Molecular Pre-trained Models

Wu Hualong

HARBIN INSTITUTE OF TECHNOLOGY,SHEN ZHEN

Outline

SMILES

MULTIMODAL MOLECULAR PRETRAINING VIA MODALITY BLENDING

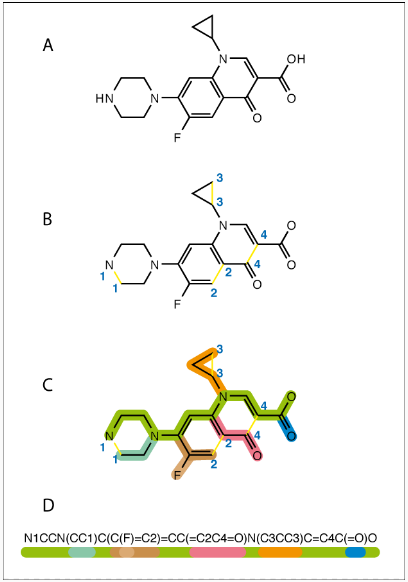
SMILES

SMILES

The **simplified molecular-input line-entry system (SMILES)** is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.

- Atoms are represented by the standard abbreviation of the chemical elements, in square brackets, such as [Au] for gold.
    - Brackets may be omitted in the common case of atoms: B, C, N, O, P, S, F, Cl, Br, or I
- single bond Hydrogen atoms are often omitted. For instance, the SMILES for water is written as either O.
- Double bonds are represented by ”=”; The three keys are represented by ”♯”.Carbon dioxide containing double bonds is expressed as O=C=O

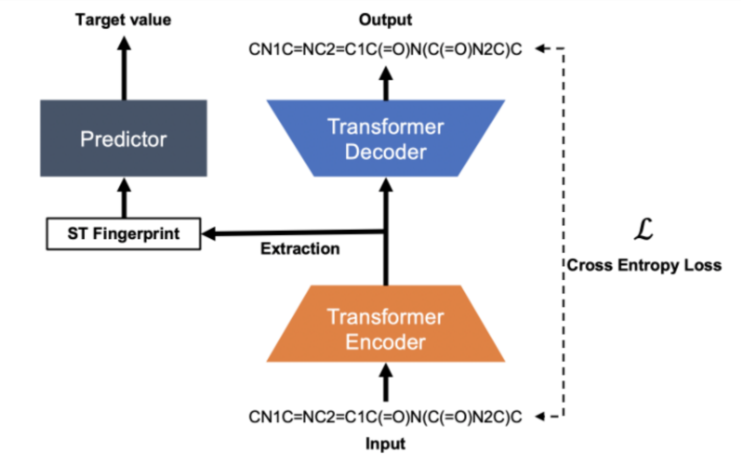
SMILES



N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

SMILES

SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery(2019)



MULTIMODAL MOLECULAR PRETRAINING VIA MODALITY BLENDING

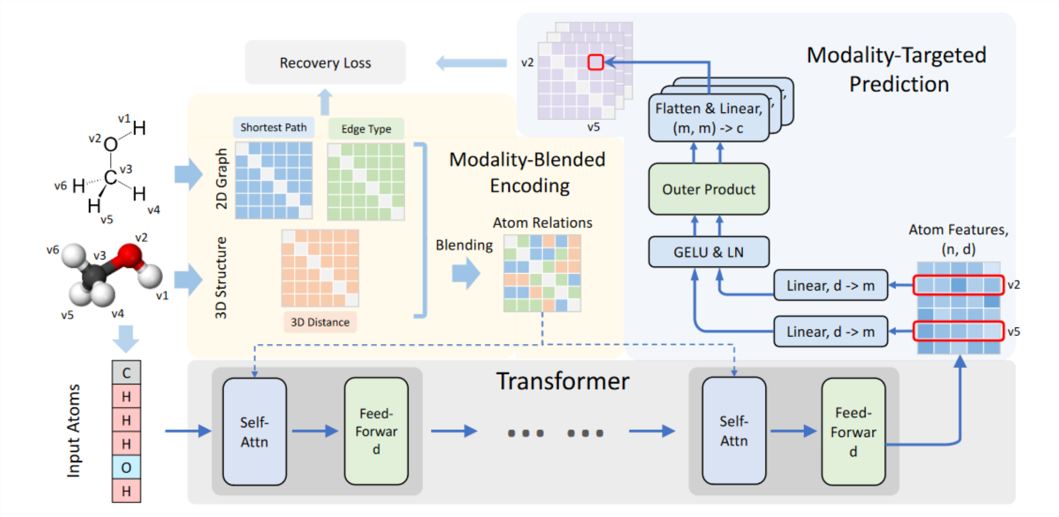
MULTIMODAL MOLECULAR PRETRAINING VIA MODALITY BLENDING

ICLR 2024

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou,Jingjing Liu

Tsinghua University,Harbin Institute of Technology, Chinese Academy of Sciences

Illustration of unified molecular representation learning process



Modality-blended Encoding

Author adopts three appearances of relations across 2D and 3D modalities following **Transformer-M**.

- Ψ^{ij}_{SPD} represents the shortest path distance between atom i and j
- $\Psi^{ij}_{\mathrm{Edge}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{w}_n^\top \mathbf{e}_n$
- $\Psi^{ij}_{\mathsf{Distance}}$ is the encoding of Euclidean distances of an atom pair (i, j)

The blended matrix is determined as follows:

$$\Psi^{ij}_{\mathsf{2D\&3D}} = \Psi^{ij}_{\mathsf{SPD}} \mathbb{1}_1 + \Psi^{ij}_{\mathsf{Edge}} \mathbb{1}_2 + \Psi^{ij}_{\mathsf{Distance}} \mathbb{1}_3, \text{ where } \mathbb{1}_k = \begin{cases} 1 & \text{if } s^{ij} = k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where each position (i, j) randomly selects its element from one of the $\Psi^{ij}_{\mathsf{SPD}}, \Psi^{ij}_{\mathrm{Edge}}, \Psi^{ij}_{\mathsf{Distance}}$

Modality-blended Encoding

Then inject this modality-blended relation $\Psi_{\mathbf{2D\&3D}}$ into the self-attention module:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}} + \Psi_{\mathbf{2D\&3D}}\right)\mathbf{V} \tag{2}$$

Modality-targeted Prediction

The model recovers the full $\mathcal{R}_{\mathrm{spd}}, \mathcal{R}_{\mathrm{edge}}, \mathcal{R}_{\mathrm{distance}}$.

- atom representations $\mathbf{X}^{L+1} \in \mathbb{R}^{n \times d}$
- two independent Linear layers $\mathbf{W}_L, \mathbf{W}_r \in \mathbb{R}^{m \times d}$
- modality-targeted head:$\mathbf{W}_{\mathrm{head}} \in \mathbb{R}^{c \times m^2}$
- $\mathrm{G}(\cdot) = \mathrm{LayerNorm}(\mathrm{GELU}(\cdot))$

$$\begin{aligned} \mathbf{o}_{ij} &= \mathrm{G}(\mathbf{W}_l \mathbf{X}_i^{L+1}) \otimes \mathrm{G}(\mathbf{W}_r \mathbf{X}_j^{L+1})^\top \in \mathbb{R}^{m \times m} \\ \mathbf{z}_{ij} &= \mathbf{W}_{\mathrm{head}} \cdot \mathrm{Flatten}(\mathbf{o}_{ij}) \in \mathbb{R}^c \end{aligned} \tag{3}$$

We now obtain the modality-targeted relation matrix $\mathbf{Z} \in \mathbb{R}^{n \times n \times c}$, where c depends on the targeted task

Fine-tuning

For scenarios where a large amount of 2D molecular graphs is available while 3D conformations are too expensive to obtain:

$$\mathcal{L}_{\mathrm{2D}} = \frac{1}{K}\sum_{k=1}^{K} \ell\left(f(\mathcal{R}^{k}_{\mathrm{spd}}, \mathcal{R}^{k}_{\mathrm{edge}}, \mathcal{V}^{k}), \mathrm{y}^{k}_{\mathrm{2D}}\right) \tag{4}$$

When it comes to scenarios where 3D information is obtained:

$$\begin{gathered}
\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\top}}{\sqrt{d}} + \Psi_{\mathrm{SPD}} + \Psi_{\mathrm{Edge}} + \Psi_{\mathrm{Distance}}\right)\mathrm{V} \\
\mathcal{L}_{3D} = \frac{1}{K}\sum_{k=1}^{K} \ell\left(f(\mathcal{R}^{k}_{\mathrm{spd}}, \mathcal{R}^{k}_{\mathrm{edge}}, \mathcal{R}^{k}_{\mathrm{distance}}, \mathcal{V}^{k}), \mathrm{y}^{k}_{\mathrm{3D}}\right)
\end{gathered} \tag{5}$$

Insights

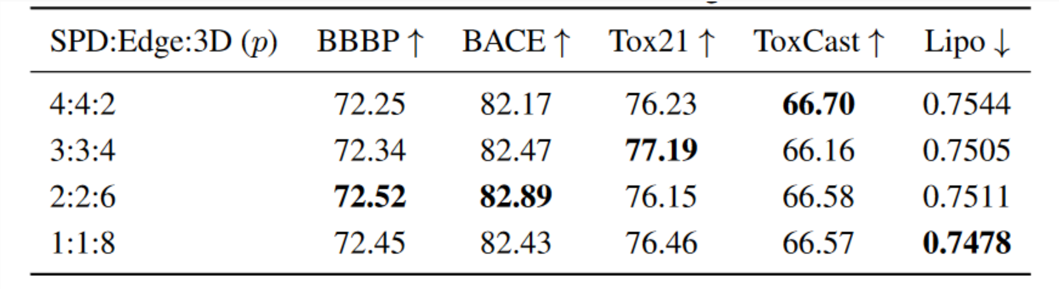
Consider two relations, denoted as $\mathcal{R}_{\text{2D}} = (a_{ij})_{n\times n}$ and $\mathcal{R}_{\text{3D}} = (b_{ij})_{n\times n}$,Their elements are randomly partitioned into two parts, represented as $\mathcal{R}_{\text{2D}} = [A_1, A_2], \mathcal{R}_{\text{3D}} = [B_1, B_2]$ The blended matrix is denoted as $\mathcal{R}_{\text{2D\&3D}} = [A_1, B_2]$.

- The training process with modality-blending maximizes the lower bound of the following mutual information:$I(A_2; A_1, B_2) + I(B_1; A_1, B_2)$
- The mutual information $I(A2; A1, B2) + I(B1; A1, B2)$ can be decomposed into two components below:

$$I(A_2;A_1,B_2)+I(B_1;A_1,B_2) = \frac{1}{2}[\underbrace{I(A_1;B_1)+I(A_2;B_2)}_{\textit{contrastive and generative}} + \underbrace{I(A_1;B_1|B_2)+I(A_2;B_2|A_1)}_{\textit{conditional contrastive and generative}}]$$

$$+\frac{1}{2}[\underbrace{I(A_1;A_2)+I(B_1;B_2)}_{\textit{mask-then-predict}} + \underbrace{I(A_1;A_2|B_2)+I(B_1;B_2|A_1)}_{\textit{multimodal mask-then-predict}}]$$

Experiments



SPD:Edge:3D (p)	BBBP ↑	BACE ↑	Tox21 ↑	ToxCast ↑	Lipo ↓
4:4:2	72.25	82.17	76.23	**66.70**	0.7544
3:3:4	72.34	82.47	**77.19**	66.16	0.7505
2:2:6	**72.52**	**82.89**	76.15	66.58	0.7511
1:1:8	72.45	82.43	76.46	66.57	**0.7478**

Figure 2: Ablations on the blending ratio