Knowledge-aware Parsimony Learning: A Perspective from Relational Graphs

Quanming Yao[1]*, Yongqi Zhang[2], Yaqing Wang[3], Nan Yin[4], James Kwok[4], Qiang Yang[4]
[1]Tsinghua University, [2]Hong Kong University of Science and Technology (Guangzhou),
[3]Beijing Institute of Mathematical Sciences and Applications,
[4]Hong Kong University of Science and Technology
qyaoaa@tsinghua.edu.cn, yongqizhang@hkust-gz.edu.cn, wangyaqing@bimsa.cn,
nanyin@ust.hk, {jamesk, qyang}@cse.ust.hk,

The scaling law, which involves the brute-force expansion of training datasets and learnable parameters, has become a prevalent strategy for developing more robust learning models. However, due to bottlenecks in data, computation, and trust, the sustainability of the scaling law is a serious concern for the future of deep learning. In this paper, we address this issue by developing next-generation models in a parsimonious manner (i.e., achieving greater potential with simpler models). The key is to drive models using domain-specific knowledge, such as symbols, logic, and formulas, instead of relying on the scaling law. This approach allows us to build a framework that uses this knowledge as "building blocks" to achieve parsimony in model design, training, and interpretation. Empirical results show that our methods surpass those that typically follow the scaling law. We also demonstrate the application of our framework in AI for science, specifically in the problem of drug-drug interaction prediction. We hope our research can foster more diverse technical roadmaps in the era of foundation models.

1. Introduction

The learning techniques have progressed from manual feature engineering to shallow models, then to deep networks, and now to foundation models, achieving great success in the field of computer vision, natural language understanding and speech processing. Specifically, large language models, like ChatGPT [1], as representatives of foundation model, has shown strong performance in versatile learning, which can adopted in many different tasks. The belief is that larger models can be more expressive, thus are likely to generalize better given sufficient training data [2, 3]. This gives birth to the current roadmap, i.e., achieving stronger performance by aggressively scaling up the size of data and models, which is also observed as scaling law [4].

However, such a roadmap potentially leads to serious problems (as shown in the left of Figure 1): *Data bottleneck*—the scaling law relies on vast amounts of high-quality data, yet all available online corpora are projected to be exhausted by 2028 [5]; *Computational bottleneck*—the exponentially growing number of parameters demands substantial high-performance computing power, yet the pace of hardware development struggles to keep up with this rapid increase [6]; and *Trust bottleneck*—the scaling law follows the data driven path, disregarding internal logical relationships, which leads to opaque reasoning processes and severe hallucination issues [7]. These indicate that the current roadmap is not sustainable, and motivates us to ask: *where is the way if the brute-force scaling up fails?*

To address this question, we look back to the fundamental principles that drive machine learning. Albert Einstein famously stated, "Everything should be made as simple as possible, but not simpler." This insight has inspired many AI researchers to develop learning techniques that embrace parsimony, aiming to achieve "maximum output with minimal input" and "leveraging small inputs for significant effects". Traditional approaches to achieving parsimony learning have typically followed

*Corresponding author.

Second Conference on Parsimony and Learning (CPAL 2025).

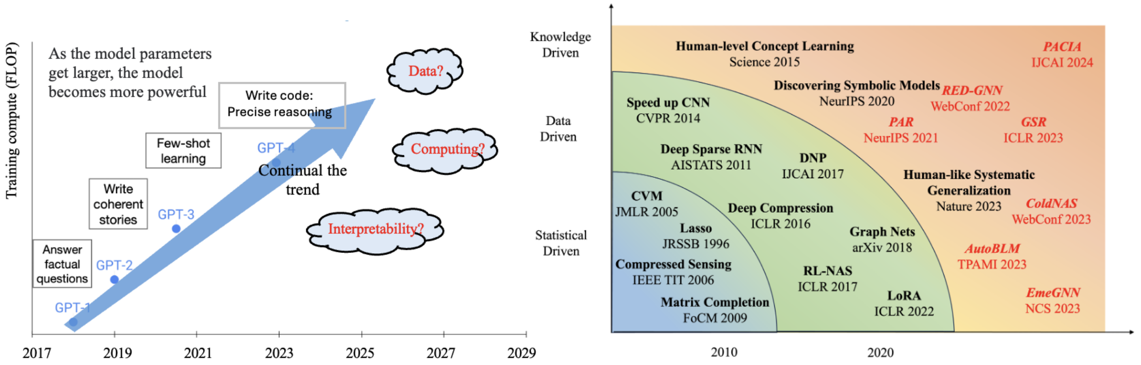


Figure 1: The data, computational and trust bottlenecks of LLMs (left), and the development of parsimony learning (right).

statistical-driven and data-driven methods (as shown in the right of Figure 1). Statistical driven methods are rooted in well-established statistical theories and principles. Example works are core vector machines [8], compressed sensing [9], matrix completion [10] and Lasso [11]. These methods penalize complexity and encourage models to focus on essential features. Later on, data driven methods gradually developed. Example works are Deep Sparse CNN[12], Speedup CNN [13], Deep Compression [14], RL-NAS [15], DNP [16] and LoRA [17]. These methods model and make predictions based on the data, rather than relying on explicit statistical assumptions. However, the development of data driven approaches would still lead to the way of scaling law, which suffers from the bottlenecks mentioned above.

Intuitively, humans accumulate knowledge in the form of symbols, concepts, rules, and principles, which allows us to learn a wide range of subjects or skills, apply them across various tasks, quickly adapt to new tasks with few or no demonstrations, and uncover the underlying reasons behind different phenomena. Inspired by [18–20], which shows that neural networks can achieve human-like systematicity with smaller models and achieve performance comparable to that of larger models, we propose knowledge driven methods that emphasize guiding machine learning models with domain-specific knowledge, using it as "building blocks" to achieve parsimony in learning. This approach allows us to tackle the bottlenecks of scaling law, focusing instead on efficient model design, training, and interpretation. We demonstrate the effectiveness of this framework in AI for science, particularly in addressing the problem of drug-drug interaction prediction. Empirical results show that our methods can outperform those limited by the scaling law.

2. Research Landscape

Our research landscape starts from relational graphs. Different from images, natural language and speech, relational graphs are a way to represent knowledge using nodes and edges, which symbolizes human knowledge in a structured form. The application of graphs are widely spread in real-life scenarios, such as designing and planning of urban networks, prediction of molecular properties, reasoning from knowledge graphs, and recommendation systems [21, 22].

Our key innovation lies in enabling parsimony learning through knowledge-aware approaches. Specifically, knowledge, i.e., symbolic logic and physical laws, is ubiquitous in real-world scenarios and coexists with data. As illustrated on the left side of Figure 2, the diverse colors found in nature can be derived from three primary colors (i.e., red, blue, and green). Building on this, we propose the concept of the "duality of knowledge and data", which posits that data is both numerical and knowledge-based. Based on this concept, we design a framework capable of learning from both data and knowledge simultaneously, thereby effectively achieving parsimony in learning. As shown on the right side of Figure 2, the framework first identifies symbolic logic and physical laws at the knowledge level, then performs combinatorial generalization of this knowledge at the data

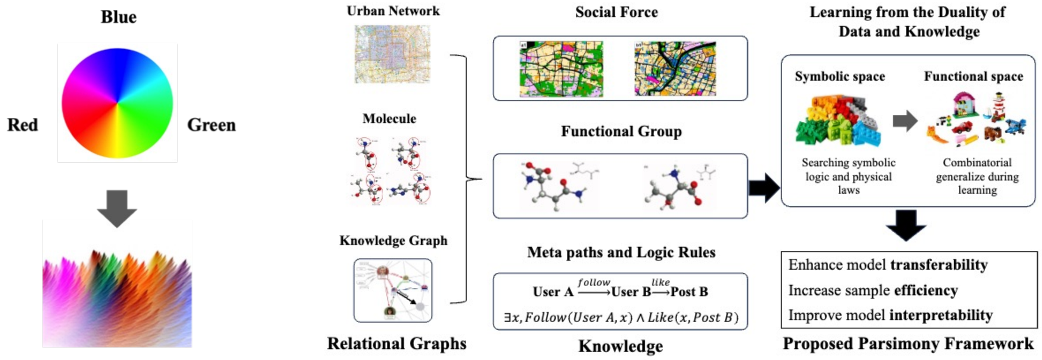

Blue
Red
Green
Urban Network
Molecule
Knowledge Graph
Relational Graphs
Social Force
Functional Group
Meta paths and Logic Rules
User A $\xrightarrow{follow}$ User B $\xrightarrow{like}$ Post B
$\exists x, Follow(User\ A, x) \wedge Like(x, Post\ B)$
Knowledge
Learning from the Duality of Data and Knowledge
Symbolic space
Functional space
Searching symbolic logic and physical laws
Combinatorial generalize during learning
Enhance model transferability
Increase sample efficiency
Improve model interpretability
Proposed Parsimony Framework

Figure 2: The three primary colors (left) and the knowledge-aware parsimony learning framework (right).

level. By treating knowledge as "building blocks" and focusing on its learning and generalization, the framework leverages simple knowledge to solve complex problems.

To efficiently address these bottlenecks, we introduce domain-specific knowledge into our framework to achieve greater potential with smaller models. Specifically, we implement model parsimony, leveraging simple architectures to overcome computational bottleneck. Additionally, training parsimony is employed to reduce training costs, tackling the data bottleneck. To address the trust bottleneck, interpretive parsimony is used to identify key evidence. Finally, we apply this proposed framework to the field of AI for science, demonstrating its significant potential.

- *Parsimony on Model*. The goal of architectural parsimony is to utilize simple architecture to achieve the comparable performance of complex models. As introduced in [20], simple neural networks can achieve human-like systematicity when optimized for compositional skills, offering the potential for smaller models to achieve performance comparable to that of larger ones. Following this intuition, we propose AutoBLM [23] and ColdNAS [24], which utilize prior knowledge to streamline architectures. By achieving this, we can maintain high performance with simpler structures, thereby efficiently addressing computational bottlenecks.

- *Parsimony on Training*. The purpose of training parsimony is to reduce the amount of training data. Inspired by [25], which shows that task-related knowledge help to determine parameters by simple arithmetic operations without training, we propose new approaches for learning parsimony [26–28]. These approaches leverage knowledge to guide the fine-tuning process on specific tasks under few-shot circumstances. By achieving this, we can efficiently utilize knowledge to guide the optimization of related tasks with limited training data, thereby efficiently addressing data bottlenecks.

- *Parsimony on Interpretation*. The parsimony on interpretation is to identify important evidence in face to massive connections on graphs. Our primary approach is to encapsulate logical rules within graphs as supporting evidence, utilizing the logical interpretation of knowledge to elucidate the reasoning processes of models [29, 30]. By achieving this, we can efficiently interpret the model's result with subgraph, thereby efficiently addressing trust bottlenecks.

- *Potential in AI for Science*. In above, we have elaborate how parsimony in model, training and interpretation can be achieved individually following a knowledge-driven way. We will combine these idea and further show how they can help AI for science. Specifically, we will take drug-drug interaction as an application, which predicts interactions between emerging and existing drugs. By extracting path-based subgraphs and learning subgraph representation, we can efficiently reduce the data and computational requirements and identify evidence with subgraphs.

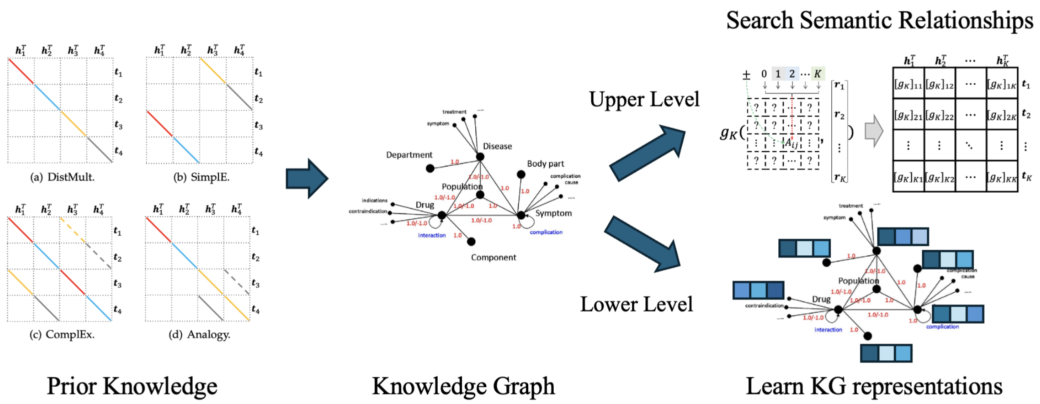


Figure 3: AutoBLM first sets up a search space by analyzing existing scoring functions and then utilizes the bi-level optimization to extract semantics and relationships simultaneously.

3. Parsimony on Model

The goal of parsimony on model is to match the performance of complex models using a simpler architecture. To achieve this, it is necessary to prune the network architecture or search the simple structure to guide the model simplification or semantic reorganization. Inspired by [20], which highlights the potential of achieving superior performance with standard neural network architectures through "knowledge-aware" meta-learning, we propose approaches to architectural parsimony by leveraging knowledge as prior information to guide the search for alternative but simpler structures in learning.

3.1. Automated Bi-linear Scoring Function Design

In knowledge graph learning, the scoring function is a key component that measures the plausibility of edges [21]. Human experts often design various complex scoring functions to evaluate edge plausibility, which can be redundant and inefficient in model design. To achieve parsimony across different scoring functions, we propose AutoBLM [23], which focuses on semantics and leverages computational power to reconfigure simple architectural elements, enabling efficient learning across diverse tasks. Specifically, AutoBLM employs a bi-level framework to simultaneously extract semantic representations (i.e., the embeddings of entities and relations in knowledge graph) and search the relations of entities, which is shown in Figure 3. The lower level learns representations from the training dataset, while the upper level searches the knowledge space to discover semantic relationships in a unified search space, using the validation dataset. Additionally, AutoBLM enables the learning of new models that are better adapted to specific datasets through an efficient evolutionary search algorithm, enhancing transferability.

The performance of AutoBLM is evaluated with knowledge graph reasoning benchmarks. The results, evaluated with mean reciprocal ranking (MRR) metric, are shown in Table 1. Compared with the complex neural network models Interstellar [31] and CompGCN [32], AutoBLM gains significant improvement on the three datasets. In particular, it recombines the simple relation to simulate complex structure and outperforms the deep neural network models, which often have better expressiveness with higher computational costs.

Table 1: MRR performance comparison of AutoBLM and neural models (Interstellar and CompGCN).

Model	WN18RR	FB15k237	YAGO3-10
Interstellar	0.48	0.32	0.51
CompGCN	0.479	0.355	0.421
AutoBLM	0.490	0.360	0.571

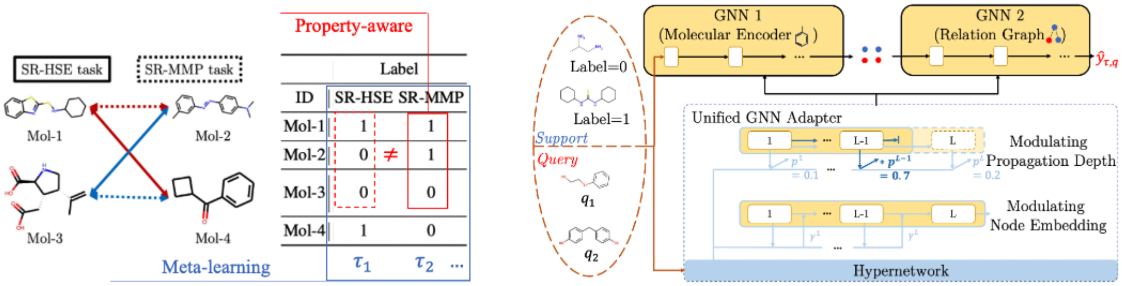
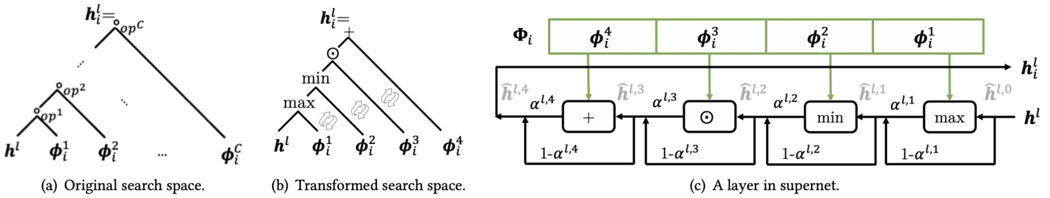


Figure 4: PAR meta-learns a property-aware network to improve sample efficiency (left). PACIA introduces GNN adapter to achieve efficient hierarchical adaptation (right).



(a) Original search space. (b) Transformed search space. (c) A layer in supernet.

Figure 5: ColdNAS uses a hypernetwork to map each user's history interactions to user-specific parameters which are then used to modulate the predictor, and formulate how to modulate and where to modulate as a NAS problem.

3.2. Symbolized Architecture Search for Recommendation

User cold-start recommendation problem targets at quickly generalize to new tasks (i.e. personalized recommendation for cold-start users) with a few training samples (i.e. a few interaction histories). A number of works [33–35] adopt the classic gradient-based meta-learning strategy called model-agnostic meta learning (MAML) [36], which learns a good initialized parameter from a set of tasks and adapts it to a new task by taking a few steps of gradient descent updates on a limited number of labeled samples. This line of models has demonstrated high potential of alleviating user cold-start problem. However, gradient-based meta-learning strategy require expertise to tune the optimization procedure to avoid over-fitting. Besides, the inference time can be long. To address these challenges, we propose ColdNAS [24], which searches for proper modulation structures to adapt user-specific recommendations as shown in Figure 5. The core of ColdNAS is to utilize a hypernetwork that maps user interaction history to personalized parameters, which are then used to modulate the predictor model. Specifically, ColdNAS formulates a unified search space and optimizes the modulation functions as part of a differentiable neural architecture search (NAS). By adopting a symbolic search approach instead of deep learning-based fitting methods, ColdNAS significantly reduces the search space and improves efficiency. Furthermore, models that are better suited for specific datasets can be learned through an efficient and robust search algorithm, achieving adaptability across datasets.

Table 2: Test performance (%) obtained on benchmark datasets. The best results are highlighted in bold and the second-best in italic. A smaller value is better.

Dataset	Metric	MAMO	TaNP	ColdNAS
MovieLens	MSE	$90.20_{(0.22)}$	$89.11_{(0.18)}$	$\mathbf{87.96}_{(0.12)}$
	MAE	$75.34_{(0.26)}$	$74.78_{(0.14)}$	$\mathbf{74.29}_{(0.20)}$
Book Crossing	MSE	$14.82_{(0.05)}$	$14.75_{(0.05)}$	$\mathbf{14.15}_{(0.08)}$
	MAE	$3.51_{(0.02)}$	$3.48_{(0.01)}$	$\mathbf{3.40}_{(0.01)}$
Last.fm	MSE	$21.64_{(0.10)}$	$21.58_{(0.20)}$	$\mathbf{20.91}_{(0.05)}$
	MAE	$42.30_{(0.28)}$	$42.15_{(0.56)}$	$\mathbf{41.78}_{(0.24)}$

The performance of ColdNAS is evaluated on multiple benchmark datasets (i.e., MovieLens [37], Book-Crossing [38] and Last.fm [39]), and measured with metrics such as MSE and MAE, are shown in the Table 2. Compared with existing state-of-the-art cold-start models (i.e., MAML and

TaNP [39]) demonstrates significant improvements on all datasets. In particular, it automatically identifies the optimal modulation structures that outperform fixed modulation strategies, yielding superior performance while maintaining computational efficiency.

4. Parsimony on Training

In AI-assisted scientific research, particularly in molecules and biomedicine, the scarcity of labeled data presents significant challenges. Motivated by [40], which utilizes the symbolic structure between parameters obtained from different tasks for downstream tasks prediction, we propose approaches to learning parsimony by leveraging task-related knowledge to guide the fine-tuning on downstream tasks. In molecules properties prediction, we have developed meta-learning learning techniques that enforce parsimony on learning. These approaches ensure that parameters can be efficiently adapted in relation to functional groups, optimizing their use and enhancing model adaptability.

4.1. Property-Aware Relation Networks

Our first work towards this problem is the property-aware relation networks (PAR) [26, 27]. PAR uses a property-aware molecular encoder to transform the generic molecular embeddings to property-aware ones. To fully leverage the supervised learning signal, PAR learns to estimate the molecular relation graph by a query-dependent relation graph learning module, in which molecular embeddings are refined w.r.t. the target property. Thus, the facts that both property-related information and relationships among molecules change across different properties are utilized to better learn and propagate molecular embeddings. Besides, we propose a selective update strategy for handling generic and property-aware information. In the inner-loop update, only the property-aware information is updated, while both generic and property-aware information are updated simultaneously in the outer-loop. We use gradient descents to update parameters in both loops. Through the selective update strategy, the model can capture generic and property-aware information separately in the training procedure.

The results are reported in Table 3. From the results, we observe that PAR obtains the best performance among methods using graph-based molecular encoders learned from scratch. The outperforming results can be attributed to the combination of metric-based and optimization-based method in the design of PAR method. In terms of average improvement, PAR obtains significantly better performance than the best baseline learned from scratch (e.g. EGNN) by 1.59%, showing enhanced performance.

Method	Tox21		SIDER		MUV		ToxCast	
	10-shot	1-shot	10-shot	1-shot	10-shot	1-shot	10-shot	1-shot
MAML	$79.59_{(0.33)}$	$75.63_{(0.18)}$	$70.49_{(0.54)}$	$68.63_{(1.51)}$	$68.38_{(1.27)}$	$65.82_{(2.49)}$	$68.43_{(1.85)}$	$66.75_{(1.62)}$
IterRefLSTM	$81.10_{(0.10)}$	$80.97_{(0.06)}$	$69.63_{(0.16)}$	$71.73_{(0.06)}$	$49.56_{(2.32)}$	$48.54_{(1.48)}$	-	-
PAR	$82.13_{(0.26)}$	$\underline{80.02}_{(0.30)}$	$\underline{75.15}_{(0.35)}$	$\underline{72.33}_{(0.47)}$	$68.08_{(2.23)}$	$65.62_{(3.49)}$	$70.01_{(0.85)}$	$\underline{68.22}_{(1.34)}$
ADKF-IFT	$\underline{82.43}_{(0.60)}$	$77.94_{(0.91)}$	$67.72_{(1.21)}$	$58.69_{(1.44)}$	$\mathbf{98.18}_{(3.05)}$	$\underline{67.04}_{(4.86)}$	$\underline{72.07}_{(0.81)}$	$67.50_{(1.23)}$
PACIA	$\mathbf{84.25}_{\mathbf{(0.31)}}$	$\mathbf{82.77}_{\mathbf{(0.15)}}$	$\mathbf{82.40}_{\mathbf{(0.26)}}$	$\mathbf{77.72}_{\mathbf{(0.34)}}$	$\underline{72.58}_{(2.23)}$	$\mathbf{68.80}_{(4.01)}$	$\mathbf{72.38}_{\mathbf{(0.96)}}$	$\mathbf{69.89}_{\mathbf{(1.17)}}$

Table 3: Test ROC-AUC (%) obtained on MoleculeNet. The best results are bolded, second-best results are underlined.

4.2. Parameter-Efficient GNN Adapter

We further introduce parameter-efficient graph neural network (GNN) adapter (PACIA) [28]. By adopting this approach, PACIA significantly reduces the risk of overfitting. Moreover, it offers the advantage of faster inference speeds, as the adapted parameters are generated through a single forward pass rather than through iterative optimization steps. Additionally, initializing the function $g(\cdot)$ with neural networks allows for more flexible forms of updates compared to traditional gradient descent methods. Further, we design a hierarchical adaptation mechanism in the framework: Task-level adaptation is achieved in the encoder since the structural features in molecular graphs needs to be captured in a property-adaptive manner, while query-level adaptation is achieved in the predictor based on the property-adaptive representations. To adapt GNN's parameter-efficiently, we design a hypernetwork-based GNN adapter to generate a few adaptive parameters to modulate the node embedding and propagation depth, which are essential in message passing process. No further fine-tuning is required.

Table 3 shows the prediction performance. Our analysis reveals that methods utilizing pretrained graph-based molecular encoders generally outperform those with encoders learned from scratch. This underscores the effectiveness of pretrained encoders in capturing rich, generic molecular information, subsequently providing superior molecular embeddings. In further evaluations, meta-learning methods that learn relational graphs—specifically EGNN, PAR, and PACIA—demonstrate enhanced performance. Notably, PACIA consistently achieves the highest ROC-AUC scores, followed closely by PAR. In summary, PACIA establishes itself as the new state-of-the-art for predicting molecular properties. It not only delivers superior predictive accuracy but also features significantly faster adaptation speed and requires fewer adaptive parameters.

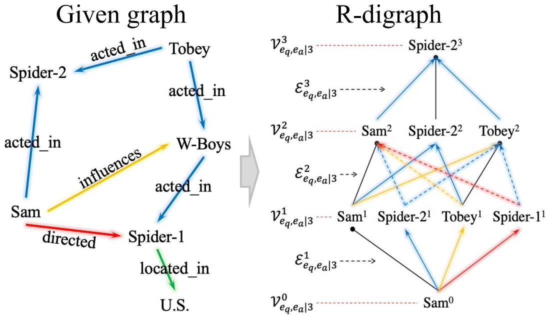


Figure 6: RED-GNN makes use of dynamic programming to recursively encodes multiple r-digraphs with shared edges, and utilizes query-dependent attention mechanism to select the strongly correlated edges.

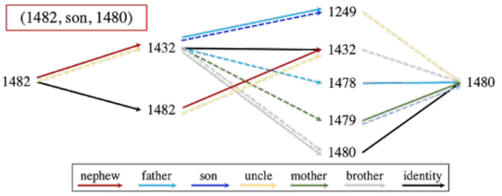


Figure 7: Visualization of the learned structures. Dashed lines mean inverse relations. The query triples are indicated by the red rectangles.

5. Parsimony on Interpretation

Interpretability is important for understanding the result. In many fields, experts need to interpret and understand the results of models in order to make informed decisions. GNNs, though effective in learning from relational graphs, is very challenging to provide interpretable inference when facing a massive amount of associated information on graphs. For graph learning, the core problem in achieving model interpretability lies in accurately capturing strong logical relationships and exhibiting inference process with evidence. To achieve interpretability on graphs, our key idea is to capture the logical rules inside graphs as supporting evidence and use the logical interpretation of knowledge to clarify models' reasoning process. In the following, we introduce the idea of subgraph learning for interpreting and show the interpretable inference process with the learned subgraphs.

5.1. Interpreting with Subgraph Learning

The relational graph structures are complex and hard to understand directly. In comparison, relational paths with multiple connected edges are more interpretable [41, 42]. Based on this observation, RED-GNN [30] introduces a new relational structure, called relational di-graph (r-digraph) as illustrated in Figure 6. The r-digraphs generalize relational paths to subgraphs by preserving

Table 4: Performance comparison on knowledge graph reasoning tasks. The best results are bolded and the second-best underlined.

Model	WN18RR			FB15k-237			YAGO3-10		
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10
DRUM	0.486	42.5	58.6	0.343	25.5	51.6	0.531	45.3	67.6
RNNLogic	0.483	44.6	55.8	0.344	25.2	53.0	0.554	**50.9**	67.3
CompGCN	0.487	44.3	54.6	0.355	26.4	53.5	0.421	39.2	57.7
NBFNet	0.551	49.7	66.6	0.415	32.1	**59.9**	0.550	47.9	68.6
RED-GNN	**0.564**	**50.2**	**67.8**	**0.418**	**32.9**	59.0	**0.584**	**50.9**	71.3

the overlapped relational paths and structures of relations for reasoning. By leveraging the GNN model with attention mechanism to propagate information over the subgraph, significant performance improvement has been achieved and logical paths in r-digraphs can be captured.

The effectiveness of subgraph learning methods are evaluated on general knowledge graph reasoning benchmarks. As shown in Table 4, the subgraph learning methods NBFNet [43] and RED-GNN [30] show significant advantage over the embedding-based methods. We visualize an exemplar learned r-digraphs by RED-GNN on the Family dataset. Figure 7 shows one triple that DRUM fails. Chain-like structures alone, such as `id-1482` $\xrightarrow{\text{nephew}}$ `id-1432` $\xrightarrow{\text{brother}}$ `id-1480` $\xrightarrow{\text{identify}}$ `id-1480` or `id-1482` $\xrightarrow{\text{nephew}}$ `id-1432` $\xrightarrow{\text{father}}$ `id-1249` $\xrightarrow{\text{uncle}}$ `id-1480`, can only imply that `id-1482` is the son or nephew of `id-1480`. These two chain-like structures together provide the evidence that `id-1432` and `id-1480` are the only brother of each other, which is crucial in inferring that `id-1482` is the son of `id-1480`.

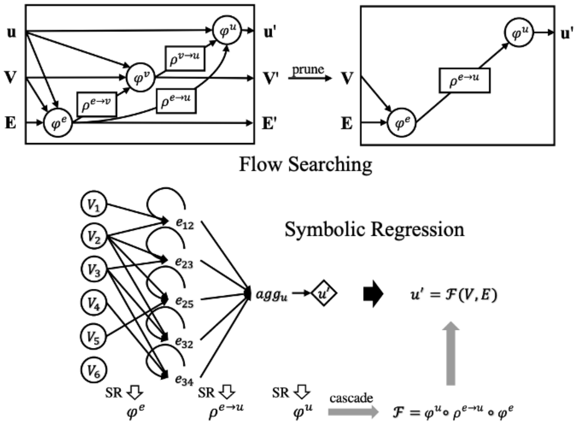


Figure 8: GSR models the formula skeleton with a message-passing flow, which helps transform the discovery of the skeleton into the search for the message-passing flow. Then, the formulas can be identified by interpreting component functions of the searched message-passing flow, reusing classical symbolic regression methods.

5.2. Symbolic Regression on Graphs

Graph-structured physical mechanisms are commonly found in various scientific domains, where variables such as mass, force, and energy interact through relationships on a graph. Traditional symbolic regression (SR) methods have been used to discover formulas from input-output data pairs but struggle to handle graph-structured inputs. We present the GSR [29] as shown in Figure 8, which is designed to extend symbolic regression to graph-structured physical mechanisms. By integrating prior knowledge in the form of symbolic logic and physical laws, the proposed approach transforms the discovery of formula skeletons into a search for efficient message-passing flows in GNNs. The framework enables the model to balance accuracy and simplicity by identifying Pareto-optimal flows that generalize well across different physical domains. Through knowledge-guided symbolic learning, the model captures essential relationships within graph-structured data, achieving parsimony on interpretation.

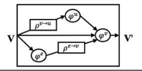
Scenario	**Unidirectional flow in corridor**
Searched Message-passing Flow	 V V' $\rho^{v\to u}$ φ^u φ^v φ^e $\rho^{e\to v}$
Searched Skeleton	$\varphi^v\left(x_i, \sum_{j\in N(i)} \varphi^e(x_i, x_j), \sum_{j\in V} \varphi^u(x_j)\right)$
Learned Formula (Difference with Social Force)	φ^e: $ke^{-\lambda r}$ sign(v_n) (normal component) φ^e: δ (tangential component) φ^u: $v_{mean} = \text{mean}(v_i)$ φ^v: $\sum_{j\in N(i)} \varphi^e + (v - \lambda_1 v_0 - \lambda_2 \varphi^u)/\tau$

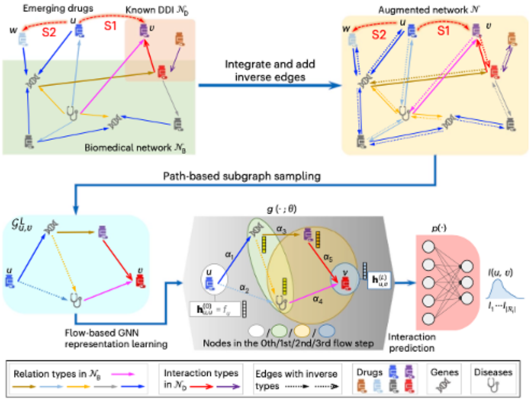
Figure 9: Learned formulas for pedestrian dynamics.

The learned formulas and the corresponding physical meanings are reported in Figure 9, which demonstrates that our model can learn different skeletons and formulas that are more precise than the social force model with explicit physical meanings.

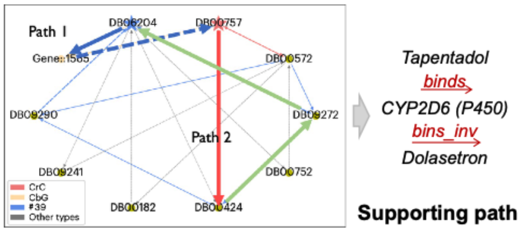
6. Potential in Drug Development

In the ever-changing world of pharmaceuticals, the intersection of scientific advancements and regulatory changes has led to a significant breakthrough, particularly in the rapid development of new drugs aimed at treating rare, severe, or life-threatening diseases [44, 45]. Machine learning models have become powerful tools for predicting drug interactions, capable of comprehensively capturing intricate relationships on the interaction graphs. Our work EmerGNN [46] takes this a step further by introducing a graph learning framework that predicts interactions between emerging and existing drugs by integrating drug-drug interaction (DDI) and biomedical networks, extracting path-based subgraphs, and learning subgraph representations.

The general idea of EmerGNN is illustrated in Figure 10. Emerging drugs typically have limited interactions with existing drugs. To address this issue, EmerGNN utilizes biomedical network and extracts subgraph $\mathcal{G}^L_{u,v}$ from it to connect emerging drug u with existing drug v. Drawing inspiration from the subgraph learning methods [30, 43], EmerGNN proposes a flow-based GNN $g(\mathcal{G}^L_{u,v}; \boldsymbol{\theta})$ with an attention mechanism to encode pair-wise subgraph representations for drug pairs. These subgraph representations are then used to directly predict the interaction of the drug pair. This approach enables a more nuanced understanding of interactions involving emerging drugs, incorporating knowledge from the interconnected network of biomedical entities and their relationships, and providing interpretable insights.


Emerging drugs
Known DDI $\mathcal{N}_D$
w
u
v
S2
S1
Biomedical network $\mathcal{N}_B$
Integrate and add inverse edges
Augmented network $\mathcal{N}$
Path-based subgraph sampling
$\mathcal{G}^L_{u,v}$
Flow-based GNN representation learning
$g(\cdot;\theta)$
α_1
α_2
α_3
α_4
α_5
$\mathbf{h}^{(0)}_{u,v} = f_u$
$\mathbf{h}^{(L)}_{u,v}$
Nodes in the 0th/1st/2nd/3rd flow step
$p(\cdot)$
Interaction prediction
$l(u, v)$
$l_1 \cdots l_{|\mathcal{R}_I|}$
Relation types in $\mathcal{N}_B$
Interaction types in $\mathcal{N}_D$
Edges with inverse types
Drugs
Genes
Diseases

(a) Illustration of EmerGNN



(b) Visualization of learned by EmerGNN

Datasets	DrugBank (S1)		TWOSIDES (S1)	
Methods	F1-score	Accuracy	PR-AUC	ROC-AUC
CSMDDI	<u>19.8</u>±3.1	<u>37.3</u>±4.8	55.8±4.9	57.0±6.1
HIN-DDI	8.8±1.0	27.6±2.4	<u>64.8</u>±2.3	<u>58.5</u>±1.6
KG-DDI	1.1±0.1	32.2±3.6	53.9±3.9	47.0±5.5
DeepLGF	4.8±1.9	31.9±3.7	59.4±8.7	54.7±5.9
EmerGNN	**25.0**±2.8	**46.3**±3.6	**81.4**±7.4	**79.6**±7.9

(c) Performance comparison on DDI

Figure 10: (a) EmerGNN learns pairwise representations of drugs by extracting the paths between drug pairs, propagating information from one drug to the other, and incorporating the relevant biomedical concepts on the paths. (b) Visualization of the structure learned by EmerGNN. DB006204 (Tapentadol) is an existing drug, and DB00757 (Dolasetron) is an emerging drug. (c) Performance comparison on DDI prediction tasks. The best results are bolded and the second-best underlined.

The results are shown in Table 10(c), from the results, we find that the proposed EmerGNN outperforms CSMDDI [47] that directly uses drug features to predict, HINDDI [48] that counts meta-paths from biomedical network for prediction, KG-DDI [49] that learns drug embeddings, and DeepLGF [50] that models with GNN in both benchmarks and both settings. Furthermore, we show a case study with the visualized subgraph in Figure 10(b) by selecting important paths according to the attention values. The path connecting two drugs through the binding protein `Gene::1565 (CYP2D6)`, which is a `P450` enzyme that plays a key role in drug metabolism, for example, shows a way to interpret the inference process with supporting evidence. From the empirical comparison and case studies, we conclude that subgraph learning can not only achieve improvements in representation learning, but also interpret the inference process with supporting evidence.

7. Future Works

Finally, we talk about our ongoing future works from theory, method and application.

- *Theory: Knowledge quantization.* One key area for further research is the development of a theoretical foundation for knowledge quantization. Current machine learning frameworks heavily depend on large-scale data and complex models; however, incorporating mechanisms to distill knowledge into more compact and efficient forms can result in models that are both more efficient and parsimonious. Future work will focus on quantifying how much domain-specific knowledge is necessary to drive model performance while reducing the need for large datasets. This will involve formalizing the concept of "knowledge as a resource" and developing algorithms that can optimize the utilization of knowledge while maintaining performance across various tasks. A theoretical framework for knowledge quantization will pave the way for more efficient integration of symbolic logic, physical laws, and other forms of structured knowledge into learning models.
- *Methods: Integrating with LLMs.* As large language models (LLMs) continue to dominate AI research, finding efficient ways to integrate them with specialized knowledge and smaller models is crucial. Our future work will focus on creating methods that allow seamless integration of LLMs with domain-specific frameworks. Instead of scaling up models indefinitely, the focus will be on using LLMs as modular components that can be fine-tuned or adapted based on task-specific knowledge. This approach will involve the development of techniques for aligning the outputs of LLMs with symbolic and logical reasoning systems, enhancing the ability of these models to perform in specialized fields such as bioinformatics or physics. Furthermore, exploring ways to reduce the computational footprint of LLMs while maintaining their versatility will be a key area of focus.
- *Applications: AI for Science.* The application of AI for scientific discovery remains one of the most promising areas for further exploration. Future research will focus on leveraging the proposed parsimonious learning framework to tackle more complex challenges in drug discovery, molecular property prediction, and other areas within AI for science. In drug-drug interaction prediction, the intricate nature of molecular interactions typically demands significant computational resources. By applying parsimony to the model, neural networks can be simplified, enabling efficient predictions while reducing computational load. In protein structure prediction, the scarcity of labeled data often makes training challenging. Parsimony in training leverages domain knowledge, thereby reducing the reliance on large datasets and enabling accurate predictions even with limited data. In molecular property prediction, understanding the reasoning behind model predictions is crucial for advancing scientific research. Parsimony in interpretation allows for the extraction of key insights from complex data, improving both the transparency of the model and the trustworthiness of its predictions.

8. Conclusion

This paper introduces an alternative way to develop next-generation learning techniques instead of scaling law. By leveraging the duality between data and knowledge, our method extracts symbolic logic and physical laws during the learning process and applies combinatorial generalization to various tasks. This approach effectively overcomes the limitations of traditional scaling methods. Experimental results demonstrate that our framework significantly improves model performance, showcasing its ability to achieve parsimony on model, training and interpretation. These findings underscore the potential of integrating knowledge into machine learning models, offering a promising direction for future research and applications.

References

[1] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

[2] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 2015.

[3] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 2017.

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint*, 2020.

[5] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *International Conference on Machine Learning*, 2024.

[6] Christopher Wolters, Xiaoxuan Yang, Ulf Schlichtmann, and Toyotaro Suzumura. Memory is all you need: An overview of compute-in-memory architectures for accelerating large language model inference. *arXiv preprint*, 2024.

[7] Hanyu Duan, Yi Yang, and Kar Yan Tam. Do LLMs know about hallucination? an empirical investigation of LLM's hidden states. *arXiv preprint*, 2024.

[8] Ivor W Tsang, James T Kwok, Pak-Ming Cheung, and Nello Cristianini. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 2005.

[9] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 2006.

[10] Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 2012.

[11] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996.

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[13] M Jaderberg, A Vedaldi, and A Zisserman. Speeding up convolutional neural networks with low rank expansions. In *The British Machine Vision Conference*, 2014.

[14] Song Han, Huizi Mao, and William Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations*, 2016.

[15] Barret Zoph and Quoc Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*, 2017.

[16] Bo Liu, Ying Wei, Yu Zhang, and Qiang Yang. Deep neural networks for high dimension, low sample size data. In *International Joint Conference on Artificial Intelligence*, 2017.

[17] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[18] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015.

[19] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems*, 2020.

[20] Brenden M. Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 2023.

[21] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[22] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[23] Yongqi Zhang, Quanming Yao, and James T Kwok. Bilinear scoring function search for knowledge graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[24] Shiguang Wu, Yaqing Wang, Qinghe Jing, Daxiang Dong, Dejing Dou, and Quanming Yao. ColdNAS: Search to modulate for user cold-start recommendation. In *The ACM Web Conference*, 2023.

[25] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[26] Yaqing Wang, Abulikemu Abuduweili, Quanming Yao, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. In *Advances in Neural Information Processing Systems*, 2021.

[27] Quanming Yao, Zhenqian Shen, Yaqing Wang, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[28] Shiguang Wu, Yaqing Wang, and Quanming Yao. PACIA: Parameter-efficient adapter for few-shot molecular property prediction. In *International Joint Conference onArtificial Intelligence*, 2024.

[29] Hongzhi Shi, Jingtao Ding, Yufan Cao, Quanming Yao, Li Liu, and Yong Li. Learning symbolic models for graph-structured physical mechanism. In *International Conference on Learning Representations*, 2023.

[30] Yongqi Zhang and Quanming Yao. Knowledge graph reasoning with relational digraph. In *The ACM Web Conference*, 2022.

[31] Yongqi Zhang, Quanming Yao, and Lei Chen. Interstellar: Searching recurrent architecture for knowledge graph embedding. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

[32] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. In *International Conference on Learning Representations*, 2020.

[33] Manqing Dong, Feng Yuan, Lina Yao, Xiwei Xu, and Liming Zhu. MAMO: Memory-augmented meta-optimization for cold-start recommendation. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[34] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. MeLU: Meta-learned user preference estimator for cold-start recommendation. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[35] Yuanfu Lu, Yuan Fang, and Chuan Shi. Meta-learning on heterogeneous information networks for cold-start recommendation. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[36] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

[37] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 2015.

[38] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *The ACM Web Conference*, 2005.

[39] Xixun Lin, Jia Wu, Chuan Zhou, Shirui Pan, Yanan Cao, and Bin Wang. Task-adaptive neural process for user cold-start recommendation. In *The ACM Web Conference*, 2021.

[40] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

[41] Wenhan Xiong, Thien Hoang, and William Yang Wang. DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Conference on Empirical Methods in Natural Language Processing*, 2017.

[42] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. In *International Conference on Learning Representations*, 2018.

[43] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural Bellman-Ford Networks: A general graph neural network framework for link prediction. In *Advances in Neural Information Processing Systems*, 2021.

[44] Xian Su, Haixue Wang, Nan Zhao, Tao Wang, and Yimin Cui. Trends in innovative drug development in china. *Nature Reviews Drug Discovery*, 2022.

[45] Heidi Ledford. Hundreds of COVID trials could provide a deluge of new drugs. *Nature*, 2022.

[46] Yongqi Zhang, Quanming Yao, Ling Yue, Xian Wu, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. Emerging drug interaction prediction enabled by a flow-based graph neural network with biomedical network. *Nature Computational Science*, 2023.

[47] Zun Liu, Xing-Nan Wang, Hui Yu, Jian-Yu Shi, and Wen-Min Dong. Predict multi-type drug–drug interactions in cold start scenario. *BMC Bioinformatics*, 2022.

[48] Farhan Tanvir, Muhammad Ifte Khairul Islam, and Esra Akbas. Predicting drug-drug interactions using meta-path based similarities. In *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 2021.

[49] Md Rezaul Karim, Michael Cochez, Joao Bosco Jares, Mamtaz Uddin, Oya Beyan, and Stefan Decker. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. In *ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019.

[50] Zhong-Hao Ren, Zhu-Hong You, Chang-Qing Yu, Li-Ping Li, Yong-Jian Guan, Lu-Xiang Guo, and Jie Pan. A biomedical knowledge graph-based method for drug–drug interactions prediction through combining local and global features with deep neural networks. *Briefings in Bioinformatics*, 2022.