Global Concept Explanations for Graphs by Contrastive Learning

Jonas Teufel[1[0000−0002−9228−9395]] and Pascal Friederich[1,2†[0000−0003−4465−1465]]

[1] Institute of Theoretical Informatics, Karlsruhe Insitute of Technology, Kaiserstr. 12, 76131 Karlsruhe, Germany
jonas.teufel@kit.edu
[2] Institute of Nanotechnology, Karlsruhe Insitute of Technology, Kaiserstr. 12, 76131 Karlsruhe, Germany
pascal.friederich@kit.edu

Abstract. Beyond improving trust and validating model fairness, xAI practices also have the potential to recover valuable scientific insights in application domains where little to no prior human intuition exists. To that end, we propose a method to extract global concept explanations from the predictions of graph neural networks to develop a deeper understanding of the task's underlying structure-property relationships. We identify concept explanations as dense clusters in the self-explaining MEGAN model's subgraph latent space. For each concept, we optimize a representative prototype graph and optionally use GPT-4 to provide hypotheses about why each structure has a certain effect on the prediction. We conduct computational experiments on synthetic and real-world graph property prediction tasks. For the synthetic tasks we find that our method correctly reproduces the structural rules by which they were created. For real-world molecular property regression and classification tasks, we find that our method rediscovers established rules of thumb. More specifically, our results for molecular mutagenicity prediction indicate more fine-grained resolution of structural details than existing explainability methods, consistent with previous results from chemistry literature. Overall, our results show promising capability to extract the underlying structure-property relationships for complex graph property prediction tasks.

Keywords: Graph Neural Network · Global Concept Explanations · Self Explaining Model · Molecular Property Prediction

1 Introduction

The practice of explainable artificial intelligence (xAI) is often expected to improve trust during human-AI interactions, provide tools for model debugging, validate model fairness, and comply with anti-discrimination laws [11]. These merits are of high interest in application domains where humans already have

† Corresponding author.

solid intuitions and expectations about how a model *should* behave, such as image and natural language processing tasks.

We argue that in applications in which little to no prior intuition about a task exists, xAI practices can be used to extract knowledge from a high-performing AI model and to gain new insights about the underlying tasks. Especially important application domains in which this is the case are chemistry and material science. In recent years, graph neural networks (GNNs) have proven to be a valuable tool for producing accurate property predictions from molecular graph representations [34]. Despite access to accurate predictive models, the underlying working mechanisms of many of these tasks remain open questions.

In most application domains, including graph processing, there already exist various xAI methods that help to make AI predictions more transparent. However, most of these methods are focused on local explainability, which aims to supply additional information and explanations for each individual prediction. As previous authors have stated already [43,21], there is a lot of additional potential in generating global explanations, rather than local instance-level explanations. In contrast to purely local explanations, global explanations aim to provide additional information about the behavior of the model as a whole. Here, we present a method to generate global concept explanations for graph property prediction tasks. The aim of these global explanations is to approximately extract the general structure-property relationships that govern the model's decisions. For this purpose, we propose MEGAN2—an extension of the recently proposed multi-explanation graph attention network (MEGAN) [39] architecture. Our modifications include an additional contrastive representation learning objective which promotes the model's latent space of subgraph embeddings to be aligned according to the structural similarity of the corresponding motifs. Explanations are then generated by applying clustering methods to this latent space. The extracted clusters consisting of similar subgraph motifs can be interpreted as overarching *concepts* discovered and exploited by the model. By analyzing the prediction contributions of the individual cluster members, it is also possible to determine each concept's average influence on the prediction outcome. Furthermore, we use a genetic algorithm optimization to determine a small and representative prototype graph for each cluster. Additionally, a suitable representation of this prototype graph can then be given to a large language model such as GPT-4 [1,6] to provide an initial hypothesis about possible causal relationships behind the identified structure-property correlation.

We experimentally validate the proposed framework for global explainability on several synthetic graph prediction tasks and real-world molecular property prediction tasks. On synthetic datasets for graph regression and classification, we show that our method can correctly reconstruct the structure-property relationships based on which the datasets were created. Furthermore, we validate our method on two molecular property prediction datasets for the prediction of water solubility and the classification of mutagenicity. For these real-world datasets, our method re-discovers known rules of thumb about the underlying molecular properties. Specifically for the mutagenicity prediction we find that

our method produces significantly more fine-grained explanations than previously published methods for global graph explainability which are consistent with previously published work from the chemistry literature. This opens wide applicability to many tasks in graph classification and regression, specifically in chemistry and materials science, to potentially derive unknown structure-property relationships.

Contribution. Compared to the previously published literature on the topic of global concept explainability for graph neural networks our main contributions are the following:

- Previous work focuses entirely on graph classification tasks. However, regression tasks are equally important—especially considering the various material property prediction applications in chemistry and material science. We show that our method can produce reasonable explanations for classification and regression tasks.
- Previous work either uses partitioning methods [28] or a concept-bottleneck architecture [3], which requires the number of concepts to be configured beforehand. We argue that this presents a limitation, since a good estimation of the number of concepts may not be possible for unknown tasks. By building on a density-based clustering method, our method can more naturally discover the appropriate number of explanations for each task dynamically.
- We present an entirely self-explaining method, where the local, as well as global explanations, are derived directly from a single attention-based model. A self-explaining approach aims to alleviate the recently brought-up conceptual problem of trying to "explain a black box with another black box" which increasingly complex post-hoc approaches are facing [35,4].

2 Related Work

Graph Explainability. Graph neural networks have proven to be useful tools to handle predictive tasks on graph-structured data, such as molecular and material property prediction [34]. In parallel, xAI methods are developed to make the predictions of these complex models more transparent and interpretable to their human users. Yuan *et al.* [43], and more recently Kakkad *et al.* [21], provide a comprehensive overview of the existing literature on explainability methods specifically for graph neural networks. The majority of this literature is focused on local attributional explanation methods, that aim to provide additional information about the reasoning for each prediction. Many existing methods for local explainers, such as GraphLIME [17], GNN GradCAM [33] and GraphShap [32], are adaptations of previously existing work from the image and text processing domains. Other notable local graph explanation methods include GNNExplainer [41], PGExplainer [26], SubgraphX [44], and more recently the self-explaining MEGAN model architecture [39]. Besides attributional explanation masks, counterfactual explanations are another popular modality for local explanations of

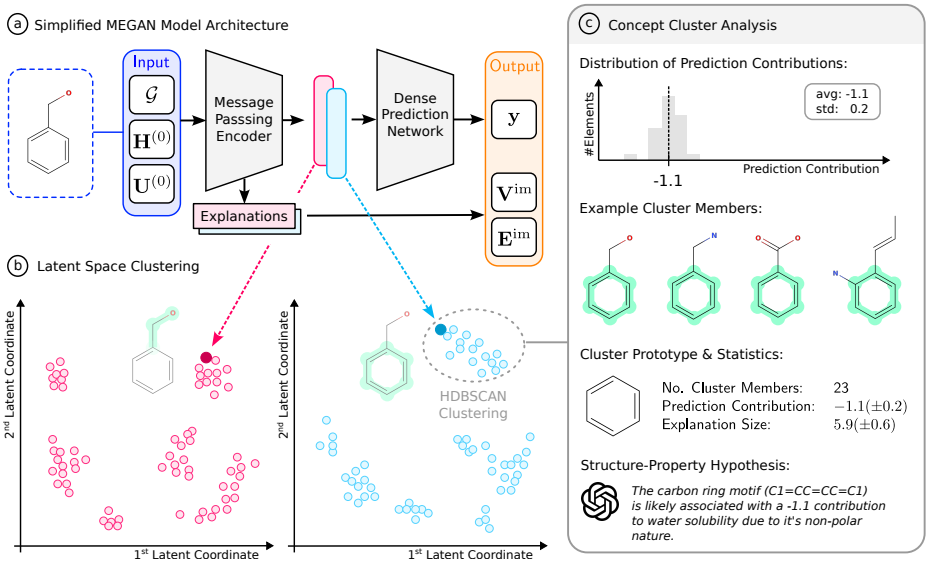

Fig. 1. Visual overview of the proposed global explanation method. ⓐ Simplified MEGAN model architecture. The message-passing encoder creates explanation masks along multiple channels (pink/blue), which then result in individual subgraph latent representations for each channel. ⓑ In each channel's subgraph latent space, a clustering algorithm is used to find elements in the dataset that exhibit structurally similar explanation masks. ⓒ Each cluster is analyzed regarding its members and results are presented to the user in the form of an automatically generated report.

graph neural networks [25,38]. Contrary to local explainers, methods of global explainability aim to provide information about large sections of the model's decision space at once—largely decoupled from individual predictions. Although global explainability is less intensively explored, existing work can roughly be split into generative explanations [42] prototype-based [9,46] and concept-based explanations [28,3].

In parallel to the development of the core explainability mechanisms, another area of ongoing research is that of AI visual analytics systems [16]. In this line of work, GNNLens [18] and CorGIE [24] are recently proposed visual analytics systems specifically created for the exploration of GNN predictions.

Concept Explanations. An early notion of concept explanations is presented by Kim *et al.* [23] who developed their framework for testing with concept activation vectors (TCAV) as an alternative to saliency maps for visual explanations. However, for this initial formulation, the concepts have to be manually defined to test their relevance to the prediction outcome. Ghorbani *et al.* [14] extend this method with their automated concept explanation (ACE) algorithm. The authors automatically assemble concept clusters by clustering image fragments according to their similarity in the model's latent space. These methods are

subsequently extended and refined in various image-processing [12,45] and text-processing applications [36,19].

Graph Concept Explanations. Even though concept-based explanations have mainly been applied to the image- and text-processing domains, there already exists some previously published work on concept-based explainability for graph neural networks as well. To our knowledge, Magister *et al.* [27] first introduced concept explanations to the graph processing domain with their work on the Graph Concept Explainer (GCExplainer). More recently, their work was extended toward hierarchical concepts [20] and shared concept spaces for multimodal learning [10]. In another line of work, Azzolin *et al.* [3] introduce the Global Logic-based Graph Explainer (GLGExplainer), which uses logic explainable networks (LEN) [8] to form predictions by combining concepts via logical formulas.

3 Background

3.1 MEGAN: Multi-Explanation Graph Attention Network

The multi-explanation graph attention network (MEGAN) was recently introduced as a self-explaining graph neural network architecture, which directly produces local node and edge attributional explanations alongside its main target predictions. The following section summarizes the main features of the MEGAN architecture (see Figure 2 for a visual overview), while a detailed description can be found in the original work by Teufel *et al.* [39].

The model's attributional explanation masks are based on the model's internal attention mechanism and serve to highlight graph substructures which are particularly important for the model's prediction. Specifically, the model produces a multiple of K attributional explanation masks for each element of the input graph structure, where the number of explanations is an architectural hyperparameter independent of the task specifications.

Formally, The input data structure $\mathcal{X} = \left(\mathcal{G}, \mathbf{H}^{(0)}, \mathbf{U}^{(0)}\right)$ of a MEGAN model consists of a graph structure $\mathcal{G}$, a tensor $\mathbf{H}^{(0)} \in \mathbb{R}^{V \times N_0}$ of initial node features, and a tensor $\mathbf{U}^{(0)} \in \mathbb{R}^{E \times M_0}$ of initial edge features. The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as a set $\mathcal{V} = \{0, \ldots, V\}$ of nodes and a set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ of node pairs. As the corresponding output $\mathcal{Y} = \left(\mathbf{y}^{\text{pred}}, \mathbf{V}^{\text{im}}, \mathbf{E}^{\text{im}}\right)$, the model produces the tensor of target value predictions $\mathbf{y}^{\text{pred}} \in \mathbb{R}^C$, the multi-channel node importance tensor $\mathbf{V}^{\text{im}} \in [0,1]^{V \times K}$, and the multi-channel edge importance tensor $\mathbf{E}^{\text{im}} \in [0,1]^{E \times K}$. The model can therefore be described as a function

$$f_{\boldsymbol{\theta}} : \left(\mathcal{G}, \mathbf{H}^{(0)}, \mathbf{U}^{(0)}\right) \mapsto (\mathbf{y}^{\text{pred}}, \mathbf{V}^{\text{im}}, \mathbf{E}^{\text{im}}) \tag{1}$$

with trainable parameters $\boldsymbol{\theta}$.

The message passing part of MEGAN's architecture is based on GATv2 [5] attention layers as its main attention mechanism. The first part of the network consists of L attention layers, where each layer consists of K parallel attention heads.

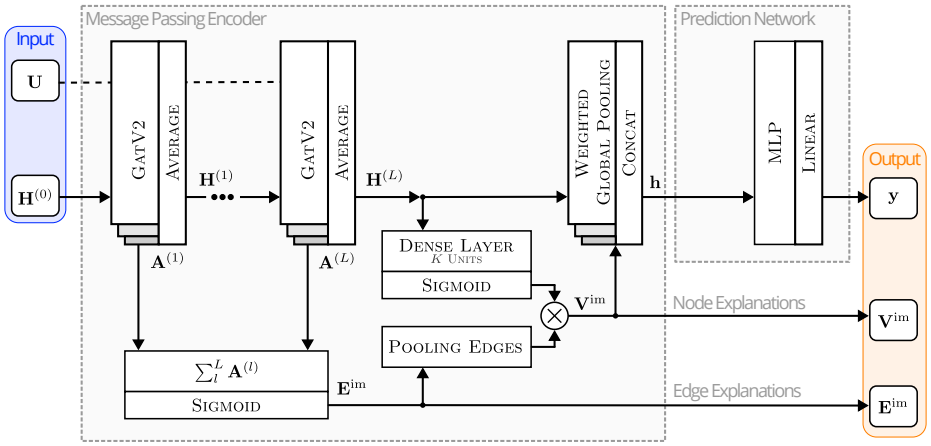

Fig. 2. Simplified overview of the multi-explanation graph attention network (MEGAN) architecture. Modified illustration based on the original work of Teufel *et al.* [39]. Rounded boxes represent input/output tensor structures of the network. Square boxes represent network layers and arrows indicate layer interconnections.

The l-th attention layer—consisting of K individual attention heads—outputs a node embedding representation $\mathbf{H}^{(l)}$ based on the graphs edge features $\mathbf{U}^{(0)}$ and the previous layers node features $\mathbf{H}^{(l-1)}$. In addition to the node feature updates, each attention head also produces an edge attention tensor $\mathbf{A}^{(l)} \in \mathbb{R}^{E \times K}$. These edge attention values are then aggregated across all layers to produce edge importance tensor $\mathbf{E}^{\text{im}} \in \mathbb{R}^{E \times K}$, representing the edge explanations. Additionally, the node importance tensor $\mathbf{V}^{\text{im}} \in \mathbb{R}^{V \times K}$ is obtained from a combination of edge importances and node embeddings of the final layer. After this message-passing part of the network, a graph pooling operation is used to create the global graph embedding $\mathbf{h}$. The graph pooling is realized as an attention-weighted sum over the final node embeddings—done for each channel independently—resulting in the channel-specific embedding

$$\mathbf{h}^{(k)} = \sum_{i}^{V} \mathbf{H}_{i,:}^{(L)} \cdot \mathbf{V}_{i,k}^{\text{im}} \tag{2}$$

which are then concatenated into the overall graph embedding $\mathbf{h} = \mathbf{h}^{(0)} || \cdots || \mathbf{h}^{(K)}$. This overall graph embedding is then passed through an MLP with C output units, which results in the target prediction $\mathbf{y}^{\text{pred}}$.

To promote the model's explanation masks to behave according to their predetermined interpretations, the explanation co-training method is introduced. In addition to the main prediction loss $\mathcal{L}^{\text{pred}}$, during training, the model is also subject to an explanation loss $\mathcal{L}^{\text{expl}}$, as well as a sparsity regularization term $\mathcal{L}^{\text{spar}}$. The total training loss is therefore given as

$$\mathcal{L} = \mathcal{L}^{\text{pred}} + \beta \cdot \mathcal{L}^{\text{expl}} + \gamma \cdot \mathcal{L}^{\text{spar}} \tag{3}$$

where the weights β and γ are model hyperparameters.
The explanation co-training loss trains the network to approximate a solution to the primary target prediction problem purely based on the explanation masks $\mathbf{V}^{\text{im}}$ across the K explanation channels. The sparsity loss applies an L1 regularization on the explanation masks to promote sparser explanations that focus only on a subset of nodes and edges.

3.2 Definition of Graph Concept Explanations

Concept-based explanations are a type of global explanations that aim to provide a set of generalizable rules about a model's overall behavior. In this context, a *concept* is a pattern that can be present and shared among multiple concrete instances. We follow the convention introduced by Ghorbani *et al.* [14] and represent a concept $\mathcal{C} = \{\hat{\mathcal{X}}\}$ as a set of input fragments $\hat{\mathcal{X}} \subset \mathcal{X}$ of an input data structure $\mathcal{X}$. What exactly constitutes such an input fragment, and therefore a concept itself, depends on the application domain. In the image processing domain, for example, valid input fragments would be superpixels or small image segments, and the concept be defined as the common motif these image snippets share.
For the graph processing domain, we define a concept

$$\mathcal{C} = \{\hat{\mathcal{G}}, \ \dots\} = \left\{(\mathcal{G}, \hat{\mathbf{V}}, \hat{\mathbf{E}}), \ \dots\right\} \tag{4}$$

as a set of subgraph motifs $\hat{\mathcal{G}} \subset \mathcal{G}$. More specifically, we consider a *soft* subgraph definition, where subgraphs are delimited by continuous node masks $\hat{\mathbf{V}} \in [0,1]^V$ and edge masks $\hat{\mathbf{E}} \in [0,1]^E$, rather than hard binary masks.
A concept explanation then seeks to associate concepts with specific influences on the model prediction to act as a set of generic rules about the model's behavior. Formally, Ghorbani *et al.* require a valid concept explanation to fulfill the three properties of *meaningfulness*, *coherency*, and *importance*. This means that input fragments have to be large enough to independently make sense, the fragments of a concept have to share an identifiable commonality and the concept has to have demonstrable importance for the model's prediction.

4 Methods

We propose a framework for the generation of concept explanations for graph property prediction tasks, in which the concepts are derived as clusters in a latent space of subgraph embeddings (see Fig.1). In this context, we define a subgraph latent space such that the constituting vector representations only embed information about graph substructures rather than the full graph.

Assumption 1 *There exists a function $f^\dagger : \mathcal{G} \mapsto \mathbf{z}$ that projects a graph structure $\mathcal{G} \in \mathbb{G}$ to a latent vector representation $\mathbf{z} \in \mathbb{R}^D$ such that the embedding only captures information about a specific subgraph structure $\hat{\mathcal{G}} \subseteq \mathcal{G}$. In other words, that $f^\dagger(\mathcal{G}^1) = f^\dagger(\mathcal{G}^2)$ when $\hat{\mathcal{G}} \subseteq \mathcal{G}^1, \mathcal{G}^2$ even if $\mathcal{G}^1 \neq \mathcal{G}_2$*

Additionally, we expect such a latent space representation to reflect structural similarities between the relevant graph structures. In other words, if two graphs contain substructures that are only slightly different, then their corresponding embeddings should lie close together.

Assumption 2 *May $f^{\dagger} : \mathcal{G} \mapsto \mathbf{z}$ be a subgraph projection, then there exists at least one distance measure $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ for the embedding space that is proportional to a measure $s : \mathbb{G} \times \mathbb{G} \to \mathbb{R}$ of structural similarity between the original graphs: $d(\mathbf{z}^1, \mathbf{z}^2) \propto s(\mathcal{G}^1, \mathcal{G}^2)$.*

Based on these assumptions, dense clusters in such a subgraph latent space represent original input graphs that are not necessarily similar overall, but which contain similarly structured subgraph motifs. Therefore one can say that the cluster members are graphs that share a common structural *concept*.
We argue that the architecture of the MEGAN graph neural network already approximately fulfills Assumption 1. In the model, the global graph pooling operation is realized as an attention-weighted sum, where the individual node embeddings are multiplied by their corresponding importance values (see Sec. 3.1). The pooled embedding vector is mostly comprised of information from high-importance nodes, while information from low-importance nodes is suppressed. Therefore, the pooled embedding $\mathbf{h}^{(k)}$ from the k-th explanation channels mostly represents the subgraph structure defined by that channel's explanation mask.
In general, Assumption 2 will not be fulfilled, since the latent space will be structured according to the network's primary target predictions. In the case of a three-class classification task, for example, the latent space will be structured into three clusters that maximize the decision boundary between the classes. Therefore, additional measures are necessary to ensure Assumption 2 to be approximately fulfilled. For this purpose, we propose the updated MEGAN2 version of the original model with the specific intent of facilitating the extraction of global concept explanations.

4.1 Extended Network Architecture

In the original MEGAN model architecture, each channel's embedding $\mathbf{h}^{(k)}$ is created as the attention-weighted sum of the graph's node embeddings, then immediately concatenated and used as input to the final property prediction network. We introduce additional projection networks $g^{(k)} : \mathbb{R}^D \to \mathbb{R}^P,\ \mathbf{h}^{(k)} \mapsto \mathbf{z}^{(k)}$ for each channel separately, which transform the original embedding $\mathbf{h}^{(k)} \in \mathbb{R}^D$ into a projected embedding $\mathbf{z}^{(k)} \in \mathbb{R}^P$. This additional projection serves two main purposes: Firstly, it ensures that each channel's embeddings can develop independently. Without the additional projection, the embeddings directly result from a summation of the node embeddings. A change of the pooled embedding $\mathbf{h}^{(k)}$ would be necessarily coupled to a change of the node embeddings. However, since the embeddings of all other channels also directly result from those same node embeddings, each channel's representations are strongly coupled to each other. With the additional projection, this is no longer necessary and channel

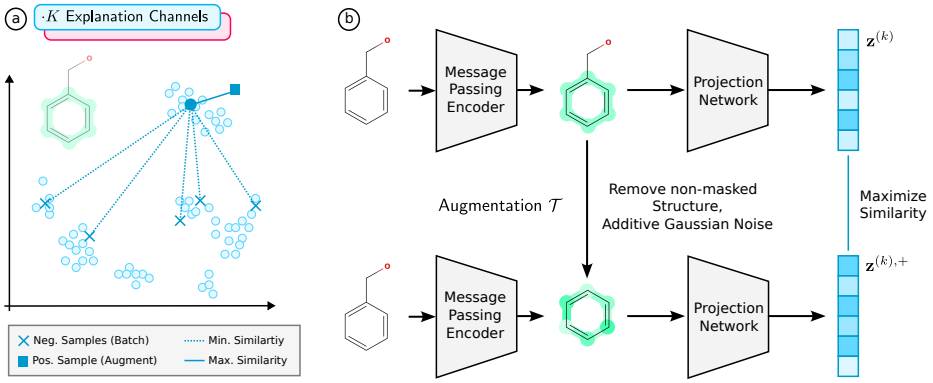

Fig. 3. Visualization of the contrastive learning method. (a) The contrastive learning is applied for each explanation channel's latent space individually by maximizing an embedding's similarity to the positive sample and minimizing similarity to negative samples. Other elements in the training batch are used for the negative samples and the positive samples are derived by data augmentation. (b) The augmentation is applied in the final stage of the message-passing encoder, during the global graph pooling operation. The augmented view removes all graph structures outside the explained areas and applies additive gaussian noise to the explanation masks.

embeddings can develop independently from each other. The second purpose of the projection networks is to increase the general encoding capabilities of the latent spaces. In other words, to make sure that sufficiently complex functions can be represented—which can capture greater details about subgraph similarity relationships. Specifically, by increasing the dimensionality $P \gg D$ of the embeddings through the projection, we make sure that more information can be encoded in the representations.
Additionally, L2-normalization is applied as the activation of the projection network's output which causes all the projected embeddings $\mathbf{z}^{(k)}$ to lie on the P-dimensional unit sphere. This normalization makes the cosine similarity an especially suitable distance measure for the latent space.

4.2 Contrastive Explanation Learning

For the proposed concept clustering scheme to yield meaningful results, Assumption 2 has to be approximately fulfilled, which means that the distance in the model's latent space should be proportional to the structural similarity of the corresponding graph motifs. To promote this property, we add a contrastive representation learning objective $\mathcal{L}^{\text{contr}}$ to the overall training loss

$$\hat{\mathcal{L}} = \mathcal{L}^{\text{pred}} + \beta \cdot \mathcal{L}^{\text{expl}} + \gamma \cdot \mathcal{L}^{\text{spar}} + \mu \cdot \mathcal{L}^{\text{contr}} \tag{5}$$

of the MEGAN model, where μ is a hyperparameter of the model that determines the weight of this new loss term during training.
For the contrastive learning scheme, we follow the general SimCLR framework [7]

(see Fig. 3), for which the basic idea is to maximize an embedding's similarity with a set of positive samples (similar input elements) while minimizing the similarity with a set of negative samples (dissimilar input elements). We apply the InfoNCE loss

$$\mathcal{L}^{\text{contr},(k)} = -\log\left(\frac{\exp(\mathbf{z}^{(k)} \cdot \mathbf{z}^{(k),+}/\tau)}{\exp(\mathbf{z}^{(k)} \cdot \mathbf{z}^{(k),+}/\tau) + \sum_b^B \exp(\mathbf{z}^{(k)} \cdot \mathbf{z}_b^{(k),-}/\tau)}\right) \tag{6}$$

separately to the projected embeddings $\mathbf{z}^{(k)}$ of each explanation channel. Since the projected embeddings are L2-normalized (see Sec. 4.1), the inner product between the embedding vectors is equivalent to their cosine similarity. For each element in a training batch with batch size B, all other elements of the batch are considered negative samples $\{\mathbf{z}_b^{(k),-}\}$. The positive samples $\mathbf{z}^{(k),+}$ are derived through data augmentations. Specifically, two different kinds of augmentations are applied to the channel's explanation mask $\mathbf{V}_{:,k}^{\text{im}}$ during the forward pass of the model (see Fig 3). First, we use a threshold to turn the still-continuous explanation into a binary mask. During the model forward passes, this binary mask is then applied to the node embedding during the global pooling such that all nodes outside of the mask are ignored. As a second augmentation, we apply additive Gaussian noise to the explanation mask itself. Together, these augmentations promote the embeddings to mainly represent the explained substructures, while being invariant towards small changes in the mask's specific importance values.

4.3 Concept Clustering

By training a MEGAN2 model with an additional contrastive loss term (see Section 4.2) acting on the projected channel embeddings $\mathbf{z}^{(k)}$, Assumption 1 and Assumption 2 are approximately fulfilled. Consequently, the channel projections $\mathbf{z}^{(k)}$ primarily encode information about the subgraph highlighted in the corresponding explanation. In addition, the distance between two latent projections approximately corresponds to the structural similarity of their motifs. Therefore, overarching structural concepts can be identified as dense clusters within this projected latent space.

To find these clusters we use the HDBSCAN [29] clustering algorithm on the channel projections $\mathbf{z}^{(k)}$ for each of the K channels independently (see Figure 1). HDBSCAN is a density-based clustering algorithm that doesn't require the number of clusters to be known beforehand. Instead, an appropriate number of clusters is dynamically discovered from the data.

The clustering is done for each of the model's K explanation channels separately. Therefore, a concept cluster $\mathcal{C}^{(k,q)}$ with index $q \in \{0, \ldots, Q\}$ is also uniquely associated with the explanation channel $k \in \{0, \ldots, K\}$ from which it was extracted. The members of these concept clusters consist of various soft graph fragments $\hat{\mathcal{G}}$ (see Section 3.2).

Each concept cluster can additionally be associated with an average contribution to the outcome of the prediction—thereby forming an explanation of the model's

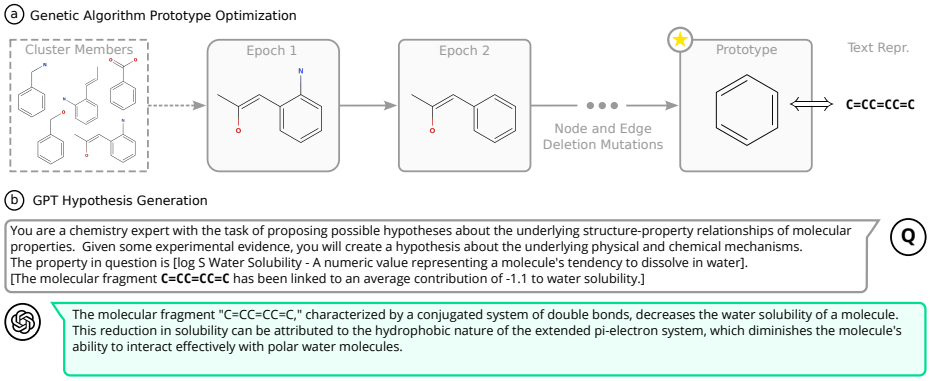

Fig. 4. (a) Illustration of the prototype optimization process. Starting from a population consisting of a concept's member graphs, node and edge deletion mutations are applied in each epoch to find a suitable prototype graph. (b) Example for the GPT-based hypothesis generation. The text representation of the concept prototype graph is included in a query that prompts the model to generate a causal explanation for the concept's structure-property relationship.

behavior that associates a structural pattern (cluster members) with an effect on the predicted property (contribution).
For each concept cluster, this contribution is calculated as an average over the contributions of the cluster members. Each member of a cluster $\mathcal{C}^{(k,q)}$ is the combination of an original graph element $\mathcal{G}$ from the dataset its corresponding local explanation mask $\mathbf{V}^{\text{im}}_{:,k}$ for the k-th explanation channel. Here, we define one such element's contribution as it's leave-one-out deviation $\Delta^{(k)}$ as it was originally introduced by Teufel *et al.* [39] to quantify the fidelity of MEGAN's explanation channels.

4.4 Prototype Optimization

To increase the interpretability of the concept representations, we additionally generate a prototype graph $\mathcal{G}^{(k,q)}_*$ for each concept cluster. In this context, we define a prototype as a graph consisting of a minimal number of nodes that is still sufficiently representative of the concept's underlying pattern. Formally, we optimize these prototype graphs

$$\begin{aligned} \mathcal{G}^{(k,q)}_* = & \underset{\mathcal{G} \in \mathcal{C}^{(k,q)}}{\arg\min} \ |\mathcal{G}| \\ & \text{s.t.} \quad \mathbf{z}^{(k)} \cdot \overline{\mathbf{z}}^{(k,q)} \leq \epsilon \\ & \text{where } \mathbf{z}^{(k)} \text{ embedding of } \mathcal{G} \end{aligned} \tag{7}$$

for each of the concept clusters separately. The main objective function of this optimization is to minimize the graph size in terms of the number of nodes. Additionally, the optimization is subject to the constraint of maintaining at

least a minimum of structural similarity to the underlying pattern of the concept cluster. This constraint is realized as a minimum threshold ϵ of cosine similarity between the
constraining the corresponding channel projection $\mathbf{z}^{(k)}$ to maintain at least a minimal cosine similarity ϵ to the concept cluster centroid $\overline{\mathbf{z}}^{(k,q)}$. Here, the cluster centroid $\overline{\mathbf{z}}^{(k,q)}$ vector is computed as the average vector over all embeddings corresponding to the concept's cluster members.
We use a genetic algorithm (GA) with a fixed number of epochs for the optimization of the concept prototype graphs (see Figure 4). The initial population of the GA is chosen as the concept members' full graph structures. In each epoch, exclusively random node and edge deletion mutations are applied to all elements of the population. The GA uses a tournament selection strategy, while a small fraction of the best-performing individuals are guaranteed to be transferred to the next epoch.

4.5 Hypothesis Generation

If applicable, we also generate hypotheses about the causal reasons behind the extracted structure-property correlations. For this purpose, we construct a prompt containing the string representation of the prototype graph and the average contribution of the concept toward the prediction outcome. We query OpenAI's GPT-4 API to generate a causal hypothesis about why the given pattern—represented by the prototype graph—might have the associated impact on the prediction (see Figure 4 for an example).
However, this hypothesis generation is only available for tasks about which the language model contains prior domain knowledge and for which meaningful text representations exist. Consequently, the method is not applicable to custom synthetic graph property prediction tasks. However, it can be used for most molecular property prediction tasks, for which the graph structure can be given the model in SMILES representation.

5 Computational Experiments

To illustrate the effectiveness of our proposed global explanation framework, we conduct two parts of computational experiments. In the first part, we apply our method to synthetic datasets. The labels of these datasets were created based on pre-determined structure-property relationships, where the existence of certain subgraph motifs results in corresponding graph labels. For such datasets, global explanations should be able to discover these basic structure-property relationships that were initially used to construct the dataset. In the second part, we apply our method to real-world molecular property prediction datasets. Due to the complexity of real-world tasks, there is no certainty about the true underlying structure-property relationships. Consequently, there exists no hard ground truth in terms of the generated explanations. However, for the

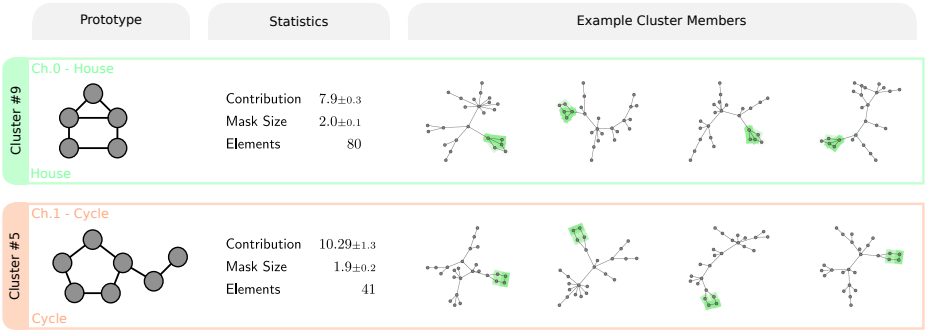

Fig. 5. Selected results of the concept clustering for a MEGAN2 model trained on the BA2Motifs dataset. Each row represents one concept cluster for either the house class (green) or the cycle class (orange). Columns from left to right show the result of the prototype optimization, aggregated statistics for the cluster members, and the 4 members closest to the cluster centroid. Prototype graphs are automatically optimized during the report generation, but have been re-drawn for the clarity of the visualization.

chosen datasets, there either exist commonly known rules of thumb about certain structure-property relationships or previously published hypotheses in the scientific literature, which the generated explanations are compared with.

5.1 Synthetic Datasets

BA2Motifs - Graph Classification. The BA2Motifs dataset was introduced by Luo *et al.* [26] and is considered a standard dataset to evaluate graph explainability [21]. The dataset consists of 1k randomly generated Barabási-Albert (BA) graphs. Each graph is either seeded with either a "house" motif or a five-node "cycle" motif which determines the graph's class for the underlying classification task. Therefore, the house motif and the cycle motif are considered the perfect ground truth structure-property explanations for this task.
In this experiment, we train a MEGAN2 model on the BA2Motifs dataset to generate global concept explanations and concept prototypes. The model is evaluated for its target prediction performance and explanation accuracy regarding the known ground truth explanation masks. For the performance evaluation, the model uses a 90/10 train-test split while the concept clustering is done on the full dataset.
The results firstly show that the model can perfectly solve the simple classification task with the expected test-set accuracy of 100%. The model also scores a high explanation accuracy (node AUC=0.96, edge AUC=0.97) comparable to state-of-the-art local explainers. More importantly, regarding the global concept explanations (see Fig. 5), the house and the cycle pattern are correctly identified as the relevant motifs for the task. The prototype optimization is also able to recover almost the exact motifs for both cases, with only some additional nodes attached to the cycle pattern. It is also important to point out that our method returns 14 concept clusters for this dataset, indicating some redundancy in the

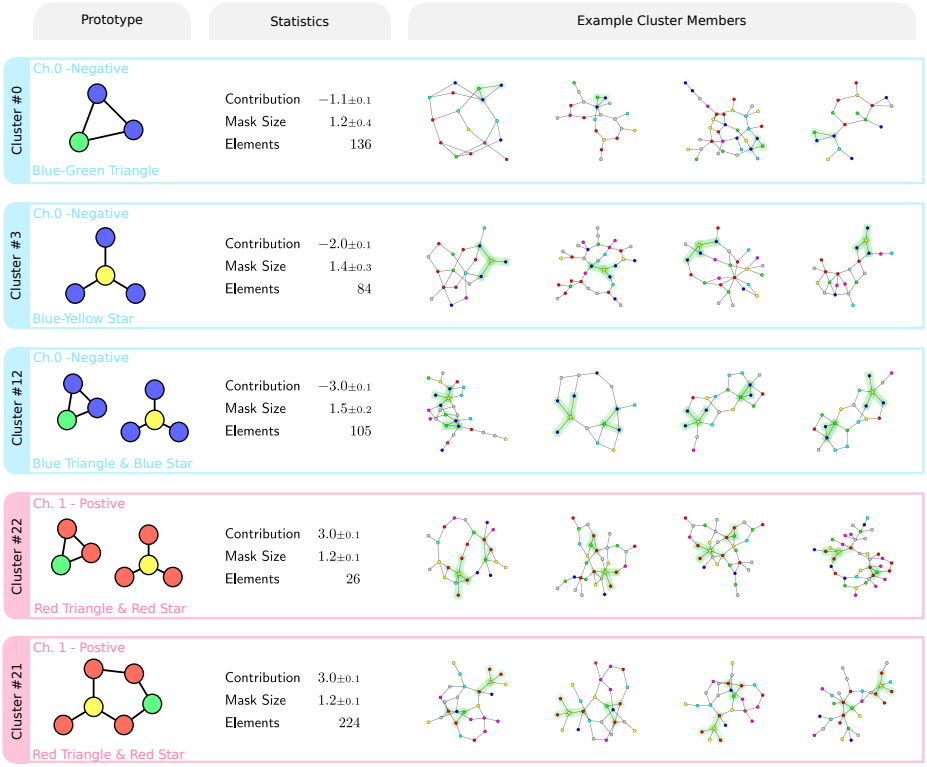

Fig. 6. Selected results of the concept clustering for a MEGAN2 model trained on the RbMotifs dataset. Each row represents one concept cluster for either the negative (blue) or positive (red) explanation channel. Columns from left to right show the result of the prototype optimization, aggregated statistics for the cluster members, and the 4 members closest to the cluster centroid. Prototype graphs are automatically optimized during the report generation, but have been re-drawn for the clarity of the visualization.

clustering. However, all concepts exclusively represent either the house or the cycle pattern—differing only in the local neighborhood of the BA graphs where these motifs are attached.

RbMotifs - Graph Regression. The RbMotifs dataset was introduced by Teufel *et al.* [39] as an example of a regression task that is influenced by substructures of opposing influence. The dataset consists of 10k randomly generated color graphs, where the feature vector for each node represents an RGB color code. The graphs can additionally be seeded with multiple subgraph motifs—two mainly blue motifs that contribute negatively and two mainly red motifs that contribute positively to the overall graph label.
In contrast to other attributional explanation methods, the MEGAN model is shown able to correctly capture the polarity of the opposing influences presented by the negative and positive motifs for this regression task [39]. However,

previous work only investigates local explainability in the form of attributional masks for each individual prediction. While this already reveals some information about the underlying task, purely local explanations still require a manual review of many individual examples to be able to spot the underlying motifs that govern the task as a whole.

In this experiment, we train a MEGAN2 model on the RbMotifs dataset to generate global concept explanations and concept prototypes. Additionally, we train a model with the original unmodified MEGAN architecture to compare against. Models are evaluated on their target prediction performance and the explanation accuracy with respect to the ground truth explanation masks. For the performance evaluation, both models use the same 90/10 train-test split while the concept clustering is done on the full dataset.

We most importantly find that the MEGAN2 model shows comparable prediction performance ($R^2 \approx 0.98$), and explanation accuracy (node AUC ≈ 0.99, edge AUC ≈ 0.95) as the original MEGAN model (see Tab. 1). This indicates that our proposed modifications do not interfere with the model's primary predictive and explanatory function. Regarding the concept clustering, MEGAN2 results in significantly more clusters (18 vs 7) with the same clustering parameters.

Regarding the concept clustering report that is generated for the MEGAN2 model (see Fig. 6), we find that all of the structure-property relationships from which the dataset was initially constructed are reflected in the concept explanations, almost perfectly matching each motif's ground truth contribution to the overall graph target value. For example, there exist concept clusters that correctly identify the blue-green triangle (true contribution -1, predicted -1.1) and the blue-yellow star (true contribution -2, predicted -2) pattern as the relevant motifs related to this task. Furthermore, a different cluster can be found which features both patterns (e.g. Fig. 6, Cluster#12) with a correctly predicted contribution of -3—therefore correctly representing the additive nature of the task. Overall, the method shows coverage of the ground truth explanatory motifs, which is *complete* and *comprehensive*, meaning that the report contains all of the relevant motifs and no additional irrelevant ones. However, we find that the method tends to *reduncancy* in the sense that the same motif can be the subject of multiple clusters. For example, there exist multiple clusters related to the combined red-triangle and red-star pattern, as illustrated in Figure 6 for Cluster#21 and Cluster#22. This example also shows that duplicate clusters can

Table 1. Evaluation results for computational experiments different model versions on the RbMotifs dataset. Experiments were repeated 5 independent times with the same train-test split. We report average results and standard deviation for 5 independent repetitions of the dataset.

Model	$R^2 \uparrow$	Node AUC ↑	Edge AUC ↑	#Clusters ↑
MEGAN	$0.97_{\pm 0.01}$	$0.99_{\pm 0.01}$	$0.95_{\pm 0.01}$	$7.8_{\pm 0.75}$
MEGAN2	$0.98_{\pm 0.01}$	$0.99_{\pm 0.01}$	$0.95_{\pm 0.03}$	$18.6_{\pm 4.13}$

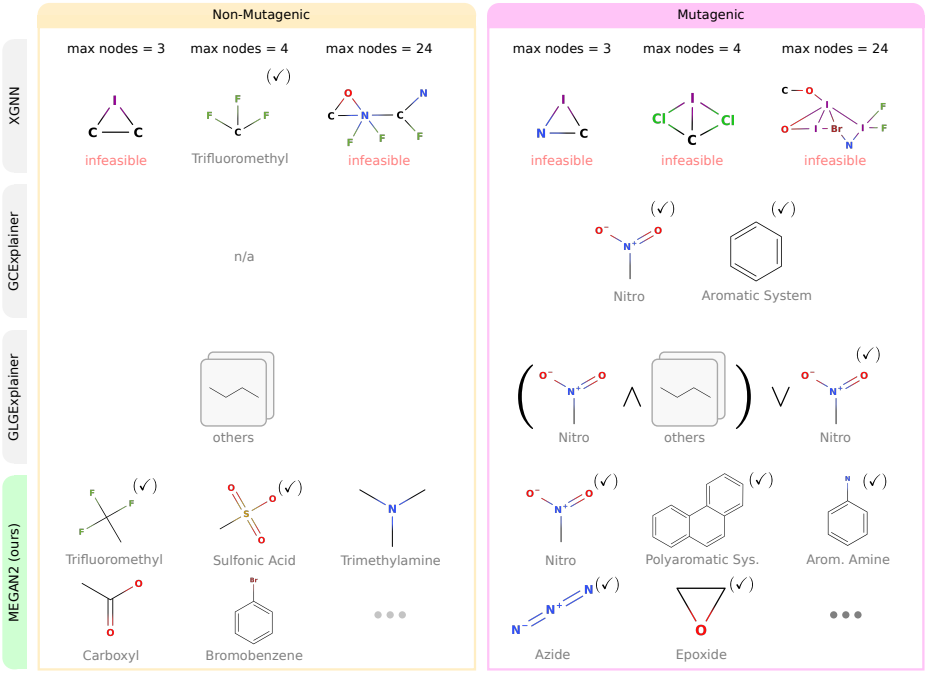

Fig. 7. Comparison of global graph explainability methods for the Mutagenicity dataset. The two columns represent the two possible classes non-mutagenic (yellow) and mutagenic (purple). Each row shows the global explanations of a different method from the literature, while the last row shows explanations obtained by our proposed concept clustering method based on the MEGAN2 model. Explanations for XGNN and GLGExplainer are taken from Azzolin *et al.* [3] and explanations for GCExplainer are taken from Magister *et al.* [28].
(✓) Symbol indicates consistency with hypotheses from chemistry literature [22]
.

have slightly different prototypes. While Cluster#21's prototype shows the correct motifs, Cluster#21's prototype shows a slightly merged version of the two motifs.

5.2 Real-World Datasets

Mutagenicity - Graph Classification. The mutagenicity dataset [15,30] consists of roughly 6500 molecular graphs annotated with a binary classification label of being either mutagenic or non-mutagenic. These labels have been experimentally determined by test results of Ames mutagenicity assays [2,31], which measure a substance's capability to induce mutations to the DNA of bacteria.
In this experiment, we train a MEGAN2 model on the Mutagenicity dataset to generate global concept explanations and concept prototypes. Training and evaluation of the model are done on a random 90/10 train-test split of the

dataset, while concepts are extracted on the full dataset. We also compare the extracted concepts with explanations from previously published work on global graph explainability—namely XGNN [42], GCExplainer [27], and GLGExplainer [3]. XGNN is a generative explanation method, which uses a reinforcement learning approach to generate graphs that maximize classification probability. GCExplainer and GLGExplainer are concept-based explainability methods. For GLGExplainer specifically, individual motifs can additionally be joined to express explanations as logic formulas.

When comparing the different explainability methods for the mutagenicity prediction (see Fig. 7), the previously published global graph explainers are subject to some limitations. Firstly we find XGNN's graph generation procedure does not take into account the chemical constraints on the graph structure, such as the maximum number of possible bonds for certain atom types. This generally leads to chemically infeasible graph structures—limiting the overall usability of the generated explanations. Since GCExplainer and GLGExplainer are concept-based explainability methods, they are not subject to the same problem. Both methods successfully recover the nitro (NO_2) group as the most relevant motif. In addition, GCExplainer also indicates aromatic carbon ring systems as possible explanations for the mutagenic class. Although these are among the most commonly considered explanations in the xAI literature [41], previous work from chemistry literature suggests that many more structural explanations for mutagenicity exist [22]. Kazius *et al.* [22] propose 39 structural motifs linked to mutagenic behavior, including but not limited to the often cited nitro (NO_2) group. Interestingly, the authors also mention the possibility of detoxifying structures, which can inhibit mutagenic behavior by reducing DNA interactions of otherwise mutagenic substances. We argue that these detoxifying substructures can be interpreted as explanatory motifs for the non-mutagenic class. This perspective of detoxifying structural influences was largely neglected by previous methods.

In contrast to existing methods, we find that our proposed concept extraction yields a substantially more diverse set of explanatory concepts for both the mutagenic and the non-mutagenic classes. For the non-mutagenic class, our method highlights, among other things, the trifluoromethyl (CF_3) and sulfonic acid (SO_3H) groups, which were specifically mentioned by Kazius *et al.* as possible detoxifying influences. For the mutagenic class, our method also identifies the nitro group and the polyaromatic systems as strong influences for mutagenic behavior. In addition, our method also identifies aromatic amine groups (NH_2), azide groups (N_3) and epoxide (C_2O) groups as mutagenic influences—all of which are mentioned by Kazius *et al.* as well. Besides the groups indicated in the literature, our method yields several additional concepts. However, we also identified some faults in the explanations. For example, while haloalkenes are correctly identified as a relevant concept for the mutagenic class, there also exists a high-fidelity haloalkene concept for the non-mutagenic class as well. We suspect that such cases of double attribution may be an artifact of the approximate assumption of linearly contributing evidence for tasks in which this is not

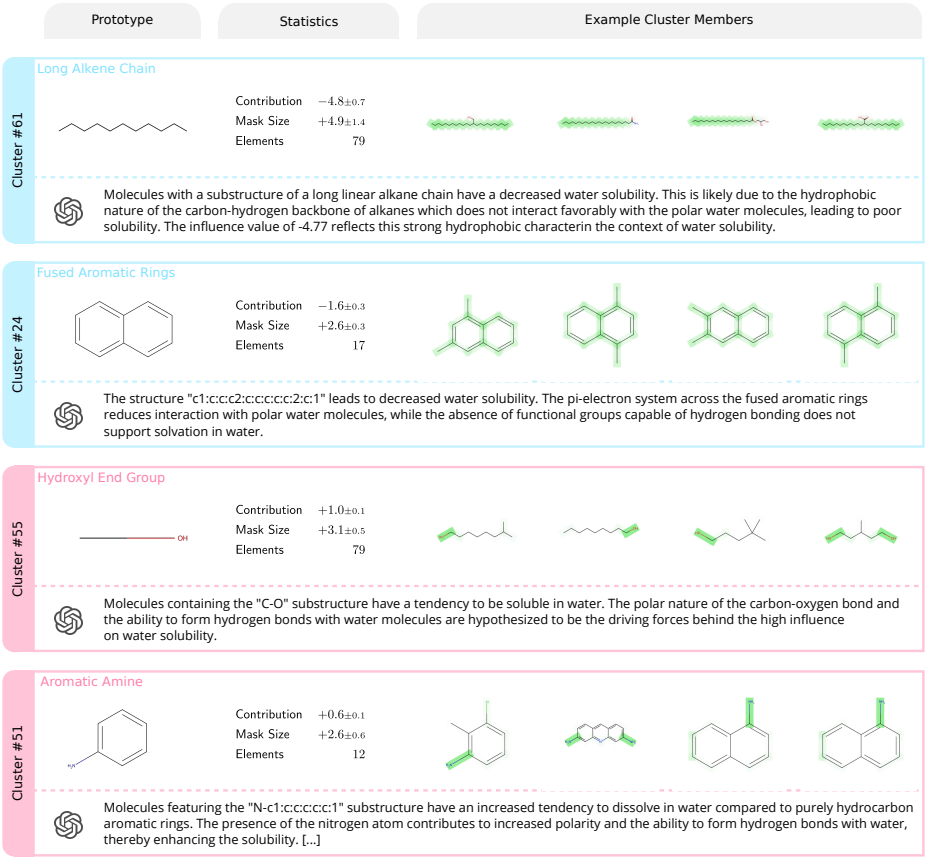

Fig. 8. Selected results of the concept clustering for a MEGAN2 model trained on the AqSolDB dataset. Each row represents one extracted concept cluster for either the negative (blue) or positive (red) explanation channel. Columns from left to right show the result of the prototype graph optimization, aggregated statistics for the cluster members, and the 4 members closest to the cluster centroid. In addition, the GPT-generated hypothesis is displayed underneath each concept. The prototype graphs in the first column have been automatically generated.

entirely the case.

AqSolDB - Graph Regression. AqSolDB is a dataset that was introduced by Sorkun *et al.* [37] which consists of roughly 10k molecular graph structures annotated with experimentally determined logS values for water solubility. The values approximately lie in the range between -14 and +3, where higher values indicate a higher solubility in water.

For the prediction of water solubility there generally exists no exact ground truth explanations. However, since water solubility is a relatively well-studied phenomenon overall, there exist several rules of thumb about what kinds of sub-

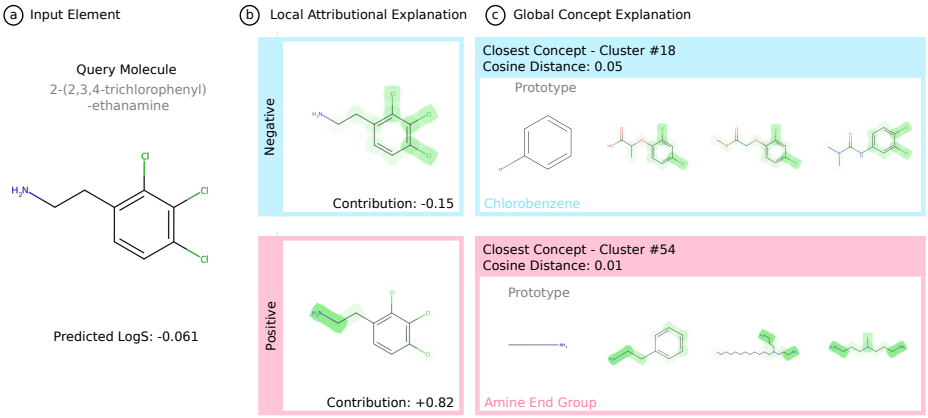

Fig. 9. Visualization of how the global concept explanations can be used to improve local explainability as well. (a) The model outputs a prediction for an individual query element. (b) In addition to the target value prediction itself, the model also outputs the local explanation masks for the different channels. (c) Each local explanation mask can be associated with a concept by finding the concept whose cluster centroid is the closest to the corresponding explanation embedding.

structures influence this property. In a simplified manner, water solubility is largely influenced by the polarity of a molecule and its capability to form hydrogen bonds. This is because water molecules themselves are strong dipoles and largely bind via hydrogen bonds. As a consequence, large non-polar structures such as long carbon chains and carbon rings are generally considered to reduce water solubility, while polar functional groups such as those including oxygen and nitrogen are considered to increase water solubility.

In this experiment, we train a MEGAN2 model on the AqSolDB dataset to predict the continuous logS values. The model is trained and evaluated on a 90/10 train-test split, while the concept clustering is done on the full dataset.

We find that our concept clustering results for the solubility prediction (see Fig 8) are generally in agreement with the previously mentioned rules of thumb. For the negative channel, the extracted concept clusters primarily represent different variants of aromatic structures, but several clusters indicate carbon chains in various sizes as another negative influence as well. For the positive channel, the extracted concepts span a variety of different functional groups mainly based on oxygen and nitrogen atoms. For example, we find the hydroxyl (OH) and amine (NH_2) groups as two primary motifs for positive contributions to water solubility. One interesting observation regarding these results is that concepts that are described by the clusters seem to be neighborhood-dependent. An illustrative example for this is that two separate clusters describe the amine (NH_2) group. In one cluster the amines are end pieces of carbon chains (see Fig. 8, Cluster#54) while in the other cluster, the amines are connected to aromatic systems (see Fig. 8, Cluster#51). This behavior contradicts Assumption 2 and most likely stems from the fact that the thresholding augmentation for the con-

trastive learning (see Sec. 4.2) is applied only after neighborhood information was already spread by the message-passing. However, in the context of molecular property prediction, this neighborhood dependency is a desirable property. While the basic explanation itself is the amine group in this case, the exact way in which such a functional group affects the overall property of the molecule is also dependent on the local environment where that functional group is attached.
For the water solubility experiments, we furthermore use GPT-4 to generate concept hypotheses (see Fig. 8). For each concept cluster, the model is queried with a custom prompt that provides information about the identified concept and prompts the language model to form a hypothesis for the causal reasoning about why the explained substructure may be associated with a particular positive or negative influence. In some cases, these explanations seemingly reflect the known rules of thumb quite well such as for the non-polar alkene chain (Fig. 8, Concept#61) and the polar hydroxyl group (Fig. 8, Concept#55). However, we also observe cases where the generated hypotheses are incorrect. One possible reason for this is that the language model does not correctly identify the given motif. An example of this is Cluster#52, whose prototype clearly shows an aromatic amine group (nitrogen attached outside of carbon ring) but is incorrectly identified as a pyridine (nitrogen part of carbon ring) group. Another possible source of error is cases where either the concept or the prototype themselves are incorrect. In these cases, the language model still tries to come up with a hypothesis about the behavior, hallucinating sentences that may sound like valid explanations but convey incomplete or incorrect content [40].

Besides shedding light on the model's overall decision-making behavior, the extracted concept explanations can also be used to improve local explainability as well. When a MEGAN model is queried with a specific input graph, the model outputs the primary target prediction and the local explanations in the form of node and edge importance masks for all the explanation channels. Additionally, each explanation can be associated with one of the extracted global concepts by finding the concept cluster centroid closest to each explanation channel's corresponding embedding vector. For the water solubility prediction of the exemplary 2-(2,3,4-trichlorophenyl)ethanamine molecule (see Fig. 9) the local explanations already indicate the ring structure of the graph as a negative influence and the nitrogen group as a positive influence. However, only the association with the corresponding concept explanations provides the validation that the highlighted chlorophenyl and amine functional groups are recurring patterns that are generally associated with negative/positive influences across the whole dataset. This provides further insight and helps to better understand the underlying structure-property relations

6 Limitations

Despite the encouraging results obtained by the proposed method, it is important to point out its limitations as well. Most importantly, since the method is based on the MEGAN model, it inherits all prior limitations thereof. MEGAN's explanation co-training procedure approximately assumes global linearity of explanations—meaning that for an explanation to be recognized as such, a non-zero statistical correlation with the explained property has to exist. For instance, for a structure to be recognized as a "negative" explanation, it has to appear more frequently in elements with relatively low target values. While this is a reasonable assumption for many real-world tasks, it is not necessarily applicable to all kinds of graph property prediction tasks.
Other limitations are related to the HDBSCAN clustering algorithm that is used to find the concept clusters. Since the method is based on the density of the latent space, a certain minimum dataset size is required to estimate this density, rendering the method less suitable for small datasets. While the method is seemingly able to capture more fine-grained structural details in the identified concept clusters, it also tends to produce redundant clusters, where we sometimes observe multiple clusters that seemingly represent the same motif.
Another important limitation relates to the language model-based automatic generation of causal hypotheses. While these hypotheses seem to work reasonably well for some tasks, it is fundamentally restricted to tasks about which the language model has prior knowledge. For example, a model may provide useful information about relatively well-studied tasks, such as water solubility, which were likely well represented in its training corpus. However, the models will provide decreasingly useful information about niche topics or custom prediction tasks. Additionally, the accumulation of errors from prior processing steps may lead the language model to hallucinate a correct-sounding yet factually incorrect explanation. This presents a major danger, especially to non-expert users who are unlikely to recognize the errors.

7 Conclusion

In this work, we introduce a framework for the generation of global concept-based explanations for graph regression and classification tasks. In our framework, concepts are identified by clustering in a model's latent space of subgraph embeddings. We show that the recently proposed multi-explanation graph attention network (MEGAN) approximately creates such a substructure latent space. We introduce an updated MEGAN2 model version which includes modifications to the original architecture and training procedure to improve this latent space clustering capability.
We conduct computational experiments on multiple synthetic and real-world graph property prediction datasets to validate the effectiveness of the proposed explanation procedure. For the synthetic tasks, our method can correctly reproduce the basic structure-property relationships from which the datasets were

originally constructed. For the real-world datasets, we find that our method rediscovers multiple known rules of thumb about molecular property regression and classification tasks. Specifically for mutagenicity prediction, our method produces substantially more fine-grained explanations than existing global graph explainability methods. Existing methods primarily focus on the nitro group as an explanation for mutagenic behavior. Contrary to that, our method recovers multiple additional structural motifs for both classes which are consistent with previous work in chemistry literature.
Overall, we argue that global explainability is an important step in extracting human-interpretable knowledge from complex graph property prediction models. The proposed method shows promising potential to reconstruct a task's underlying structure-property relationships. We also find that recent advances in large language models offer intriguing possibilities to communicate the extracted knowledge in natural language—which could especially benefit non-expert users and educational settings. However, in this direction, more work is needed to ensure the integrity of the generated text and prevent the spread of misinformation. Based on the promising results, future work will need to investigate the method's applicability to more complex real-world graph property prediction tasks, as well as a user study to evaluate the effectiveness of the generated explanations.

Reproducibility Statement. All code is publicly available on GitHub: `https://github.com/aimat-lab/megan_global_explanations`. The code is implemented in the Python 3.10 programming language and builds on the Pytorch Geometric [13] graph neural network library. Besides the experiment modules, the repository contains the automatically generated full concept clustering reports for all of the experiments that were referenced in this publication.

Acknowledgements. This work was supported by funding from the pilot program Core-Informatics of the Helmholtz Association (HGF).

Conflict of Interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Introducing ChatGPT. https://openai.com/blog/chatgpt
2. Ames, B.N., Lee, F.D., Durston, W.E.: An Improved Bacterial Test System for the Detection and Classification of Mutagens and Carcinogens. Proceedings of the National Academy of Sciences **70**(3), 782–786 (Mar 1973). `https://doi.org/10.1073/pnas.70.3.782`
3. Azzolin, S., Longa, A., Barbiero, P., Liò, P., Passerini, A.: Global Explainability of GNNs via Logic Combination of Learned Concepts (2022). `https://doi.org/10.48550/ARXIV.2210.07147`
4. Bordt, S., Finck, M., Raidl, E., Von Luxburg, U.: Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. 2022 ACM Conference on Fairness, Accountability, and Transparency pp. 891–905 (Jun 2022). `https://doi.org/10.1145/3531146.3533153`

5. Brody, S., Alon, U., Yahav, E.: How Attentive are Graph Attention Networks? In: International Conference on Learning Representations (Feb 2022). https://doi.org/10.48550/arXiv.2105.14491
6. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of Artificial General Intelligence: Early experiments with GPT-4 (Apr 2023). https://doi.org/10.48550/arXiv.2303.12712
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607. PMLR (Nov 2020). https://doi.org/10.5555/3524938.3525087
8. Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Liò, P., Maggini, M., Melacci, S.: Logic Explained Networks. Artificial Intelligence **314**, 103822 (Jan 2023). https://doi.org/10.1016/j.artint.2022.103822
9. Dai, E., Wang, S.: Towards Prototype-Based Self-Explainable Graph Neural Network (Oct 2022). https://doi.org/10.48550/arXiv.2210.01974
10. Dominici, G., Barbiero, P., Magister, L.C., Liò, P., Simidjievski, N.: SHARCS: Shared Concept Space for Explainable Multimodal Learning (Jul 2023). https://doi.org/10.48550/arXiv.2307.00316
11. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [cs, stat] (Mar 2017). https://doi.org/10.48550/arXiv.1702.08608
12. Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadénc, R., Serre, T.: CRAFT: Concept Recursive Activation FacTorization for Explainability. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2711–2721 (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00266
13. Fey, M., Lenssen, J.E.: Fast Graph Representation Learning with Pytorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds (2019)
14. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards Automatic Concept-based Explanations. In: Neural Information Processing Systems (Feb 2019). https://doi.org/10.48550/arXiv.1902.03129
15. Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., Müller, K.R.: Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. Journal of Chemical Information and Modeling **49**(9), 2077–2081 (Sep 2009). https://doi.org/10.1021/ci900161g
16. Hohman, F., Kahng, M., Pienta, R., Chau, D.H.: Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. IEEE Transactions on Visualization and Computer Graphics **25**(8), 2674–2693 (Aug 2019). https://doi.org/10.1109/TVCG.2018.2843369
17. Huang, Q., Yamada, M., Tian, Y., Singh, D., Yin, D., Chang, Y.: GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. arXiv:2001.06216 [cs, stat] (Sep 2020)
18. Jin, Z., Wang, Y., Wang, Q., Ming, Y., Ma, T., Qu, H.: GNNLens: A Visual Analytics Approach for Prediction Error Diagnosis of Graph Neural Networks. IEEE Transactions on Visualization and Computer Graphics **29**(6), 3024–3038 (Jun 2023). https://doi.org/10.1109/TVCG.2022.3148107
19. Jourdan, F., Picard, A., Fel, T., Risser, L., Loubes, J.M., Asher, N.: COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for ex-

plaining neural net classifiers on NLP tasks. arXiv (2023). https://doi.org/10.48550/ARXIV.2305.06754
20. Jürß, J., Magister, L.C., Barbiero, P., Liò, P., Simidjievski, N.: Everybody Needs a Little HELP: Explaining Graphs via Hierarchical Concepts (Dec 2023). https://doi.org/10.48550/arXiv.2311.15112
21. Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C., Medya, S.: A Survey on Explainability of Graph Neural Networks (Jun 2023). https://doi.org/10.48550/arXiv.2306.01958
22. Kazius, J., McGuire, R., Bursi, R.: Derivation and Validation of Toxicophores for Mutagenicity Prediction. Journal of Medicinal Chemistry **48**(1), 312–320 (Jan 2005). https://doi.org/10.1021/jm040835a
23. Kim, B., Wattenberg, M., Gilmer, J., Cai, C.J., Wexler, J., Viégas, F., Sayres, R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: International Conference on Machine Learning (Nov 2017). https://doi.org/10.48550/arXiv.1711.11279
24. Liu, Z., Wang, Y., Bernard, J., Munzner, T.: Visualizing Graph Neural Networks with CorGIE: Corresponding a Graph to Its Embedding (Nov 2021). https://doi.org/10.48550/arXiv.2106.12839
25. Lucic, A., Hoeve, M., Tolomei, G., Rijke, M., Silvestri, F.: CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. ArXiv (Feb 2021). https://doi.org/10.48550/arXiv.2102.03322
26. Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., Zhang, X.: Parameterized explainer for graph neural network. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 19620–19631. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (Dec 2020). https://doi.org/10.5555/3495724.3497370
27. Magister, L.C., Barbiero, P., Kazhdan, D., Siciliano, F., Ciravegna, G., Silvestri, F., Jamnik, M., Lio, P.: Encoding Concepts in Graph Neural Networks (Aug 2022). https://doi.org/10.48550/arXiv.2207.13586
28. Magister, L.C., Kazhdan, D., Singh, V., Liò, P.: GCExplainer: Human-in-the-Loop Concept-based Explanations for Graph Neural Networks (Jul 2021). https://doi.org/10.48550/arXiv.2107.11889
29. Malzer, C., Baum, M.: A Hybrid Approach To Hierarchical Density-based Cluster Selection. In: 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). pp. 223–228 (Sep 2020). https://doi.org/10.1109/MFI49285.2020.9235263
30. Martínez, M.J., Sabando, M.V., Soto, A., Roca, C., Requena-Triguero, C., Campillo, N., Páez, J., Ponzoni, I.: Ames Mutagenicity Dataset for Multi-Task Learning **1** (Aug 2022). https://doi.org/10.17632/ktc6gbfsbh.1
31. Mortelmans, K., Zeiger, E.: The Ames Salmonella/microsome mutagenicity assay. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **455**(1-2), 29–60 (Nov 2000). https://doi.org/10.1016/S0027-5107(00)00064-6
32. Perotti, A., Bajardi, P., Bonchi, F., Panisson, A.: GRAPHSHAP: Explaining Identity-Aware Graph Classifiers Through the Language of Motifs (Jul 2023). https://doi.org/10.48550/arXiv.2202.08815
33. Pope, P.E., Kolouri, S., Rostami, M., Martin, C.E., Hoffmann, H.: Explainability Methods for Graph Convolutional Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10772–10781 (2019)

34. Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., vanHoesel, C., Schopmans, H., Sommer, T., Friederich, P.: Graph neural networks for materials science and chemistry. Communications Materials **3**(1), 1–18 (Nov 2022). https://doi.org/10.1038/s43246-022-00315-6
35. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (May 2019). https://doi.org/10.1038/s42256-019-0048-x
36. Shi, T., Zhang, X., Wang, P., Reddy, C.K.: Corpus-level and Concept-based Explanations for Interpretable Document Classification. ACM Transactions on Knowledge Discovery from Data **16**(3), 1–17 (Jun 2022). https://doi.org/10.1145/3477539
37. Sorkun, M.C., Khetan, A., Er, S.: AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. Scientific Data **6**(1), 143 (Aug 2019). https://doi.org/10.1038/s41597-019-0151-1
38. Tan, J., Geng, S., Fu, Z., Ge, Y., Xu, S., Li, Y., Zhang, Y.: Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning. In: Proceedings of the ACM Web Conference 2022. pp. 1018–1027. WWW '22, Association for Computing Machinery, New York, NY, USA (Apr 2022). https://doi.org/10.1145/3485447.3511948
39. Teufel, J., Torresi, L., Reiser, P., Friederich, P.: MEGAN: Multi-explanation Graph Attention Network. In: Longo, L. (ed.) Explainable Artificial Intelligence. pp. 338–360. Communications in Computer and Information Science, Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-44067-0_18
40. Tian, K., Mitchell, E., Yao, H., Manning, C., Finn, C.: Fine-tuning Language Models for Factuality (2023). https://doi.org/10.48550/ARXIV.2311.08401
41. Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J.: GNNExplainer: Generating Explanations for Graph Neural Networks. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019). https://doi.org/10.5555/3454287.3455116
42. Yuan, H., Tang, J., Hu, X., Ji, S.: XGNN: Towards Model-Level Explanations of Graph Neural Networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 430–438. KDD '20, Association for Computing Machinery, New York, NY, USA (Aug 2020). https://doi.org/10.1145/3394486.3403085
43. Yuan, H., Yu, H., Gui, S., Ji, S.: Explainability in Graph Neural Networks: A Taxonomic Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–19 (2022). https://doi.org/10.1109/TPAMI.2022.3204236
44. Yuan, H., Yu, H., Wang, J., Li, K., Ji, S.: On Explainability of Graph Neural Networks via Subgraph Explorations. In: Proceedings of the 38th International Conference on Machine Learning. pp. 12241–12252. PMLR (Jul 2021). https://doi.org/10.48550/arXiv.2102.05152
45. Zhang, R., Madumal, P., Miller, T., Ehinger, K.A., Rubinstein, B.I.P.: Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11682–11690 (May 2021). https://doi.org/10.1609/aaai.v35i13.17389
46. Zhang, Z., Liu, Q., Wang, H., Lu, C., Lee, C.: ProtGNN: Towards Self-Explaining Graph Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence **36**(8), 9127–9135 (Jun 2022). https://doi.org/10.1609/aaai.v36i8.20898