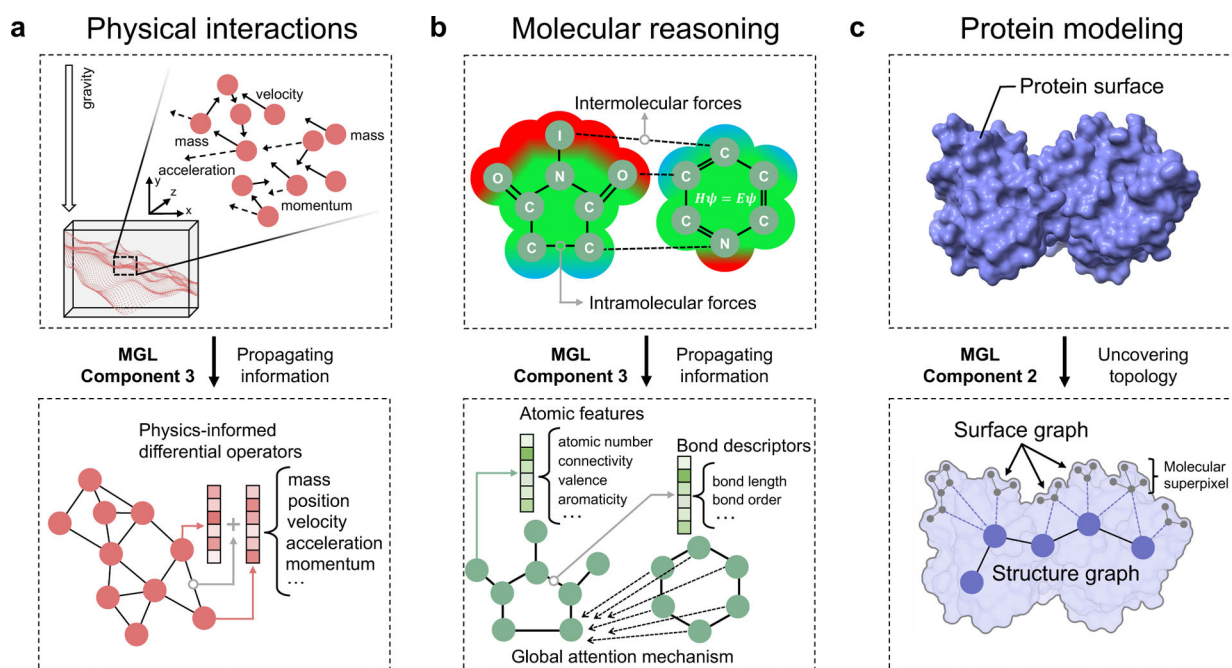
Fig. 5. Applications of multimodal graph learning to natural sciences.
a, Information propagation in physical interactions where physics-informed neural message passing is used to update the states of particles in a system due to inter-particle interactions and other forces. **b,** Information propagation in molecular reasoning where a global attention mechanism is used to model the potential interaction between atoms in two molecules to predict whether two molecules will react. **c,** Topology uncovering in protein modeling where a multiscale graph representation is used to integrate primary, secondary, and tertiary structures of a protein with higher-level protein motifs summarized in molecular superpixels to represent a protein [26]. This robust topology provides a better prediction on tasks such as protein-ligand binding affinity prediction.