A Dataset documentation

A.1 Motivation

Q1 For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

To assess how well LLMs can understand graph patterns, we create a series of datasets for evaluation. The specific tasks include pattern translation, pattern detection, pattern modification, isomorphic pattern matching, densely connected subgraph detection such as k-core, frequent subgraph extraction, pattern discrimination, and classification. Additionally, we provide molecule pattern detection and graph classification tasks to evaluate whether LLMs can extend their understanding to real-world graph data.

Q2 Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

It violates the double-blind policy, and we will release it after the paper is accepted.

Q3 Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

It violates the double-blind policy, and we will release it after the paper is accepted.

Q4 Any other comments? No.

A.2 Composition

Q5 What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

We provide pickle files for graph pattern information, and JSON files for the prompts and responses of LLMs during pattern discrimination and downstream tasks. The graph pattern files provide a list of graph patterns using NetworkX graph format. The discrimination files provide input texts. Meanwhile, we also provide a pickle file with index when we need sampling ids from the dataset.

Q6 How many instances are there in total (of each type, if appropriate)?

The instance numbers for synthetic data and real-world data are summarized in Table 2 and Table 1,respectively.

Q7 Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances from a larger set. For graphs, the larger set consists of all nonisomorphic simple graphs with a given number of nodes. For node pair, the larger set consists of all node pairs that have the same connectivity type. Prompts are unique, so it contains all possible instances. The instances are representative of the larger set because we have balanced the distribution of the various connectivity types and domains in the real-world.

Q8 What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

The instances include data in its raw form, presented as graphs, node attributes, edge features, node pairs, or molecular SMILES. For all instances, we utilize Python NetworkX library to depict both undirected (Graph) and directed (DiGraph) graphs, employing integer tuples to denote node pairs. For the real-world tasks related to molecular chemistry, we leverage Python rdkit library to transform the SMILES of compounds into the graph data in NetworkX format. In addition, prompts are in a standard textual format.

Table 1: Statistics of real-world datasets.

Task	Domain	Name	Phrase	Num	AVG. node	AVG. edge	AVG. density
Bi-Class.	Molecule	MUTAG	summary	150	15.67	16.79	0.0725
			classification	38	15.68	16.76	0.0723
			overall	188	15.67	16.78	0.0725
		ogbg-molhiv	summary	200	31.77	34.82	0.0445
			classification	40	30.50	33.45	0.0467
			overall	240	31.65	34.69	0.0447
		BBBP	summary	500	21.99	23.40	0.0595
			classification	50	28.24	30.64	0.0429
			overall	550	22.56	24.06	0.0580
	Social Network	IMDB-BINARY	summary	500	19.56	96.18	0.2457
			classification	50	19.73	98.43	0.2466
			overall	550	19.57	96.39	0.2458
Pattern Detection	Chemical	Benzene	overall	200	20.49	21.75	0.0547
		Alkane-Carbonyl	overall	200	41.54	42.72	0.0259
		Fluoride-Carbonyl	overall	200	21.46	22.65	0.0508
Multi-Class.	Bioinformatics	ENZYMES	summary	240	33.40	63.91	0.0731
			classification	60	31.93	62.78	0.0774
			overall	300	33.15	63.72	0.0738
	Computer Vision	Fingerprint	summary	300	2.92	2.13	0.2428
			classification	60	2.93	2.20	0.2560
			overall	360	2.92	2.14	0.2450
	Social Network	IMDB-MULTI	summary	300	12.95	67.21	0.3503
			classification	60	12.62	52.90	0.3279
			overall	360	12.89	64.83	0.3466

Q9 Is there a label or target associated with each instance? If so, please provide a description.

In real-world datasets, the label descriptions are listed as follows:

- MUTAG: There are binary labels that indicate the mutagenicity of nitroaromatic compounds on Salmonella typhimurium. Positive samples correspond to compounds being mutagenic.
- OGBG-HIV: The primary objective is to predict whether molecules inhibit HIV.
- OGBG-BBBP: This dataset includes binary labels for 2,050 compounds on their permeability properties of blood–brain barriers.
- IMDB-BINARY: The target is to predict whether a movie graph is an Action or Romance network.
- IMDB-MULTI: Its task involves predicting whether a movie graph corresponds to a Comedy, Romance, or Sci-Fi network.
- Fingerprint: he Fingerprint dataset is a multi-classification dataset and the goal is to determine which person the fingerprint belongs to.
- ENZYMES: Its labels feature distinct enzymes.
- Benzene: The data are classified into two classes to represent whether a Benzene ring is existed in each molecule or not.
- Alkane-Carbonyl: It aims to identify whether a molecule includes both alkane and carbonyl functional groups.
- Fluoride-Carbonyl: It aims to identify whether a molecule includes both fluoride atoms and carbonyl functional groups

Q10 Is any information missing from individual instances? If so, please provide a description.

No.

Task	Dataset type		difficulty	Num	AVG. node	AVG. edge	AVG. density
Pattern detection	Undirected graph	Training	-	1,893	17.78	86.49	0.43
		Evaluation	Small	250	9.50	22.80	0.52
			Medium	250	19.50	96.20	0.52
			Large	250	29.50	247.96	0.58
	Directed graph	Training	-	1,313	18.05	103.89	0.24
		Evaluation	Small	250	9.50	23.44	0.26
			Medium	250	19.50	96.98	0.26
			Large	250	29.50	223.58	0.26
Modification	Undirected graph	Square → House	Small	166	9.71	25.07	0.56
			Medium	347	14.67	22.09	0.33
			Large	476	18.54	24.89	0.26
		Square → Diamond	Small	144	9.91	10.96	0.27
			Medium	332	15.32	17.95	0.19
			Large	484	19.54	23.45	0.16
		Diamond → Square	Small	111	8.95	10.98	0.34
			Medium	180	12.59	13.92	0.26
			Large	205	14.52	16.03	0.24
	Directed graph	FFL → FBL	Small	227	9.63	14.64	0.18
			Medium	396	13.69	19.08	0.13
			Large	493	16.60	23.48	0.12
Frequent subgraph	Undirected graph	Triangle	Small	231	9.87	17.61	0.39
			Medium	248	19.46	56.04	0.31
			Large	247	29.46	149.27	0.35
		Square	Small	217	10.14	19.35	0.40
			Medium	249	19.49	56.32	0.31
			Large	249	29.49	152.85	0.35
		Diamond	Small	214	10.19	20.12	0.42
			Medium	244	19.44	63.30	0.35
			Large	246	29.45	168.59	0.39
		House	Small	205	10.37	20.50	0.41
			Medium	250	19.50	60.47	0.33
			Large	247	29.46	156.12	0.37
	Directed graph	FFL	Small	238	9.71	17.93	0.20
			Medium	248	19.48	59.90	0.17
			Large	250	29.50	154.67	0.18
		FBL	Small	208	10.20	20.79	0.21
			Medium	244	19.41	64.15	0.18
			Large	248	29.48	156.56	0.18
		D-Diamond	Small	187	10.60	22.33	0.21
			Medium	248	19.48	62.01	0.17
			Large	247	29.47	150.57	0.18
Discriminative pattern learning	Discrimination	-	-	900	25	25.5	0.09
	Classification	-	-	100	25	25.5	0.09

Table 2: Details of Synthetic dataset

Q11 Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The connections between individual instances are clearly defined. Within each file, all information pertaining to the same question is grouped together. Pickle files and JSON files have a one-to-many matching relationship based on their file names.

Q12 Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

It is recommended to split the instances into training instances and test instanes. The training split is built to extract patterns and test instanecs are for evaluations.

Q13 Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No.

Q14 Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Our data for real-world graph pattern tasks is sourced from ten consistent open-source datasets. These datasets and their resources are listed as follows: MUTAG [2], OGBG-HIV [3, 8], OGBG-BBBP [3, 4], IMDB-BINARY [9], IMDB-MULTI [9], Fingerprint [5], ENZYMES [1], Benzene [6, 7], Alkane-Carbonyl [6], and Fluoride-Carbonyl [6]. All these datasets can be accessed through a Python library called PyGeometric [1], except for the Benzene, Alkane-Carbonyl, and Fluoride-Carbonyl datasets, which are available for download from a GitHub repository at the link [2].

Q15 Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No.

Q16 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

A.3 Collection process

Q17 How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Python programs generate graphs, query pairs, prompts, and answers. All the data is readily observable.

[1] `https://www.pyg.org/`

[2] `https://github.com/realMoana/ProxyExplainer/tree/master`

Q18 What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

Software programs are utilized for data collection. Python packages, such as PyGeometric, NetworkX, and rdkit, are included to generate graphs, query pairs, prompts, and answers. To guarantee accuracy and thoroughness, the generated patterns and prompts undergo manual inspection. Moreover, the creation of algorithm prompting examples serves to validate the accuracy of answers concerning query information.

Q19 If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

When working with real-world datasets, we ensure sample balance by randomly selecting an equal number of graphs from each class.

Q20 Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

No crowdworkers were used in the curation of the dataset. Author details will be released after the paper is accepted.

Q21 Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was collected during the period from August 1, 2024 to October 1, 2024. Dataset is irrelevant with time.

Q22 Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Q23 Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

A.4 Preprocessing/Cleaning/Labeling

Q24 Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No.

A.5 Uses

Q28 Has the dataset been used for any tasks already?

The dataset is not used except in our research ”How Do Large Language Models Understand Graph Patterns? A Benchmark for Graph Pattern Comprehension”.

Q29 Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. The link will be released after the paper is accepted.

Q30 What (other) tasks could the dataset be used for?

No.

Q31 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

We believe that neither the composition nor the collection method of the dataset would affect its future applications.

Q32 Are there tasks for which the dataset should not be used? If so, please provide a description.

No.

Q33 Any other comments? No.

A.6 Distribution

Q34 Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset will be open-source.

Q35 How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? The data will be available after the paper is accepted.

Q36 When will the dataset be distributed?

The dataset will be distributed after the paper is accepted.

Q37 Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be distributed after the paper is accepted.

Q38 Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Q39 Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Q40 Any other comments?

No.

A.7 Maintenance

Q41 Who will be supporting/hosting/maintaining the dataset?

Release after the paper is accepted.

Q42 How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Release after the paper is accepted.

Q43 Is there an erratum? If so, please provide a link or other access point.

There is no erratum for our initial release.

Q44 Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Release after the paper is accepted.

Q45 If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No, the dataset does not relate to people.

Q46 Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We will continue to support the older versions

Q47 If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We encourage everyone to share their ideas on extending our dataset to cover more compression cases and provide more reliable results. We invite anyone interested to reach out and contribute to this effort.

Q48 Any other comments?

No.

References

[1] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1):i47–i56, 2005.

[2] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

[3] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

[4] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697, 2012.

[5] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

[6] Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Wang, Wesley Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating attribution for graph neural networks. *Advances in neural information processing systems*, 33:5898–5910, 2020.

[7] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

[8] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[9] Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1365–1374, 2015.