DACON
DATA TO VALUE

Dacon Competition
2022 Samsung AI Challenge

Algorithm of team [cgu]

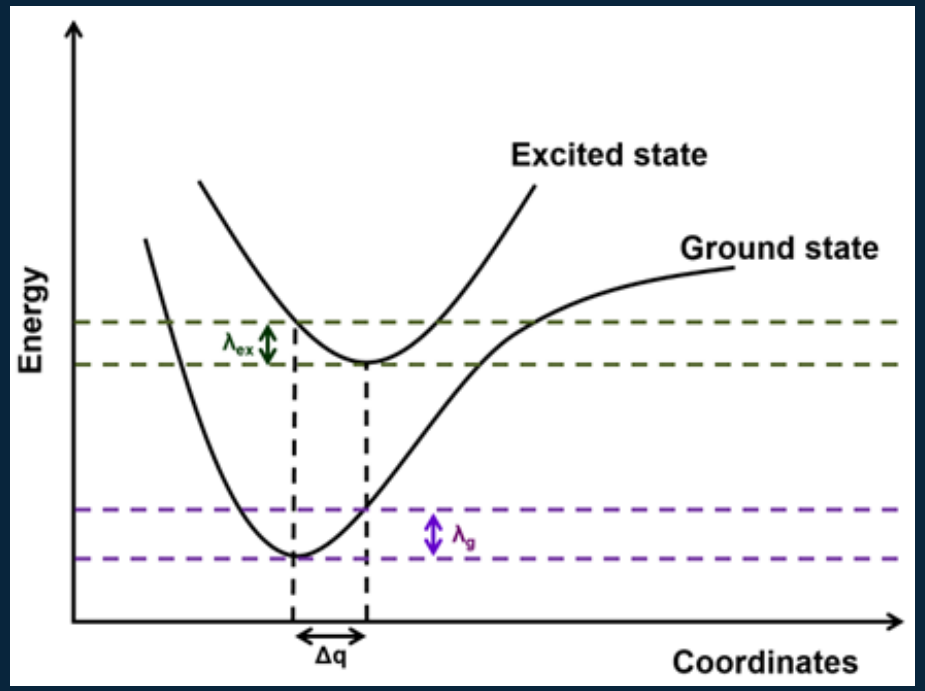
SAMSUNG

SAMSUNG ADVANCED INSTITUTE OF TECHNOLOGY

Seoul National University
Bio & Health Informatics Lab (Prof. Sun Kim)
JeongHyeon Gu, Danyeong Lee, Sangyeup Kim

Problem definition

"Development of an AI algorithm that predicts **reorganization energy** from 3d molecular structure"



$\lambda_g = E_g^* - E_g$, $\lambda_{ex} = E_{ex}^* - E_{ex}$

- E_g : Ground state energy in the ground state geometry
- E_g^* : Ground state energy in the excited state geometry
- E_g : Excited state energy in the excited state geometry
- E_{ex}^* : Excited state energy in the ground state geometry

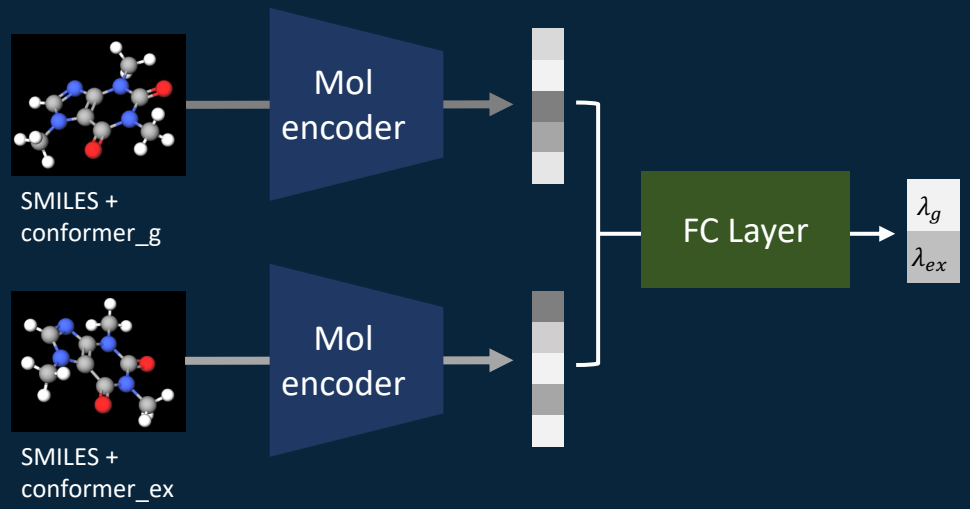
Approaches

Data
- Input: SMILES, conformer_g, conformer_ex
- Output: Reorg_g, Reorg_ex

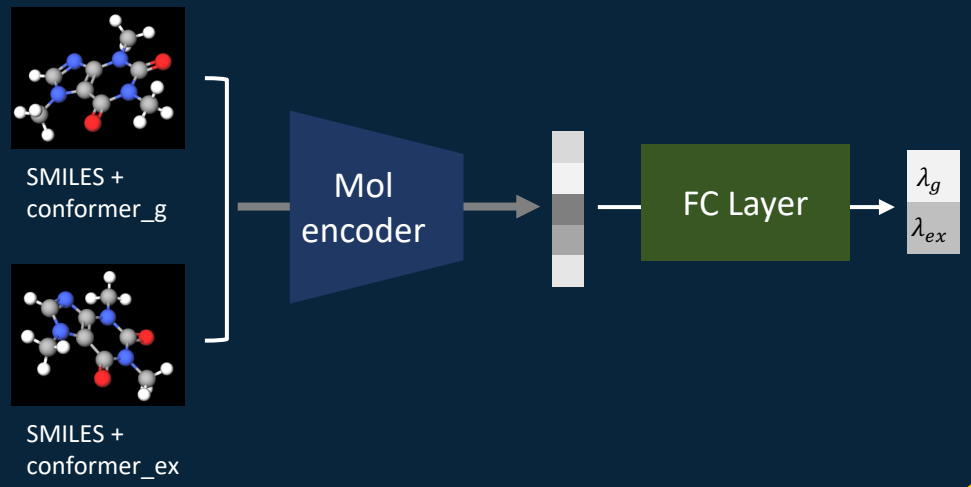
Molecule encoder candidates
- GNN models
- Transformer models

Framework: Molecule encoder + FC Layer

Framework Candidate 1

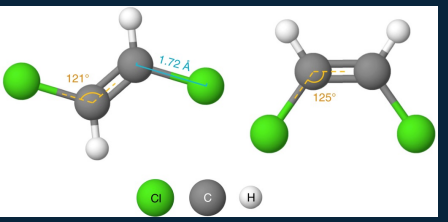


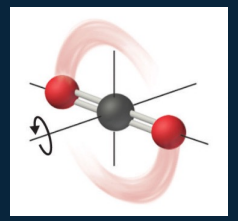
Framework Candidate 2

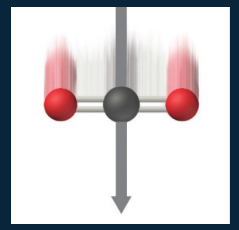


: chosen one

How to utilize 3D coordinate information?







	Rotation	Translation	Articles
Raw coordinates (x, y, z)	variant	variant	
Bond lengths & Bond-bond angles	invariant	invariant	DimeNet (ICLR, ‘20) HMGNN (ICDM, ‘20) **GeoGNN (Nature Machine Intelligence, ‘22)**
Atom-atom distance	invariant	invariant	PhysChem (NeurIPS, ‘21) GEM-2 (arxiv, ‘22)

Candidate baselines

GNN-based

- GIN-virtual (ICLR, 2019)
- **GeoGNN (Nature Machine Intelligence, 2022)**

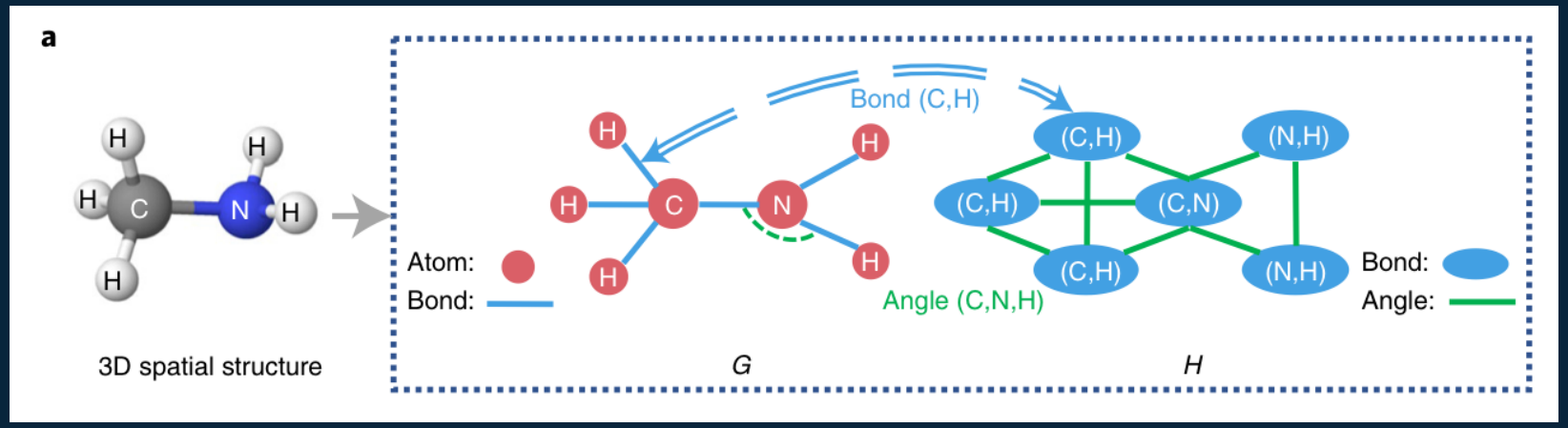
Transformer-based

- EGT (KDD, 2022)
- GRPE (ICLR, 2022)
- GEM-2 (arXiv, 2022)

Pre-training datasets

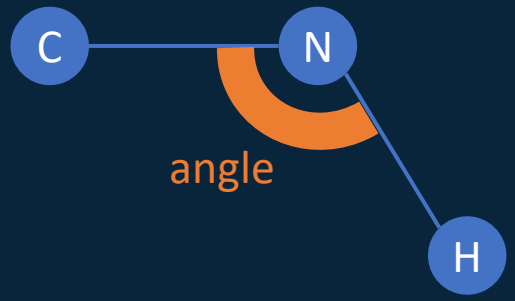
- GEOM
- PCQM4Mv2

Baseline: GeoGNN



- Atom-Bond graph G

- Bond-Angle graph H



- Bond lengths
  RBF-embedded & added to bond features
- Bond-bond angles
  RBF-embedded & used as angle features

- Iteratively update
  1. Bond features on H
  2. Atom features on G

- Used GIN (Xu et al., 2019)

cf) Radial basis function (RBF)

$$e_m(x) = \exp(-\gamma \parallel x - \mu_m \parallel^2)$$

Used to expand each continuous value x **into a vector**

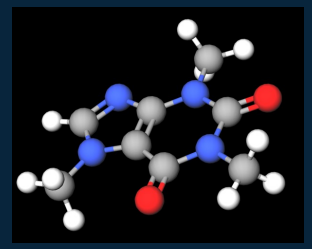
- x: input continuous value
- μ_m: centers ranging from minimum value to maximum value of corresponding features with certain stride (0.01 in GeoGNN)
- γ: controls the shape of radial kernel

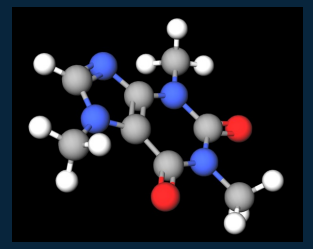
Fang, X., Liu, L., Lei, J. et al. Geometry-enhanced molecular representation learning for property prediction. Nat Mach Intell ***4****, 127–134 (2022).*

Our Model

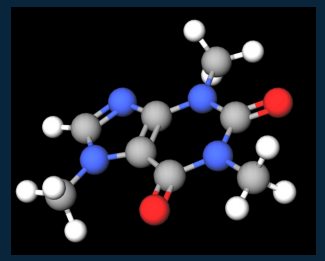
- In our challenge dataset, we have **two different conformers** representing two states (ground, excited)

- **Bond lengths**
  - Bond lengths of 2 states are RBF-embedded & concatenated
  - Added to bond features
- **Bond-bond angles**
  - Bond-bond angles of 2 states are RBF-embedded & concatenated
  - Used as angle features
- **+ Atom position gaps between two states**
  - Euclidean distances between atoms of two states are RBF-embedded
  - Concatenated to atom features

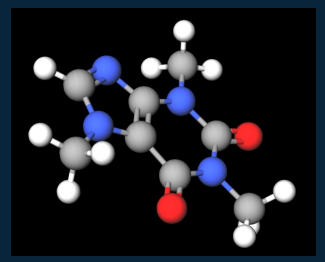




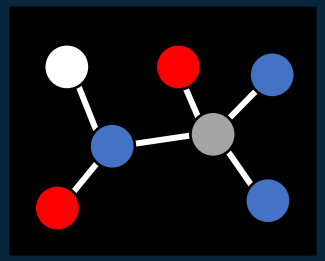
Our Model



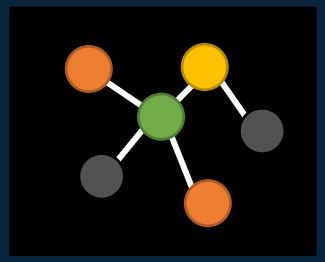
SMILES +
conformer_g



SMILES +
conformer_ex



atom-bond graph



bond-angle graph

GeoGNN

GNN :
- Layer type: GINEConv
- # of layers: 5
- Pooling: Add pooling

Molecular
Representation

Fully Connected Layers

λ_g

λ_{ex}

Encode bond lengths & bond angles & atom position gaps

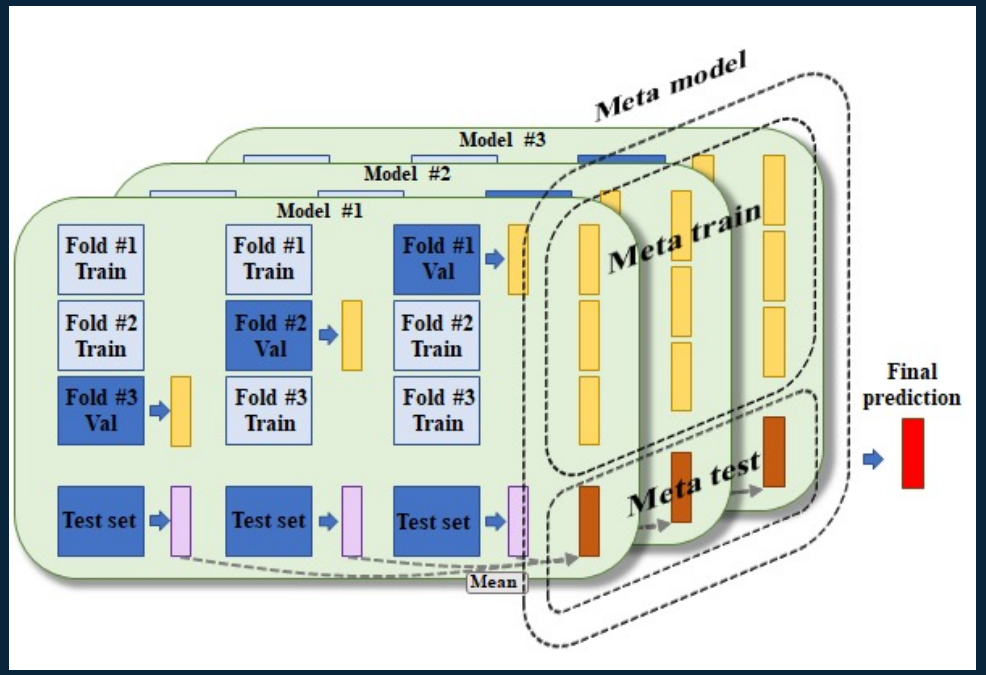
Training & Evaluation

- Implementation
  - PyTorch & PyTorch Lightning
  - Configuration with Hydra

- Training
  - AdamW & Cosine Annealing LR scheduler
  - Gradient Clipping

- Evaluation (10-fold CV)
  - Metrics: RMSE, Pearson Corr., R2 score
  - Tracked with weight & biases
  - Hyperparameter tuning with Optuna

- 10-fold CV stacking ensemble
  - 4 different hyperparameters set / 3 different seeds (0, 1, 2)
    ⇒ 10-fold CV stacking of 12 different models
    ⇒ XGBoost 7 fold CV ensemble.



Stacking Ensemble

PyTorch · PyTorch Lightning · Hydra · W&B · Optuna · XGBoost

Discussion

- Integration of features of 2 states
  - Separately encoding each conformer did not work well. (Framework candidate 1)
  - Using concatenated features of two states showed better performance. (Framework candidate 2)

- CV stacking ensemble
  - Predictions of several models are used as new features to train new model.
  - Two labels (λ_g, λ_{ex}) are moderately correlated -> More informative new features

- Future works
  - Self-supervised pre-training with large-scale datasets
    - GEOM, PCQM4Mv2, ...
  - Transformer models
    - Graphormer, EGT, GRPE..

Thank you