Aalto University

Deep representation Learning for Small Molecules

Juho Rousu
Department of Computer Science
Aalto University

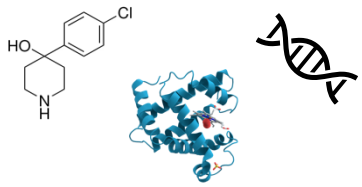
juho.rousu@aalto.fi

KEPACO group: Kernel Methods, Pattern Analysis and Computational biology

Department of Computer Science, Aalto University

Machine Learning methods

Structured input and output

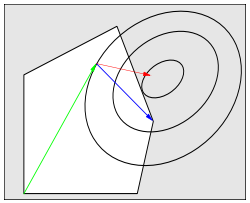


Embedding/kernel learning

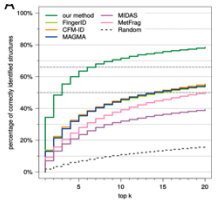
Regularized modeling

$$\max_{\boldsymbol{u}_x,\boldsymbol{u}_y} \quad \rho(\boldsymbol{u}_x,\boldsymbol{u}_y) = \frac{\mathbf{k}_x(\cdot,\boldsymbol{u}_x)^T \mathbf{k}_y(\cdot,\boldsymbol{u}_y)}{\|\mathbf{k}_x(\cdot,\boldsymbol{u}_x)\|_2 \|\mathbf{k}_y(\cdot,\boldsymbol{u}_y)\|_2}$$
$$\text{s.t.} \|\boldsymbol{u}_x\|_{P_u} \leq s_u \text{ and } \|\boldsymbol{u}_y\|_{P_v} \leq s_v$$

Efficient training algorithms

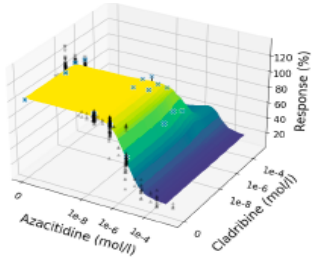


Prediction and interpretation

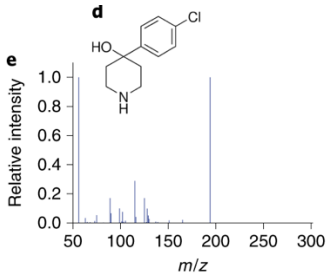


Life Science Applications

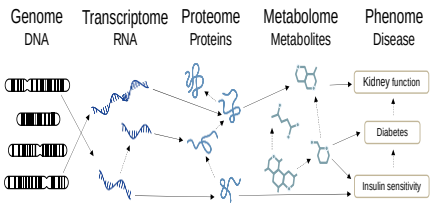
Predicting drug responses



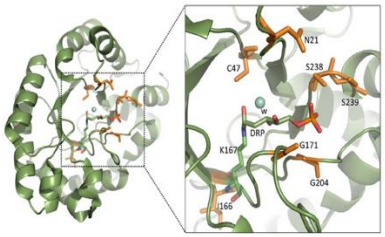
Small molecule identification



Multiomic biomarker discovery



Synthetic biology



Contact: Prof. Juho Rousu, juho.rousu@aalto.fi

WWW: https://research.cs.aalto.fi//kepaco/

Aalto University

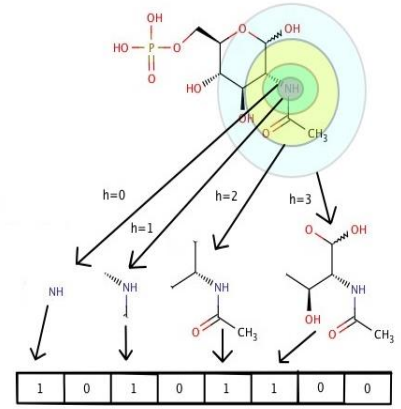
Representation learning for molecular data

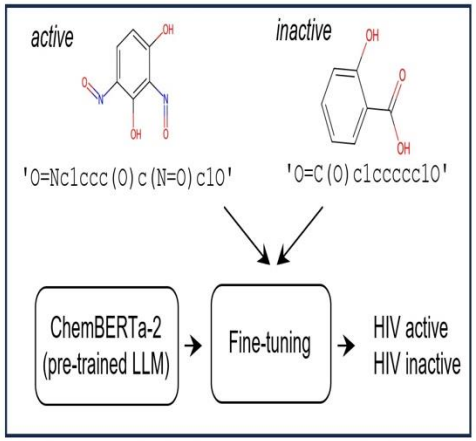
Challenge: data-hungriness of deep learning

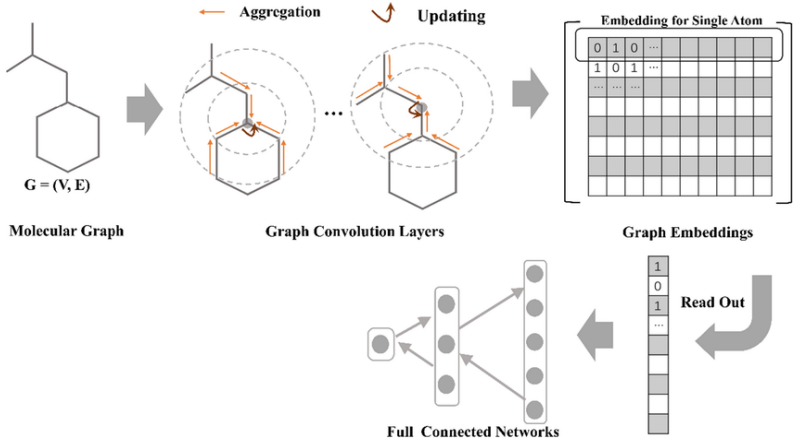
Dataset/Tool	Application	Datatype	Size (examples)	Year
MovieLens	Movie reviews	Text	22M	2016
ImageNet	Object recognition	Image	14M	2014
AlphaFold 2	Protein structure	Protein 3D structures	100K	2020
Moleculenet	Drug discovery	Molecules	700K	2018
UPTSO-50K	Chem. Synthesis	Chemical reactions	50K	2016

Molecular representations for ML

- **Molecular fingerprints/descriptors (right): fixed sets of features, based either on chemical knowledge or combinatorial algorithms**
- **Large language model (LMM) representations – learned from SMILES strings of large set of molecules**
- **Graph Neural Networks – embeddings learned from molecular graphs**



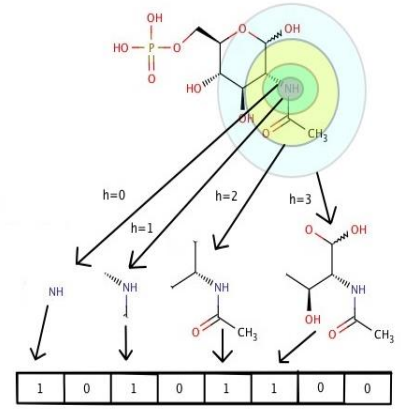


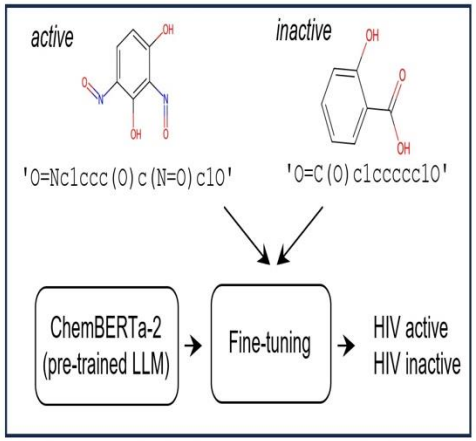


ChemBERTa (LLM) **Graph Neural Networks**

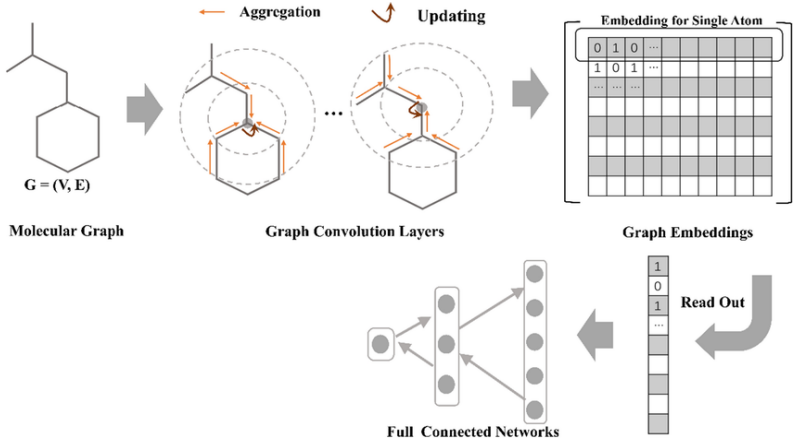
Molecular representations for ML

- Two general ways in making use of representation learning (MRL)
  - Use pretrained features as a plugin to your task-specific model (e.g. ChemBERTa)
  - End-to-end learning: integrate MRL into your task-specific model





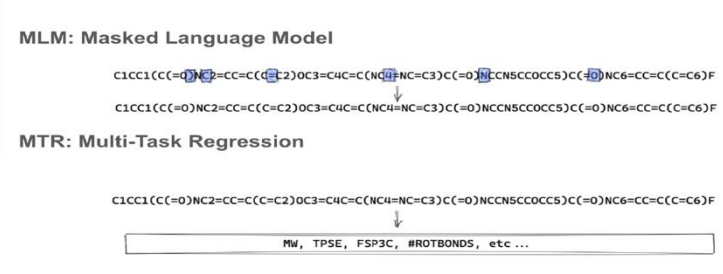
ChemBERTa (LLM)



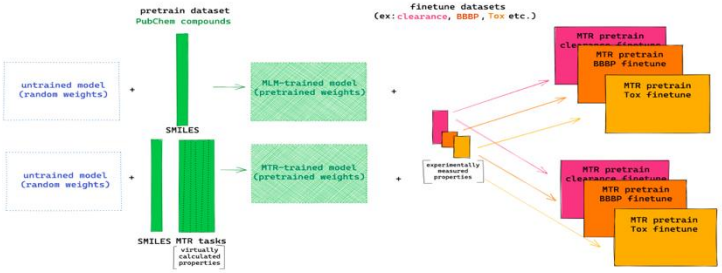
Graph Neural Networks

Molecular representations for ML

- **ChemBERTA-2 is a transformer model based on the BERT architecture**
- **It uses SMILES strings as the input representation of molecules**
- **Uses masked training and multi-task regression**
- **Trained with 77M molecular structures from PubChem**
- **Theoretically limited by the many-to-one mapping from strings to graphs – but big training data may diminish this risk**



(a) MLM vs. MTR

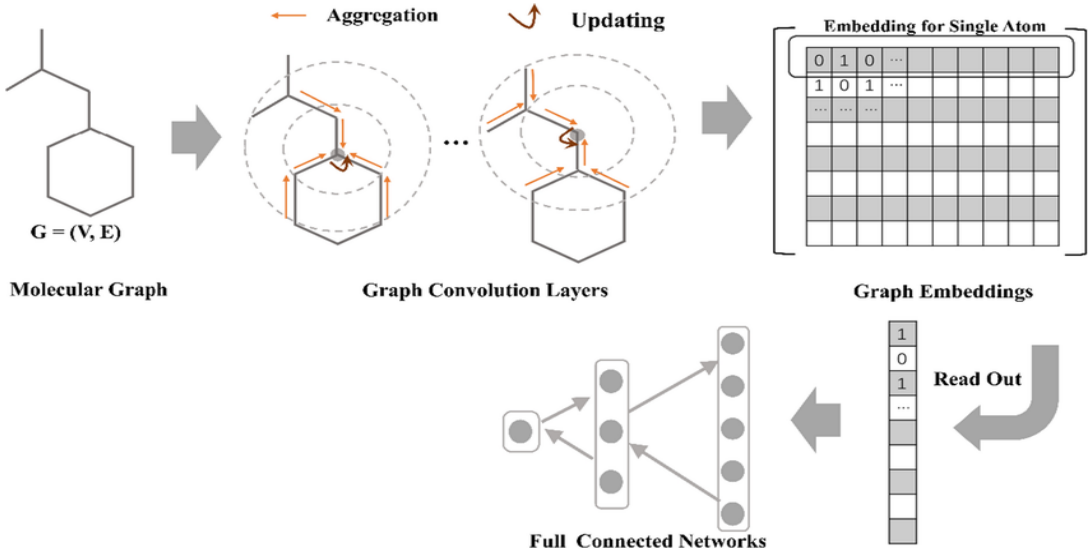


(b) Training pipeline

Ahmad, W., Simon, E., Chithrananda, S., Grand, G. and Ramsundar, B., 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Molecular representations for ML

- **Graph Neural Networks (GNN– embeddings learned from molecular graphs)**
- **Theoretically superior to sequence-based representations, also practical evidence of the same**
- **Wide variety GNNs exist for molecular applications**

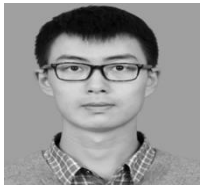


Wang, Y., Li, Z. and Barati Farimani, A., 2023. Graph neural networks for molecules. In *Machine Learning in Molecular Sciences* (pp. 21-66). Cham: Springer International Publishing.

Chemical Reaction Enhanced Graph Learning for Molecule Representation

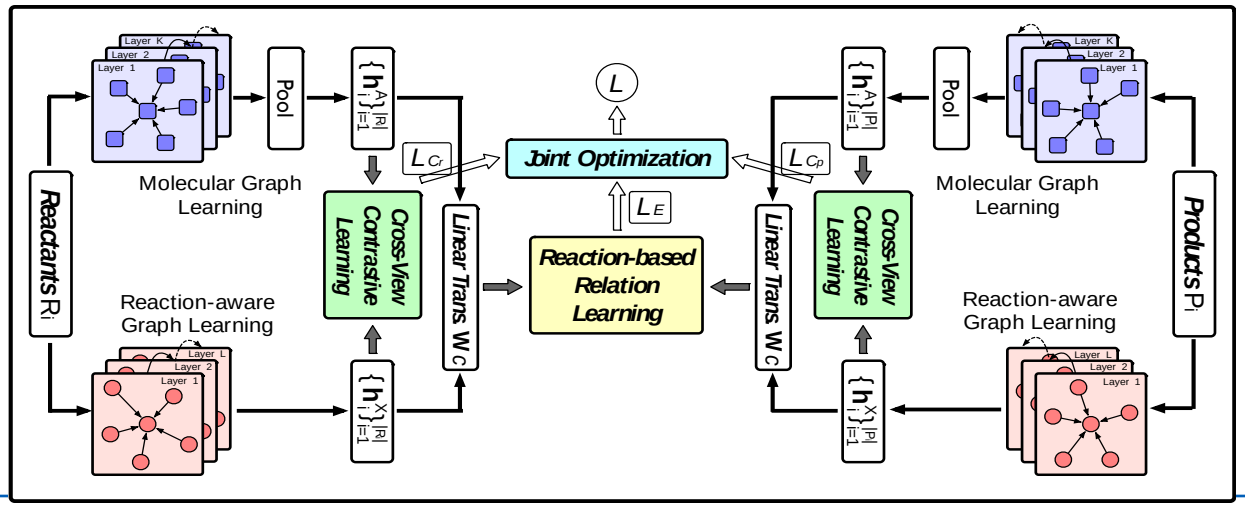
Aalto University

Reaction-aware molecular representation learning by RXGL



Anchen Li

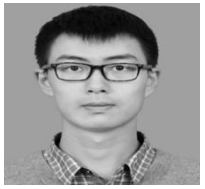
- **Motivation: molecular representation learning is typically focussing on the molecular structure**
- **We wish to leverage additional data sourse, in particular the context given by the chemical reactions the molecules are known to participate**



Aalto University

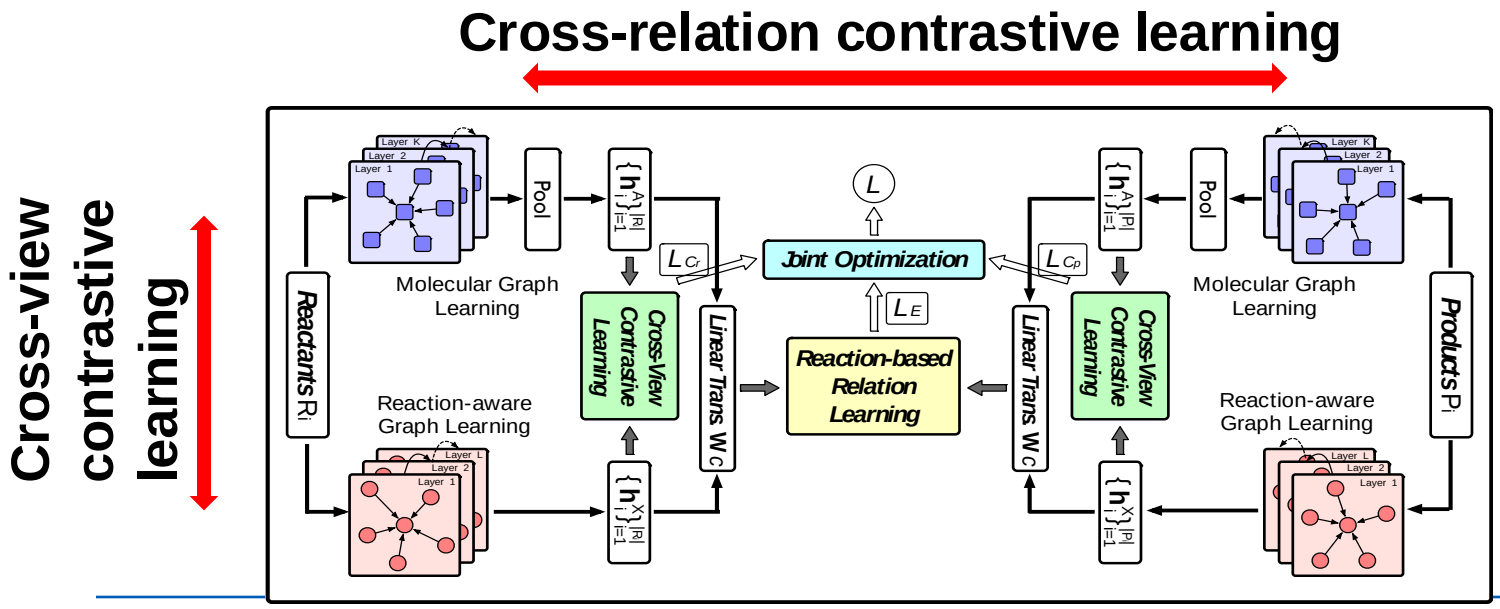
Anchen Li, Elena Casiraghi, Juho Rousu. Chemical Reaction Enhanced Graph Learning for Molecule Representation. Bioinformatics, 2024, to appear

Reaction-aware molecular representation learning by RXGL



Anchen Li

- **Cross-view contrastive learning: Atom-level and Network-level representations of molecules should be in agreement**
- **Cross-relation: representations of reactants should be similar to there presentations of the products**

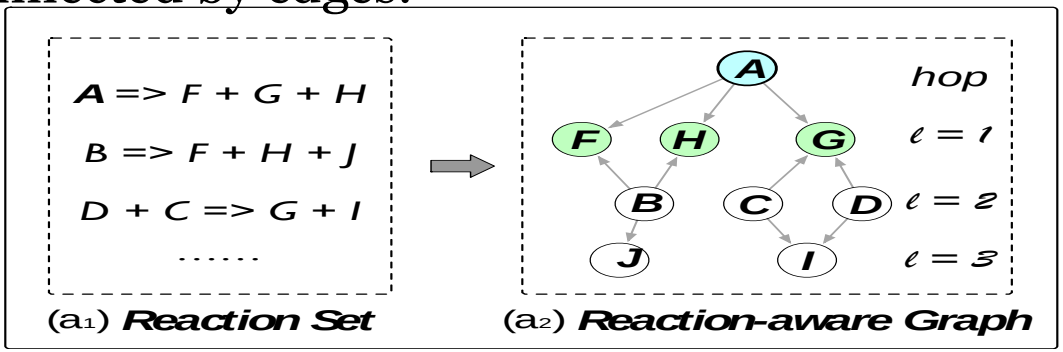


Anchen Li, Elena Casiraghi, Juho Rousu. Chemical Reaction Enhanced Graph Learning for Molecule Representation. Bioinformatics, 2024, to appear

I. Cross-view contrastive learning

Cross-view contrastive learning relies on two representations (views) of molecules

- **Atom-level representation as a molecular graph**
  - similar molecular structures give similar representations
- **Reaction-aware representation computed from the reaction network composed of a large set of reactions (below)**
  - molecules are nodes and reactants and products of a reaction are connected by edges.



Anchen Li, Elena Casiraghi, Juho Rousu. Chemical Reaction Enhanced Graph Learning for Molecule Representation. Bioinformatics, 2024, to appear

I. Cross-view contrastive learning

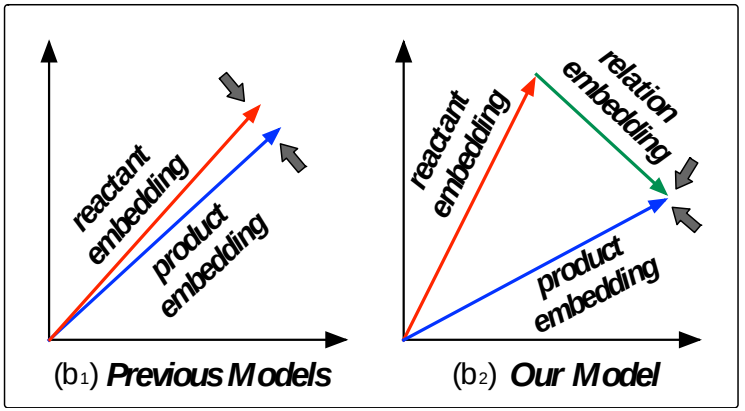
- **Learning accross two levels of representation is achieved by contrastive learning**
- **Principle: latent representation learned from atom-level ($x^A_{R_i}$) and network level representations ($x^X_{R_i}$) of the same reactant (R_i) should be more similar than the representations of a pair of two different molecules (R_i and R_j)**
- **This is encoded in the loss function below**

$$\mathcal{L}_{C_r} = \sum_{x_i \in \mathcal{X}_B} -\log \frac{\exp\left(c(\mathbf{x}^A_{R_i}, \mathbf{x}^X_{R_i})/\tau\right)}{\sum_{x_j \in \mathcal{X}_B} \exp\left(c(\mathbf{x}^A_{R_i}, \mathbf{x}^X_{R_j})/\tau\right)},$$

Anchen Li, Elena Casiraghi, Juho Rousu. Chemical Reaction Enhanced Graph Learning for Molecule Representation. Bioinformatics, 2024, to appear

II. Cross-reaction contrastive learning

- **The representation of the reactant set of a reaction should be *similar, but not the same* as the representation of the product set**
- **The term $e_{R_i \to P_i}$ represents the transformation made by the reaction**
- **We achieve this by another contrastive learning task where true reactant product pairs (R_i,P_j) are to be similar and arbitrary pairs (R_i,P_j), i≠j are to be dissimilar**



$$\mathrm{S}(R_i, P_i) = \|\mathbf{x}_{R_i} + \mathbf{e}_{R_i \to P_i} - \mathbf{x}_{P_i}\|_2.$$

$$\mathcal{L}_E = \frac{1}{|\mathcal{X}_\mathcal{B}|} \sum_{x_i \in \mathcal{X}_\mathcal{B}} \mathrm{S}(R_i, P_i) + \frac{1}{|\mathcal{X}_\mathcal{B}|^2 - |\mathcal{X}_\mathcal{B}|} \sum_{x_i \in \mathcal{X}_\mathcal{B}} \sum_{x_j \in \mathcal{X}_\mathcal{B}} \max\Big(\gamma - \mathrm{S}(R_i, P_j), 0\Big),$$

Anchen Li, Elena Casiraghi, Juho Rousu. Chemical Reaction Enhanced Graph Learning for Molecule Representation. Bioinformatics, 2024, to appear

Results: product prediction

- **Prediction task: Choose a product from a set of candidate products for a reaction**
- **TXGL is competitive against state-of-the-art MolR**

Datasets: USPTO chemical reactions benchmarks

- USPTO-15K
- USPTO-50K

Evaluation metrics:

- MRR – mean reciprocal rank
- Hit@1 – how often the correct product has the highest predicted score

Table 1. Results of the product prediction task on the two datasets. The numbers in brackets are the standard deviations.

Methods	USPTO-15K		USPTO-50K	
	MRR	Hit@1	MRR	Hit@1
Mol2vec	0.519	0.468	0.835	0.801
MolBERT	0.790	0.734	0.913	0.874
MolR-GCN	$0.883_{(0.005)}$	$0.847_{(0.007)}$	$0.958_{(0.003)}$	$0.944_{(0.006)}$
MolR-GAT	$0.881_{(0.003)}$	$0.846_{(0.002)}$	$0.952_{(0.002)}$	$0.931_{(0.004)}$
MolR-SAGE	$0.932_{(0.006)}$	$0.905_{(0.003)}$	$0.972_{(0.005)}$	$0.960_{(0.005)}$
MolR-TAG	$0.925_{(0.005)}$	$0.898_{(0.004)}$	$0.974_{(0.010)}$	$0.965_{(0.009)}$
RXGL-GCN	$0.927_{(0.003)}$	$0.899_{(0.005)}$	$0.965_{(0.007)}$	$0.954_{(0.002)}$
RXGL-GAT	$0.925_{(0.007)}$	$0.894_{(0.006)}$	$0.967_{(0.009)}$	$0.958_{(0.011)}$
RXGL-SAGE	$\mathbf{0.956}_{(0.004)}$	$\mathbf{0.936}_{(0.007)}$	$\mathbf{0.982}_{(0.005)}$	$0.973_{(0.003)}$
RXGL-TAG	$0.941_{(0.006)}$	$0.919_{(0.009)}$	$0.979_{(0.004)}$	$\mathbf{0.974}_{(0.009)}$

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}.$$

Anchen Li, Elena Casiraghi, Juho Rousu. Chemical Reaction Enhanced Graph Learning for Molecule Representation. Bioinformatics, 2024, to appear

Results: molecular property prediction

- **Prediction task: classify molecules based on their molecular properties**
- **RXGL compares favourably against existing deep learning methods**

- **Benchmark datasets from MoleculeNet:**
  - **BBBP. The Blood–brain barrier penetration (BBBP): >2000 compounds classified by permeability**
  - **BACE: binding of molecules to human beta-secretase 1**
  - **Tox21: 8014 compounds on 12 different toxicological target classes**
  - **ClinTox: FDA approved drugs vs. failed in clinical trials, ca 1500 compounds**

Table 4. Property prediction results (split type: *scaffold split*). The numbers in brackets are the standard deviations. The results of symbols ★ and ♠ are taken from GraphMVP and ReaKE.

Methods	BBBP	BACE	Tox21	ClinTox
EdgePred★	$0.645_{(0.031)}$	$0.646_{(0.047)}$	$0.745_{(0.004)}$	$0.558_{(0.062)}$
AttrMask★	$0.702_{(0.005)}$	$0.772_{(0.014)}$	$0.742_{(0.008)}$	$0.686_{(0.096)}$
GPT-GNN★	$0.645_{(0.011)}$	$0.776_{(0.005)}$	$0.753_{(0.005)}$	$0.578_{(0.031)}$
InfoGraph★	$0.692_{(0.008)}$	$0.739_{(0.025)}$	$0.730_{(0.007)}$	$0.751_{(0.050)}$
ContextPred★	$0.712_{(0.009)}$	$0.786_{(0.014)}$	$0.733_{(0.005)}$	$0.737_{(0.040)}$
G-Contextual★	$0.703_{(0.016)}$	$0.792_{(0.003)}$	$0.752_{(0.003)}$	$0.599_{(0.082)}$
G-Motif★	$0.664_{(0.034)}$	$0.734_{(0.040)}$	$0.732_{(0.008)}$	$0.778_{(0.020)}$
JOAO★	$0.660_{(0.006)}$	$0.729_{(0.020)}$	$0.744_{(0.007)}$	$0.663_{(0.039)}$
GraphMVP★	$0.724_{(0.016)}$	$0.812_{(0.009)}$	$\mathbf{0.744}_{(0.002)}$	$0.775_{(0.042)}$
MolR♠	–	0.774	0.670	0.830
ReaKE♠	–	0.781	0.713	0.862
RXGL	$\mathbf{0.729}_{(0.015)}$	$\mathbf{0.825}_{(0.006)}$	$0.736_{(0.003)}$	$\mathbf{0.912}_{(0.011)}$

Anchen Li, Elena Casiraghi, Juho Rousu. Chemical Reaction Enhanced Graph Learning for Molecule Representation. Bioinformatics, 2024, to appear

Predicting Atom-Atom mappings in Chemical Reactions

Aalto University

Predicting Atom-Atom mappings in Chemical Reactions

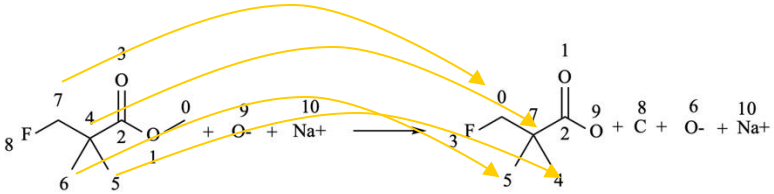
Maryam Astero

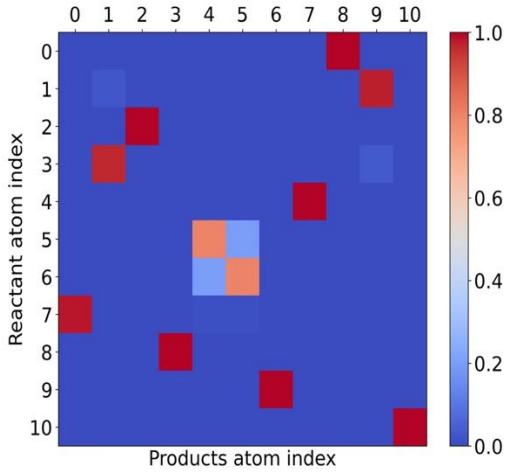
Task: predict the correspondence of reactant and product atoms in chemical reactions

Data: 15000 chemical reactions from USPTO

Input: the reactants and products of a reaction

Output: a mapping matrix predicting for each reaction the corresponding pairs of atoms





Aalto University

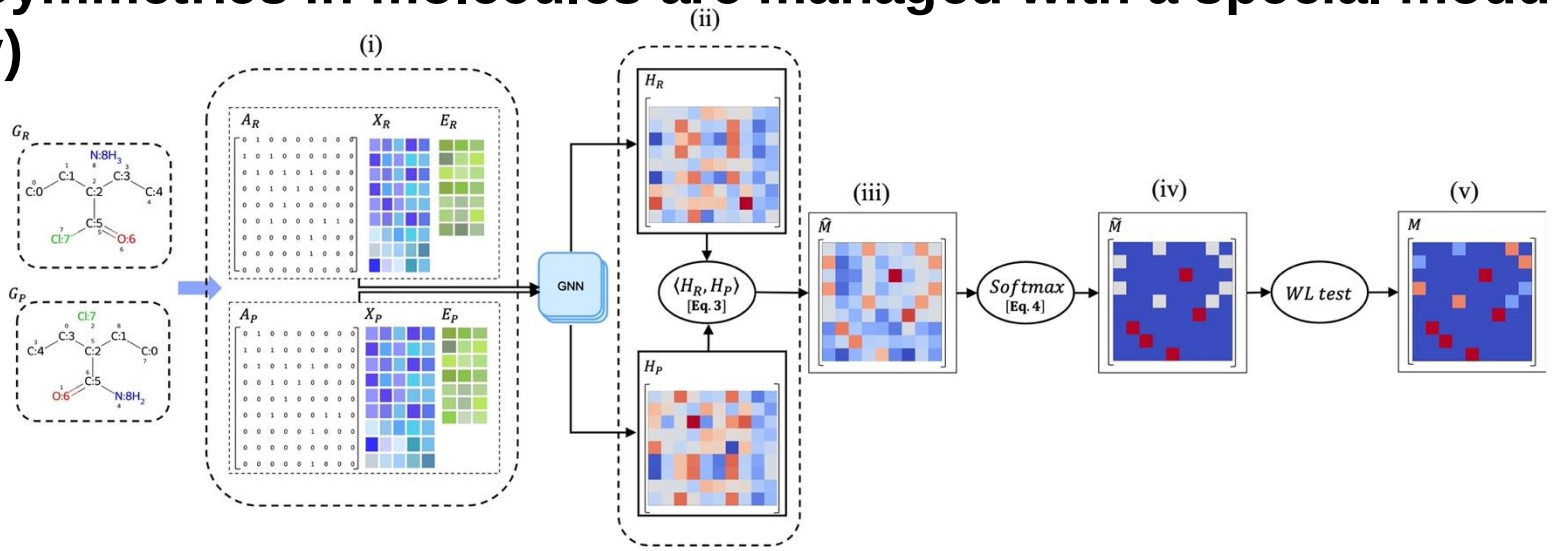
Astero, M. and Rousu, J., 2024. Learning symmetry-aware atom mapping in chemical reactions through deep graph matching. *Journal of Cheminformatics*, *16*(1), p.46.

Atom-mapping Network (AMNet)

AMNet is composed of two graph neural networks (resp. for reactants and products) (i)

The outputs of the two networks are trained so that they give high similarity to the corresponding atoms and low similarity for other pairs (ii-iii)

Symmetries in molecules are managed with a special module (iv-v)



Astero, M. and Rousu, J., 2024. Learning symmetry-aware atom mapping in chemical reactions through deep graph matching. *Journal of Cheminformatics*, *16*(1), p.46.

Symmetry-aware training

- **Symmetries in molecules is a potential difficulty for predictive models**
- **AMNet recognizes topologically equivalent atoms and uses this information to train more accurate models**
- **Significant improvements are obtained in overall accuracy (% correctly mapped atoms) and %top@k ranking (how often ground truth at given rank of better)**

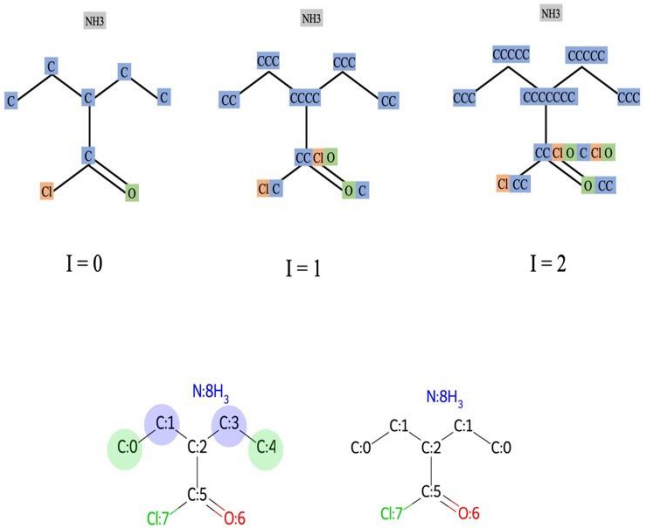


Table 3 Performance of the AMNet with and without molecule symmetry identification

Symmetry	Avg. Acc. (%) ± std	%Top@1 (%) ± std	%Top@3 (%) ± std	%Top@5 (%) ± std	%Top@10 (%) ± std
Yes	**97.3** ± 0.1	**66.2** ± 0.1	**96.6** ± 0.0	**99.3** ± 0.0	**99.7** ± 0.0
No	83.7 ± 0.2	43.8 ± 0.2	79.9 ± 0.1	96.2 ± 0.0	98.7 ± 0.0

The highest average accuracy and Top@k are highlighted in bold font

Astero, M. and Rousu, J., 2024. Learning symmetry-aware atom mapping in chemical reactions through deep graph matching. *Journal of Cheminformatics*, *16*(1), p.46.

Summary

Summary

- Public molecular datasets have grown in recent years to a size that enables deep representation learning on them
- Molecular representation learning uses these data to arrive at accurate predictive models
- RXGL is our recent MLR method that leverages large reaction datasets such as USPTO data
- AMNet is a method for predicting atom-atom mappings in chemical reactions based on graph matching