# Molecular Pre-trained Models

Wu Hualong

HARBIN INSTITUTE OF TECHNOLOGY,SHEN ZHEN

# Start

**Problem:** 3D geometric information is scarce in downstream tasks due to its expensive cost to obtain.
**Goal:** Propose a Molecular Pre-trained model exploiting both 2D and 3D information of molecules.
**Requirement:**

- The model should be pre-trained with the knowledge of 3D geometry and then fine-tuned on downstream tasks with or without 3D information.
- Consider the symmetry of molecules.

# Dataset

Datasets:

- Pre-training
  - GEOM, a dataset with over 37 million molecular conformations annotated by energy and statistical weight for over 450,000 molecules.
  - PCQM4Mv2, including 3.37 million molecules with both 2D graphs and 3D geometric structures.(3D available only for training molecules)
- Downstream
  - MoleculeNet:one of the most widely used benchmarks for 2D molecular property prediction
    - 2D, 8 classification tasks
  - PCQM4Mv2 humo-lumo gap (2D)
  - QM9 quantum properties
    - 3D, 12 regression tasks

# Pre-traing Strategies

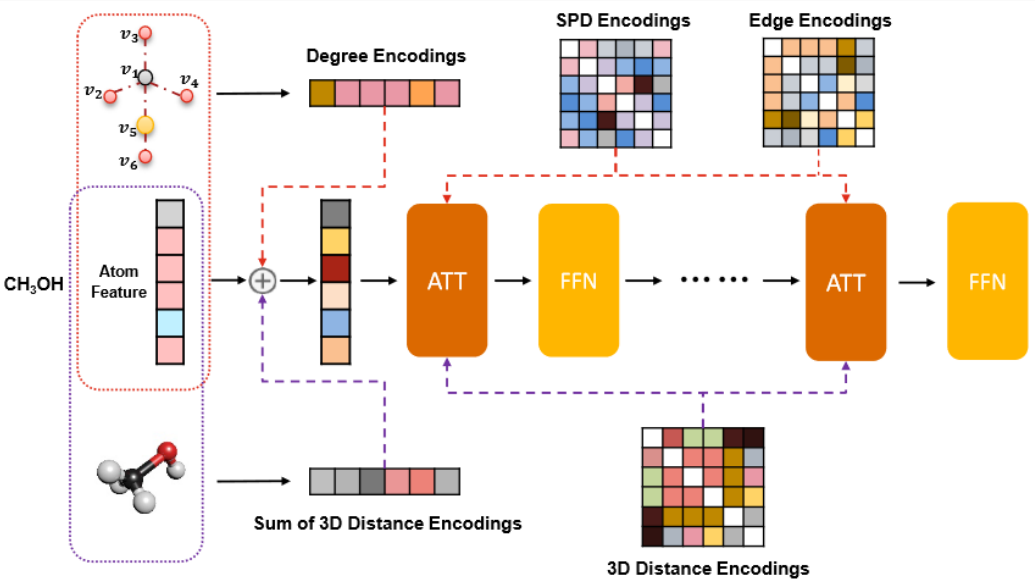
- AutoEncoding (AE)
- Masked Component Modeling
- Contrastive Learning (CL)
  - Same-Scale Contrast (SSC)
- DeNoising (DN)

# Baseline

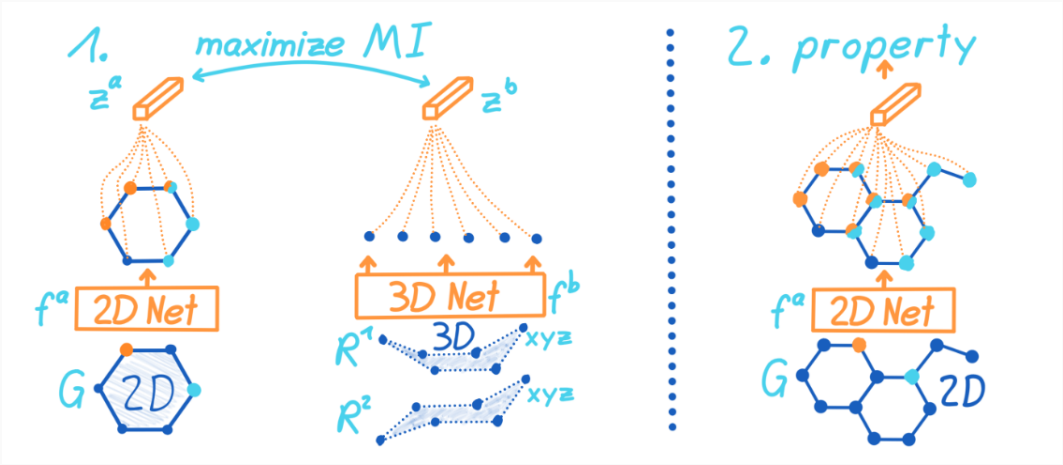
| Model | Input | Backbone architecture | Pre-training task | Pre-training database |
|---|---|---|---|---|
| GraphCL | Graph | 5-layer GIN | SSC | ZINC15 (2M) |
| Graphormer | Graph | Graphormer | Supervised | PCQM4M-LSC (3.8M) |
| Mole-BERT | Graph | 5-layer GIN | MCM + SSC | GEOM (100K) |
| GraphMVP | Graph+Geometry | 5-layer GIN + SchNet | SSC + AE | GEOM (50K) |
| 3D Informax | Graph+Geometry | PNA | SSC | QM9 (50K) + GEOM (140K) + QMugs (620K) |
| Transformer-M | Graph+Geometry | Transformer-M | Supervised + DN | PCQM4Mv2 |
| Denoising | Geometry | GNS | DN | PCQM4Mv2 (3.4M) |

**Table 1:** A summary of Molecular Pre-trained Models

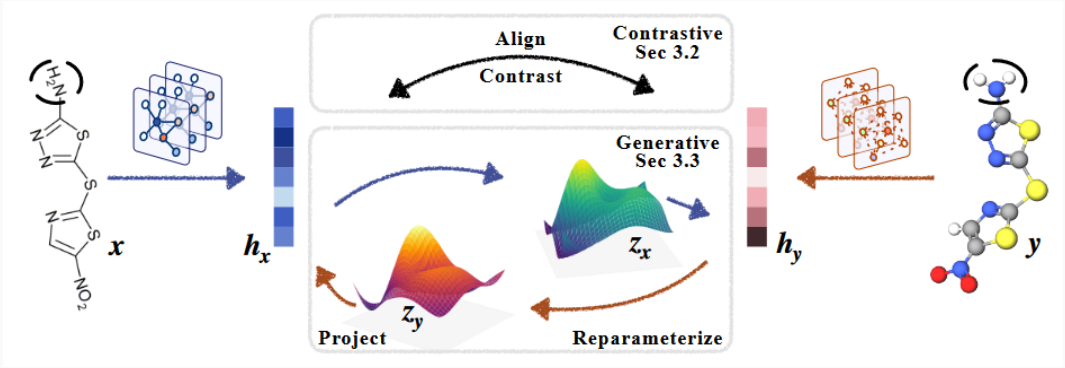
# Transformer-M



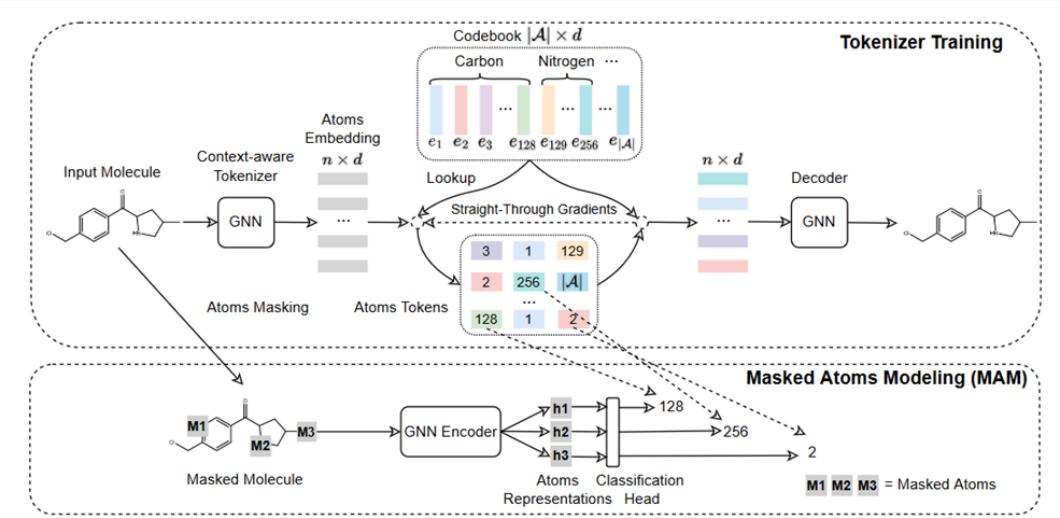
# 3D Informax


1.
maximize MI
$z^a$
$z^b$
$f^a$
2D Net
G
2D
3D Net
$f^b$
$R^1$
3D
xyz
$R^2$
xyz
2. property
$f^a$
2D Net
G
2D

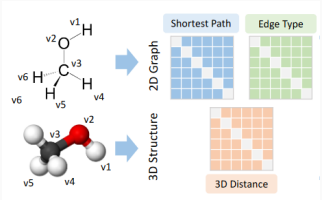
# GraphMVP


Align
Contrastive
Sec 3.2
Contrast
Generative
Sec 3.3
x
$h_x$
$z_x$
$z_y$
$h_y$
y
Project
Reparameterize

# Mole-BERT


Tokenizer Training
Codebook $|\mathcal{A}| \times d$
Carbon
Nitrogen
$e_1$ $e_2$ $e_3$ $e_{128}$ $e_{129}$ $e_{256}$ $e_{|\mathcal{A}|}$
Input Molecule
Context-aware Tokenizer
GNN
Atoms Embedding $n \times d$
Lookup
Straight-Through Gradients
$n \times d$
Decoder
GNN
3 1 129
2 256 $|\mathcal{A}|$
128 1 2
Atoms Masking
Atoms Tokens
Masked Atoms Modeling (MAM)
M1
M2
M3
Masked Molecule
GNN Encoder
h1
h2
h3
Atoms Representations
Classification Head
128
256
2
M1 M2 M3 = Masked Atoms

# Initial idea




We have two channels of input:

$$
\begin{aligned}
X_{2D}^{0} &= X + \Psi^{Degree} \\
X_{3D}^{0} &= X + \Psi^{Sum \quad of \quad Distance}
\end{aligned}
\tag{1}
$$

where $\Psi^{Degree}$ denote the degree encoding of the atom, $\Psi_{i}^{\text{Sum of 3D Distance}} = \sum_{j \in [n]} \psi_{(i,j)} W_{D}^{3}$

Then in self-attention:

$$A_{2D}(X) = \text{softmax}\left(\frac{X_{2D}W_Q(X_{2D}W_K)^\top}{\sqrt{d}} + \underbrace{\Phi^{\text{SPD}} + \Phi^{\text{Edge}}}_{\text{2D pair-wise channel}}\right) \tag{2}$$

$$A_{3D}(X) = \text{softmax}\left(\frac{X_{3D}W_Q(X_{3D}W_K)^\top}{\sqrt{d}} + \underbrace{\Phi^{\text{3D Distance}}}_{\text{3D pair-wise channel}}\right) \tag{3}$$

Then we have two types of representation $z_{3D}, z_{2D}$,

$$\begin{aligned} z_{2D} &= A_{2D}V_{2D} \\ z_{3D} &= A_{3D}V_{3D} \end{aligned} \tag{4}$$

these can be atom-wise or molecular-wise. And we can design pre-training task via these representations.