화학 데이터 인공지능을 위한 Multimodal Learning 라이브러리 및 ChemAI

나경석
한국화학연구원 (KRICT)

Contents

01 Heterogeneous Data Types on Chemical Data

Numerical Value

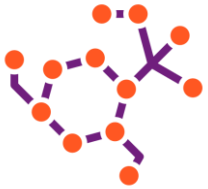
Numerical data represented by the scalar values and the feature vectors (engineering conditions, experimental values)

String Data

A sequence of characters to encode chemical structures and attributes of chemical compounds (chemical formula, SMILES)

Image Data

Image-like data to represent physical patterns and analysis results of chemical compounds (diffraction patterns)



Attributed Graph

Mathematical graphs with node and edge features to describe the atomic structures of molecules and crystalline systems

01 Heterogeneous Data Types on Chemical Data

Numerical Value

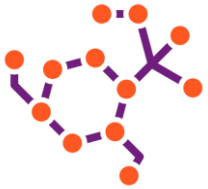
Numerical data represented by the scalar values and the feature vectors (engineering conditions, experimental values)

String Data

A sequence of characters to encode chemical structures and attributes of chemical compounds (chemical formula, SMILES)

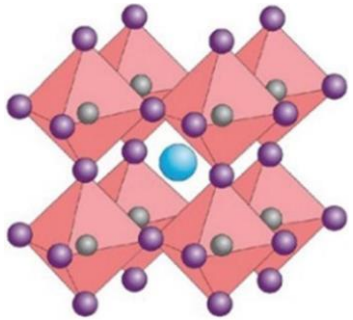
Image Data

Image-like data to represent physical patterns and analysis results of chemical compounds (diffraction patterns)



Attributed Graph

Mathematical graphs with node and edge features to describe the atomic structures of molecules and crystalline systems



Hybrid perovskite

01 Heterogeneous Data Types on Chemical Data

Numerical Value

Numerical data represented by the scalar values and the feature vectors (engineering conditions, experimental values)

String Data

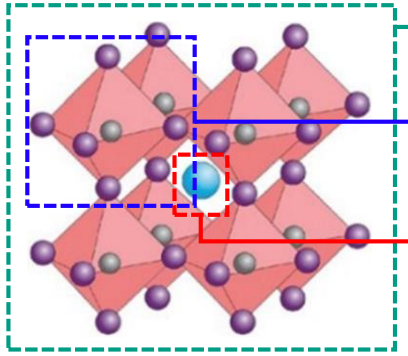
A sequence of characters to encode chemical structures and attributes of chemical compounds (chemical formula, SMILES)

Image Data

Image-like data to represent physical patterns and analysis results of chemical compounds (diffraction patterns)

Attributed Graph

Mathematical graphs with node and edge features to describe the atomic structures of molecules and crystalline systems



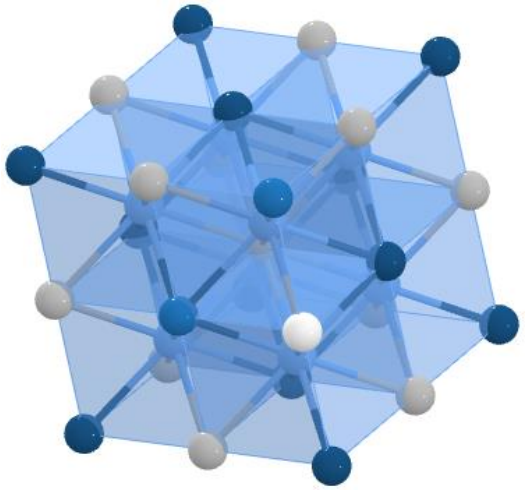
Hybrid perovskite

Band gap, Efficiency, Analysis results

Chemical formula ($CsPbI_3$), Crystal structure (CIF)

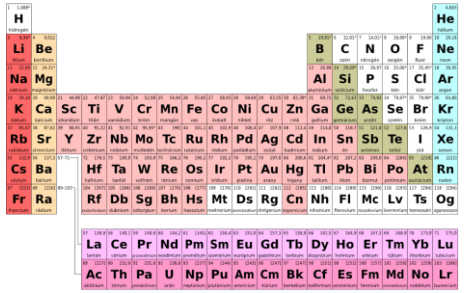
Chemical formula ($C_3H_5N_2$), Molecular structure (SMILES)

01 Heterogeneous Data Types on Chemical Data



Metadata

Chemical formula (Ac_2AgIr), Space group ($Fm\bar{3}m$)



Interpretations

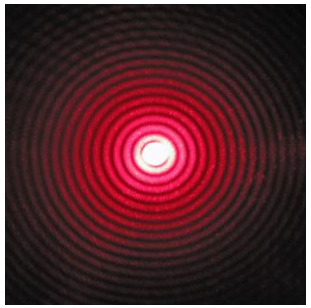
Refined information by physical and chemical domain knowledge, such as Cartesian coordinates

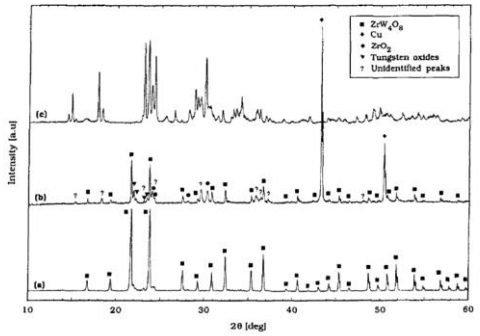
Synthesis Conditions

Experimental conditions in the synthesis process.

Analysis Data

Diffraction pattern image and spectrum, ···





Reaction conditions and results

Experimental conditions of the chemical systems based on atomic and substructure interactions

High-Level Information

Physical and chemical information about composite and heterogeneous structures, such as device structures and device properties

02 Multimodal Learning for Chemical Data

- Modality [Wikipedia]
  - The channel by which signs are transmitted (linguistics)
  - A path of communication between the human and the computer (computer science)

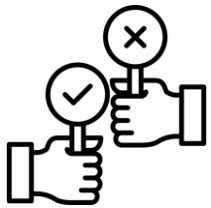
- Multimodal Learning
  - Different modality → Different statistics
  - Joint representation learning from the data sources of different modalities

Metadata (domain knowledge, experience, ···)

Vision Data (image, chart, gesture, ···)

Sound Data (audio, voice, ···)

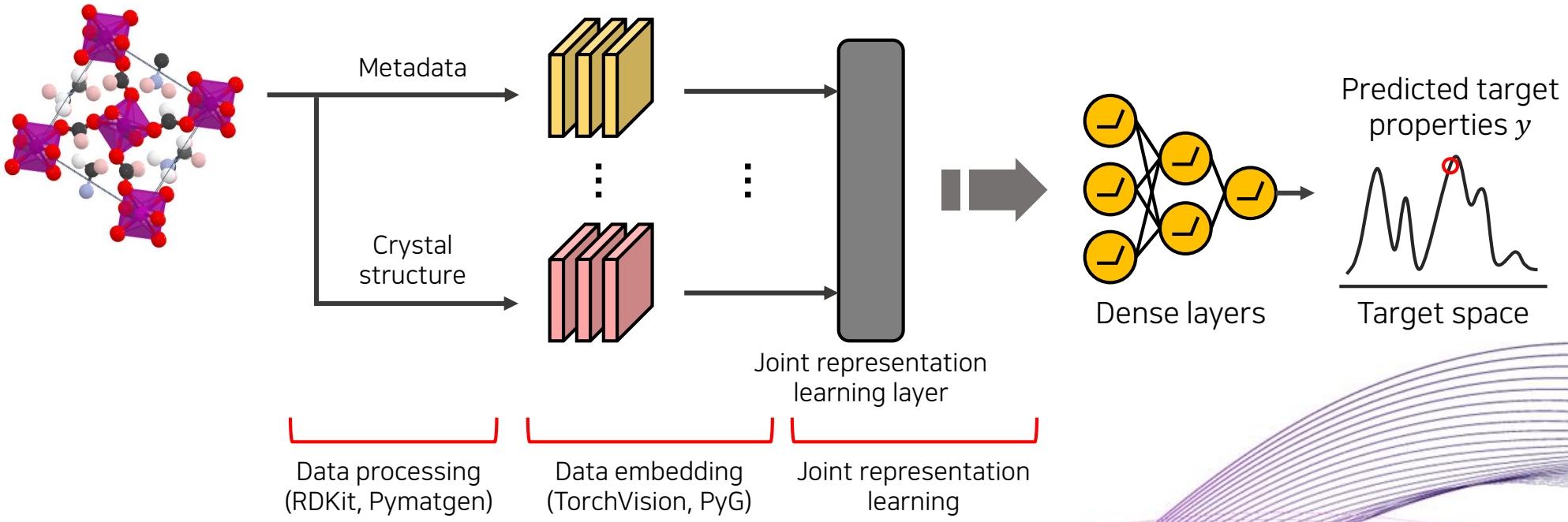
⋮



Decision

02 Multimodal Learning for Chemical Data

- **Different modality → Different prediction model**
  - **Vector-shaped data:** feedforward neural networks, gradient boosting tree
  - **Image data:** convolutional neural networks
  - **Sequential data:** recurrent neural networks, transformer models
  - **Graph data:** graph neural networks, graph kernel methods

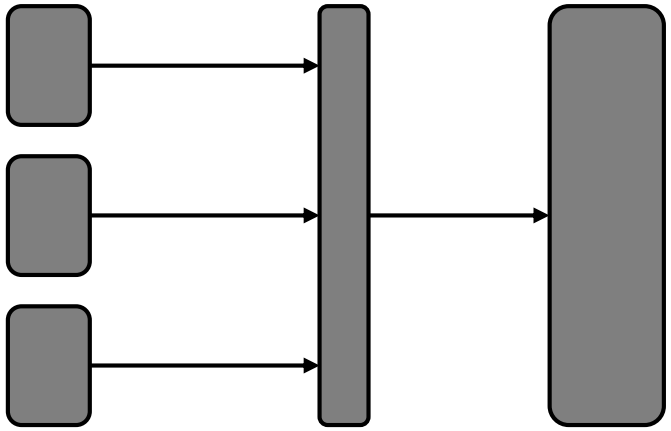


02 Multimodal Learning for Chemical Data

```
(environment) path/user> pip install ailca
```

Step 1: load dataset and construct a multimodal dataset

```
dataset_crystal = crystal.load_dataset("path_metadata", "path_structs", load_information)
dataset_image = image.load_dataset("path_metadata", "path_imgs", load_information)
dataset = MultimodalDataset(datasets=[dataset_crystal, dataset_image], load_information)
dataset_train, dataset_test = dataset.split(train_ratio)
data_loader = get_data_loader(dataset_train, batch_size, shuffle=True)
```



Step 2: define the embedding networks for each data type

```
emb_net_crystal = CGCNN(dataset_crystal, dim_out, readout_method)
emb_net_image = ResNet34(dim_out)
model = MultimodalNet(emb_nets=[emb_net_crystal, emb_net_image], dim_out, rep_method).cuda()
```

02 Multimodal Learning for Chemical Data

Step 3: optimize model parameters of the multimodal prediction model

```
optimizer = get_optimizer(model, gradient_method)
loss_func = get_loss_func(loss_name)

for epoch in range(0, epochs):
    train_loss = model.fit(data_loader, optimizer, loss_func)
```

Step 4: evaluate the trained model on test dataset

```
eval_results = MLResult(model, dataset_train, dataset_test)
eval_results.save("path_result_file")
model.save("path_model_file")
print(eval_results)
```

03 ChemAI: AI for Chemical Applications

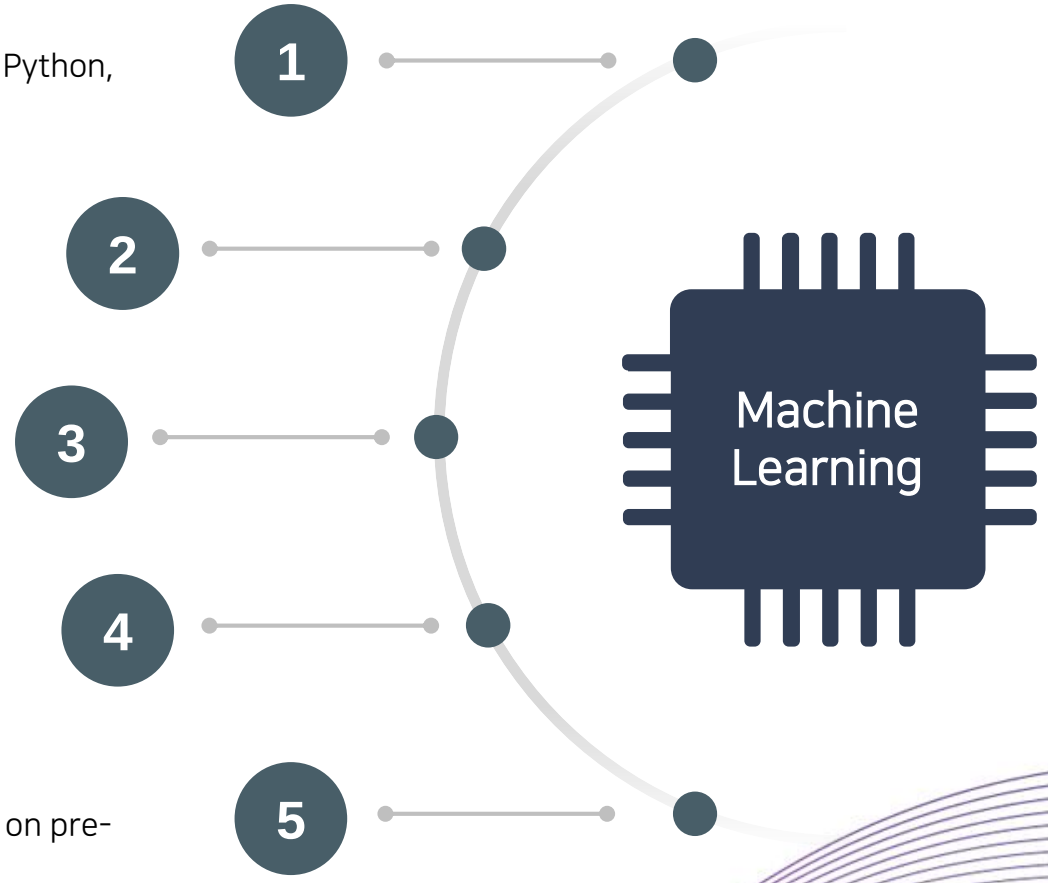
Installation
Install development environments, such as Python, GPU drive, external frameworks

Data Preprocessing
Convert raw chemical data of unstructured formats into the machine-readable formats

Model Configuration
Select appropriate machine algorithms and initial configurations according to the input data

Hyperparameter Optimization
Search optimal hyperparameters to generate an accurate prediction model

Model Evaluation
Evaluate the trained prediction models on pre-defined test datasets



03 ChemAI: AI for Chemical Applications

Installation | Data Preprocessing | Model Configuration | Hyperparameter Optimization | Model Evaluation

Web implementation

ChemAI is available on desktop, tablet, and smartphone without GPU machines and software installations

Multimodal Learning

ChemAI supports multimodal learning for heterogeneous data formats in chemical applications

Auto-Configuration

Appropriate machine learning algorithms are automatically trained according to data formats of the input data

Transfer Learning

We provide several pre-trained models to support transfer learning on user-defined problems and datasets

03 ChemAI: AI for Chemical Applications

Installation | Data Preprocessing | Model Configuration | Hyperparameter Optimization | Model Evaluation

Web implementation

ChemAI is available on desktop, tablet, and smartphone without GPU machines and software installations

Multimodal Learning

ChemAI supports multimodal learning for heterogeneous data formats in chemical applications

Auto-Configuration

Appropriate machine learning algorithms are automatically trained according to data formats of the input data

Transfer Learning

We provide several pre-trained models to support transfer learning on user-defined problems and datasets

03 ChemAI: AI for Chemical Applications

Installation | Data Preprocessing | Model Configuration | Hyperparameter Optimization | Model Evaluation

Web implementation

ChemAI is available on desktop, tablet, and smartphone without GPU machines and software installations

Multimodal Learning

ChemAI supports multimodal learning for heterogeneous data formats in chemical applications

Auto-Configuration

Appropriate machine learning algorithms are automatically trained according to data formats of the input data

Transfer Learning

We provide several pre-trained models to support transfer learning on user-defined problems and datasets

03 ChemAI: AI for Chemical Applications

Installation | Data Preprocessing | Model Configuration | Hyperparameter Optimization | Model Evaluation

Web implementation

ChemAI is available on desktop, tablet, and smartphone without GPU machines and software installations

Multimodal Learning

ChemAI supports multimodal learning for heterogeneous data formats in chemical applications

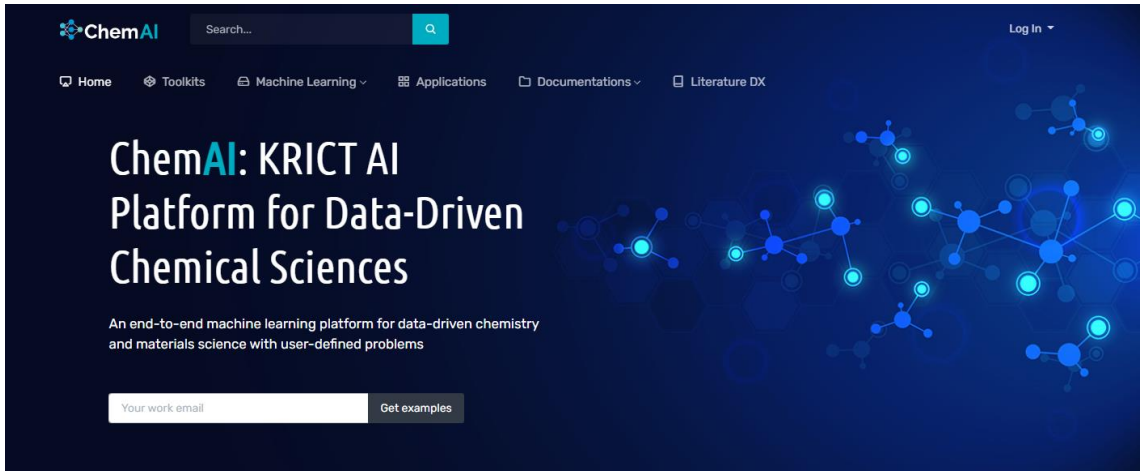
Auto-Configuration

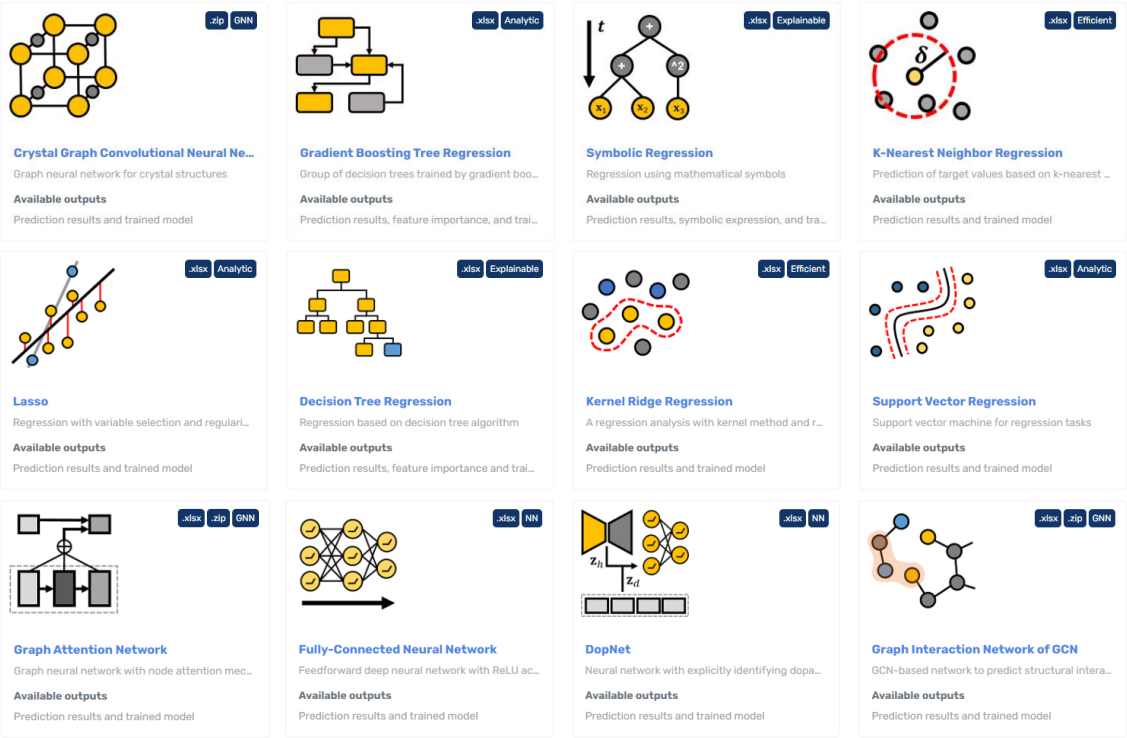
Appropriate machine learning algorithms are automatically trained according to data formats of the input data

Transfer Learning

We provide several pre-trained models to support transfer learning on user-defined problems and datasets

03 ChemAI: AI for Chemical Applications





Toolkits

Pre-trained prediction models and data-preprocessing algorithms for chemical applications

Machine Learning

Machine learning to generate the prediction models based on user-defined problems

Applications

Model store of the users to share their machine learning tasks and prediction models

03 ChemAI: AI for Chemical Applications

Toolkit	Input Type	Target Property (Unit)	R^2 Score	Source Files
Band Gap Prediction Prediction of expeirmental band gap [more]	Composition	Experimental band gap (eV)	0.909	Trained model Source dataset
Formation Energy Prediction Prediction of experimental formation energy [more]	Composition	Experimental formation energy (eV/atom)	0.907	Trained model Source dataset
Thermoelectricity Prediction Prediction of ZT (figure of merit) for a given temperature [more]	• Composition • Temperature (K)	ZT (Figure of merit)	0.867	Trained model
Band Gap Prediction of Perovskites Band gap prediction of organic-inorganic perovskites [more]	Crystal structure (.cif)	HSE band gap (eV)	0.901	Trained model Source dataset
Band Gap Correction Prediction of G_0W_0 band gap from naive GGA band gap [more]	• Crystal structure (.cif) • GGA band gap (eV)	G_0W_0 band gap (eV)	0.951	Trained model Source dataset Reference
Prediction of Absorption Max Predicting absorption max of organic chromophores [more]	• Chromophore structure (SMILES) • Solvent structure (SMILES)	Absorption max (nm)	0.902	Trained model Source dataset
Data Clustering Grouping data points based on density of data [more]	• Data points (.xlsx) • Distance threshold (epsilon) • Quantity threshold (MinPts)	None	Not available	Reference
Outlier Detection Detection of abnormal data based on data distribution [more]	• Data points (.xlsx) • Number of neighbors (NumNN)	None	Not available	Reference

04 Future Open Source Projects

Model or Method Name	Year	Input Data	Description
Extended GNN (EGNN)	2020	▪ Molecular structure ▪ Metadata of molecule	Joint representation learning of attributed graph and global information
Tuplewise Graph Neural Network (TGNN)	2020	▪ Crystal structure ▪ Low-level band gap	Correction model from low-cost band gaps to expensive but accurate band gaps
Unsupervised Subspace Extraction (USE)	2021	▪ Unlabelled dataset	Feature extraction to find optimal subspace
DopNet	2021	▪ Chemical formula	Prediction model for doped materials
Automated Nonlinearity Encoder (ANE)	2022	▪ Not specified	A generalized method to generate latent data representations for extrapolation
System-Identified Material Descriptor (SIMD)	2022	▪ Chemical formula ▪ Set of materials	A material descriptor based on latent representation of the material systems
Conditional Graph Information Bottleneck (CGIB)	2022	▪ Molecular structures	Graph neural network to predict the target properties from molecular interactions
Substructure Interaction Graph Network with Node Augmentation (SIGNNA)	2022	▪ Molecular structures ▪ Crystal structures	Graph neural network to predict the target properties of heterogeneous systems

04 Future Open Source Projects

Model or Method Name	Year	Input Data	Description
Extended GNN (EGNN)	2020	▪ Molecular structure ▪ Metadata of molecule	Joint representation learning of attributed graph and global information
Tuplewise Graph Neural Network (TGNN)	2020	▪ Crystal structure ▪ Low-level band gap	Correction model from low-cost band gaps to expensive but accurate band gaps
Unsupervised Subspace Extraction (USE)	2021	▪ Unlabelled dataset	Feature extraction to find optimal subspace
DopNet	2021	▪ Chemical formula	Prediction model for doped materials
Automated Nonlinearity Encoder (ANE)	2022	▪ Not specified	A generalized method to generate latent data representations for extrapolation
System-Identified Material Descriptor (SIMD)	2022	▪ Chemical formula ▪ Set of materials	A material descriptor based on latent representation of the material systems
Conditional Graph Information Bottleneck (CGIB)	2022	▪ Molecular structures	Graph neural network to predict the target properties from molecular interactions
Substructure Interaction Graph Network with Node Augmentation (SIGNNA)	2022	▪ Molecular structures ▪ Crystal structures	Graph neural network to predict the target properties of heterogeneous systems

04 Future Open Source Projects

Model or Method Name	Year	Input Data	Description
Extended GNN (EGNN)	2020	▪ Molecular structure ▪ Metadata of molecule	Joint representation learning of attributed graph and global information
Tuplewise Graph Neural Network (TGNN)	2020	▪ Crystal structure ▪ Low-level band gap	Correction model from low-cost band gaps to expensive but accurate band gaps
Unsupervised Subspace Extraction (USE)	2021	▪ Unlabelled dataset	Feature extraction to find optimal subspace
DopNet	2021	▪ Chemical formula	Prediction model for doped materials
Automated Nonlinearity Encoder (ANE)	2022	▪ Not specified	A generalized method to generate latent data representations for extrapolation
System-Identified Material Descriptor (SIMD)	2022	▪ Chemical formula ▪ Set of materials	A material descriptor based on latent representation of the material systems
Conditional Graph Information Bottleneck (CGIB)	2022	▪ Molecular structures	Graph neural network to predict the target properties from molecular interactions
Substructure Interaction Graph Network with Node Augmentation (SIGNNA)	2022	▪ Molecular structures ▪ Crystal structures	Graph neural network to predict the target properties of heterogeneous systems

04 Future Open Source Projects

Model or Method Name	Year	Input Data	Description
Extended GNN (EGNN)	2020	▪ Molecular structure ▪ Metadata of molecule	Joint representation learning of attributed graph and global information
Tuplewise Graph Neural Network (TGNN)	2020	▪ Crystal structure ▪ Low-level band gap	Correction model from low-cost band gaps to expensive but accurate band gaps
Unsupervised Subspace Extraction (USE)	2021	▪ Unlabelled dataset	Feature extraction to find optimal subspace
DopNet	2021	▪ Chemical formula	Prediction model for doped materials
Automated Nonlinearity Encoder (ANE)	2022	▪ Not specified	A generalized method to generate latent data representations for extrapolation
System-Identified Material Descriptor (SIMD)	2022	▪ Chemical formula ▪ Set of materials	A material descriptor based on latent representation of the material systems
Conditional Graph Information Bottleneck (CGIB)	2022	▪ Molecular structures	Graph neural network to predict the target properties from molecular interactions
Substructure Interaction Graph Network with Node Augmentation (SIGNNA)	2022	▪ Molecular structures ▪ Crystal structures	Graph neural network to predict the target properties of heterogeneous systems

감사합니다

화학 데이터 인공지능을 위한 Multimodal Learning 라이브러리 및 ChemAI