A Graph to Graphs Framework for Retrosynthesis Prediction

Chence Shi   Minkai Xu   Hongyu Guo   Ming Zhang   Jian Tang

May 6, 2021

- Introduction
- Reaction Center Identification
- Reactants Generation via Variational Graph Translation

Introduction

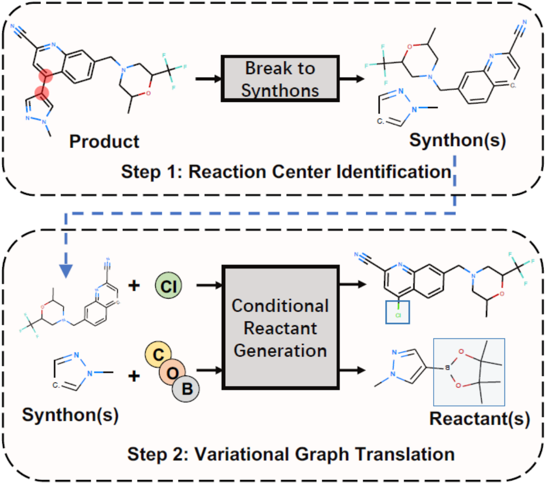

Figure 1: The overall framework of the proposed method.

Notation

- $G = (A, X)$, a labeled graph for a molecule, where $A \in \{0,1\}^{n \times n \times b}$ is the adjacency matrix and $X \in \{0,1\}^{n \times d}$ the matrix of node features;
- n, the number of atoms;
- b, the number of bond types;
- d, the dimension of node features;
- $(\{G_i\}_{i=1}^{N_1}, G_p)$, a reaction where G_i denotes a reactant graph and G_p a product graph

- Introduction
- Reaction Center Identification
- Reactants Generation via Variational Graph Translation

Reaction Center

Reaction center means an atom pair (i, j) that satisfies:

- there is a bond between the i-th and j-th nodes in the **product** graph;
- there is **no** bond between the i-th and j-th nodes in the **reactant** graph.

Reaction Center

Reaction center means an atom pair (i, j) that satisfies:

- there is a bond between the i-th and j-th nodes in the **product** graph;
- there is **no** bond between the i-th and j-th nodes in the **reactant** graph.

Synthons are subgraphs extracted from the products by breaking the bonds in the reaction centers.

Reaction Center Identification

The reactivity score of the atom pair (i, j) is calculated as:

$$s_{ij} = \sigma(m_r(e_{ij})) \tag{1}$$

where m_r is a feedforward network that maps e_{ij} to a scalar, $\sigma(\cdot)$ denotes the Sigmoid function, and e_{ij} the edge embedding.

Reaction Center Identification

The reactivity score of the atom pair (i, j) is calculated as:

$$s_{ij} = \sigma(m_r(e_{ij})) \tag{1}$$

where m_r is a feedforward network that maps e_{ij} to a scalar, $\sigma(\cdot)$ denotes the Sigmoid function, and e_{ij} the edge embedding.

$$\mathcal{L}_1 = -\sum_{r}\sum_{i \neq j} \lambda Y_{ij} \log(s_{ij}) + (1 - Y_{ij}) \log(1 - s_{ij}) \tag{2}$$

where $Y \in \{0, 1\}^{n \times n}$ indicates the reaction centers.

The Edge Embedding

$$e_{ij} = H_i^L \parallel H_j^L \parallel A_{ij} \parallel h_{G_p} \tag{3}$$

where $H^L \in \mathbb{R}^{n \times k}$ denotes the node embeddings, and h_{G_p} denotes the entire graph embedding of product G_p.

The Node Embedding

R-GCN, a variant of Relational GCN,

$$H^L = \text{R-GCN}(G_p) \tag{4}$$

For the l-th layer,

$$H^l = \text{Agg}(\text{ReLu}(\{E_i H^{l-1} W_i^l\} | i \in (1, \ldots, b))) \tag{5}$$

where $E_i = A_{[:,:,i]} + I$ denotes the adjacency matrix of the i-th edge type, W_i^l is the trainable weight matrix for the i-th edge type, and $\text{Agg}(\cdot)$ denotes an aggregation function.

- Introduction
- Reaction Center Identification
- Reactants Generation via Variational Graph Translation

Reactants Generation via Variational Graph Translation

$$(\{G_i\}_{i=1}^{N_1}, G_p) \rightarrow (\{G_i\}_{i=1}^{N_1}, \{S_i\}_{i=1}^{N_1})$$

Reactants Generation via Variational Graph Translation

$$(\{G_i\}_{i=1}^{N_1}, G_p) \to (\{G_i\}_{i=1}^{N_1}, \{S_i\}_{i=1}^{N_1})$$

$$p(G|S)$$

$$(\{G_i\}_{i=1}^{N_1}, G_p) \rightarrow (\{G_i\}_{i=1}^{N_1}, \{S_i\}_{i=1}^{N_1})$$

$$p(G|S)$$

$$p(G|z, S)$$

where z denotes a low-dimensional latent vector.

The Generative Model

Let $t = (a_1, \ldots, a_T) \in \mathcal{T}$ be a sequence of graph transformation actions, and $\mathcal{T}$ be the collection of all trajectories that can translate synthons S to target reactants G,

$$P(G|z, S) \to P(t|z, S)$$

The Generative Model

Let S^i denote the graph after applying the sequence of actions $a_{1:i}$ to S,

$$P(S^i|S^{i-1}, z) = P(a_i|S^{i-1}, z)$$

The Generative Model

Let S^i denote the graph after applying the sequence of actions $a_{1:i}$ to S,

$$P(S^i|S^{i-1}, z) = P(a_i|S^{i-1}, z)$$

Use the assumption of a Markov Decision Process (MDP),

$$p(t|z, S) = p(a_{1:T}|z, S) = \prod_{i=1}^{T} p(a_i|S^{i-1}, z) \tag{6}$$

The Defination of an Action

$$a_i = (a_i^1, a_i^2, a_i^3, a_i^4) \tag{7}$$

- $a_i^1 \in \{0,1\}^2$ predicts the termination of the graph translation procedure;
- $a_i^2 \in \{0,1\}^n$ indicates the first node;
- $a_i^3 \in \{0,1\}^{n+m}$ indicates the second node;
- $a_i^4 \in \{0,1\}^b$ predicts the bond type between two nodes.

Termination Prediction

$$H = \mathcal{R}(S^{i-1}),\ h_S = \text{Readout}(H)$$
$$p(a_i^1 | z, S^{i-1}) = \tau(m_t(h_S, z))$$

where $\tau(\cdot)$ denotes the softmax function, and $m_t(\cdot)$ is a feedforward network.

Nodes Selection

$$p(a_i^2|z, S^{i-1}, a_i^1) = \tau(\beta_1 \odot m_f(\mathcal{R}(\tilde{S}^{i-1}), z))$$
$$a_i^2 \sim p(a_i^2|z, S^{i-1}, a_i^1)$$
$$p(a_i^3|z, S^{i-1}, a_i^{1:2}) = \tau(\beta_2 \odot m_s(\mathcal{R}(\tilde{S}^{i-1}), z, a_i^2))$$
$$a_i^3 \sim p(a_i^3|z, S^{i-1}, a_i^{1:2})$$

where $\tilde{S}^{i-1} = S^{i-1} \cup V$, V is the set of possible atoms to be added during graph translation. $m_f(\cdot)$ and $m_s(\cdot)$ are feedforward networks. β_1 and β_2 are masks to zero out the probability of certain atoms being selected.

Edge Labeling

$$p(a_i^4|z, S^{i-1}, a_i^{1:3}) = \tau(m_e(\mathcal{R}(\tilde{S}^{i-1}), z, a_i^{2:3}))$$
$$a_i^4 \sim p(a_i^4|z, S^{i-1}, a_i^{1:3})$$

where $m_e(\cdot)$ is a feedforward networks.

Reactants Generation via Variational Graph Translation

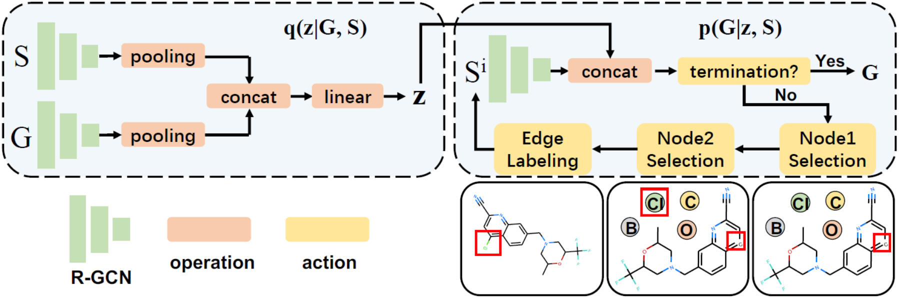

Figure 2: Illustration of the proposed variational graph translation module.