Maximum Common Subgraph Guided Graph Retrieval: Late and Early Interaction Networks

Indradyumna Roy **Soumen Chakrabarti** **Abir De**
Indian Institute of Technology Bombay
{indraroy15, soumen, abir}@cse.iitb.ac.in

Abstract

The graph retrieval problem is to search in a large corpus of graphs for ones that are most similar to a query graph. A common consideration for scoring similarity is the maximum common subgraph (MCS) between the query and corpus graphs, usually counting the number of common edges (i.e., MCES). In some applications, it is also desirable that the common subgraph be connected, i.e., the maximum common connected subgraph (MCCS). Finding exact MCES and MCCS is intractable, but may be unnecessary if ranking corpus graphs by relevance is the goal. We design fast and trainable neural functions that approximate MCES and MCCS well. Late interaction methods compute representations for the query and corpus graphs separately, and compare these representations using simple similarity functions at the last stage, leading to highly scalable systems. Early interaction methods combine information from both graphs right from the input stages, are usually considerably more accurate, but slower. We propose both late and early interaction neural MCES and MCCS formulations. They are both based on a continuous relaxation of a node alignment matrix between query and corpus nodes. For MCCS, we propose a novel differentiable 'gossip' network for estimating the size of the largest connected common subgraph. Extensive experiments with seven data sets show that our proposals are superior among late interaction models in terms of both accuracy and speed. Our early interaction models provide accuracy competitive with the state of the art, at substantially greater speeds.

1 Introduction

Given a query graph, the graph retrieval problem is to search for *relevant* or *similar* response graphs from a corpus of graphs [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Depending on the application, the notion of relevance may involve graph edit distance (GED) [12, 13, 14], the size of the maximum common subgraph (MCS) [5, 6, 7, 8, 9], full graph or subgraph isomorphism [5, 10, 11], *etc*. In this work, we focus on two variations of MCS-based relevance measures: (i) maximum common edge subgraph (MCES) [15], which has applications in distributed computing [16, 15] and molecule search [17, 18, 19, 20] and (ii) maximum common connected subgraph (MCCS) [21], which has applications in keyword search over knowledge graphs [22, 23], software development [24, 25, 26], image analysis [27, 28, 29], *etc*.

In recent years, there has been an increasing interest in designing neural graph retrieval models [5, 6, 7, 8, 9, 10]. However, most of them learn black box relevance models which provide suboptimal performance in the context of MCS based retrieval (Section 4). Moreover, they do not provide intermediate matching evidence to justify their scores and therefore, they lack interpretability. In this context, Li et al. [5] proposed a graph matching network (GMN) [5] based on a cross-graph attention mechanism, which works extremely well in practice (Section 4). Nevertheless, it suffers from three key limitations, leaving considerable scope for the design of enhanced retrieval models. (i) Similar to other graph retrieval models, it uses a general-purpose scoring layer, which renders it suboptimal in the context of MCS based graph retrieval. (ii) As acknowledged by the authors, GMN is slow in both training and inference, due to the presence of the expensive cross-attention mechanism. (iii) MCS or

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

any graph similarity function entails an *injective* mapping between nodes and edges across the graph pairs. In contrast, cross-attention induces potentially inconsistent and non-injective mappings, where a given query node can be mapped to multiple corpus nodes and vice-versa.

1.1 Our contributions

We begin by writing down (the combinatorial forms of) MCES and MCCS objectives in specific ways that facilitate subsequent adaptation to neural optimization using both late and early interaction networks. Notably, these networks are trained end-to-end, using only the distant supervision by MCES or MCCS *values* of query-corpus graph pairs, without explicit annotation of the structural mappings identifying the underlying common subgraph.

Late interaction models. We use a graph neural network (GNN) to first compute the node embeddings of the query and corpus graphs *independently* of each other and then deploy an interaction network for computing the relevance scores. This decoupling between embedding computation and interaction steps leads to efficient training and inference. We introduce LMCES and LMCCS, two late interaction neural architectures for MCES and MCCS based graph retrieval respectively. The interaction model is a differentiable graph alignment planner. It learns a Gumbel-Sinkhorn (GS) network to provide an *approximate alignment plan* between the query and corpus graphs. In contrast to GMN [5], it induces an approximately injective mapping between the nodes and edges of the query and corpus graphs. The MCES objective is then computed as a differentiable network applied to this mapping. For MCCS, we further develop a novel *differentiable gossip network* that computes the size of the largest connected component in the common subgraph estimated from the above mapping. These neural gadgets may be of independent interest.

Early interaction model. In the early interaction model, we perform the interaction step during the node embedding computation phase, which makes the query and corpus embeddings dependent on each other. This improves predictive power, at the cost of additional training and inference time. Here, we propose XMCS (cross-MCS), an early interaction model that works well for both MCES and MCCS based graph retrieval. At each propagation layer of the GNN, we first refine the alignment plan using the embeddings computed in the previous layer, then update the underlying coverage objective using the refined alignment and finally use these signals to compute the node embeddings of the current layer.

Comprehensive evaluation. We experiment with seven diverse datasets, which show that: (i) our late and early interaction models outperform the corresponding state-of-the-art methods in terms of both accuracy and inference time; (ii) in many cases, LMCES and LMCCS outperform the early interaction model of GMN [5]; and (iii) GMN's accuracy can be significantly improved by substituting its final layer with our MCS-specific neural surrogate.

1.2 Related work

Combinatorial algorithms for MCS. Both MCES and MCCS are NP-complete [15]. Several works designed heuristics for computing MCES for specific types of graphs [18, 17, 30]. Bahiense et al. [15] formulated MCES as an integer programming problem, provided a polyhedral analysis of the underlying formulation, and finally designed a branch and cut algorithm to solve it. Such polyhedral study for MCES was also performed by others [31, 32]. Combinatorial methods for different variants of MCCS have also been thoroughly studied. Some of them provide exact MCCS [33, 34, 35]. McCreesh et al. [33] proposed McSplit, a branch and bound algorithm for maximum common induced and connected graph, which is efficient in terms of time and memory. Other works provide effective heuristics to find approximate MCCS [36, 37]. However, these methods are not differentiable and therefore not suitable for data-driven MCS estimation.

Learning models for MCS. There are some recent efforts to design machine learning models for graph similarity and search [38, 7]. Among them, Bai et al. [38, GLsearch] compute MCS between two graphs using a reinforcement learning setup. In contrast, we consider a supervised learning setting for graph retrieval. Although Bai et al. [7, GraphSim] focus on the supervised learning scenario, their relevance scoring model performs poorly for MCS based retrieval.

Graph matching for computer vision. Neural models for graph matching are used in applications of computer vision for computing image similarity, object detection, etc. However, these applications permit explicit supervision of the underlying node alignments [39, 40, 41, 42, 43, 44, 45, 46, 47, 48]. They adopt different types of losses which include permutation loss [41, 42, 43, 44], Hungarian loss [46], and displacement loss [40]. In our problem, we only use distant supervision in the form of size of the underlying MCS. Moreover, these work mostly consider graph matching problems,

whereas we consider maximum common subgraph detection using distant supervision from MCS score alone.

Graph representation learning. Representation learning on graphs has been widely researched in the last decade [49, 50, 51, 52, 53, 54]. Among them, graph neural networks (GNN) are the most popular node embedding models [49, 50, 51, 52, 55, 56]. Given a node u, a GNN collects information from K-hop neighbors of the node u and applies a symmetric aggregator on top of it to obtain the representation vector of the nodes. In the context of graph retrieval, the node embeddings are used in two ways. In the first approach, they are further aggregated into graph embeddings, which are then used to compute the similarity between query and corpus graphs, by comparing the embeddings in the vector space [10, 5]. The second approach consists of SimGNN [6], GOTSim [8], GraphSim [7], GEN [5] and GMN [5], which compare the node embeddings and find suitable alignments between them. Here, GMN applies cross attention based mechanism on the node embeddings given a graph neural network [49]. Recently, Chen et al. [57] designed a structure aware transformer architecture, which can represent a subgraph around a node more effectively than several other representation models.

Differentiable solvers for combinatorial algorithms. Our work attempts to find a neural surrogate for the combinatorial challenge of maximizing objective scores over a permutation space. In effect, we are attempting to solve an Optimal Transport problem, where the central challenge is to present a neural gadget which is differentiable and backpropagable, thus enabling end-to-end training. Cuturi [58] utilized iterative row and column normalization, earlier proposed by Sinkhorn [59], to approximately solve the transportation problem subject to marginal constraints. In another approach, Vlastelica et al. [60] attempted to solve combinatorial problems *exactly* using available black-box solvers, by proposing to use the derivative of an affine surrogate of the piece-wise constant function in the backward pass. Rolínek et al. [61] leverage this to perform deep graph matching based on explicit supervision of the ground truth node alignments. In another approach, Berthet et al. [62] perturb the inputs to the discrete solvers with random noise, so as to make them differentiable. Karalias and Loukas [63] design probabilistic loss functions for tackling the combinatorial objective of selecting a subset of nodes adhering to some given property. Finally, Kotary et al. [64] present a detailed survey of the existing neural approaches for solving constrained optimization problems on graphs.

2 Late interaction models for MCES and MCCS

In this section, we first write down the exact objectives for MCES and MCCS. These expressions are partly based upon a pairing of nodes between query and corpus graphs, and partly on typical graph algorithms. They lead naturally to our subsequent development of two late interaction models, LMCES and LMCCS. We begin with formal definitions of MCES and MCCS.

Definition 1 (MCES and MCCS) *Given query and corpus graphs G_q=(V_q, E_q) and G_c=(V_c, E_c).*

(1) The maximum common edge subgraph $\mathrm{MCES}(G_q, G_c)$ is the common (not necessarily induced nor connected) subgraph between G_q and G_c, having the maximum number of edges [15].

(2) The maximum common connected subgraph $\mathrm{MCCS}(G_q, G_c)$ is the common connected subgraph with the maximum number of nodes [21].

2.1 Combinatorial formulations for MCES and MCCS

As mentioned in Section 1.2, combinatorial algorithms for MCES and MCCS abound in the literature [34, 33, 15, 18, 17, 30]. However, it is difficult to design neural surrogates for these algorithms. Therefore, we come up with the new optimization objectives for MCES and MCCS, which allow us to design neural surrogates by gradually replacing its different components with differentiable parameterized components.

Exact MCES. Given a query graph $G_q = (V_q, E_q)$ and a corpus graph $G_c = (V_c, E_c)$, we pad the graph having fewer vertices (typically, G_q), with $||V_c| - |V_q||$ disconnected nodes. This ensures that the padded graphs have an equal number of nodes N. Let us denote the adjacency matrices of G_q and G_c (after padding) as $\boldsymbol{A}_q \in \{0,1\}^{N\times N}$ and $\boldsymbol{A}_c \in \{0,1\}^{N\times N}$. To find $\mathrm{MCES}(G_q, G_c)$ from the adjacency matrices $\boldsymbol{A}_q$ and $\boldsymbol{A}_c$, we first obtain the candidate common subgraph under some proposed node alignment given by permutations $\boldsymbol{P}$ of $\boldsymbol{A}_c$, which can be characterized by the adjacency matrix $\min(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^T)$. This matrix shows the overlapping edges under the proposed node alignment. Subsequently, we choose the permutation which maximizes the total number of edges in this subgraph.

Formally, we compute the MCES score by solving a coverage maximization problem, as follows:

$$\max_{\boldsymbol{P}\in\mathcal{P}} \sum_{i,j} \min\left(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top\right)_{i,j} \tag{1}$$

where $\mathcal{P}$ is the set of permutation matrices of size $N \times N$ and the $\min$ operator is applied elementwise.

Exact MCCS. MCES does not require the common subgraph to be connected, which may be desirable in some applications. For example, in keyword search and question answering over knowledge graphs (KGs) [22, 65, 23], one may wish to have the entity nodes, forming the response, to be connected to each other. In molecule search, one may require a connected functional group to be present in both query and corpus graphs [19]. In such situations, MCCS may be more appropriate.

Given a query graph G_q and a corpus graph G_c and their adjacency matrices $\boldsymbol{A}_q$ and $\boldsymbol{A}_c$ after padding, we first apply a row-column permutation on $\boldsymbol{A}_c$ using the permutation matrix $\boldsymbol{P}$ and obtain the candidate common subgraph with adjacency matrix $\min(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top)$. Then we apply Tarjan's algorithm [66] to return the size of the largest connected component, which we maximize w.r.t. $\boldsymbol{P}$:

$$\max_{\boldsymbol{P}\in\mathcal{P}} \ \textsc{Tarjan}\mathrm{SCC}\left(\min\left(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top\right)\right). \tag{2}$$

Here $\textsc{Tarjan}\mathrm{SCC}$ takes the adjacency matrix as input and returns the size of the largest connected component of corresponding graph.

Bottleneck of approximating the optimization problems (1) **and** (2)**.** One way of avoiding the intractability of searching through $\boldsymbol{P}$, is to replace the hard permutation matrix with a differentiable soft surrogate via the Gumbel-Sinkhorn (GS) network [67, also see Appendix B]. However, such a relaxation is not adequate on its own.

1. Most elements of $\min(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top)$ are binary, which deprives the learner of gradient signals.
2. In practice, the nodes or edges may have associated (noisy) features, which play an important role in determining similarity between nodes or edges across query and corpus graphs. For example, in scene graph based image retrieval, "`panther`" may be deemed similar to "`leopard`". The objectives in Eqs. (1) and (2) do not capture such phenomenon.
3. Tarjan's algorithm first applies DFS on a graph to find the connected components and then computes the size of the largest among them, in terms of the number of nodes. Therefore, even for a fixed $\boldsymbol{P}$, it is not differentiable.

Next, we address the above bottlenecks by replacing the objectives (1) and (2) with two neural surrogates, which are summarized in Figure 1.

2.2 Design of LMCES

We design the neural surrogate of Eq. (1) by replacing the adjacency matrices with the corresponding continuous node embeddings computed by a GNN, and the hard permutation matrix with a soft surrogate—a doubly stochastic matrix—generated by the Gumbel-Sikhorn network [67]. These node embeddings allow us to compute non-zero gradients and approximate similarity between nodes and their local neighborhood in the continuous domain. Specifically, we compute this neural surrogate in the following two steps.

Step 1: Computing node embeddings. We use a message passing graph neural network [49, 68] GNN_θ with R propagation layers and trainable parameters θ, to compute the node embeddings $\boldsymbol{h}_u(1), \ldots, \boldsymbol{h}_u(R) \in \mathbb{R}^d$, for each node u in the query and corpus graphs, to which GNN_θ is applied separately. Finally, we build two matrices $\boldsymbol{H}_q(r), \boldsymbol{H}_c(r) \in \mathbb{R}^{N\times d}$ for $r \in [R]$ by stacking the node embedding vectors for query and corpus graphs. Formally, we have

$$\boldsymbol{H}_q(1), \ldots, \boldsymbol{H}_q(R) = \mathrm{GNN}_\theta(G_q), \qquad \boldsymbol{H}_c(1), \ldots, \boldsymbol{H}_c(R) = \mathrm{GNN}_\theta(G_c) \tag{3}$$

Step 2: Interaction between $\boldsymbol{H}_q(r)$ and $\boldsymbol{H}_c(r)$. In principle, the embeddings $\boldsymbol{h}_u(1), \ldots, \boldsymbol{h}_u(R)$ of a node u capture information about the neighborhood of u. Thus, we can view the set of embedding matrices $\{\boldsymbol{H}_q(r) \,|\, r \in [R]\}$ and $\{\boldsymbol{H}_c(r) \,|\, r \in [R]\}$ as a reasonable representation of the query and corpus graphs, respectively. To compute a smooth surrogate of the adjacency matrix of the common subgraph, i.e., $\min(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top)$, we seek to find the corresponding alignment between $\boldsymbol{H}_q(r)$ and $\boldsymbol{H}_c(r)$ using soft-permutation (doubly stochastic) matrices $\boldsymbol{P}^{(r)}$ generated through a Gumbel-Sikhorn network GS_ϕ. Here, we feed $\boldsymbol{H}_q(r)$ and $\boldsymbol{H}_c(r)$ into GS_ϕ and obtain a doubly stochastic matrix $\boldsymbol{P}^{(r)}$:

$$\boldsymbol{P}^{(r)} = \mathrm{GS}_\phi(\boldsymbol{H}_q(r), \boldsymbol{H}_c(r)) \quad \forall\, r \in [R] \tag{4}$$

Finally, we replace the true relevance scoring function in Eq. (1) with the following smooth surrogate:

$$s(G_q, G_c) = \sum_{r\in[R]} w_r \sum_{i,j} \min(\boldsymbol{H}_q(r), \boldsymbol{P}^{(r)}\boldsymbol{H}_c(r))_{i,j} \tag{5}$$

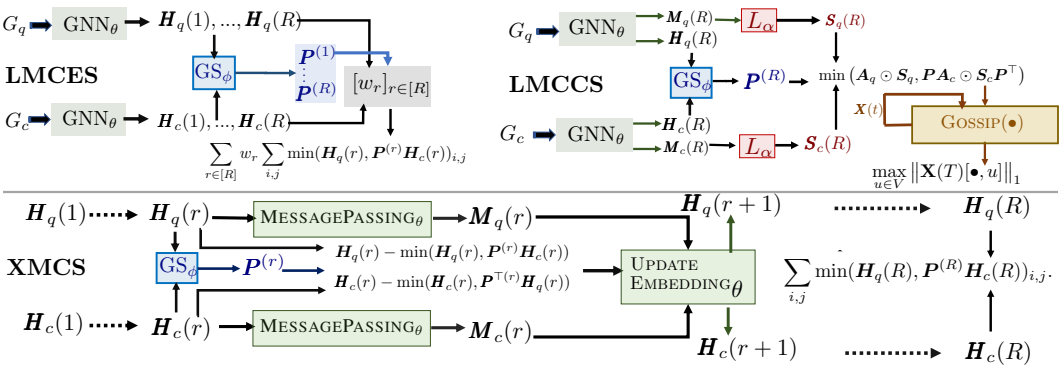

Figure 1: Proposed late (top row) and early (bottom row) interaction models. **Top Left**: LMCES encodes the given query-corpus graphs separately, using R layers of GNN_θ to compute node embedding matrices $\boldsymbol{H}_q(r), \boldsymbol{H}_c(r)$ with $r = \{1, \ldots, R\}$. For each layer r, the Gumbel-Sinkhorn network GS_ϕ computes the optimal alignment $\boldsymbol{P}^{(r)}$ between $\boldsymbol{H}_q(r)$ and $\boldsymbol{H}_c(r)$, which is finally used compute the relevance score in Eq. (5). **Top Right**: LMCCS uses GNN_θ to encode the R-hop node and edge embeddings, $\boldsymbol{H}_\bullet(R)$ and $\boldsymbol{M}_\bullet(R)$. The edge embeddings $\boldsymbol{M}_\bullet(R)$ are fed into L_α to predict edge score matrix $\boldsymbol{S}_\bullet(R)$ whose valid edge entries are encoded in $\boldsymbol{A} \odot \boldsymbol{S}$. As before, $\text{GS}_\phi(\boldsymbol{H}_q(R), \boldsymbol{H}_c(R))$ generates alignment $\boldsymbol{P}^{(R)}$ which is used to compute candidate MCS graph $\min(\boldsymbol{A}_q \odot \boldsymbol{S}_q, \boldsymbol{P}(\boldsymbol{A}_c \odot \boldsymbol{S}_c)\boldsymbol{P}^\top)$, which is then fed into the neural GOSSIP module, to predict the final MCCS score. **Bottom**: At each layer r, XMCS uses the alignment matrix $\boldsymbol{P}^{(r)}$ to factor the cross graph influences using $\boldsymbol{H}_q(r) - \min(\boldsymbol{H}_q(r), \boldsymbol{P}^{(r)}\boldsymbol{H}_c(r))$ for query, and $\boldsymbol{H}_c(r) - \min(\boldsymbol{H}_c(r), {\boldsymbol{P}^{(r)}}^T\boldsymbol{H}_q(r))$ for corpus. This is used to update the node embeddings $\boldsymbol{H}_\bullet(r) \to \boldsymbol{H}_\bullet(r+1)$. Embeddings from the final layer are used to compute the final relevance score.

Here $\{w_r \geq 0 : r \in [R]\}$ are trainable parameters, balancing the quality of signals over all message rounds r. Note that the R message rounds execute on the query and corpus graphs separately. The interaction between corresponding rounds, happens at the very end.

2.3 Design of LMCCS

In case of MCES, our key modification was to replace the adjacency matrices with node embeddings and design a differentiable network to generate a soft-permutation matrix. In the case of MCCS, we have to also replace the non-differentiable step of finding the size of the largest connected component of the common subgraph with a neural surrogate, for which we design a novel *gossip* protocol.

Differentiable gossip protocol to compute the largest connected component. Given any graph $G = (V, E)$ with the adjacency matrix $\boldsymbol{B}$, we can find its largest connected component (LCC) by using a gossip protocol. At iteration $t = 0$, we start with assigning each node a *message* vector $\boldsymbol{x}_u(0) \in \mathbb{R}^N$, which is the one-hot encoding of the node u, *i.e.*, $\boldsymbol{x}_u(0)[v] = 1$ for $v = u$ and 0 otherwise. In iteration $t > 0$, we update the message vectors $\boldsymbol{x}_u(t+1)$ as $\boldsymbol{x}_u(t+1) = \sum_{v \in \text{nbr}(u) \cup \{u\}} \boldsymbol{x}_v(t)$. Here, $\text{nbr}(u)$ is the set of neighbors of u. Initially we start with sparse vector $\boldsymbol{x}_u$ with exactly one non-zero entry. As we increase the number of iterations, u would gradually collect messages from the distant nodes which are reachable from u. This would increase the number of non-zero entries of $\boldsymbol{x}_u$. For sufficiently large value of iterations T (diameter of G), we will attain $\boldsymbol{x}_u(T)[v] \neq 0$ whenever u and v lie in the same connected component and $\boldsymbol{x}_u(T)[v] = 0$, otherwise. Once this stage is reached, one can easily compute the number of nodes in the largest connected component of G as $\max_u ||\boldsymbol{x}_u(T)||_0$, *i.e.*, the maximum number of non-zero entries in a message vector. Algorithm 1 shows the gossip protocol, with the adjacency matrix $\boldsymbol{B}$ as input and the size of the largest connected component as output.

Algorithm 1 GOSSIP($\boldsymbol{B}$)

1: $\mathbf{X}(0) = \mathbb{I}$ `# identity`
2: **for** $t = 1, \ldots T-1$ **do**
3: $\quad \mathbf{X}(t+1) \leftarrow \mathbf{X}(t)(\boldsymbol{B} + \mathbb{I})$
4: **Return** $\max_{u \in V} ||\mathbf{X}(T)[\bullet, u]||_0$

Exact MCCS computation using the gossip protocol. Given the query and corpus graphs G_q and G_c with their adjacency matrices $\boldsymbol{A}_q$ and $\boldsymbol{A}_c$, we can rewrite Eqn. (2) in the equivalent form

$$\max_{\boldsymbol{P} \in \mathcal{P}} \text{GOSSIP}(\min(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top)). \tag{6}$$

Recall that $\mathcal{P}$ is the set permutation matrices of size $N \times N$.

Neural surrogate. One can use a GS network to obtain a permutation matrix $\boldsymbol{P}$, which in principle, can support backpropagation of $\min(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top)$. However, as mentioned in Section 2.1 (items 1 and 2), $\boldsymbol{A}_\bullet$ is 0/1, and $\boldsymbol{P}$ often saturates, losing gradient signal for training. In response, we will design a neural surrogate for the true MCCS size given in Eq. (2), using three steps.

Step 1: Computation of edge embeddings. To tackle the above challenge, we introduce a parallel back-propagation path, in order to allow the GNNs to learn which edges are more important than others. To this end, we first use the GNNs to compute edge embedding vectors $\boldsymbol{m}_e(r) \in \mathbb{R}^{d_E}$ for edges $e \in E$, in addition to node embeddings at each propagation layer $r \in [R]$. Then, we gather the edge embeddings from the final layer R, into the matrices $\boldsymbol{M}_q(R), \boldsymbol{M}_c(R) \in \mathbb{R}^{|E| \times d_E}$ for query and corpus pairs. Next, we feed each separately into a neural network L_α with trainable parameters α, which predicts the importance of edges based on each edge embedding. Thus, we obtain two matrices $\boldsymbol{S}_q \in \mathbb{R}^{N \times N}$ and $\boldsymbol{S}_c \in \mathbb{R}^{N \times N}$, which are composed of the edge scores between the corresponding node pairs. Formally, we have:

$$\boldsymbol{H}_q(R), \boldsymbol{M}_q(R) = \mathrm{GNN}_\theta(G_q), \quad \boldsymbol{H}_c(R), \boldsymbol{M}_c(R) = \mathrm{GNN}_\theta(G_c) \tag{7}$$

$$\boldsymbol{S}_q = L_\alpha\left(\boldsymbol{M}_q(R)\right), \quad \boldsymbol{S}_c = L_\alpha\left(\boldsymbol{M}_c(R)\right) \tag{8}$$

In our implementation, L_α consists of one linear layer, a ReLU layer and another linear layer.

Step 2: Continuous approximation of MCCS. In MCES, we approximated the MCES score in Eq. (1) directly, using the neural coverage function defined in Eq. (5) in terms of the node embeddings. Note that, here, we did not attempt to develop any continuous approximation of $\min\left(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top\right)$—the adjacency matrix of the candidate common subgraph. However, in order to apply our proposed gossip protocol (or its neural approximation), we need an estimate of the adjacency matrix of the common subgraph. Therefore, we compute the noisy estimator as follows:

$$\boldsymbol{P} = \mathrm{GS}_\phi(\boldsymbol{H}_q(R), \boldsymbol{H}_c(R)), \qquad \boldsymbol{B} = \min\left(\boldsymbol{A}_q \odot \boldsymbol{S}_q, \boldsymbol{P}(\boldsymbol{A}_c \odot \boldsymbol{S}_c)\boldsymbol{P}^\top\right) \tag{9}$$

In the above expression, we replace all the non-zero entries of $\boldsymbol{A}_q$ and $\boldsymbol{A}_c$ with the corresponding edge importance scores of $\boldsymbol{S}_q$ and $\boldsymbol{S}_c$. This facilitates backpropagation more effectively, as opposed to the original 0/1 adjacency representation of the common subgraph (item 1 in Section 2.1). Here, we generate the permutation matrix $\boldsymbol{P}$ using Gumbel-Sinkhorn network similar to MCES, except that here we generate only one permutation matrix based on the embeddings on the final layer, whereas in MCES, we generated permutations $\boldsymbol{P}^{(r)}$ for each layer r.

Step 3: Neural implementation of gossip. Our gossip protocol in Algorithm 1 is already differentiable. However, by the virtue of the way it is designed, it would give accurate results only if the input matrix consists of 0/1 values. However, our neural estimate $\boldsymbol{B}$ contains mostly non-zero continuous values and many of them can go beyond ± 1. As a result, the resultant matrix $\mathbf{X}(t) = \mathbf{X}(0)(\boldsymbol{B} + \mathbb{I})^t$ in Algorithm 1 may suffer from extremely poor conditioning. To tackle this problem, we use a *noise filter* network at the final step T. We first estimate a dynamic threshold $\tau \in [0, \infty)$ and then set the values of $\mathbf{X}(T)$ below that threshold to zero. Finally, we scale the non-zero entries between $(0, 1)$ using a sigmoid activation $\sigma(\cdot)$. Formally, we define:

$$\tau = \textsc{Thresh}_\beta(\mathbf{X}(T)); \qquad \hat{\mathbf{X}}(T) = 2\sigma\left(\mathrm{ReLU}[\mathbf{X}(T) - \tau]/\lambda\right) - 1, \tag{10}$$

where λ is a temperature hyperparameter. $\textsc{Thresh}_\beta$ consists of a linear, a ReLU, then another linear layer. Note that $\hat{\mathbf{X}}(T) \in [0, 1]$ by design, which lets us replace the L_0 norm in the final score (Algorithm 1, line 4) with the more benign L_1 norm, followed by a maxpool over nodes:

$$s(G_q, G_c) = \max_{u \in V} \left\|\hat{\mathbf{X}}(T)[\bullet, u]\right\|_1 \tag{11}$$

Note that the interaction steps 2 and 3 above are completely decoupled from step 1 and do not have any contribution in computing embeddings.

3 Early/cross interaction model: XMCS

Although late interaction models offer efficient training and fast inference, prior work [5] suggests that early interaction models, while slower, may offer better accuracy. Motivated by such successes, we propose a unified early interaction model, called XMCS, which works for both MCES and MCCS. XMCS is slower than LMCES and LMCCS, but provides significant accuracy gains (Section 4). Moreover, it is significantly faster than GMN [5]. As before, the relevance score $s(G_q, G_c)$ is computed using a graph neural network (GNN) with R propagation layers, but each graph influences embeddings of the other graph in each layer.

Initialization of node embeddings. We start with the raw node features $\boldsymbol{z}_u$ for each node $u \in V_q \cup V_c$ and use them to initialize the node embeddings $\boldsymbol{h}_u$.

$$\boldsymbol{h}_u(0) = \text{FeatureEncoder}_\theta(\boldsymbol{z}_u) \tag{12}$$

Embedding computation via interaction between G_q and G_c. Given the node embeddings $\boldsymbol{h}_u(r)$ for some propagation layer $r < R$, we first encode the intra-graph influences across all edges in both query and corpus graphs. Accordingly, we obtain directed message vectors, for each pair of nodes $(u, v) \in E_q \cup E_c$, which are then aggregated using a simple sum aggregator.

$$\boldsymbol{m}_{uv}(r) = \text{MessagePassing}_\theta(\boldsymbol{h}_u(r), \boldsymbol{h}_v(r)) \quad \forall\, (u, v) \in E_q \cup E_c \tag{13}$$

$$\overline{\boldsymbol{m}}_u(r) = \textstyle\sum_{v \in \text{nbr}(u)} \boldsymbol{m}_{uv}(r) \quad \forall u \in V_q \cup V_c \tag{14}$$

Next, we perform the interaction step across the query and corpus graphs G_q and G_c, using a graph alignment network, which is modeled using GS_ϕ, similar to the late interaction models in Eq. (4). We build embedding matrices $\boldsymbol{H}_q(r)$ and $\boldsymbol{H}_c(r)$ by stacking $\boldsymbol{h}_u(r)$ from G_q and G_c respectively. Then, we feed them into GS_ϕ to generate an alignment matrix, and finally compute the difference of the query and the corpus graphs from the underlying MCS in the continuous embedding space:

$$\boldsymbol{P}^{(r)} = \text{GS}_\phi(\boldsymbol{H}_q(r), \boldsymbol{H}_c(r)); \qquad \boldsymbol{\Delta}_q(r) = \boldsymbol{H}_q(r) - \min(\boldsymbol{H}_q(r), \boldsymbol{P}^{(r)}\boldsymbol{H}_c(r));$$

$$\boldsymbol{\Delta}_c(r) = \boldsymbol{H}_c(r) - \min(\boldsymbol{H}_c(r), \boldsymbol{P}^{(r)\top}\boldsymbol{H}_q(r)). \tag{15}$$

Note that $\boldsymbol{\Delta}_q(r)$ in the above can also be written as $\text{ReLU}[\boldsymbol{H}_q(r) - \boldsymbol{P}^{(r)}\boldsymbol{H}_c(r)]$ (because $\min(a, b) = a - \text{ReLU}(a - b)$) and thus $\boldsymbol{\Delta}_q(r)$ (similarly, $\boldsymbol{\Delta}_c(r)$) captures the representation of a subgraph present in G_q (G_c), which is not present in G_c (G_q). Here, $\boldsymbol{P}$ provides an injective mapping from V_c to V_q, in contrast to attention-based GMN [5], which is non-injective—it assigns one corpus node to one query node but one query node to possibly multiple corpus nodes.

Next, the node embeddings $\boldsymbol{h}_u$ are updated using aggregated intra-graph and cross-graph influences.

$$\boldsymbol{h}_u(r+1) = \text{UpdateEmbedding}_\theta\big(\boldsymbol{h}_u(r), \overline{\boldsymbol{m}}_u(r), \textstyle\sum_j \boldsymbol{\Delta}(r)[u, j]\big) \quad \forall\, u \in V_q \cup V_c \tag{16}$$

The node embeddings of G_q explicitly depend on G_c via the alignment $\boldsymbol{P}^{(\bullet)}$ and vice-versa.

Relevance score computation. Finally, we compute the relevance score using the neural surrogate:

$$s(G_q, G_c) = \sum_{i,j} \min(\boldsymbol{H}_q(R), \boldsymbol{P}^{(R)}\boldsymbol{H}_c(R))_{i,j}. \tag{17}$$

Clearly, the above scoring function directly approximates the MCES objective (1) similar to the score given by LMCES in Eq. (5), except that here we use the embeddings at the last layer to compute the score. Although one can subsequently a combine gossip network with the above model, we found that it does not improve accuracy significantly and moreover, results in extreme slowdown.

4 Experiments

In this section, we provide a comprehensive evaluation of our models across seven datasets and show that they outperform several competitors [6, 7, 8, 10, 11, 5]. Our code is in https://tinyurl.com/mccs-xmcs

4.1 Experimental setup

Datasets. We experiment with seven datasets, *viz.*, MSRC-21 (MSRC), PTC-MM (MM), PTC-FR (FR), PTC-MR (MR), PTC-FM (FM), COX2 (COX) and DD. The details about them are described in Appendix C. Among these datasets, we report the results of the first six datasets in the main paper and the DD dataset in Appendix D. For every dataset, we have a corpus graph database with 800 graphs and a set of 500 query graphs, leading to 400000 query-corpus pair of graphs.

State-of-the-art methods compared. We compare our method against six state-of-the-art late interaction models, *viz.*, (i) SimGNN [6], (ii) GraphSim [7], (iii) GOTSim [8], (iv) NeuroMatch [10], (v) IsoNet [11] and (vi) Graph embedding network (GEN) [5]; and one early interaction model, *viz.*, (vii) Graph Matching Network (GMN) [5]. All methods use a general purpose scoring layer, except for NeuroMatch and IsoNet, which are specifcially designed for subgraph isomorphism.

Training and evaluation. Given corpus graphs $C = \{G_c\}$, query graphs $Q = \{G_q\}$ and their gold MCES and MCCS values $\{y_{\text{MCES}}(G_q, G_c)\}$ and $\{y_{\text{MCCS}}(G_q, G_c)\}$, for $G_q \in Q, G_c \in C$, we partition the query set into 60% training, 20% validation and 20% test folds. We train all methods by minimizing mean square error (MSE) loss between the predicted output and gold MCS (MCES or MCCS) values on the training split.

$$\min_\Lambda \textstyle\sum_{G_q \in Q, G_c \in C} (s_\Lambda(G_q, G_c) - y(G_q, G_c))^2 \tag{18}$$

MCES		MSE (lower is better)						KTau (higher is better)					
		MSRC	MM	FR	MR	FM	COX	MSRC	MM	FR	MR	FM	COX
Late	SimGNN	0.910	0.302	0.355	0.337	0.331	0.281	0.232	0.368	0.358	0.354	0.372	0.394
	GraphSim	0.629	0.274	0.282	0.274	0.261	0.249	0.461	0.432	0.458	0.454	0.500	0.403
	GOTSim	0.496	0.343	0.326	0.320	0.359	0.328	0.564	0.464	0.448	0.516	0.496	0.374
	NeuroMatch	0.582	0.308	0.282	0.795	0.604	0.269	0.632	0.488	0.516	0.548	0.535	0.514
	IsoNet	0.276	0.225	0.220	0.209	0.253	0.182	0.669	0.506	0.504	0.537	0.532	0.522
	GEN	0.426	0.311	0.273	0.284	0.324	0.277	0.627	0.416	0.468	0.456	0.456	0.466
	LMCES	0.232	0.167	0.170	0.162	0.163	0.140	0.691	0.577	0.588	0.598	0.610	0.574
Early	GMN	0.269	0.184	0.181	0.178	0.189	0.155	0.670	0.544	0.567	0.568	0.569	0.555
	XMCS	0.226	0.154	0.162	0.154	0.160	0.132	0.699	0.582	0.594	0.612	0.606	0.580

MCCS		MSE (lower is better)						KTau (higher is better)					
		MSRC	MM	FR	MR	FM	COX	MSRC	MM	FR	MR	FM	COX
Late	SimGNN	0.100	0.360	0.337	0.233	0.316	0.289	0.125	0.281	0.308	0.313	0.299	0.366
	GraphSim	0.088	0.283	0.290	0.221	0.255	0.325	0.153	0.336	0.337	0.315	0.366	0.292
	GOTSim	0.165	0.416	0.340	0.330	0.321	0.318	-0.088	0.320	0.327	0.307	0.380	0.416
	NeuroMatch	0.352	0.376	0.326	0.351	0.295	0.984	0.125	0.376	0.365	0.370	0.406	0.440
	IsoNet	0.086	0.237	0.244	0.191	0.218	0.253	0.185	0.381	0.388	0.351	0.402	0.406
	GEN	0.171	0.366	0.344	0.290	0.356	0.309	0.111	0.325	0.332	0.305	0.326	0.391
	LMCCS	0.068	0.174	0.179	0.134	0.173	0.177	0.248	0.451	0.438	0.406	0.457	0.487
Early	GMN	0.101	0.200	0.216	0.156	0.193	0.176	0.174	0.416	0.405	0.379	0.431	0.479
	XMCS	0.071	0.168	0.163	0.131	0.168	0.153	0.198	0.452	0.451	0.412	0.453	0.501

Table 1: Performance measured using mean square error (MSE) (left half) and Kendall-Tau Rank Correlation (Ktau) (right half) of our models and state-of-the-art baselines, *viz.*, SimGNN [6], GraphSim [7], GOTSim [8], Neuromatch [10], IsoNet [11], GEN [5], GMN [5] on 20% test set. Top-half and bottom-half report results for MCES and MCCS respectively. Except GMN, all SOTA methods are late interaction models. Numbers in green (blue) indicate the best performers among early (late) interaction models. Numbers in yellow indicate second best performers for late interaction models.

Here y can be either y_{MCES} or y_{MCCS} and Λ is the set of trainable parameters for the relevance scoring function $s_\Lambda(G_q, G_c)$. In the context of our models, $\Lambda = \{\theta, \phi, w_{1:R}\}$ for LMCES, $\Lambda = \{\theta, \phi, \alpha, \beta\}$ for LMCCS and $\Lambda = \{\theta, \phi\}$ for XMCS model. We use the validation split to tune various hyperparameters. Subsequently, we use the trained models to predict MCS scores between the test query graphs and the corpus graphs. For each of the query graphs in the test split, we use the predicted outputs to compute the MSE and Kendall-Tau rank correlation (KTau) values. Finally, we report the average MSE and mean KTau values across all the test query graphs.

4.2 Results

Comparison with state-of-the-art methods. In Table 1, we compare the performance of LMCES and XMCS (LMCCS and XMCS) against the state-of-the-art methods on both MCES (MCCS) based retrieval tasks. We observe: **(1)** For MCES and MCCS tasks, LMCES and LMCCS outperform all late interaction models by a substantial margin in terms of MSE and KTau across all datasets. **(2)** XMCS consistently outperforms GMN, the only early interaction baseline model. Suprisingly, GMN is also outperformed by our late interaction models, except in COX for the MCCS task. The likely reason is that GMN uses a general purpose scoring function and a cross attention mechanism that induces a non-injective mapping between the nodes and edges of the query corpus pairs. **(3)** XMCS outperforms both LMCES and LMCCS, as expected. **(4)** For both MCES and MCCS tasks, IsoNet is consistently second-best with respect to MSE. IsoNet is a subgraph isomorphism based retrieval model and its scoring function is proportional to $\sum_{i,j} \text{ReLU}(\boldsymbol{M}_q(R) - \boldsymbol{P}^{(R)}\boldsymbol{M}_c(R))_{i,j}$, which can be re-written as $\sum_{i,j}[\boldsymbol{M}_q(R) - \min(\boldsymbol{M}_q(R), \boldsymbol{P}^{(R)}\boldsymbol{M}_c(R))_{i,j}]$ (since $\min(a, b) = a - \text{ReLU}(a - b)$). Thus, the second term captures MCES score. During training, IsoNet is also able to shift and scale the above score to offset the additional term involving $\boldsymbol{M}_q(R)$, which likely allows it to outperform other baselines. **(5)** Neuromatch is excellent with respect to KTau, where it is the second best performer in six out of twelve settings. This is due to NeuroMatch's order-embedding training objective, which translates well to the KTau rank correlation scores.

Effect of replacing general purpose scoring layers with MCS customized layer. The state-of-the-art methods use a general purpose scoring functions, whereas those of our models are tailored to MCS based retrieval. In order to probe the effect of such custom MCS scoring layers, we modify the three most competitive baselines, *viz.*, IsoNet, GEN, GMN, where we replace their scoring layers with a layer tailored for MCS. Specifically, for IsoNet, we set $s(G_q, G_c) = \sum_{i,j} \min(\boldsymbol{M}_q(R), \boldsymbol{P}^{(R)}\boldsymbol{M}_c(R))_{i,j}$ and for GEN and GMN, we use $s(G_q, G_c) = \sum_i \min(\boldsymbol{h}_q(R), \boldsymbol{h}_c(R))_i$. Table 2 summarizes the results, which show that: **(1)** All three baselines

	MCES	MSRC	MM	FR	MR
Late	GEN	0.426	0.311	0.273	0.284
Late	GEN (MCS)	0.284	0.181	0.179	0.169
Late	IsoNet	0.276	0.225	0.220	0.209
Late	IsoNet (MCS)	0.260	0.187	0.178	0.173
Late	LMCES	0.232	0.167	0.170	0.162
Early	GMN	0.269	0.184	0.181	0.178
Early	GMN (MCS)	0.228	0.155	0.158	0.157
Early	XMCS	0.226	0.154	0.162	0.154

	MCCS	MSRC	MM	FR	MR
Late	GEN	0.171	0.366	0.344	0.290
Late	GEN (MCS)	0.076	0.226	0.195	0.161
Late	IsoNet	0.086	0.237	0.244	0.191
Late	IsoNet (MCS)	0.088	0.230	0.225	0.161
Late	LMCCS	0.068	0.174	0.179	0.134
Early	GMN	0.101	0.200	0.216	0.156
Early	GEN (MCS)	0.070	0.178	0.173	0.125
Early	XMCS	0.071	0.168	0.163	0.131

Table 2: Effect of replacing the general-purpose scoring layers with new layers customized to MCS on most competitive baselines, *viz.*, GEN and IsoNet (late interaction models) and GMN, across first four datasets (MSE). Numbers in green (red) indicate the best (second best) performers for early interaction models. Numbers in blue (yellow) indicate the best (second best) performers for late interaction models. The proposed modification improves performance of all baselines. However our models outperform them, even after modifying their layers, in most cases.

enjoy a significant accuracy boost from the MCS-tailored scoring layer. **(2)** LMCES and LMCCS continue to outperform MCS-tailored variants of late interaction baselines. XMCS outperforms MCS-tailored GMN in a majority of cases.

Ablation study. We consider four additional variants: (i) LMCES (final layer) where the relevance score is computed using only the embeddings of the Rth layer, (ii) LMCCS (no gossip), where we remove the gossip network and compute $s(G_q, G_c) = \sum_{i,j} \min(\boldsymbol{A}_q \odot \boldsymbol{M}_q, \boldsymbol{P}^{(R)}\boldsymbol{A}_c \odot \boldsymbol{M}_c\boldsymbol{P}^{(R)})_{i,j}$, (iii) LMCCS (no NOISE FILTER) where we set $\tau_t = 0$ in Eq. (10) and (iv) XMCS (all layers) where we compute the relevance score in Eq. (17) using embeddings from all R layers. Table 3 summarizes the results where the numbers in green (red) for early interaction models, and blue (yellow) for late interaction models, indicate the best (second best) performers.

	MCES	MSRC	MM	FR
Late	LMCES (final layer)	0.237	0.175	0.170
Late	LMCES	0.232	0.167	0.170
Early	XMCS (all layers)	0.224	0.154	0.165
Early	XMCS	0.226	0.154	0.162
	MCCS	**MSRC**	**MM**	**FR**
Late	LMCCS (no gossip)	0.166	0.241	0.240
Late	LMCCS (no NOISE FILTER)	0.068	0.194	0.206
Late	LMCCS	0.068	0.174	0.179
Early	XMCS (all layers)	0.069	0.181	0.171
Early	XMCS	0.071	0.168	0.163

Table 3: Ablation study (MSE).

We observe: **(1)** The scoring function of LMCES computed using Eq. (5) improves the performance for MSRC and MM datasets. **(2)** The gossip component is the most critical part of LMCCS. Upon removal of this component, the MSE of LMCCS significantly increases—for MSRC, it more than doubles. **(3)** The noise filter unit described in Eq. (10) is also an important component of LMCCS. Removal of this unit renders noticeable rise in MSE in MM and FR datasets. **(4)** Given an already high modeling power of XMCS, we do not observe any clear advantage if we use embeddings $\{\boldsymbol{H}_q(r), \boldsymbol{H}_c(r), r \in [R]\}$ from all R layers to compute the final score in Eq. (17).

Inference times. Tables 1 and 2 suggest that GMN is the most competitive baseline. Here, we compare the inference time, on our entire test fold, taken by GMN against our late and early interaction models. Table 4 shows the results. We observe: **(1)** XMCS is $3\times$ faster than GMN. This is because GMN's cross interaction demands processing one query-corpus pair at a time to account for variable graph sizes. In contrast, our Gumbel-Sinkhorn permutation network allows us to pad the adjacency matrices for batched processing, which results in significant speedup along with accuracy improvements. **(2)** LMCES is $2\times$ slower than LMCCS (no gossip), as it has to compute the permutation matrices for all R layers. Furthermore, LMCCS and LMCCS (no gossip) require comparable inference times.

LMCCS	LMCCS (no gossip)	LMCES	XMCS	GMN
13.78	13.96	28.13	31.10	99.54

Table 4: Inference time (s), increasing from left to right.

5 Conclusion

We proposed late (LMCES, LMCCS) and early (XMCS) interaction networks for scoring corpus graphs with respect to query graphs under a maximum common (connected) subgraph consideration. Our formulations depend on the relaxation of a node alignment matrix between the two graphs, and a neural 'gossip' protocol to measure the size of connected components. LMCES and LMCCS are superior with respect to both speed and accuracy among late interaction models. XMCS is comparable to the best early interaction models, while being much faster. Our work opens up interesting avenues for future research. It would be interesting to design neural MCS models which can also factor in similarity between attributes of nodes and edges. Another interesting direction is to design neural models for MCS detection across multiple graphs.

References

[1] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[2] Miles Ohlrich, Carl Ebeling, Eka Ginting, and Lisa Sather. Subgemini: Identifying subcircuits using a fast subgraph isomorphism algorithm. In *Proceedings of the 30th International Design Automation Conference*, pages 31–37, 1993.

[3] Edward K Wong. Model matching in robot vision by subgraph isomorphism. *Pattern Recognition*, 25(3):287–303, 1992.

[4] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[5] Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. Graph matching networks for learning the similarity of graph structured objects. In *International conference on machine learning*, pages 3835–3845. PMLR, 2019.

[6] Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun, and Wei Wang. Simgnn: A neural network approach to fast graph similarity computation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 384–392, 2019.

[7] Yunsheng Bai, Hao Ding, Ken Gu, Yizhou Sun, and Wei Wang. Learning-based efficient graph similarity computation via multi-scale convolutional set matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3219–3226, 2020.

[8] Khoa D Doan, Saurav Manchanda, Suchismit Mahapatra, and Chandan K Reddy. Interpretable graph similarity computation via differentiable optimal alignment of node embeddings. pages 665–674, 2021.

[9] Yun Peng, Byron Choi, and Jianliang Xu. Graph edit distance learning via modeling optimum matchings with constraints. 2021.

[10] Zhaoyu Lou, Jiaxuan You, Chengtao Wen, Arquimedes Canedo, Jure Leskovec, et al. Neural subgraph matching. *arXiv preprint arXiv:2007.03092*, 2020.

[11] Indradyumna Roy, Venkata Sai Velugoti, Soumen Chakrabarti, and Abir De. Interpretable neural subgraph matching for graph retrieval. *AAAI*, 2022.

[12] Zhiping Zeng, Anthony KH Tung, Jianyong Wang, Jianhua Feng, and Lizhu Zhou. Comparing stars: On approximating graph edit distance. *Proceedings of the VLDB Endowment*, 2(1):25–36, 2009.

[13] Richard Myers, RC Wison, and Edwin R Hancock. Bayesian graph edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):628–635, 2000.

[14] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and applications*, 13(1):113–129, 2010.

[15] Laura Bahiense, Gordana Manić, Breno Piva, and Cid C De Souza. The maximum common edge subgraph problem: A polyhedral investigation. *Discrete Applied Mathematics*, 160(18): 2523–2541, 2012.

[16] Shahid H. Bokhari. On the mapping problem. *IEEE Trans. Computers*, 30(3):207–214, 1981.

[17] John W Raymond, Eleanor J Gardiner, and Peter Willett. Rascal: Calculation of graph similarity using maximum common edge subgraphs. *The Computer Journal*, 45(6):631–644, 2002.

[18] John W Raymond, Eleanor J Gardiner, and Peter Willett. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *Journal of chemical information and computer sciences*, 42(2):305–316, 2002.

[19] John W Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of computer-aided molecular design*, 16(7): 521–533, 2002.

[20] Hans-Christian Ehrlich and Matthias Rarey. Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(1):68–79, 2011.

[21] Philippe Vismara and Benoît Valery. Finding maximum common connected subgraphs using clique detection or constraint satisfaction algorithms. In *International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences*, pages 358–368. Springer, 2008.

[22] Spencer Bryson, Heidar Davoudi, Lukasz Golab, Mehdi Kargar, Yuliya Lytvyn, Piotr Mierzejewski, Jaroslaw Szlichta, and Morteza Zihayat. Robust keyword search in large attributed graphs. *Information Retrieval Journal*, pages 1–23, 2020. URL `https://doi.org/10.1007/s10791-020-09379-9`.

[23] Yuanyuan Zhu, Qian Zhang, Lu Qin, Lijun Chang, and Jeffrey Xu Yu. Cohesive subgraph search using keywords in large networks. *IEEE Transactions on Knowledge and Data Engineering*, 34:178–191, 2022. URL `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9028268`.

[24] Péter Englert and Péter Kovács. Efficient heuristics for maximum common substructure search. *Journal of chemical information and modeling*, 55(5):941–955, 2015.

[25] Gabriel Hjort Blindell, Roberto Castañeda Lozano, Mats Carlsson, and Christian Schulte. Modeling universal instruction selection. In *Principles and Practice of Constraint Programming - 21st International Conference, CP 2015*, pages 609–626, 2015.

[26] Michele Sevegnani and Muffy Calder. Bigraphs with sharing. *Theoretical Computer Science*, 577:43–73, 2015.

[27] Hui Jiang and Chong-Wah Ngo. Image mining using inexact maximal common subgraph of multiple args. In *International conference on visual information system*, volume 2, page 3, 2003.

[28] Avik Hati, Subhasis Chaudhuri, and Rajbabu Velmurugan. Image co-segmentation using maximum common subgraph matching and region co-growing. In *Computer Vision - ECCV 2016 - 14th European Conference*, pages 736–752, 2016.

[29] Kim Shearer, Horst Bunke, and Svetha Venkatesh. Video indexing and similarity retrieval by largest common subgraph detection using decision trees. *Pattern Recognition*, 34(5):1075–1091, 2001.

[30] Fikret Ercal, J Ramanujam, and P Sadayappan. Task allocation onto a hypercube by recursive mincut bipartitioning. *Journal of Parallel and Distributed Computing*, 10(1):35–44, 1990.

[31] Javier Marenco. A generalization of independent set inequalities for the mapping polytope. *Proceedings of CLAIO*, 2002.

[32] Javier Marenco. New facets of the mapping polytope. *Proceedings of CLAIO*, 2006.

[33] Ciaran McCreesh, Patrick Prosser, and James Trimble. A partitioning algorithm for maximum common subgraph problems. In *The Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 712–719, 2017.

[34] Ciaran McCreesh, Samba Ndojh Ndiaye, Patrick Prosser, and Christine Solnon. Clique and constraint models for maximum common (connected) subgraph problems. In *International Conference on Principles and Practice of Constraint Programming*, pages 350–368. Springer, 2016.

[35] Ruth Hoffmann, Ciaran McCreesh, and Craig Reilly. Between subgraph isomorphism and maximum common subgraph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[36] Vincenzo Bonnici, Rosalba Giugno, Alfredo Pulvirenti, Dennis Shasha, and Alfredo Ferro. A subgraph isomorphism algorithm and its application to biochemical data. *BMC bioinformatics*, 14(7):1–13, 2013.

[37] Edmund Duesbury, John D Holliday, and Peter Willett. Maximum common subgraph isomorphism algorithms. *MATCH Communications in Mathematical and in Computer Chemistry*, 77 (2):213–232, 2017.

[38] Yunsheng Bai, Derek Xu, Yizhou Sun, and Wei Wang. Glsearch: Maximum common subgraph detection via learning to search. In *The 38th International Conference on Machine Learning, ICML 2021*, pages 588–598, 2021.

[39] Alex Nowak, Soledad Villar, Afonso S Bandeira, and Joan Bruna. Revised note on learning quadratic assignment with graph neural networks. In *2018 IEEE Data Science Workshop (DSW)*, pages 1–5. IEEE, 2018.

[40] Andrei Zanfir and Cristian Sminchisescu. Deep learning of graph matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2684–2693, 2018.

[41] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3056–3065, 2019.

[42] Hao-Ru Tan, Chuang Wang, Si-Tong Wu, Tie-Qiang Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Proxy graph matching with proximal matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9808–9815, 2021.

[43] Kaixuan Zhao, Shikui Tu, and Lei Xu. Ia-gm: A deep bidirectional learning method for graph matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3474–3482, 2021.

[44] Matthias Fey, Jan E Lenssen, Christopher Morris, Jonathan Masci, and Nils M Kriege. Deep graph matching consensus. *arXiv preprint arXiv:2001.09621*, 2020.

[45] Tianshu Yu, Runzhong Wang, Junchi Yan, and Baoxin Li. Learning deep graph matching with channel-independent embedding and hungarian attention. In *International conference on learning representations*, 2019.

[46] Tianshu Yu, Runzhong Wang, Junchi Yan, and Baoxin Li. Deep latent graph matching. In *International Conference on Machine Learning*, pages 12187–12197. PMLR, 2021.

[47] Linfeng Liu, Michael C Hughes, Soha Hassoun, and Li-Ping Liu. Stochastic iterative graph matching. *arXiv preprint arXiv:2106.02206*, 2021.

[48] Jingwei Qu, Haibin Ling, Chenrui Zhang, Xiaoqing Lyu, and Zhi Tang. Adaptive edge attention for graph matching with outliers. 2021.

[49] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[50] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.

[51] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[52] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[53] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *KDD*, pages 701–710, 2014.

[54] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *SIGKDD*, 2016.

[55] Tathagat Verma, Abir De, Yateesh Agrawal, Vishwa Vinay, and Soumen Chakrabarti. Varscene: A deep generative model for realistic scene graph synthesis. In *International Conference on Machine Learning*, pages 22168–22183. PMLR, 2022.

[56] Bidisha Samanta, Abir De, Gourhari Jana, Vicenç Gómez, Pratim Kumar Chattaraj, Niloy Ganguly, and Manuel Gomez-Rodriguez. Nevae: A deep generative model for molecular graphs. *Journal of machine learning research. 2020 Apr; 21 (114): 1-33*, 2020.

[57] Dexiong Chen, Leslie O'Bray, and Karsten Borgwardt. Structure-aware transformer for graph representation learning. *ICML*, 2022.

[58] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[59] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[60] Marin Vlastelica, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolínek. Differentiation of blackbox combinatorial solvers. *arXiv preprint arXiv:1912.02175*, 2019.

[61] Michal Rolínek, Paul Swoboda, Dominik Zietlow, Anselm Paulus, Vít Musil, and Georg Martius. Deep graph matching via blackbox differentiation of combinatorial solvers. In *European Conference on Computer Vision*, pages 407–424. Springer, 2020.

[62] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020.

[63] Nikolaos Karalias and Andreas Loukas. Erdos goes neural: an unsupervised learning framework for combinatorial optimization on graphs. *Advances in Neural Information Processing Systems*, 33:6659–6672, 2020.

[64] James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, and Bryan Wilder. End-to-end constrained optimization learning: A survey. *arXiv preprint arXiv:2103.16378*, 2021.

[65] Soumajit Pramanik, Jesujoba Oluwadara Alabi, Rishiraj Saha Roy, and Gerhard Weikum. Uniqorn: Unified question answering over rdf knowledge graphs and natural language text. *ArXiv*, abs/2108.08614, 2021. URL `https://arxiv.org/pdf/2108.08614.pdf`.

[66] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, 3rd Edition*. 2009.

[67] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018. URL `https://arxiv.org/pdf/1802.08665.pdf`.

[68] Yanli Liu, Chu-Min Li, Hua Jiang, and Kun He. A learning based branch and bound for maximum common subgraph related problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2392–2399, 2020.

[69] Ramesh Hariharan, Anand Janakiraman, Ramaswamy Nilakantan, Bhupender Singh, Sajith Varghese, Gregory Landrum, and Ansgar Schuffenhauer. Multimcs: a fast algorithm for the maximum common substructure problem on multiple molecules. *Journal of chemical information and modeling*, 51(4):788–806, 2011.

[70] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[71] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020. URL `www.graphlearning.io`.

[72] Zhaoyu Lou, Jiaxuan You, Chengtao Wen, Arquimedes Canedo, Jure Leskovec, et al. Neural subgraph matching. *arXiv preprint arXiv:2007.03092*, 2020.

[73] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[74] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix A
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix and Supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix and Supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix and Supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Appendix ans Supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No such information present in the data used in this paper.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Maximum Common Subgraph Guided Graph Retrieval: Late and Early Interaction Networks (Appendix)

A Potential limitations of our work

(1) In this paper, we have considered only the structural information encoded in the graphs. Consequently, all the nodes and edges are assigned the same label. However, in practice the graph nodes and edges may contain rich features, which must be taken into account when computing MCS scores. For example, certain applications may prohibit matching of graph components with different labels, which would potentially disallow many of the alignments currently proposed by our model. Additionally, in many real world applications involving knowledge graphs, we observe hierarchical relationships amongst entity types, which can further constraint the space of possible alignments. While our current formulations do allow for the encoding of node and edge features, this is insufficient as is to ensure such constraint-based matching.

(2) We consider MCS between two graphs. Some applications in chemical science [69] involve the computation of maximum common subgraph between three or more graphs.

B Neural parameterizations of GNN and Sinkhorn

We have already described the GNN network GNN_θ which was used for early interaction model (Section 3). Here, we describe the neural architecture of GNN_θ used in our late interaction models (Section 2) and GS_ϕ modules, used in both our late and early interaction models.

B.1 Neural parameterization of GNN_θ

Our GNN GNN_θ module is used in both LMCES and LMCCS. Given a graph $G = (V, E)$ (G can be either a query G_q and G_c), it uses $R = 5$ recurrent propagation layers to encode the node and edge embeddings. It consists of three modules, as follows.

(1) $\textsc{FeatureEncoder}_\theta$: This converts the input node features $\boldsymbol{z}_u$ into the initial node embeddings $\boldsymbol{h}_u$. In this work, we have focused solely on the structural aspects of MCS scoring. Hence, all nodes are assigned the same initial feature $\boldsymbol{z}_u = [1]$ (a single-element vector with a 1 — we want it to be oblivious to local features). The initial features are mapped to a $\dim(\boldsymbol{h}_u(0)) = 10$ dimensional embedding vector using a single $\mathbb{R}^{1\times 10}$ linear layer.

$$\boldsymbol{h}_u(0) = \textsc{FeatureEncoder}_\theta(\boldsymbol{z}_u) \quad \forall\, u \in V \tag{19}$$

Having computed the initial embedding $\boldsymbol{h}_u(0)$, it computes the embeddings $\{\boldsymbol{h}_u(r) \,|\, r \in [R]\}$ described as follows:

(2) $\textsc{MessagePassing}_\theta$: Given a pair of node embedding vectors $\boldsymbol{h}_u(r), \boldsymbol{h}_v(r)$ as input, this generates a message vector $\boldsymbol{m}_{uv}(r)$ of dimension 20 using a linear layer. A simple sum aggregation is used on the incoming messages to any node, to obtain $\overline{\boldsymbol{m}}_{uv}(r)$.

$$\boldsymbol{m}_{uv}(r) = \textsc{MessagePassing}_\theta(\boldsymbol{h}_u(r), \boldsymbol{h}_v(r)) \quad \forall\, (u, v) \in E \tag{20}$$

$$\overline{\boldsymbol{m}}_u(r) = \textstyle\sum_{v \in \mathrm{nbr}(u)} \boldsymbol{m}_{uv}(r) \quad \forall u \in V \tag{21}$$

(3) $\textsc{UpdateEmbedding}_\theta$: This uses the aggregated incoming messages $\overline{\boldsymbol{m}}$, to update the current node embedding using a Gated Recurrent Unit (GRU), as proposed in [70]. Here, the current node embeddings are treated as the hidden context of the GRU, which are updated using input $\overline{\boldsymbol{m}}$.

$$\boldsymbol{h}_u(r+1) = \textsc{UpdateEmbedding}_\theta(\boldsymbol{h}_u(r), \overline{\boldsymbol{m}}_u(r)) \quad \forall\, u \in V \tag{22}$$

Furthermore, in our early cross interaction model XMCS, the cross graph signal $\boldsymbol{\Delta}$ is concatenated to the message aggregation $\overline{\boldsymbol{m}}$, while being provided as GRU input. Given this framework, we obtain the node embedding matrices $\boldsymbol{H}_\bullet(r)$, used in LMCES, by gathering all the node embeddings $\boldsymbol{h}_u(r) \forall\, u \in V$. Similarly, we obtain the edge embedding matrices $\boldsymbol{M}_\bullet(R)$, used in LMCCS, by gathering the message vectors $\boldsymbol{m}_{uv}(R) \forall\, (u, v) \in E$.

B.2 Neural parameterization of GS_ϕ

The Gumbel-Sinkhorn network is used to generate soft-permutation matrices (or doubly stochastic matrices) based on some input matrix. (One may regard the input as analogous to logits and the output as analogous to a softmax.) In principle, the input matrix is driven by some neural network and then

it is fed into an iterative GUMBELSINKHORN operator. Given a matrix $\boldsymbol{U}$, GUMBELSINKHORN$(\cdot)$ returns a soft-permutation matrix using following differentiable iterative process:

$$\text{GUMBELSINKHORN}^0(\boldsymbol{U}) = \exp(\boldsymbol{U}/\zeta) \tag{23}$$

$$\text{GUMBELSINKHORN}^{(t+1)}(\boldsymbol{U}) = \text{ColDivide}\left(\text{RowDivide}\left(\text{GUMBELSINKHORN}^t(\boldsymbol{U})\right)\right) \tag{24}$$

$$\text{GUMBELSINKHORN}(\boldsymbol{U}) = \lim_{t\to\infty} \text{GUMBELSINKHORN}^t(\boldsymbol{U}) \tag{25}$$

ColDivide and RowDivide depict the iterative (row and column) normalization across columns and rows, *i.e.*, $[\text{ColDivide}(\mathbf{X})]_{i,j} = X_{i,j}/\sum_k X_{i,k}$ and $[\text{RowDivide}(\mathbf{X})]_{i,j} = X_{i,j}/\sum_k X_{k,j}$. The final output in Eq. (25) is also given by the solution of the following optimization problem:

$$\text{GUMBELSINKHORN}(\boldsymbol{U}) = \underset{\boldsymbol{P}\in\mathcal{B}}{\text{argmax}}\, \langle \boldsymbol{P}, \boldsymbol{U}\rangle - \zeta \sum_{i,j} P_{i,j} \log P_{i,j} \tag{26}$$

where $\mathcal{B}$ is the set of doubly stochastic matrices lying in a Birkhoff polytope and ζ is the temperature parameter. As $\zeta \to 0$, one can show that GUMBELSINKHORN(C) approaches a hard-permutation matrix.

We use $\text{GS}_\phi(\boldsymbol{H}_q(r), \boldsymbol{H}_c(r)) = \text{GUMBELSINKHORN}(FF_\phi(\boldsymbol{H}_q(r))FF_\phi(\boldsymbol{H}_c(r))^T)$ in both our late (Eqs. (4), (9)) and early interaction (Eq. (9)) models. Here, FF_ϕ is linear-ReLU-linear network which consists of a 10×16 linear layer, a ReLU activation function, and a final 16×10 linear layer. We perform 20 iterations of row and column normalizations, with a temperature of $\zeta = 0.1$.

C Additional details about experimental setup

In this section, we provide further details about the experimental setup, including the baseline models, hyperparameters, datasets and computing resources.

C.1 Dataset generation

We obtain our seven datasets from the repository of graphs maintained by TUDatasets [71], for the purpose of benchmarking GNNs. Amongst them, MSRC is a computer vision dataset, DD is a Bioinformatics dataset, and the remaining are Small Molecules datasets.

We sample the corpus graphs G_c and some seed query graphs G'_q independently of each other using the Breadth First Search (BFS) sampling strategy used in previous works [72, 11]. To sample G_c or G'_q, we first randomly choose a starting node u in a graph present in the dataset, drawn uniformly at random. We implement a randomized BFS traversal to obtain node sets of size $|V| \in [10, 15]$ in the vicinity of the starting node u. Finally, the subgraph induced by this set of nodes, gives us the sampled seed query or corpus graph. Using this process we generate first generate 800 , corpus graphs. Subsequently, we sample 500 seed query graphs under the constraint that it should be subgraph isomorphic to a fraction η of the available set of corpus graphs. We set $\eta \in [0.1, 0.4]$ similar to prior works [72, 11]. Here we used the Networkx implementation of VF2 algorithm [73] for determining subgraph isomorphism. Subsequently, we augment the seed query graph G'_q, with randomly connected nodes and edges, which gives us the final query graph G_q for MCS computation. This procedure ensures that we have a significant variation in MCS sizes across the set of corpus graphs, which is a desirable condition for any retrieval setup.

The corresponding ground truth MCS values for each of the 400000 query-corpus pairs, are generated using the combinatorial formulations for exact MCCS and MCES, as described in Section 2.1. We split the set of 500 query graphs into 60% training, 20% test and 20% validation splits.

C.2 Baseline Implementations

We use the available PyTorch implementations of all the baselines, *viz.*, SimGNN[1], GraphSim[2], GOTSim[3], NeuroMatch[4], GEN[5] and GMN[6]. In order to ensure a fair comparison between our models and the baselines, we have ensured the following:

[1] https://github.com/benedekrozemberczki/SimGNN

[2] https://github.com/khoadoan/GraphOTSim

[3] https://github.com/khoadoan/GraphOTSim

[4] https://github.com/snap-stanford/neural-subgraph-learning-GNN

[5] https://github.com/Lin-Yijie/Graph-Matching-Networks

[6] https://github.com/Lin-Yijie/Graph-Matching-Networks

1. Across all models, the node embedding dimension is fixed to 10.
2. The exact same dataset is fed as input to all models, and the same early stopping mechanism used for best model selection.
3. NeuroMatch uses anchor node annotations, which are excluded so as to ensure fair comparison with other models.
4. GEN and GMN compute the Euclidean distance between the query and corpus graph vectors, which is incompatible with MCS scoring. Therefore, we add an additional linear layer on top of their outputs, so as to help them better predict the MCS ground truths.

Furthermore, SimGNN, GraphSim and GOTSim implement their neural scoring function on one graph pair at a time. This is prohibitively slow during training. Therefore, similar to previous work [11], we implement a batched version of the available implementations of these three baselines, so as to achieve the requisite speedup.

C.3 Hyperparameter details

The node embedding size is specified as 10 across all models. During training our model and all the baselines model use early stopping with patience parameter $n_p = 50$. This means that if the Mean Squared Error (MSE) loss on the validation set does not decrease for 50 epochs, then the training process is terminated and the model with the least validation MSE is returned. In all cases, we train with a batch size of 128, using an Adam optimizer with learning rate 10^{-3} and weight decay $5{\times}10^{-4}$.

For LMCCS computation as mentioned in Eq. (10), we tune the model across a range of temperature values $\lambda \in [0.05, 50]$ using cross-validation on the validation set. For each dataset, the temperatures resulting in the best performance, are reported in Tables 5.

MCCS	MM	MR	FM	FR	DD	COX2	MSRC
λ	0.7	0.1	0.8	1.4	10	1.1	1

Table 5: Values of best temperature λ for each dataset.

C.4 Evaluation Metrics

Given corpus graphs $C{=}\{G_c\}$, query graphs $Q{=}\{G_q\}$ and their gold MCES and MCCS values $\{y_{\text{MCES}}(G_q, G_c)\}$ and $\{y_{\text{MCCS}}(G_q, G_c)\}$. For each query graph, G_q, we compute three evaluation metrics based on the model predictions $\{s_\Lambda(G_q, G_c)\}$ with Λ being the set of trainable parameters and the ground truth MCS scores $\{y(G_q, G_c)\}$ (y can be either y_{MCES} or y_{MCCS}).

Mean Square Error (MSE): It evaluates how close the model predictions are to the ground truth, and a lower MSE value indicates a better performing model.

$$\text{MSE} = \frac{1}{|Q|}\sum_{G_q \in Q}\frac{1}{|C|}\sum_{G_c \in C}\left(y(G_q, G_c) - s_\Lambda(G_q, G_c)\right)^2 \tag{27}$$

Kendall-Tau correlation (Ktau) [74]: Here, we track the number of concordant pairs N_q^+ where the model and ground truth rankings agree, and the number of discordant pairs N_q^- where they disagree. A better performing model will have a larger number of concordant predictions, which results in a higher correlation score.

$$\text{KTau} = \frac{1}{|Q|}\sum_{G_q \in Q}\frac{N_q^+ - N_q^-}{\binom{|C|}{2}} \tag{28}$$

In practice we use scipy implementation of KTau to report all the numbers in our paper.

Pairwise Ranking Reward (PairRank): Here, we track the number of concordant pairs, and normalize by the maximum number of possible concordant ranking $N_{q,\max}^+$ in an ideal case. Hence, a higher value of indicates a more accurate model, with a maximum achievable value of 1. The final reported values are computed, across the set of query graphs $Q{=}\{G_q\}$, as follows:

$$\text{PairRank} = \frac{1}{|Q|}\sum_{G_q \in Q}\frac{N_q^+}{N_{q,\max}^+} \tag{29}$$

Furthermore, for each of the evaluation metrics, along with the average across query graphs, we also report the standard error.

MCES		MSE (lower is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	SimGNN	0.910±0.044	0.302±0.009	0.355±0.021	0.337±0.015	0.331±0.014	0.281±0.012	1.210±0.073
	GraphSim	0.629±0.028	0.274±0.010	0.282±0.012	0.274±0.010	0.261±0.009	0.249±0.009	0.881±0.081
	GOTSim	0.496±0.020	0.343±0.046	0.326±0.020	0.320±0.031	0.359±0.047	0.328±0.015	0.628±0.037
	NeuroMatch	0.582±0.082	0.308±0.059	0.282±0.055	0.795±0.606	0.604±0.322	0.269±0.072	2.827±2.281
	IsoNet	0.276±0.007	0.225±0.009	0.220±0.008	0.209±0.007	0.253±0.017	0.182±0.007	0.333±0.059
	GEN	0.426±0.020	0.311±0.017	0.273±0.012	0.284±0.013	0.324±0.023	0.277±0.021	0.568±0.110
	LMCES	0.232±0.008	0.167±0.005	0.170±0.005	0.162±0.005	0.163±0.004	0.140±0.004	0.223±0.006
Early	GMN	0.269±0.006	0.184±0.004	0.181±0.005	0.178±0.004	0.189±0.005	0.155±0.006	0.273±0.008
	XMCS	0.226±0.007	0.154±0.003	0.162±0.005	0.154±0.004	0.160±0.003	0.132±0.004	0.220±0.005

MCES		MSE (lower is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	SimGNN	0.100±0.020	0.360±0.032	0.337±0.036	0.233±0.026	0.316±0.033	0.289±0.021	0.136±0.023
	GraphSim	0.088±0.018	0.283±0.023	0.290±0.029	0.221±0.019	0.255±0.024	0.325±0.027	0.123±0.020
	GOTSim	0.165±0.025	0.416±0.044	0.340±0.034	0.330±0.034	0.321±0.028	0.318±0.023	0.202±0.027
	NeuroMatch	0.352±0.088	0.376±0.052	0.326±0.042	0.351±0.191	0.295±0.085	0.984±0.689	0.624±0.205
	IsoNet	0.086±0.015	0.237±0.016	0.244±0.019	0.191±0.017	0.218±0.018	0.253±0.018	0.124±0.019
	GEN	0.171±0.025	0.366±0.029	0.344±0.033	0.290±0.030	0.356±0.030	0.309±0.022	0.197±0.025
	LMCCS	0.068±0.012	0.174±0.015	0.179±0.016	0.134±0.012	0.173±0.017	0.177±0.012	0.097±0.016
Early	GMN	0.101±0.015	0.200±0.015	0.216±0.020	0.156±0.015	0.193±0.018	0.176±0.011	0.137±0.016
	XMCS	0.071±0.014	0.168±0.014	0.163±0.018	0.131±0.013	0.168±0.017	0.153±0.009	0.102±0.017

Table 6: Performance measured using **mean square error (MSE)** with standard error, of our models and state-of-the-art baselines on 20% test set, for all seven datasets. Top-half and bottom-half report results for MCES and MCCS respectively. Numbers in green (blue) indicate the best performers among early (late) interaction models. Numbers in red1 (yellow) indicate second best performers for early (late) interaction models.

C.5 Hardware and Software details

We implement our models using Python 3.8.5 and PyTorch 1.10.2. Training of our models and the baselines, was performed across servers containing Xeon E5-2620 2.10GHz CPUs, Nvidia Titan Xp-12 GB GPUs, Nvidia T4-16 GB GPUs, and Nvidia Quadro RTX 6000-48 GB GPUs. Running times are compared on the same GPU, averaged over 10 runs of each method.

C.6 License details

SimGNN repository is available under GNU license, while GEN, GMN and GOTSim are available under MIT license.

D Additional experiments

D.1 Results on comparison with SOTA methods along with standard error

In Table 1 in the main submission, we neither reported the result on DD dataset nor the standard error due to space constraints. In Tables 6– 7, we report results with standard error on all seven datsets. Here, the standard error is computed over the variation across all test queries. Moreover, in Table 8, we also report results of PairRank metric. They reveal similar observations as in Table 1.

D.2 Effect of MCS layer

In Table 2 in the main submission, we reported MSE on four datasets. In Tables 9– 10, we report results with standard error, on all seven datsets, which probe the effect of substituting the general purpose scoring layer with MCS customized scoring layer. Here, the standard error is computed over the variation across all test queries. They reveal similar insights as in Table 2.

D.3 Ablation Study

In Table 3 in the main submission, we reported MSE for an ablation study on three datasets. In Tables 11–12, we report results with standard error, on all seven datasets. Here, the standard error is computed over the variation across all test queries. We make the following observations:

1. LMCES (final layer), where the relevance score is computed using only the embeddings of the Rth layer, is outperformed by LMCES in 12 out of 14 cases across MSE and KTau.
2. LMCCS (no gossip), where we remove the gossip network and compute $s(G_q, G_c) = \sum_{i,j} \min(\boldsymbol{A}_q \odot \boldsymbol{M}_q, \boldsymbol{P}^{(R)} \boldsymbol{A}_c \odot \boldsymbol{M}_c \boldsymbol{P}^{(R)})_{i,j}$, is consistently the worst performed amongst all three MCCS late interaction variants.
3. LMCCS (no NOISE FILTER) where we set $\tau_t = 0$ in Eq. (10), is outperformed by LMCES in 12 out of 14 cases across MSE and KTau.
4. There is no clear winner between XMCS, and XMCS (all layers) where we compute the relevance score in Eq. (17) using embeddings from all R layers. We observe that the performance scores of

	MCES	KTau (higher is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	SimGNN	0.232±0.003	0.368±0.007	0.358±0.007	0.354±0.007	0.372±0.005	0.394±0.011	0.215±0.004
	GraphSim	0.461±0.004	0.432±0.008	0.458±0.006	0.454±0.005	0.500±0.006	0.403±0.009	0.570±0.003
	GOTSim	0.564±0.004	0.464±0.009	0.448±0.007	0.516±0.006	0.496±0.006	0.374±0.013	0.608±0.003
	NeuroMatch	0.632±0.005	0.488±0.010	0.516±0.007	0.548±0.008	0.535±0.007	0.514±0.011	0.667±0.008
	IsoNet	0.669±0.004	0.506±0.010	0.504±0.008	0.537±0.007	0.532±0.007	0.522±0.012	0.698±0.005
	GEN	0.627±0.005	0.416±0.013	0.468±0.011	0.456±0.012	0.456±0.015	0.466±0.014	0.635±0.010
	LMCES	0.691±0.004	0.577±0.006	0.588±0.006	0.598±0.005	0.610±0.004	0.574±0.009	0.724±0.004
Early	GMN	0.670±0.003	0.544±0.006	0.567±0.007	0.568±0.006	0.569±0.006	0.555±0.009	0.701±0.004
	XMCS	0.699±0.004	0.582±0.006	0.594±0.006	0.612±0.005	0.606±0.005	0.580±0.009	0.724±0.004

	MCES	KTau (higher is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	SimGNN	0.125±0.008	0.281±0.006	0.308±0.007	0.313±0.008	0.299±0.006	0.366±0.008	0.115±0.009
	GraphSim	0.153±0.009	0.336±0.008	0.337±0.008	0.315±0.009	0.366±0.007	0.292±0.006	0.146±0.010
	GOTSim	-0.088±0.007	0.320±0.008	0.327±0.008	0.307±0.009	0.380±0.008	0.416±0.009	-0.092±0.007
	NeuroMatch	0.125±0.011	0.376±0.010	0.365±0.009	0.370±0.012	0.406±0.010	0.440±0.009	0.142±0.014
	IsoNet	0.185±0.013	0.381±0.012	0.388±0.012	0.351±0.014	0.402±0.011	0.406±0.009	0.168±0.015
	GEN	0.111±0.013	0.325±0.011	0.332±0.012	0.305±0.011	0.326±0.011	0.391±0.011	0.137±0.015
	LMCCS	0.248±0.013	0.451±0.012	0.438±0.014	0.406±0.014	0.457±0.012	0.487±0.012	0.212±0.015
Early	GMN	0.174±0.013	0.416±0.010	0.405±0.012	0.379±0.013	0.431±0.011	0.479±0.011	0.173±0.014
	XMCS	0.198±0.014	0.452±0.012	0.451±0.014	0.412±0.014	0.453±0.012	0.501±0.011	0.201±0.015

Table 7: Performance measured using **Kendall Tau Rank Correlation (KTau)** with standard error, of our models and state-of-the-art baselines on 20% test set, for all seven datasets. Top-half and bottom-half report results for MCES and MCCS respectively. Numbers in green (blue) indicate the best performers among early (late) interaction models. Numbers in red (yellow) indicate second best performers for early (late) interaction models.

	MCES	PairRank (higher is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	SimGNN	0.644±0.002	0.768±0.005	0.756±0.005	0.754±0.005	0.764±0.004	0.796±0.007	0.631±0.003
	GraphSim	0.785±0.003	0.814±0.005	0.828±0.004	0.824±0.003	0.852±0.004	0.803±0.005	0.846±0.002
	GOTSim	0.848±0.002	0.837±0.006	0.821±0.005	0.868±0.004	0.850±0.004	0.781±0.008	0.868±0.001
	NeuroMatch	0.891±0.002	0.854±0.006	0.870±0.005	0.890±0.005	0.877±0.005	0.887±0.005	0.904±0.004
	IsoNet	0.913±0.001	0.867±0.006	0.861±0.006	0.883±0.004	0.875±0.005	0.893±0.007	0.923±0.002
	GEN	0.887±0.003	0.801±0.009	0.835±0.008	0.825±0.008	0.821±0.010	0.851±0.009	0.885±0.005
	LMCES	0.927±0.001	0.921±0.003	0.921±0.004	0.927±0.003	0.930±0.002	0.933±0.003	0.939±0.001
Early	GMN	0.914±0.001	0.896±0.003	0.906±0.004	0.905±0.003	0.901±0.004	0.919±0.004	0.925±0.001
	XMCS	0.932±0.001	0.925±0.003	0.926±0.004	0.937±0.002	0.927±0.002	0.938±0.003	0.939±0.001

	MCES	PairRank (higher is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	SimGNN	0.752±0.016	0.772±0.009	0.804±0.008	0.843±0.008	0.791±0.009	0.812±0.008	0.698±0.015
	GraphSim	0.810±0.017	0.810±0.007	0.828±0.007	0.841±0.008	0.843±0.007	0.748±0.006	0.774±0.017
	GOTSim	0.311±0.010	0.802±0.009	0.818±0.008	0.839±0.010	0.855±0.007	0.850±0.007	0.312±0.012
	NeuroMatch	0.716±0.021	0.844±0.007	0.849±0.006	0.892±0.007	0.875±0.006	0.869±0.006	0.724±0.020
	IsoNet	0.849±0.017	0.845±0.009	0.866±0.006	0.865±0.009	0.869±0.006	0.844±0.007	0.775±0.020
	GEN	0.667±0.020	0.801±0.009	0.823±0.010	0.831±0.009	0.804±0.009	0.829±0.008	0.671±0.023
	LMCCS	0.863±0.019	0.904±0.004	0.909±0.005	0.915±0.005	0.912±0.004	0.903±0.005	0.831±0.018
Early	GMN	0.818±0.017	0.878±0.006	0.881±0.006	0.895±0.006	0.893±0.005	0.898±0.005	0.805±0.017
	XMCS	0.863±0.019	0.903±0.004	0.919±0.004	0.925±0.005	0.910±0.004	0.916±0.005	0.854±0.017

Table 8: Performance measured using **Pairwise Ranking Reward (PairRank)** with standard error, of our models and state-of-the-art baselines on 20% test set, for all seven datasets. Top-half and bottom-half report results for MCES and MCCS respectively. Numbers in green (blue) indicate the best performers among early (late) interaction models. Numbers in red1 (yellow) indicate second best performers for early (late) interaction models.

both variants are quite close to each other in most cases, with XMCS gaining a slight edge over XMCS (all layers) in 14 out of 28 cases.

D.4 Recovering latent linear combinations of MCES and MCCS scores

In some applications, the ground truth relevance scores may be unknown or equal to a noisy latent function MCCS and MCES size. To simulate this scenario, here, we evaluate the performance when the ground truth relevance label is a convex combination of MCCS and MCES scores, *i.e.*, $a \cdot \text{MCCS-size} + (1-a) \cdot \text{MCES-size}$. We experiment with our late (LMCES and LMCCS) and early (XMCS) interaction models, as well as the baseline late (GEN) and early(GMN) interaction models equipped with our custom MCS layer. Additionally, we implement a new model COMBO, whose relevance score $s(G_q, G_c) = \sum w_1 s_{\text{LMCCS}}(G_q, G_c) + w_2 s_{\text{LMCES}}(G_q, G_c)$. Here,$\{w_r \geq 0 : r \in [2]\}$ are trainable parameters, which attempt to balance the signals from the LMCCS and LMCES scoring functions, in order to predict any given combination of ground truths. In Table 13, we report the

	MCES	MSE (lower is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	GEN	0.426±0.020	0.311±0.017	0.273±0.012	0.284±0.013	0.324±0.023	0.277±0.021	0.568±0.110
	GEN (MCS)	0.284±0.008	0.181±0.005	0.179±0.005	0.169±0.004	0.177±0.004	0.153±0.006	0.298±0.008
	IsoNet	0.276±0.007	0.225±0.009	0.220±0.008	0.209±0.007	0.253±0.017	0.182±0.007	0.333±0.059
	IsoNet (MCS)	0.260±0.011	0.187±0.012	0.178±0.005	0.173±0.005	0.175±0.004	0.148±0.005	0.256±0.011
	LMCES	0.232±0.008	0.167±0.005	0.170±0.005	0.162±0.005	0.163±0.004	0.140±0.004	0.223±0.006
Early	GMN	0.269±0.006	0.184±0.004	0.181±0.005	0.178±0.004	0.189±0.005	0.155±0.006	0.273±0.008
	GMN (MCS)	0.228±0.005	0.155±0.003	0.158±0.004	0.157±0.003	0.162±0.003	0.134±0.003	0.217±0.004
	XMCS	0.226±0.007	0.154±0.003	0.162±0.005	0.154±0.004	0.160±0.003	0.132±0.004	0.220±0.005

	MCES	MSE (lower is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	GEN	0.171±0.025	0.366±0.029	0.344±0.033	0.290±0.030	0.356±0.030	0.309±0.022	0.197±0.025
	GEN (MCS)	0.076±0.014	0.226±0.020	0.195±0.018	0.161±0.014	0.204±0.021	0.180±0.012	0.108±0.019
	IsoNet	0.086±0.015	0.237±0.016	0.244±0.019	0.191±0.017	0.218±0.018	0.253±0.018	0.124±0.019
	IsoNet (MCS)	0.088±0.016	0.230±0.020	0.225±0.019	0.161±0.014	0.206±0.021	0.195±0.014	0.119±0.018
	LMCCS	0.068±0.012	0.174±0.015	0.179±0.016	0.134±0.012	0.173±0.017	0.177±0.012	0.097±0.016
Early	GMN	0.101±0.015	0.200±0.015	0.216±0.020	0.156±0.015	0.193±0.018	0.176±0.011	0.137±0.016
	GMN (MCS)	0.070±0.014	0.178±0.016	0.173±0.018	0.125±0.011	0.164±0.018	0.154±0.011	0.098±0.016
	XMCS	0.071±0.014	0.168±0.014	0.163±0.018	0.131±0.013	0.168±0.017	0.153±0.009	0.102±0.017

Table 9: Performance measured using **mean square error (MSE)** with standard error, showing effect of replacing the general-purpose scoring layers with **new layers customized to MCS** on most competitive baselines, *viz.*, GEN and IsoNet (late interaction models) and GMN, across all seven datasets. Numbers in green (red) indicate the best (second best) performers for early interaction models. Numbers in blue (yellow) indicate the best (second best) performers for late interaction models. The proposed modification improves performance of all baselines. However our models outperform them, even after modifying their layers, in most cases.

	MCES	KTau (higher is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	GEN	0.627±0.005	0.416±0.013	0.468±0.011	0.456±0.012	0.456±0.015	0.466±0.014	0.635±0.010
	GEN (MCS)	0.666±0.004	0.556±0.006	0.565±0.007	0.584±0.006	0.590±0.005	0.555±0.010	0.690±0.004
	IsoNet	0.669±0.004	0.506±0.010	0.504±0.008	0.537±0.007	0.532±0.007	0.522±0.012	0.698±0.005
	IsoNet (MCS)	0.677±0.004	0.569±0.006	0.570±0.007	0.585±0.005	0.589±0.005	0.565±0.009	0.708±0.004
	LMCES	0.691±0.004	0.577±0.006	0.588±0.006	0.598±0.005	0.610±0.004	0.574±0.009	0.724±0.004
Early	GMN	0.670±0.003	0.544±0.006	0.567±0.007	0.568±0.006	0.569±0.006	0.555±0.009	0.701±0.004
	GMN (MCS)	0.693±0.004	0.583±0.006	0.592±0.006	0.598±0.005	0.606±0.005	0.573±0.010	0.727±0.004
	XMCS	0.699±0.004	0.582±0.006	0.594±0.006	0.612±0.005	0.606±0.005	0.580±0.009	0.724±0.004

	MCES	KTau (higher is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	GEN	0.111±0.013	0.325±0.011	0.332±0.012	0.305±0.011	0.326±0.011	0.391±0.011	0.137±0.015
	GEN (MCS)	0.223±0.013	0.421±0.011	0.418±0.013	0.385±0.014	0.424±0.011	0.477±0.010	0.218±0.015
	IsoNet	0.185±0.013	0.381±0.012	0.388±0.012	0.351±0.014	0.402±0.011	0.406±0.009	0.168±0.015
	IsoNet (MCS)	0.187±0.013	0.440±0.011	0.404±0.011	0.386±0.013	0.434±0.011	0.480±0.011	0.182±0.014
	LMCCS	0.248±0.013	0.451±0.012	0.438±0.014	0.406±0.014	0.457±0.012	0.487±0.012	0.212±0.015
Early	GMN	0.174±0.013	0.416±0.010	0.405±0.012	0.379±0.013	0.431±0.011	0.479±0.011	0.173±0.014
	GMN (MCS)	0.192±0.014	0.450±0.011	0.443±0.014	0.408±0.014	0.458±0.012	0.499±0.011	0.198±0.016
	XMCS	0.198±0.014	0.452±0.012	0.451±0.014	0.412±0.014	0.453±0.012	0.501±0.011	0.201±0.015

Table 10: Performance measured using **Kendall Tau Rank Correlation (KTau)** with standard error, showing effect of replacing the general-purpose scoring layers with **new layers customized to MCS** on most competitive baselines, *viz.*, GEN and IsoNet (late interaction models) and GMN, across all seven datasets. Numbers in green (red) indicate the best (second best) performers for early interaction models. Numbers in blue (yellow) indicate the best (second best) performers for late interaction models. The proposed modification improves performance of all baselines. However our models outperform them, even after modifying their layers, in most cases.

performance in terms of MSE and KTau with standard error, for two latent (hidden from the learner) values of a, *viz.*, $a = 0.3$ and $a = 0.7$. We make the following observations:

1. GEN (MCS) is consistently outperformed in all cases, by one of our late interaction variants LMCES, LMCCS, or COMBO.
2. LMCCS is seen to be very susceptible to noise. It does not perform well even for $a = 0.7$ (30% MCES noise). In fact, its performance deteriorates rapidly for MCES dataset, due to the noisy MCES signals, with the performance for $a = 0.3$ being significantly worse than $a = 0.7$.
3. LMCES is seen to be more robust to noise, and is the best performer in six out of eight cases.

	MCES	MSE (lower is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	LMCES (final layer)	0.237±0.008	0.175±0.005	0.170±0.005	0.166±0.004	0.167±0.004	0.140±0.004	0.226±0.005
Late	LMCES	0.232±0.008	0.167±0.005	0.170±0.005	0.162±0.005	0.163±0.004	0.140±0.004	0.223±0.006
Early	XMCS (all layers)	0.224±0.008	0.154±0.004	0.165±0.004	0.152±0.004	0.155±0.003	0.128±0.003	0.223±0.004
Early	XMCS	0.226±0.007	0.154±0.003	0.162±0.005	0.154±0.004	0.160±0.003	0.132±0.004	0.220±0.005

	MCES	MSE (lower is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	LMCCS (no gossip)	0.166±0.017	0.241±0.017	0.240±0.017	0.187±0.014	0.237±0.019	0.224±0.015	0.293±0.027
Late	LMCCS (no NOISE FILTER)	0.068±0.012	0.194±0.016	0.206±0.024	0.140±0.013	0.177±0.016	0.205±0.018	0.104±0.018
Late	LMCCS	0.068±0.012	0.174±0.015	0.179±0.016	0.134±0.012	0.173±0.017	0.177±0.012	0.097±0.016
Early	XMCS (all layers)	0.069±0.014	0.181±0.016	0.171±0.016	0.129±0.013	0.172±0.017	0.156±0.010	0.103±0.017
Early	XMCS	0.071±0.014	0.168±0.014	0.163±0.018	0.131±0.013	0.168±0.017	0.153±0.009	0.102±0.017

Table 11: Performance measured using **mean square error (MSE)** with standard error, on the four variants of our models considered for **Ablation study**. (i) LMCES (final layer) where the relevance score is computed using only the embeddings of the Rth layer, (ii) LMCCS (no gossip), where we remove the gossip network and compute $s(G_q, G_c) = \sum_{i,j} \min(\boldsymbol{A}_q \odot \boldsymbol{M}_q, \boldsymbol{P}^{(R)} \boldsymbol{A}_c \odot \boldsymbol{M}_c \boldsymbol{P}^{(R)})_{i,j}$, (iii) LMCCS (no NOISE FILTER) where we set $\tau_t = 0$ in Eq. (10) and (iv) XMCS (all layers) where we compute the relevance score in Eq. (17) using embeddings from all R layers. Numbers in green (red) indicate the best (second best) performers for early interaction models. Numbers in blue (yellow) indicate the best (second best) performers for late interaction models.

	MCES	KTau (higher is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	LMCES (final layer)	0.692±0.004	0.562±0.006	0.581±0.006	0.594±0.005	0.599±0.005	0.566±0.009	0.720±0.004
Late	LMCES	0.691±0.004	0.577±0.006	0.588±0.006	0.598±0.005	0.610±0.004	0.574±0.009	0.724±0.004
Early	XMCS (all layers)	0.700±0.004	0.587±0.006	0.580±0.006	0.610±0.005	0.613±0.005	0.582±0.010	0.722±0.004
Early	XMCS	0.699±0.004	0.582±0.006	0.594±0.006	0.612±0.005	0.606±0.005	0.580±0.009	0.724±0.004

	MCES	KTau (higher is better)						
		MSRC	MM	FR	MR	FM	COX	DD
Late	LMCCS (no gossip)	0.132±0.011	0.418±0.011	0.380±0.012	0.360±0.013	0.415±0.011	0.476±0.011	0.091±0.008
Late	LMCCS (no NOISE FILTER)	0.241±0.013	0.436±0.011	0.433±0.013	0.392±0.014	0.446±0.011	0.477±0.011	0.213±0.015
Late	LMCCS	0.248±0.013	0.451±0.012	0.438±0.014	0.406±0.014	0.457±0.012	0.487±0.012	0.212±0.015
Early	XMCS (all layers)	0.201±0.014	0.455±0.012	0.448±0.014	0.410±0.015	0.453±0.012	0.500±0.011	0.196±0.015
Early	XMCS	0.198±0.014	0.452±0.012	0.451±0.014	0.412±0.014	0.453±0.012	0.501±0.011	0.201±0.015

Table 12: Performance measured using **Kendall Tau correlation (KTau)** with standard error, on the four variants of our models considered for **Ablation study** mentioned in Table 11. Numbers in green (red) indicate the best (second best) performers for early interaction models. Numbers in blue (yellow) indicate the best (second best) performers for late interaction models.

4. COMBO is overall seen to be the most well-balanced in terms of performance. Although it is the best performer in two out of eight cases, it is the second best close behind LMCES, for the remaining cases.
5. XMCS is seen to be able to adapt well to this noisy setup, and performs better than GMN (MCS) in five out of eight cases.

It is encouraging to see that XMCS does not need customization based on connectedness of the MCS, but we remind the reader that late interaction methods are substantially faster and may still have utility.

D.5 Interpretabilty

For both MCES and MCES, our models propose a soft alignment between the nodes of the query-corpus pair. We use the Hungarian algorithm on top of it to obtain an injective mapping $\boldsymbol{P}$, which is depicted by matching node colors in the example graph pairs in Figure 2 and Figure 3. Subsequently, we compute the adjacency matrix of the MCS graph under the proposed alignment as $\min(\boldsymbol{A}_q, \boldsymbol{P}\boldsymbol{A}_c\boldsymbol{P}^\top)$. For LMCES, we indicate the edges of the proposed MCS graph in **thick black**. For LMCCS, we further apply TARJANSCC, to identify the largest connected component, whose edges are again indicated in **thick black**. In Figure 2, we present one example each, of the proposed alignments in the MCES and MCCS settings. For MCES, there are 10 overlapping edges under the proposed node alignment, which are in two disconnected components of 7 and 3 edges. For MCCS, the proposed node alignment identifies a set of connected 8 connected nodes, common to both graphs, which are connected by **thick black** edges.

In Figure 3, we present an example graph pair, where XMCS is able to identify a larger common connected component, as compared to LMCCS. In the alignment shown on the left, we see that there are 6 nodes in the connected component, identified under the node alignment proposed by LMCCS.

Combined		MSE (lower is better)				KTau (higher is better)			
		MSRC		MM		MSRC		MM	
		$a = 0.3$	$a = 0.7$	$a = 0.3$	$a = 0.7$	$a = 0.3$	$a = 0.7$	$a = 0.3$	$a = 0.7$
Late	GEN (MCS)	0.150±0.004	0.071±0.008	0.146±0.005	0.165±0.012	0.664±0.004	0.644±0.004	0.567±0.005	0.549±0.005
	LMCES	0.125±0.004	0.066±0.008	0.143±0.006	0.177±0.013	0.692±0.004	0.675±0.004	0.587±0.005	0.565±0.005
	LMCCS	1.044±0.069	0.230±0.011	0.172±0.007	0.166±0.011	0.553±0.010	0.528±0.011	0.561±0.006	0.554±0.006
	COMBO	0.130±0.004	0.068±0.008	0.145±0.006	0.140±0.012	0.687±0.004	0.665±0.004	0.580±0.005	0.578±0.005
Early	GMN (MCS)	0.124±0.003	0.060±0.007	0.129±0.005	0.130±0.009	0.692±0.004	0.677±0.004	0.587±0.005	0.574±0.005
	XMCS	0.121±0.003	0.063±0.008	0.124±0.005	0.129±0.009	0.695±0.004	0.671±0.004	0.593±0.005	0.572±0.005

Table 13: Performance when **the ground truth is the convex combination of MCES and MCCS sizes**, *i.e.*, $a \cdot$MCCS-size$+(1-a)\cdot$MCES-size, for two values of a *viz.*, $a \in \{0.3, 0.7\}$. Performance measured using mean square error (MSE) (left half) and Kendall-Tau Rank Correlation (Ktau) (right half) with standard error, of our models and the baselines GEN and GMN, for two datasets. Numbers in green (red) indicate the best (second best) performers for early interaction models. Numbers in blue (yellow) indicate the best (second best) performers for late interaction models.

On the other hand, on the right we observe that XMCS is able to propose a node alignment, which leads to the emergence of a connected component with 10 nodes.

(a) LMCES (b) LMCCS

Figure 2: Example graph pairs with node colors indicating the proposed alignments by our models. In Panel (a), we show MCES setup. Here, Nodes are aligned to maximize the number of overlapping edges (indicated as **thick black** edges). We observe two disconnected components in the MCES setup. In Panel (b), we show MCCS setup. Here, Nodes are aligned to identify largest connected component (identified by **thick black** edges).

(a) LMCCS (b) XMCS

Figure 3: Example graph pairs with node colors indicating the proposed alignments by our models LMCCS and XMCS, for the MCCS setup. In Panel (a), LMCCS proposes an alignment which results in an MCCS score of 6. In Panel (b), XMCS proposes an alignment on the same query-corpus pair, which results in an MCCS score of 10. The connected nodes in both cases are identified by **thick black** edges.

D.6 Training and inference times

Here, we report both the training and inference time (in seconds). The training time is computed for each batch of size 128 (which is the fixed hyperparameter used for the numbers reported in the paper). Inference time is computed for the entire test set of 100 query graphs and 800 corpus graphs, with the maximum possible batch size allowed by our GPU - Nvidia TITAN X (Pascal).

Method	Training Time per batch (in secs)	Inference time on test(in secs)
GEN	0.037	5.937
SimGNN	0.073	24.671
GraphSim	0.125	23.284
NeuroMatch	0.027	6.532
GOTSim	0.259	72.590
IsoNet	0.069	13.956
GMN	0.426	99.542
LMCES	0.109	28.129
LMCCS	0.074	13.776
XMCS	0.159	31.101

Table 14: Training and inference time

We observe that: (1) Among the late interaction models, both LMCCS and LMCES are significantly faster than GOTSim. This is because GOTSim uses a combinatorial solver, which does not allow for batched processing and results in significant slowdown. (2) LMCES is comparable to GraphSim and SimGNN, in both training and inference times, while affording significantly better performance. (3) LMCCS is comparable to IsoNet in training and inference times, and is significantly faster than SimGNN, GraphSim and GOTSim. (4) GEN and NeuroMatch are significantly faster than all late interaction models in terms of training and inference times. However, as shown in Table 1 of the main paper, both these models are outperformed by LMCCS,LMCES and IsoNet in most of the datasets. (5) Among the early interaction models, XMCS is 3X faster than GMN. The reason for this is explained in lines 338-344 in our main paper.

It is true that we performed experiments on graphs of size < 20, driven by the needs of practical graph retrieval applications like molecular fingerprint detection, object detection in images, etc. However, our method can easily scale beyond $|V| = 20$, as follows:

Inference Time (in sec)	Current Size	$\lvert V\rvert = 30$	$\lvert V\rvert = 50$	$\lvert V\rvert = 70$	$\lvert V\rvert = 100$
LMCES	0.036	0.048	0.052	0.071	0.106
LMCCS	0.022	0.031	0.039	0.042	0.064
XMCS	0.067	0.071	0.085	0.095	0.131

Table 15: Scalability for large graphs