Property-Aware Relation Networks for Few-shot Molecular Property Prediction

Dr. Quanming Yao

Assistant professor, EE Tsinghua

qyaoaa@tsinghua.edu.cn

Joint work with [*Yaqing Wang*, *Dejing Dou*] (Baidu), [*Abulikemu Abuduweili*] (CMU)

NeurIPS 2021 (Spotlight)

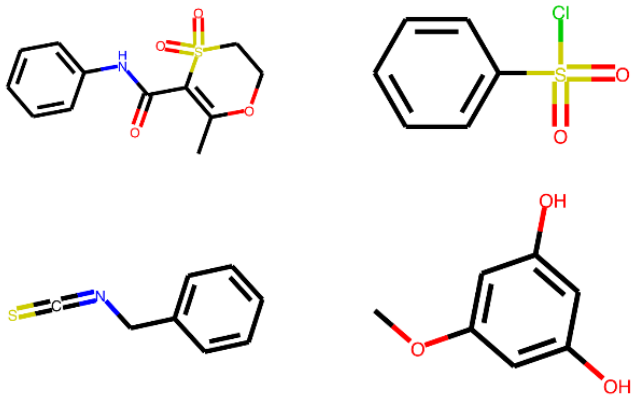
Code: https://github.com/tata1661/PAR-NeurIPS21

Outline

- Background
  - Molecular property prediction (MPP)
  - Few-shot learning (FSL)
- Existing works
- The proposed approach
- Summary

Molecular Property Prediction (MPP)

- Molecules:
  - Mainly micromolecule organics
  - Graph-structured data, Graph Neural Network (GNN) is useful in obtaining its representations
- Properties:
  - Physiology or Toxicity
  - Examples in SIDER :
    - ‘SIDER’ : [‘Hepatobiliary disorders’ , ‘Infections and infestations’ , ‘Neoplasms benign, malignant and unspecified (incl cysts and polyps) ' , …]



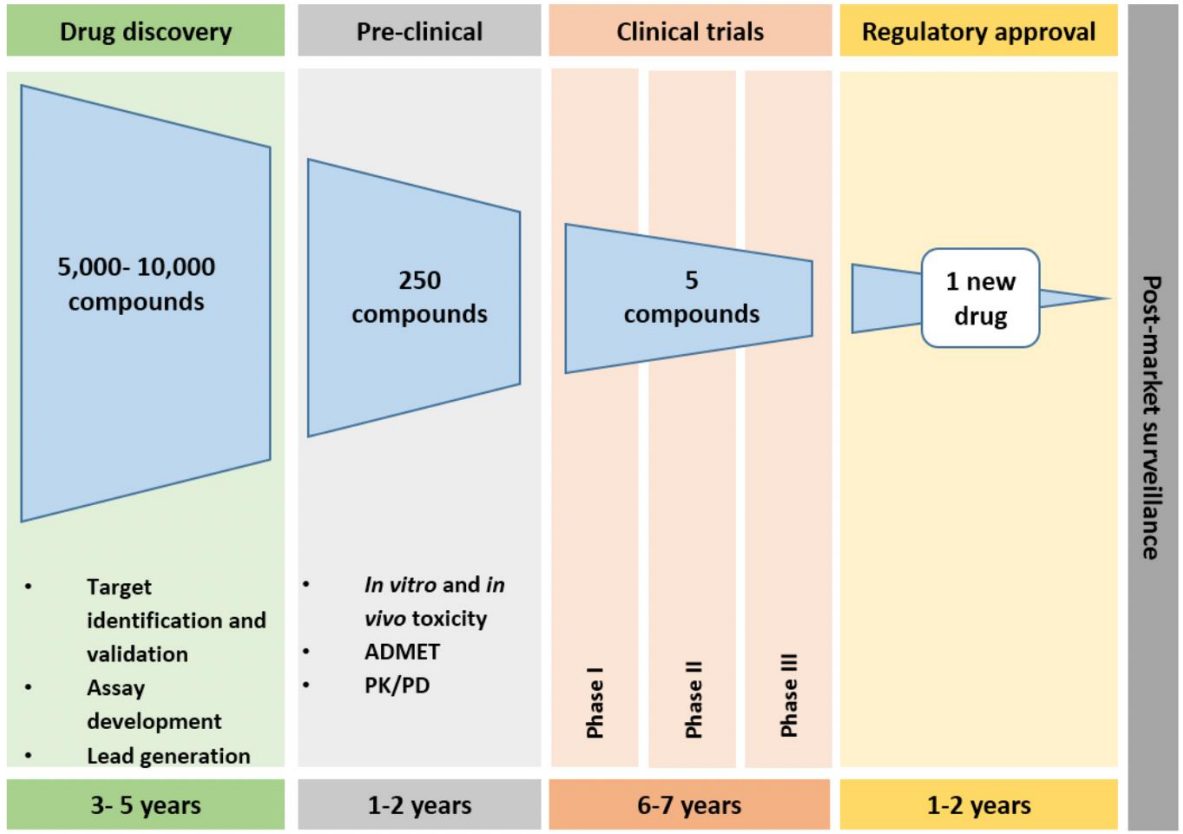
Examples of molecules

Needs for MPP

Drug discovery targets at finding new potential medical compounds with desired properties

Only a small amount of candidate molecules can pass virtual screening to be evaluated in the lead optimization stage

We only have few molecules with known pharmacological properties



Drug discovery and development timeline from [H. Matthews et al., Proteomes 2016]

Outline

- Background
  - Molecular property prediction (MPP)
  - Few-shot learning (FSL)
- Existing works
- The proposed approach
- Summary

Few-shot Learning (FSL)

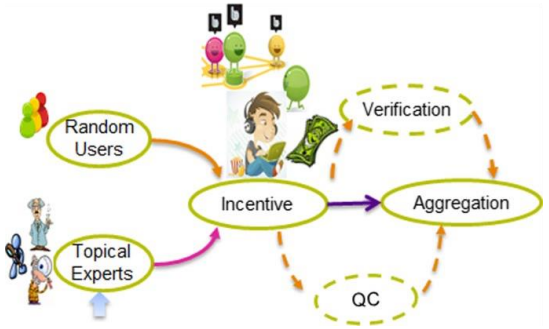
- Definition: A type of machine learning problems contains only a limited number of labeled examples

MPP: intrinsically a few-shot problem

Shot: the number of labeled examples

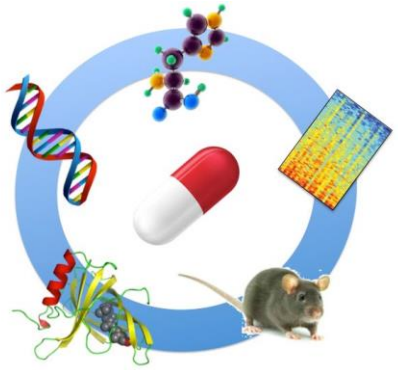
- Typical Scenarios:

Reducing data gathering effort and computational cost



Example: Image / Text Classification
Labor Intensive / Hard
(few labeled images/texts)

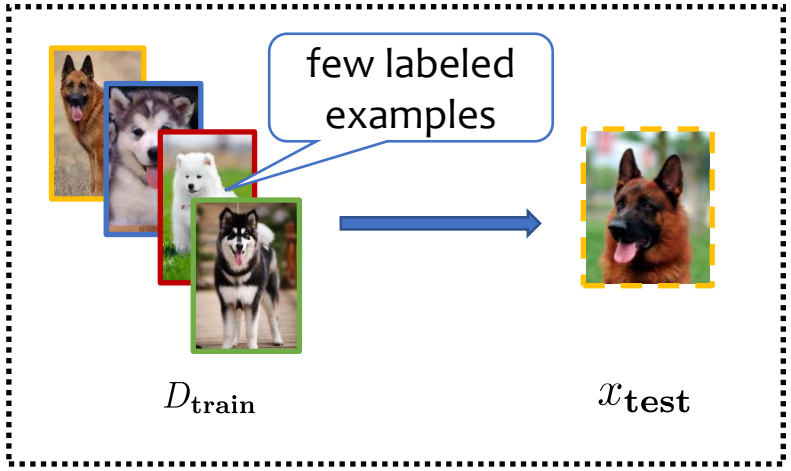
Learning for rare cases

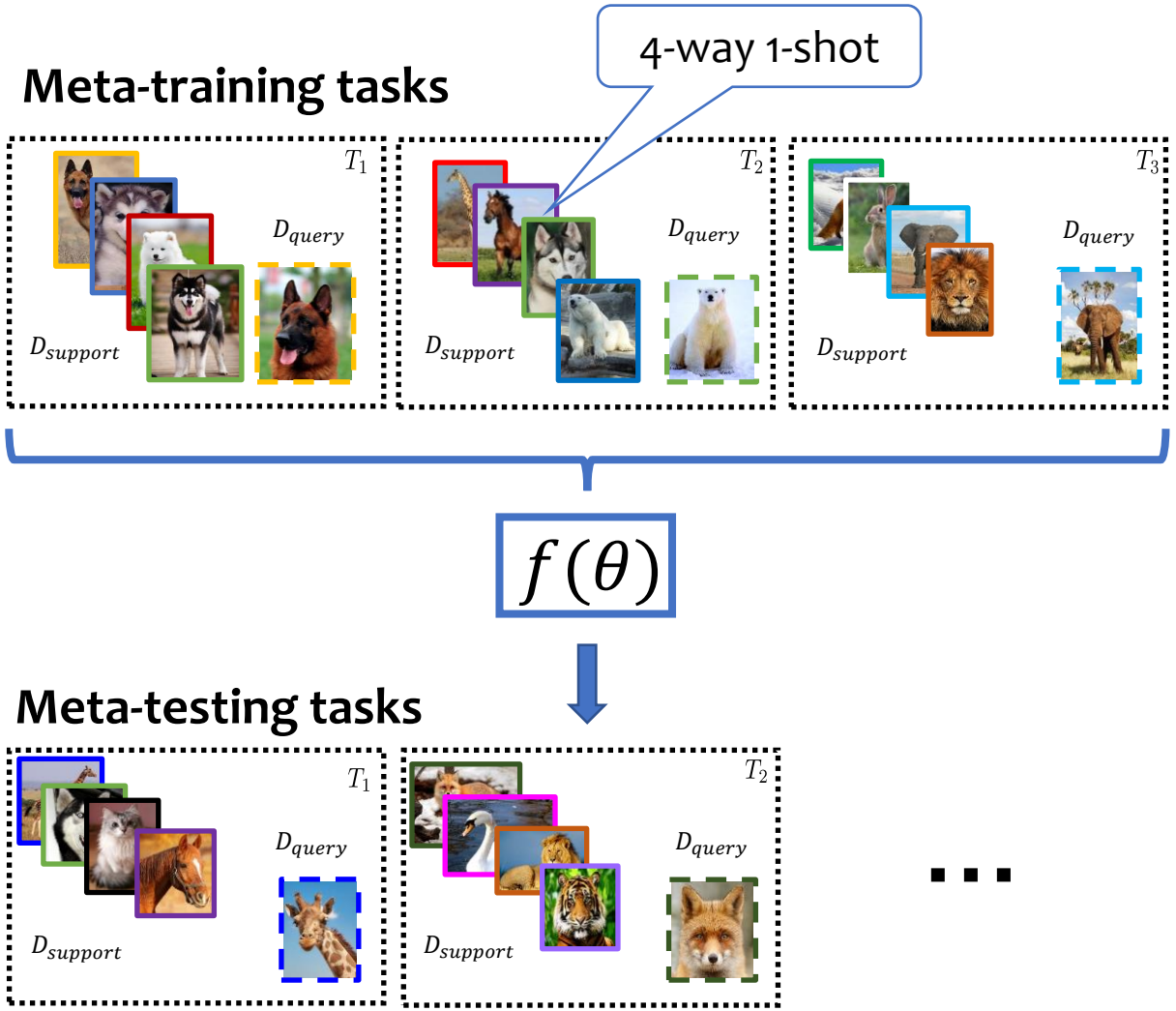


Example: Drug Discovery
Dangerous / Private / Ethical
(few labeled drug molecules)

General Idea

Obtain prior knowledge from meta-training(i.e. similar) tasks

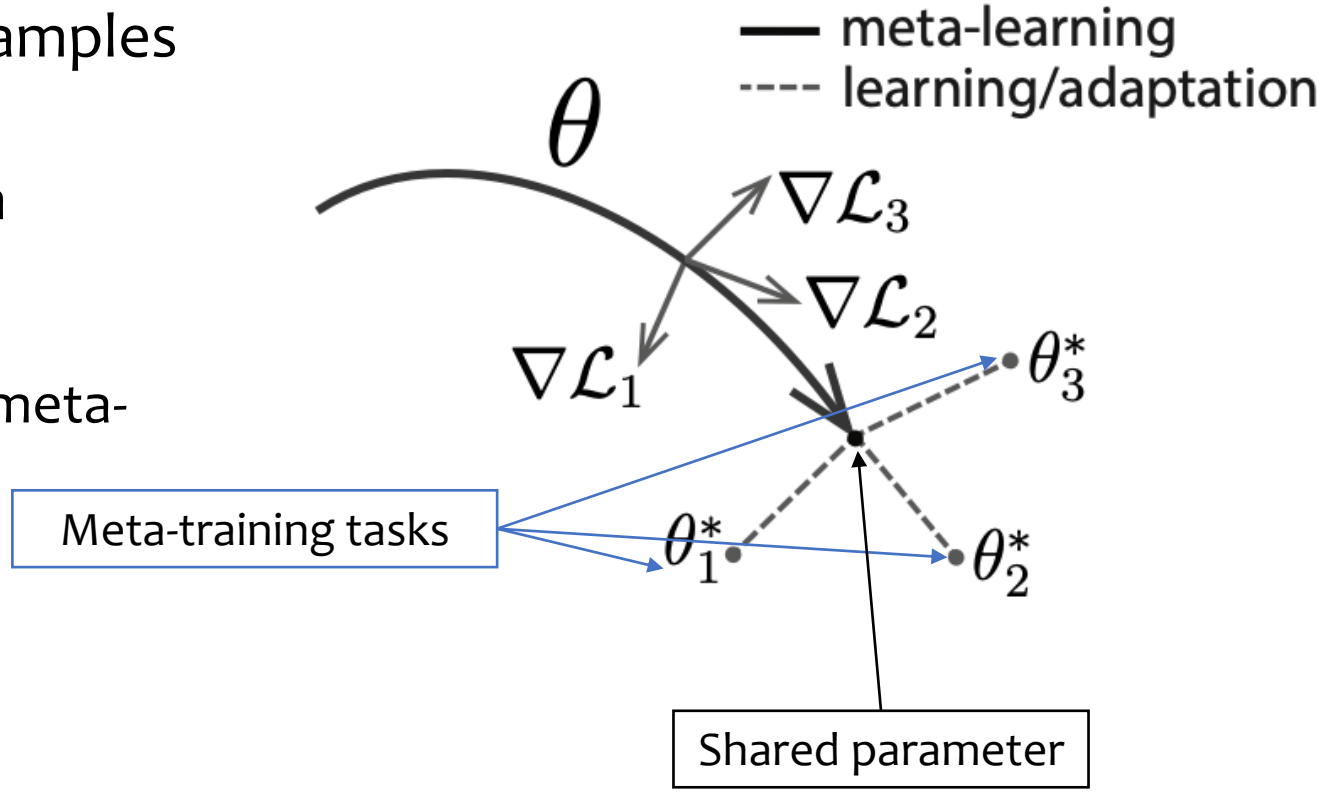




Exemplar Method: MAML

Idea: Train a model on meta-training tasks, such that it can solve new learning tasks using only a small number of training samples

- Prior knowledge: a shared θ learn from meta-training tasks
- How: using gradient descent to train all meta-training task to get θ.



Finn et al. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, ICML

Exemplar Method: MAML

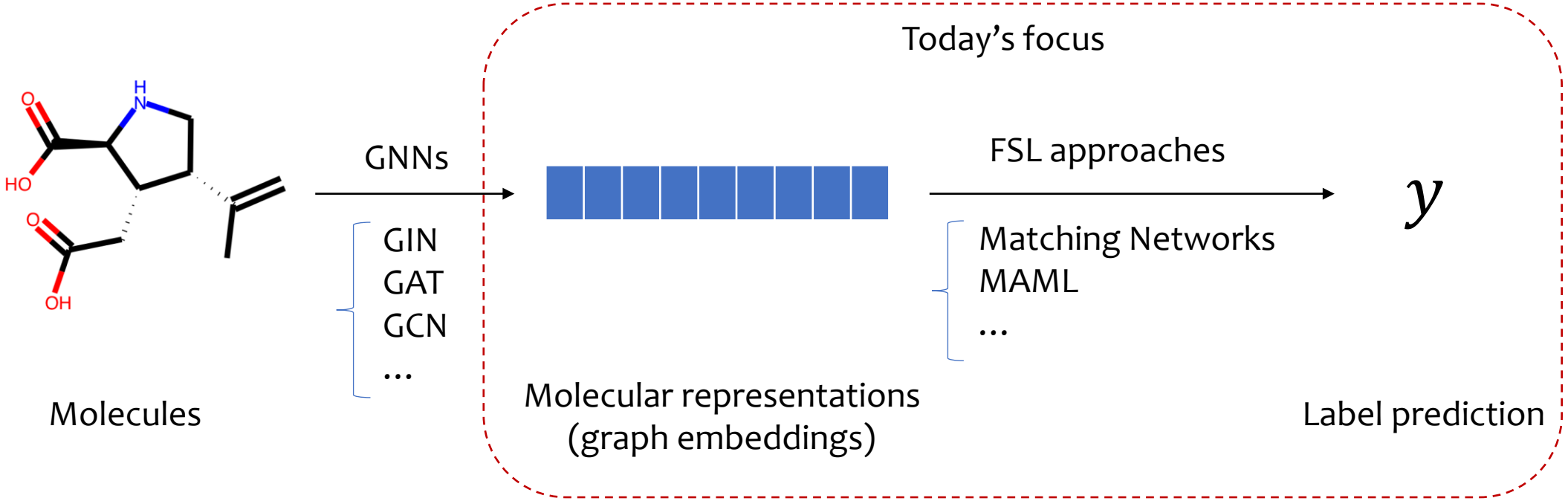
- Algorithm:
  - Step 1: Adapt the model for each meta-training tasks from the shared θ
  - Step 2: Update the share θ by the loss calculated through the models after adaptation

Algorithm 2 MAML for Few-Shot Supervised Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
randomly initialize θ
while not done **do**
  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
  for all $\mathcal{T}_i$ **do**
    Step 1:
    Sample K datapoints $\mathcal{D} = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ using $\mathcal{D}$ and $\mathcal{L}_{\mathcal{T}_i}$ in Equation (2) or (3)
    Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
    Step 2:
    Sample datapoints $\mathcal{D}_i' = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$ for the meta-update
  end for
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$ using each $\mathcal{D}_i'$ and $\mathcal{L}_{\mathcal{T}_i}$ in Equation 2 or 3
end while

Take Home Message

- Molecular Property Prediction is a Few-shot graph learning problem



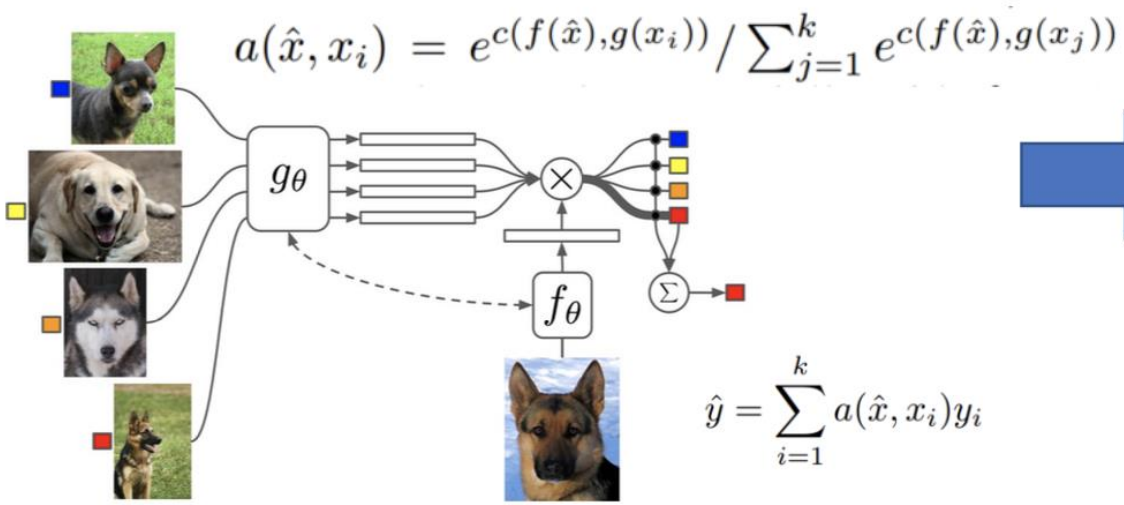
Outline

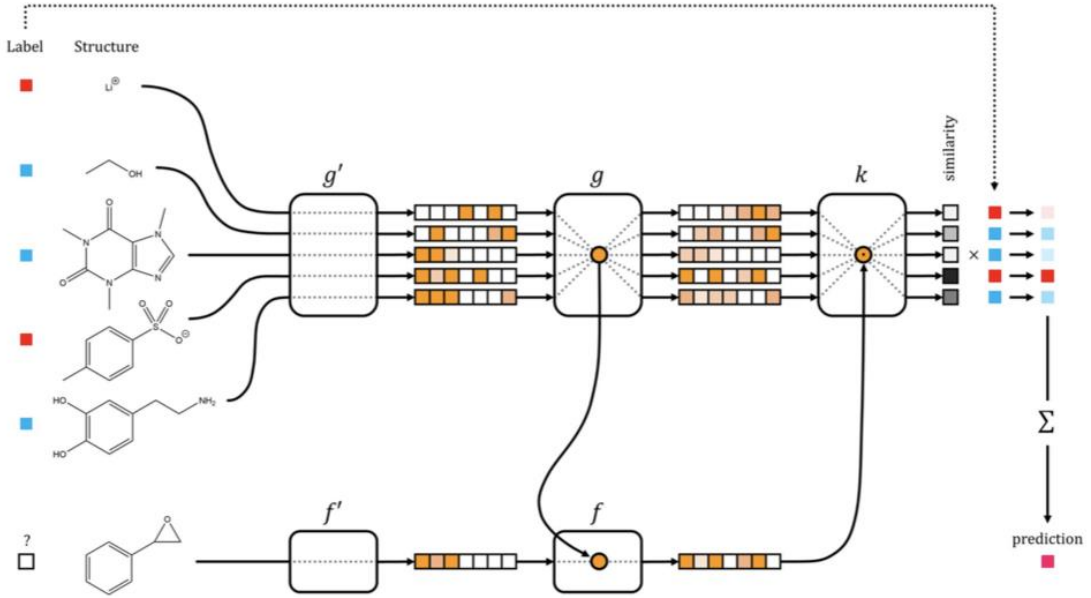
- Background
- Existing works
- The proposed approach
- Summary

Existing Work: IterRefLSTM

Motivation

- Adapt Matching Networks (one-shot learning) to handle molecular property prediction tasks with few training data
- Propose IterRefLSTM to modify Matching Networks architecture





Oriol et al. 2016. Matching Networks for One Shot Learning, NeurIPS

Han et al. 2017. Low Data Drug Discovery with One-Shot Learning, ACS Central Science

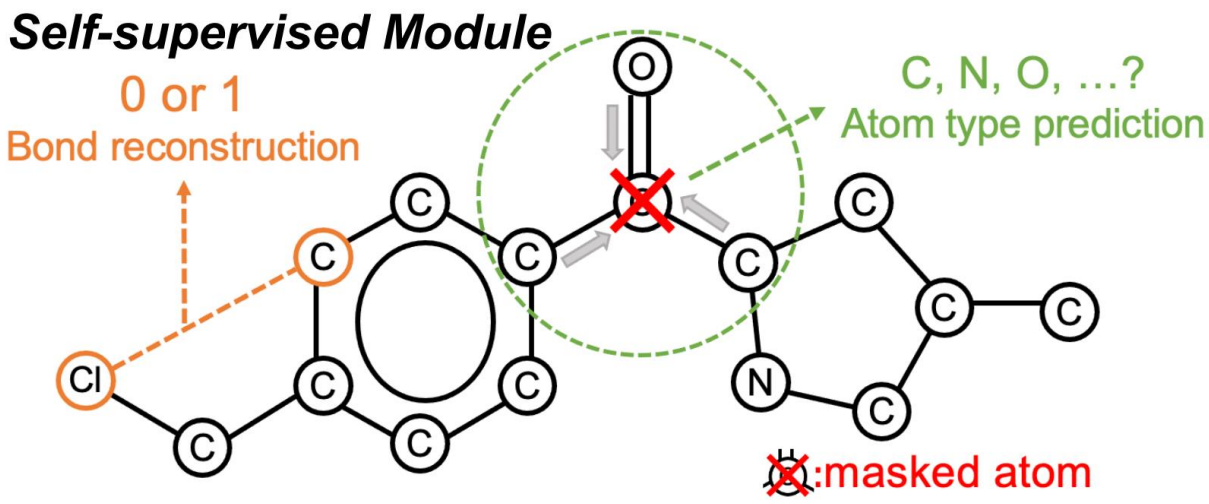
Existing Work: Meta-MGNN

Motivation

- Combine MAML and self-supervised learning (i.e., Pre-GNN) for MPP

Pre-GNN

- Exploit the useful unlabeled information in graphs



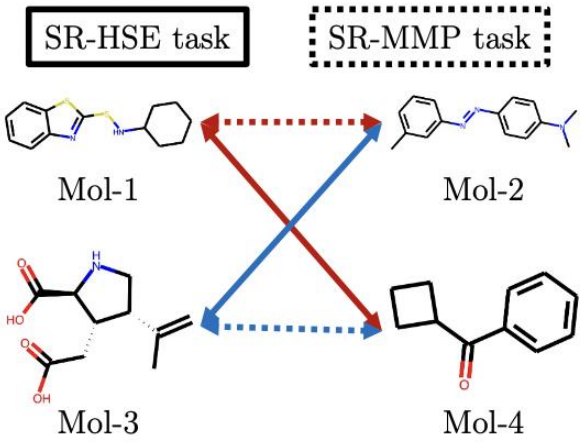
Hu et al. 2020. Strategies for Pre-training Graph Neural Networks, ICLR

Guo et al. 2021. Few-Shot Graph Learning for Molecular Property Prediction, WebConf

Outline

- Background
- Existing works
- The proposed approach
  - Property-Aware Relation networks (PAR)
  - Experiments
- Summary

Motivation



Molecules		Label	
ID	SMILES	SR-HSE	SR-MMP
Mol-1	c1ccc2sc(SNC3CCCCC3)nc2c1	1	1
Mol-2	Cc1cccc(/N=N/c2ccc(N(C)C)cc2)c1	0	1
Mol-3	C=C(C)[C@H]1CN[C@H](C(=O)O)[C@H]1CC(=O)O	0	0
Mol-4	O=C(c1ccccc1) C1CCC1	1	0

Figure 1: Examples of relation graphs for the same molecules coexisting in two tasks of Tox21. Red (blue) edges mean the connected molecules are both active (inactive) on the target property.

Existing works neglect two key facts

- The same molecule shows different properties in different MPP tasks
- Relationships among molecules is important and worth-learning

Property-Aware Relation networks (PAR)

Summary of ideas

- Property-aware molecular embedding

  The same molecule shows different properties in different MPP tasks

- Relation graph learning

  Relationships among molecules is important and worth-learning

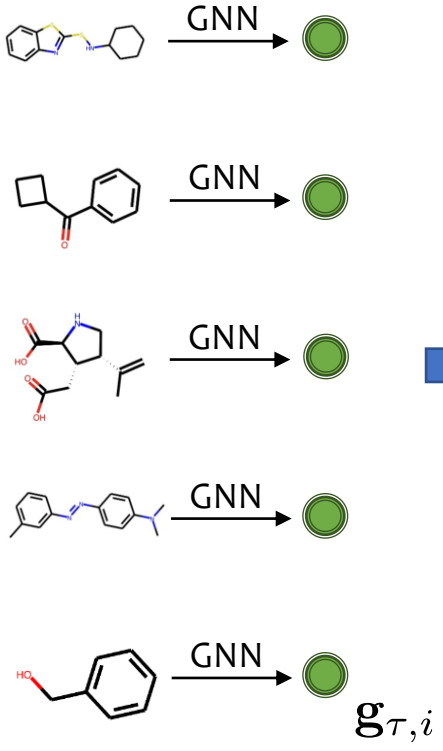
- Selective update

  Separately capture the generic knowledge shared across different tasks and property-aware knowledge within a task

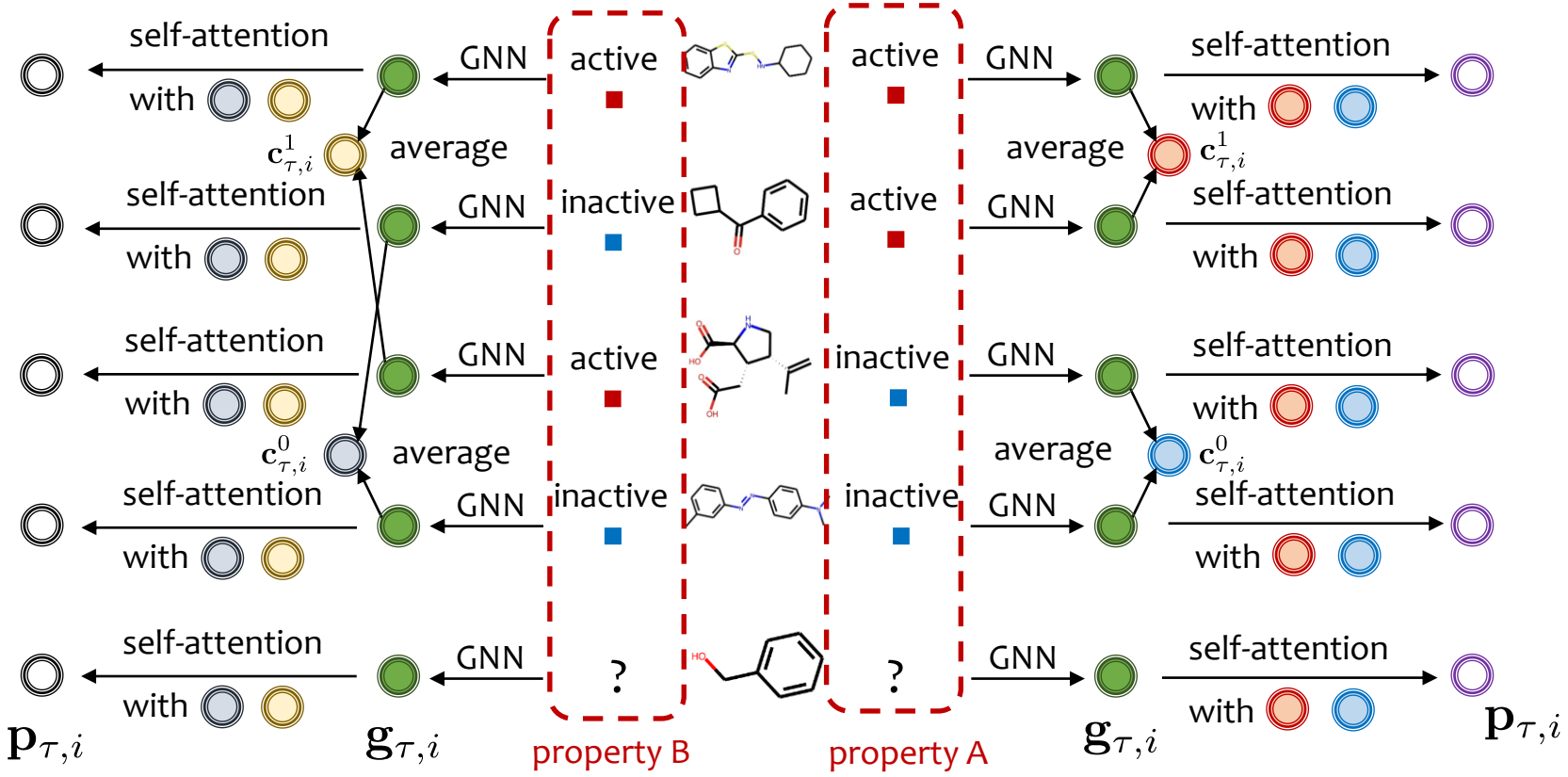
Property-aware Molecular Embedding

The same molecule shows different properties in different MPP tasks

Classical approaches

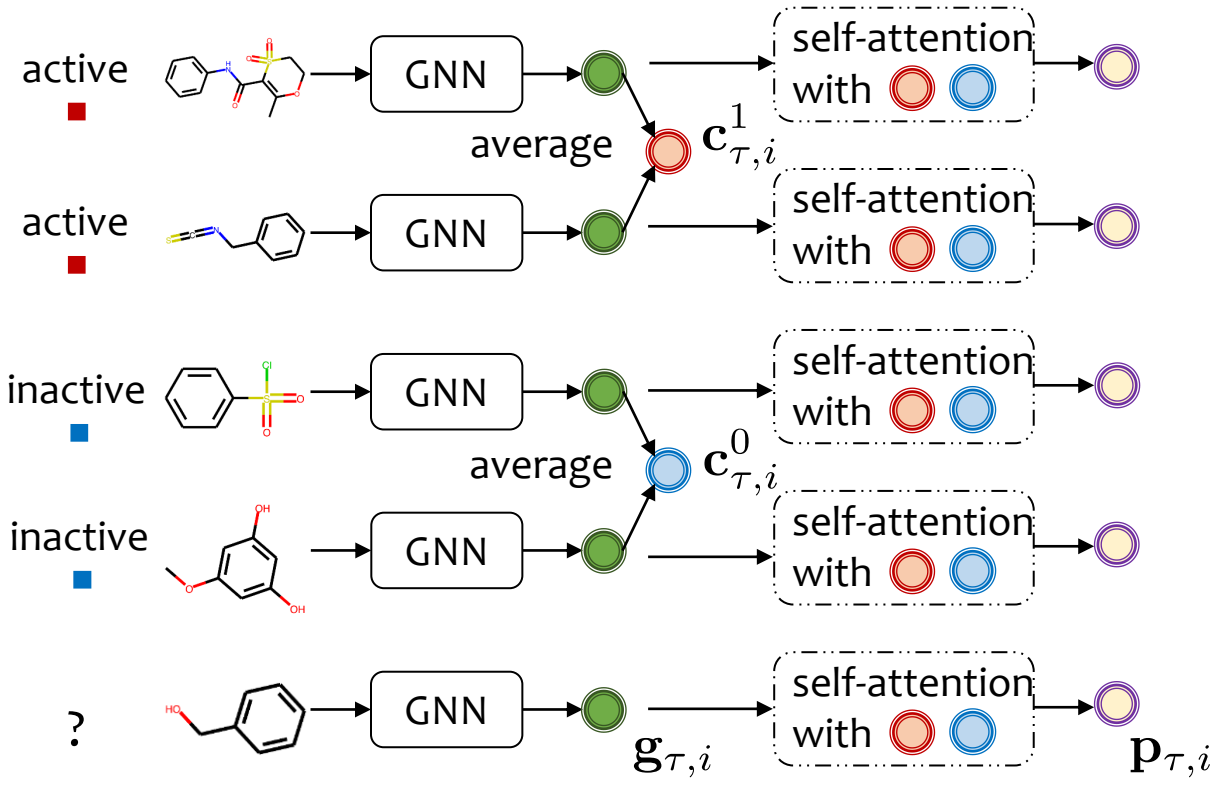


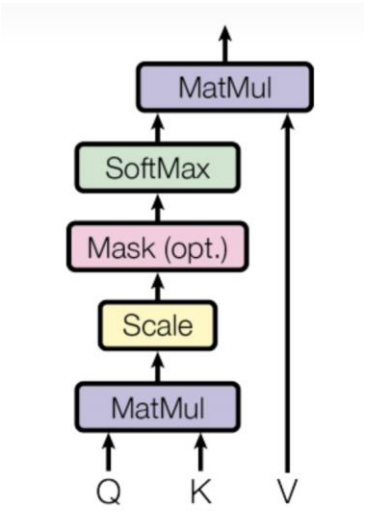
Proposed property-aware Molecular Embedding



Different labels!

Property-aware Molecular Embedding





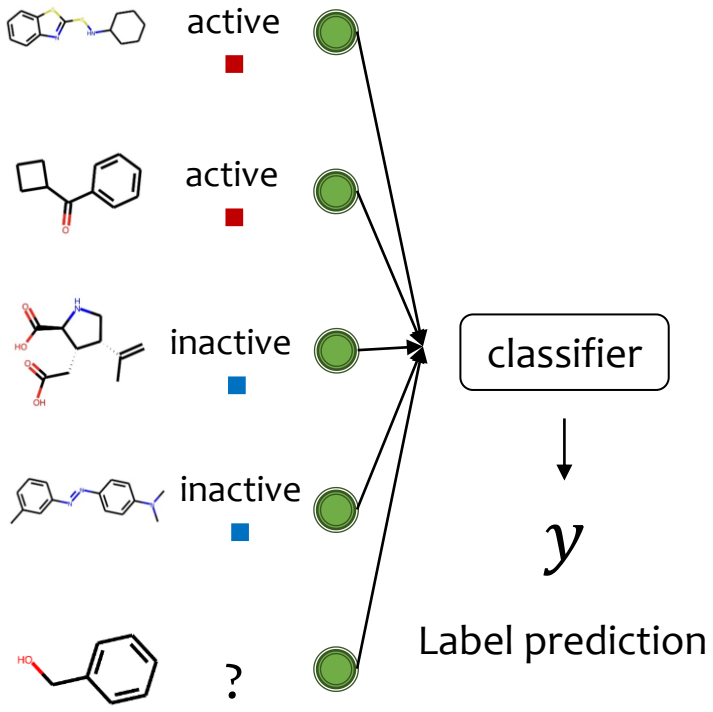
- $g_{\tau,i}$: molecule representation after GNN
- $c^1_{\tau,i}$: representative of active molecules
- $c^0_{\tau,i}$: representative of inactive molecules
- $p_{\tau,i}$: property-aware embedding
- $Q = K = V = [g_{\tau,i}, c^0_{\tau,i}, c^1_{\tau,i}]$

$$\mathbf{b}_{\tau,i} = \left[\mathtt{softmax}(\mathbf{C}_{\tau,i}\mathbf{C}_{\tau,i}^\top/\sqrt{d^g})\mathbf{C}_{\tau,i}\right]_{1:} \text{ with } \mathbf{C}_{\tau,i}^\top = [\mathbf{g}_{\tau,i}, \mathbf{c}^0_\tau, \mathbf{c}^1_\tau] \in \mathbb{R}^{d^g \times 3}$$

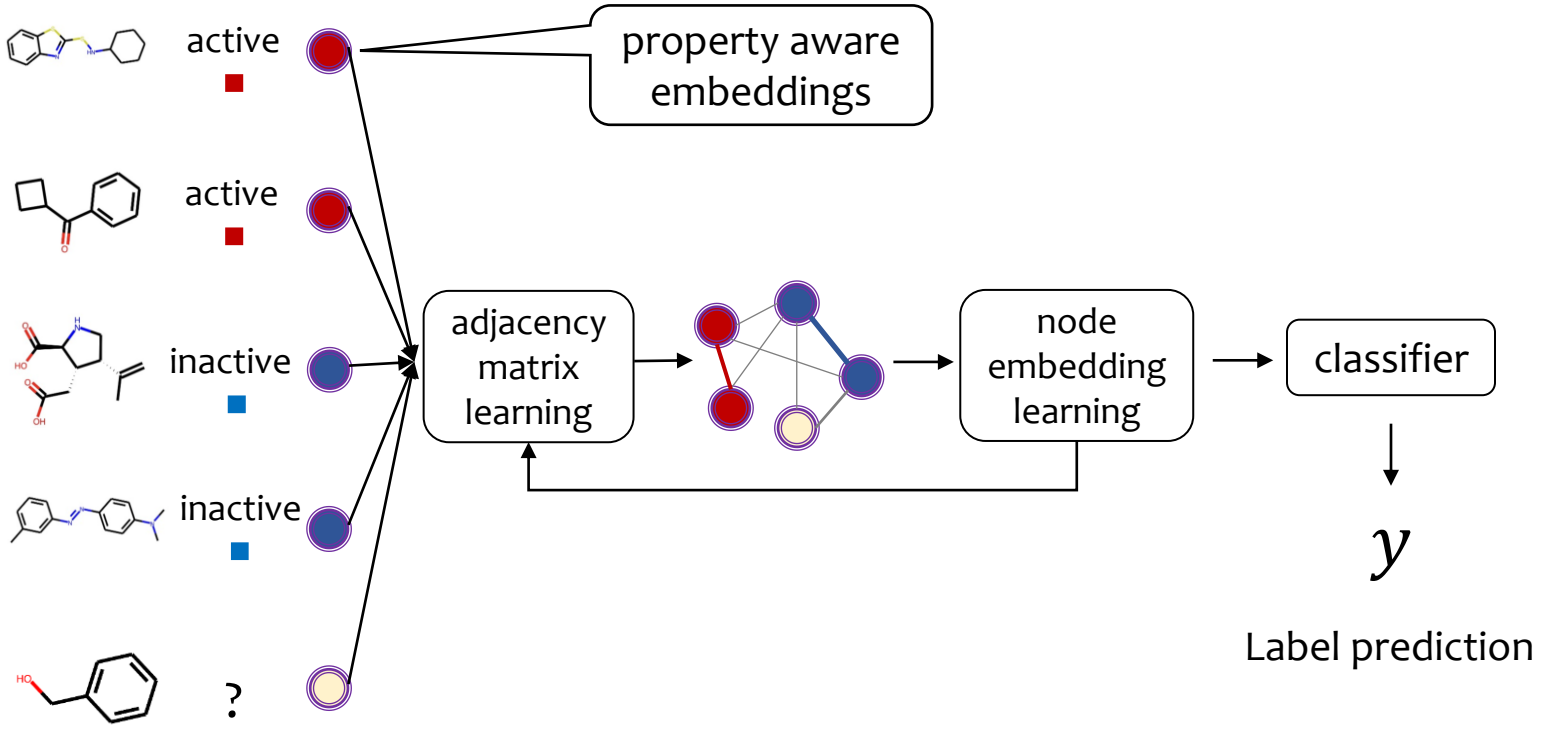
$$\mathbf{p}_{\tau,i} = \mathtt{MLP}_{\mathbf{W}_p}(\mathtt{concat}[\mathbf{g}_{\tau,i}, \mathbf{b}_{\tau,i}])$$

Relation Graph Learning

Relationships among molecules is important and worth-learning



Classical approaches

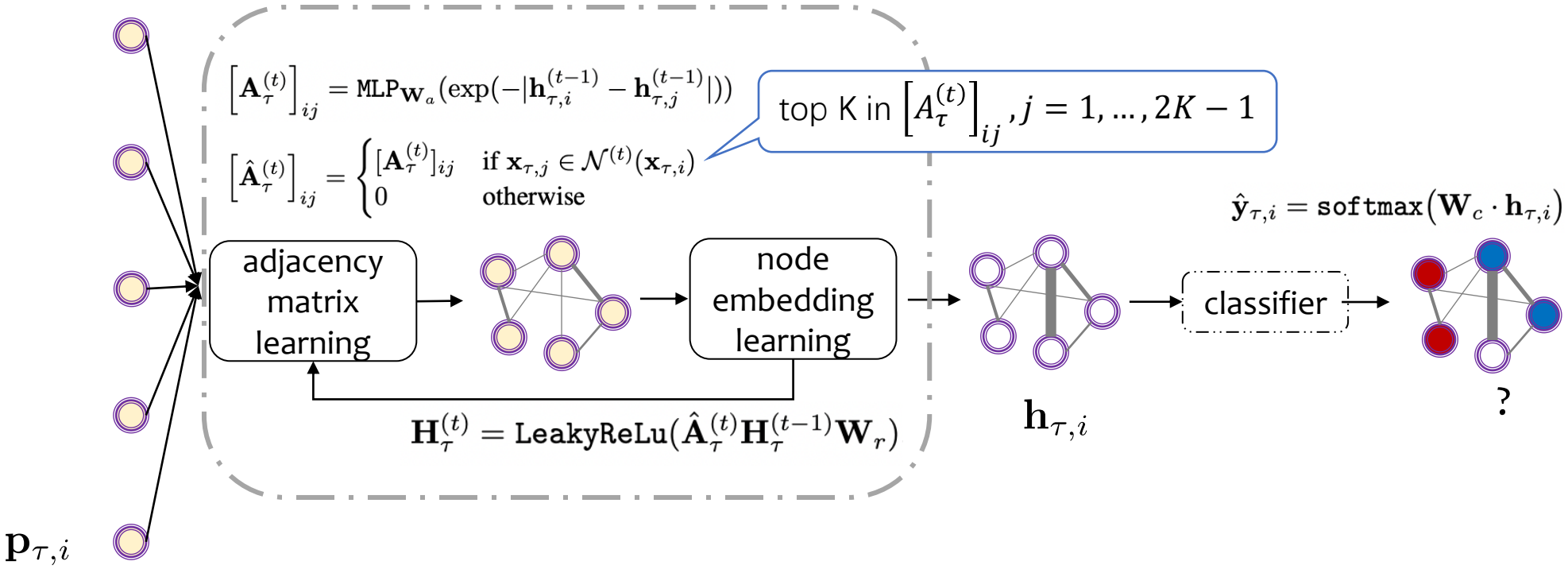


Graph Construction in PAR

Relation Graph Learning

As Relationships among molecules is important and worth-learning, we

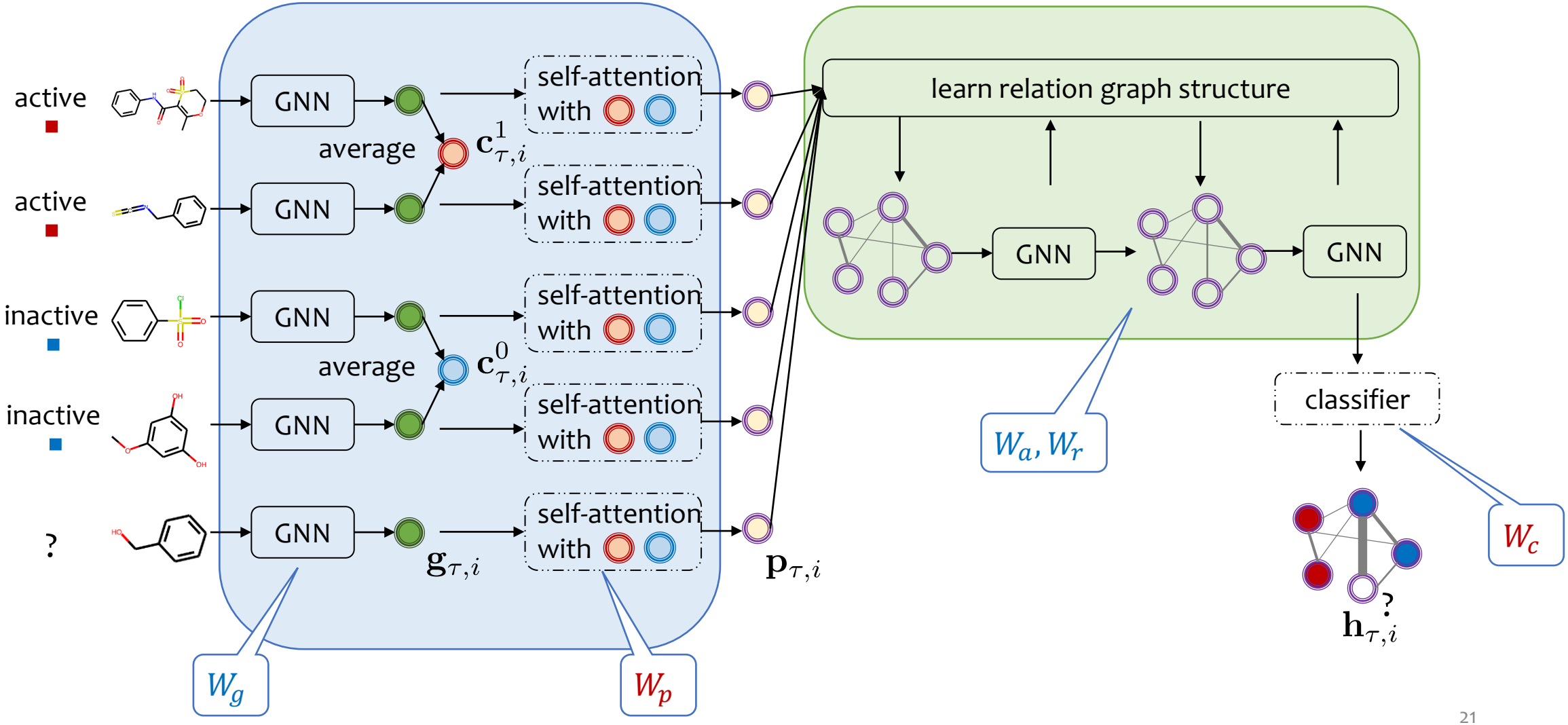
- jointly estimate molecular relation graph and refine molecular embeddings
- then can propagate limited labels efficiently between similar molecules



PAR Framework

$\theta = \{W_g, W_a, W_r\}$, generic knowledge, shared across tasks

$\Phi = \{W_p, W_c\}$, property-aware knowledge, specific to one task



Training and Inference

Denote PAR as $f_{\theta,\phi}$

- $\theta = \{W_g, W_a, W_r\}$: parameters of molecular encoder and relation graph learning module
- $\phi = \{W_p, W_c\}$: parameters of property-aware embedding function and classifier

We learn from a set of meta-training tasks a good shared parameter

$$\min_{\boldsymbol{\theta},\boldsymbol{\Phi}} \sum_{\tau=1}^{N_t} \mathcal{L}(\mathcal{Q}_\tau, f_{\boldsymbol{\theta},\boldsymbol{\Phi}_\tau})$$

Within each task, we fix θ while fine-tune Φ as Φ_τ

Ground-truth labels

$$\mathcal{L}(\mathcal{S}_\tau, f_{\boldsymbol{\theta},\boldsymbol{\Phi}}) = \sum_{(\mathbf{x}_{\tau,i}, y_{\tau,i}) \in \mathcal{S}_\tau} \underbrace{-\mathbf{y}_{\tau,i}^\top \cdot \log(\hat{\mathbf{y}}_{\tau,i})}_{\text{classification loss}} + \underbrace{\|[\mathbf{A}_\tau^*]_{i:} - [\hat{\mathbf{A}}_\tau]_{i:}\|_2^2}_{\text{neighbor alignment regularizer}} \qquad \boldsymbol{\Phi}_\tau = \boldsymbol{\Phi} - \alpha \nabla_{\boldsymbol{\Phi}} \mathcal{L}(\mathcal{S}_\tau, f_{\boldsymbol{\theta},\boldsymbol{\Phi}})$$

to separately capture the generic knowledge shared across different tasks and those property-aware

Algorithm: Selective update

Algorithm 1 Meta-training procedure for PAR.

initialize $\boldsymbol{\theta} = \{\mathbf{W}_g, \mathbf{W}_a, \mathbf{W}_r\}$ and $\boldsymbol{\Phi} = \{\mathbf{W}_p, \mathbf{W}_c\}$ randomly; if a pretrained molecular encoder is available, take its parameter as $\mathbf{W}_g$;
while not done **do**
  sample a batch of tasks $\mathcal{T}_\tau$;
  for all $\mathcal{T}_\tau$ **do**
    sample support set $\mathcal{S}_\tau$ and query set $\mathcal{Q}_\tau$ from $\mathcal{T}_\tau$;
    obtain molecular embedding $\mathbf{g}_{\tau,i}$ for each $\mathbf{x}_{\tau,i}$ by a graph-based molecular encoder;
    adapt $\mathbf{g}_{\tau,i}$ to be property-aware $\mathbf{p}_{\tau,i}$ by (5);
    initialize node embeddings as $\mathbf{h}_{\tau,i}^{(0)} = \mathbf{p}_{\tau,i}$;
    for $t = 1, \ldots, T$ **do**
      estimate adjacency matrix $\mathbf{A}_\tau^{(t)}$ of relation graph among molecules using $\mathbf{h}_{\tau,i}^{(t-1)}$ by (6);
      refine $\mathbf{h}_{\tau,i}^{(t)}$ on the updated relation graph $\mathbf{A}_\tau^{(t)}$ by (8);
    end for
    obtain class prediction $\hat{\mathbf{y}}_{\tau,i}$ using $\mathbf{h}_{\tau,i} = \mathbf{h}_{\tau,i}^{(T)}$;
    evaluate training loss $\mathcal{L}(\mathcal{S}_\tau, f_{\boldsymbol{\theta},\boldsymbol{\Phi}})$ on $\mathcal{S}_\tau$;
    fine-tune $\boldsymbol{\Phi}$ as $\boldsymbol{\Phi}_\tau$ by (11);
    evaluate testing loss $\mathcal{L}(\mathcal{Q}_\tau, f_{\boldsymbol{\theta},\boldsymbol{\Phi}_\tau})$ on $\mathcal{Q}_\tau$;
  end for
  update $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ by (12);
end while

Property-aware Molecular Embedding

Graph Construction

Selective update

MAML-based

Outline

- Background
- Existing works
- The proposed approach
  - Property-Aware Relation networks (PAR)
  - Experiments
- Summary

Experiment Setup

- Two sets of baselines
  - Methods with graph-based encoder learned from scratch including Siamese [Koch et al., 2015], ProtoNet [Snell et al., 2017], MAML [Finn et al., 2017], TPN [Liu et al., 2018], and EGNN [Kim et al., 2019], IterRefLSTM [Altae-Tran et al., 2017];
  - Methods which leverage pretained graph-based molecular encoder including Pre-GNN [Hu et al., 2019], Meta-MGNN [Guo et al., 2021], and Pre-PAR which is our PAR equipped with Pre- GNN.
- Four datasets

Dataset	Tox21	SIDER	MUV	ToxCast
# Compounds	8014	1427	93127	8615
# Tasks	12	27	17	617
# Meta-Training Tasks	9	21	12	450
# Meta-Testing Tasks	3	6	5	167

- Link of our code: https://github.com/tata1661/PAR-NeurIPS21

Overall Comparison

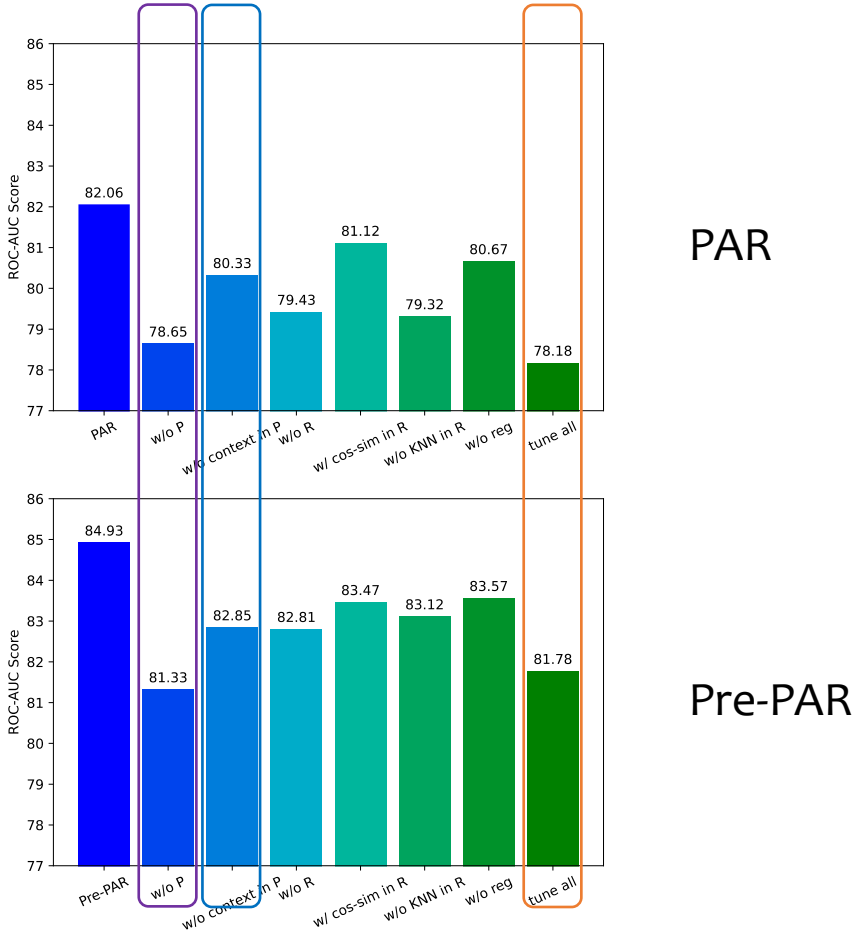
	Method	Tox21		SIDER		MUV		ToxCast	
		10-shot	1-shot	10-shot	1-shot	10-shot	1-shot	10-shot	1-shot
Train directly	Siamese	$80.40_{(0.35)}$	$65.00_{(1.58)}$	$71.10_{(4.32)}$	$51.43_{(3.31)}$	$59.96_{(5.13)}$	$50.00_{(0.17)}$	-	-
	ProtoNet	$74.98_{(0.32)}$	$65.58_{(1.72)}$	$64.54_{(0.89)}$	$57.50_{(2.34)}$	$65.88_{(4.11)}$	$58.31_{(3.18)}$	$63.70_{(1.26)}$	$56.36_{(1.54)}$
	MAML	$80.21_{(0.24)}$	$75.74_{(0.48)}$	$70.43_{(0.76)}$	$67.81_{(1.12)}$	$63.90_{(2.28)}$	$60.51_{(3.12)}$	$66.79_{(0.85)}$	$65.97_{(5.04)}$
	TPN	$76.05_{(0.24)}$	$60.16_{(1.18)}$	$67.84_{(0.95)}$	$62.90_{(1.38)}$	$65.22_{(5.82)}$	$50.00_{(0.51)}$	$62.74_{(1.45)}$	$50.01_{(0.05)}$
	EGNN	$81.21_{(0.16)}$	$79.44_{(0.22)}$	$72.87_{(0.73)}$	$70.79_{(0.95)}$	$65.20_{(2.08)}$	$62.18_{(1.76)}$	$63.65_{(1.57)}$	$61.02_{(1.94)}$
	IterRefLSTM	$81.10_{(0.17)}$	$80.97_{(0.10)}$	$69.63_{(0.31)}$	$71.73_{(0.14)}$	$49.56_{(5.12)}$	$48.54_{(3.12)}$	-	-
	PAR	$82.06_{(0.12)}$	$80.46_{(0.13)}$	$74.68_{(0.31)}$	$71.87_{(0.48)}$	$66.48_{(2.12)}$	$64.12_{(1.18)}$	$69.72_{(1.63)}$	$67.28_{(2.90)}$
Pre-train GNN	Pre-GNN	$82.14_{(0.08)}$	$81.68_{(0.09)}$	$73.96_{(0.08)}$	$73.24_{(0.12)}$	$67.14_{(1.58)}$	$64.51_{(1.45)}$	$73.68_{(0.74)}$	$72.90_{(0.84)}$
	Meta-MGNN	$82.97_{(0.10)}$	$82.13_{(0.13)}$	$75.43_{(0.21)}$	$73.36_{(0.32)}$	$68.99_{(1.84)}$	$65.54_{(2.13)}$	-	-
	Pre-PAR	$84.93_{(0.11)}$	$83.01_{(0.09)}$	$78.08_{(0.16)}$	$74.46_{(0.29)}$	$69.96_{(1.37)}$	$66.94_{(1.12)}$	$75.12_{(0.84)}$	$73.63_{(1.00)}$

- Pre-PAR consistently obtains the best performance
- PAR outperforms among methods without pretrained GNNs

Ablation Study

We further compare with

- **w/o P:** w/o property-aware embedding
- **w/o context in P:** w/o context $b_{\tau,i}$ in P
- **w/o R:** w/o adaptive relation graph learning
- **w/ cos-sim in R:** use cosine similarity to obtain the adjacency matrix
- **w/o KNN in R:** w/o reducing the learned relation graph to KNN graph
- **w/o reg:** w/o the neighbor alignment regularizer
- **tune all:** without selective update



PAR

Pre-PAR

10-shot tasks from Tox21

All components are vital to the success of PAR

Case Study on 10 Molecules

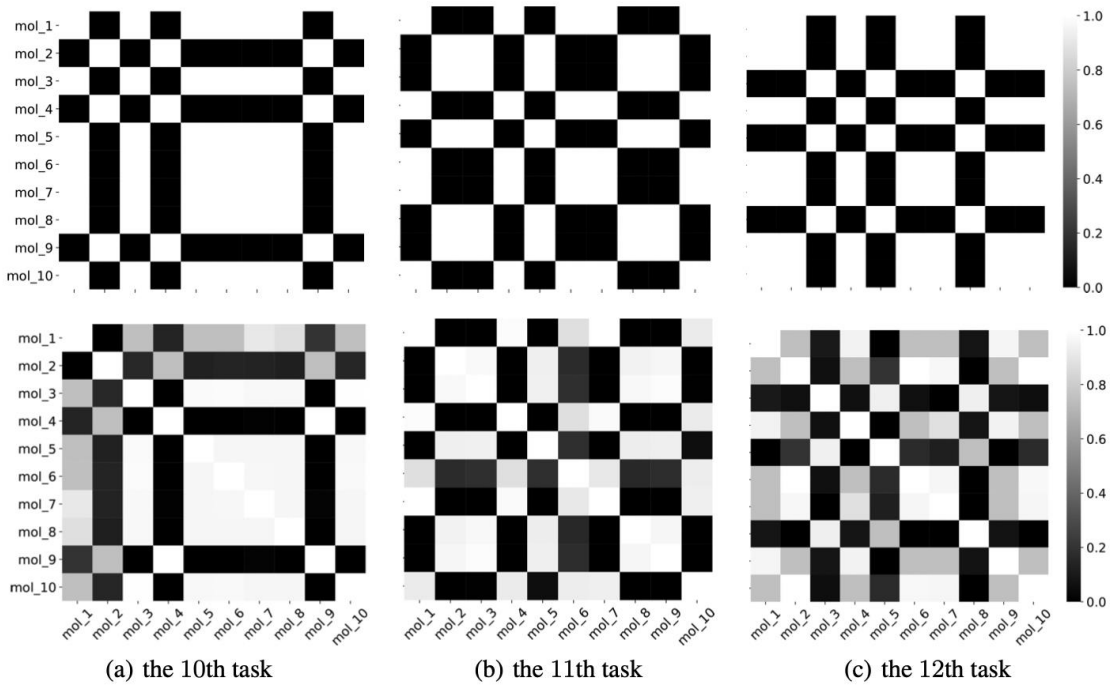
Can PAR obtain different property-aware molecular embeddings and relation graphs for tasks containing overlapping molecules but evaluating different properties?

Table 5: The 10 molecules sampled from Tox21 dataset, which coexist in the three meta-testing tasks (the 10th task for SR-HSE, the 11th task for SR-MMP, and the 12th task for SR-p53).

Molecule		Label		
ID	SMILES	SR-HSE	SR-MMP	SR-p53
Mol-1	Cc1cccc(/N=N/c2ccc(N(C)C)cc2)c1	0	1	0
Mol-2	O=C(c1ccccc1)C1CCC1	1	0	0
Mol-3	C=C(C)[C@H]1CN[C@H](C(=O)O)[C@H]1CC(=O)O	0	0	1
Mol-4	c1ccc2sc(SNC3CCCCC3)nc2c1	1	1	0
Mol-5	C=CCSSCC=C	0	0	1
Mol-6	CC(C)(C)c1cccc(C(C)(C)C)c1O	0	1	0
Mol-7	C[C@@H]1CC2(OC3C[C@@]4(C)C5=CC [C@H]6C(C)(C)C(O[C@@H]7OC[C@@H] (O)[C@H](O)[C@H]7O)CC[C@@]67C[C@@]57CC[C@]4(C)C31)OC(O)C1(C)OC21	0	1	0
Mol-8	O=C(CCCCCCC(=O)Nc1ccccc1)NO	0	0	1
Mol-9	CC/C=C\\C/C=C\\C/C=C\\CCCCCCCC(=O)O	1	0	0
Mol-10	Cl[Si](Cl)(c1ccccc1)c1ccccc1	0	1	0

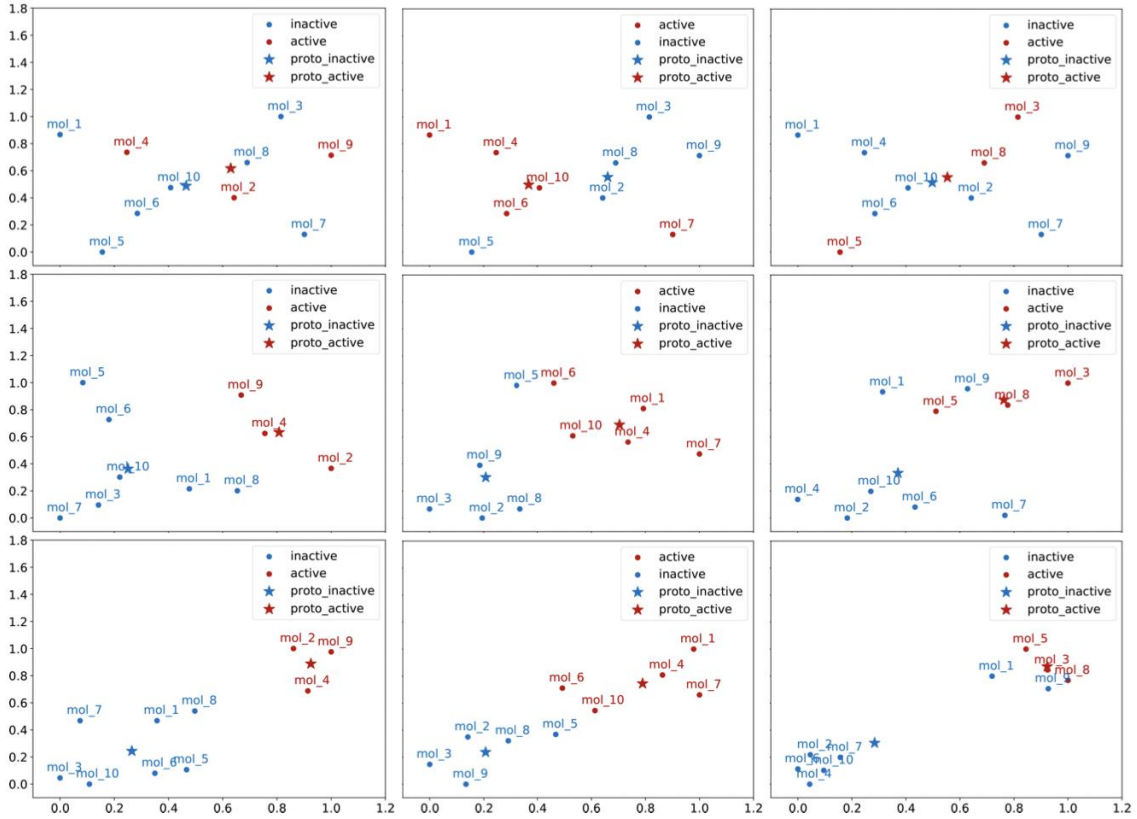
a fixed group of 10 molecules coexist in different meta-testing tasks

Visualization



(a) the 10th task (b) the 11th task (c) the 12th task

Figure 5: Comparison between $\mathbf{A}_\tau^*$ computed using ground-truth labels (the first row) and adjacency matrix $\mathbf{A}_\tau$ returned by PAR (the second row) for the ten molecules. We set $[\mathbf{A}_\tau^*]_{ij} = 1$ if molecules $\mathbf{x}_{\tau,i}$ and $\mathbf{x}_{\tau,j}$ have the same label and 0 otherwise.



(a) the 10th task (b) the 11th task (c) the 12th task

Figure 6: t-SNE visualization of $\mathbf{g}_{\tau,i}$ (the first row), $\mathbf{p}_{\tau,i}$ (the second row), and $\mathbf{h}_{\tau,i}$ (the third row) of the ten molecules. Proto_active (proto_inactive) denotes the class prototype of active (inactive) class.

PAR can model relation graphs (left) and property-aware molecular embeddings (right)

Outline

- Background
- Existing works
- The proposed approach
- Summary

Summary

- Molecular property prediction (MPP) is a few-shot learning problem on graphs
- Existing works ignore molecule's difference in different tasks and molecule's relationship inside one task
- Our Property-Aware Relation networks (PAR) solves above problems with
  - Property-aware molecular embedding / Relation graph learning / Selective update
- PAR significantly outperforms existing works

FSL: Learning resources

A CSUR Survey

- Generalizing from a Few Examples: A Survey on Few-Shot Learning. ACM CSUR. 2020
- Show you a roadmap

A GitHub Repo

- https://github.com/tata1661/FSL-Mate
- Update-to-date research papers

A Toolbox

- Support various applications and platforms
- Built with Baidu

SURVEY
June 2020

Generalizing from a Few Examples: A Survey on Few-shot Learning

Yaqing Wang, Quanming Yao, James T. Kwok, Lionel M. Ni

ACM Computing Surveys (CSUR), Volume 53, Issue 3 • May 2021, Article No.: 63, pp 1–34 • https://doi.org/10.1145/3386252

Machine learning has been highly successful in data-intensive applications but is often hampered when the data set is small. Recently, Few-shot Learning (FSL) is proposed to tackle this problem. Using prior knowledge, FSL can rapidly generalize to new ...

314 9,508 Get Access

Top 5 cited in CSUR last 5 years.

tata1661 / FSL-Mate Public

Unwatch 44 Fork 254 Starred 1.3k

Code Issues 6 Pull requests 3 Actions Projects Wiki Security Insights

master 2 branches 0 tags Go to file Add file Code

tata1661 Add PaddleFSL paper and bib, fix ... 676f5a3 27 days ago 60 commits

FewShotPapers PaddleFSL updates to v1.1.0 (add support for ... last month

PaddleFSL Add PaddleFSL paper and bib, fix bugs in exa... 27 days ago

About

FSL-Mate: A collection of resources for few-shot learning (FSL).

deep-learning paper papers one-shot-learning paddlepaddle meta-learning few-shot

			learn2learn	Torchmeta	keras-fsl	mmfewshot	Libfewshot	PaddleFSL
Application	CV	MAML	✔	✔	✗	✔	✔	✔
		ProtoNet	✔	✔	✗	✔	✔	✔
		RelationNet	✗	✗	✗	✔	✔	✔
	NLP	Siamese	✗	✗	✔	✗	✗	✔
		PET	✗	✗	✗	✗	✗	✔
		P-Tuning	✗	✗	✗	✗	✗	✔
	BIO	MAML	✗	✗	✗	✗	✗	✔
		PAR	✗	✗	✗	✗	✗	✔
Unit Test			✔	✔	✗	✔	✗	✔
Document			✔	✔	✔	✔	✔	✔
Platform		CPU	✔	✔	✔	✔	✔	✔
		GPU	✔	✔	✔	✔	✔	✔
		Cluster	✗	✗	✗	✗	✗	✔
Activity		Download	45182	131963	5484	913	N/A	24441
		Star	1725	1576	172	359	496	1247
		Fork	249	203	29	42	87	240

Table 1: Comparison of PaddleFSL with other popular FSL toolboxes. The activity statistics is collected from the respective GitHub pages on 2022/3/18.

Future Works

- Transfer learning across datasets
- Few-shot molecule regression
- Design GNNs for better molecular representations

Thanks! Questions?