

Hierarchical Generation of Molecular Graphs using Structural Motifs

ICML 2020

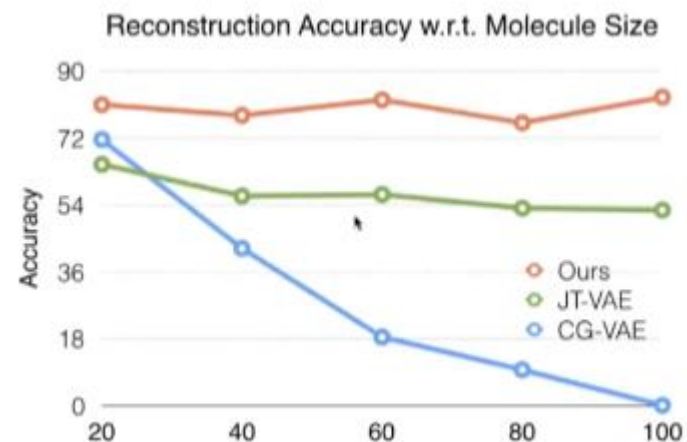
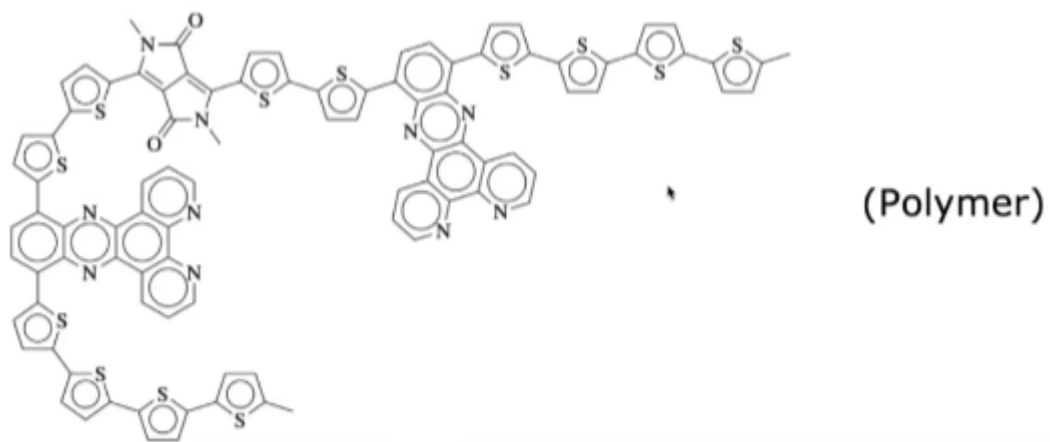
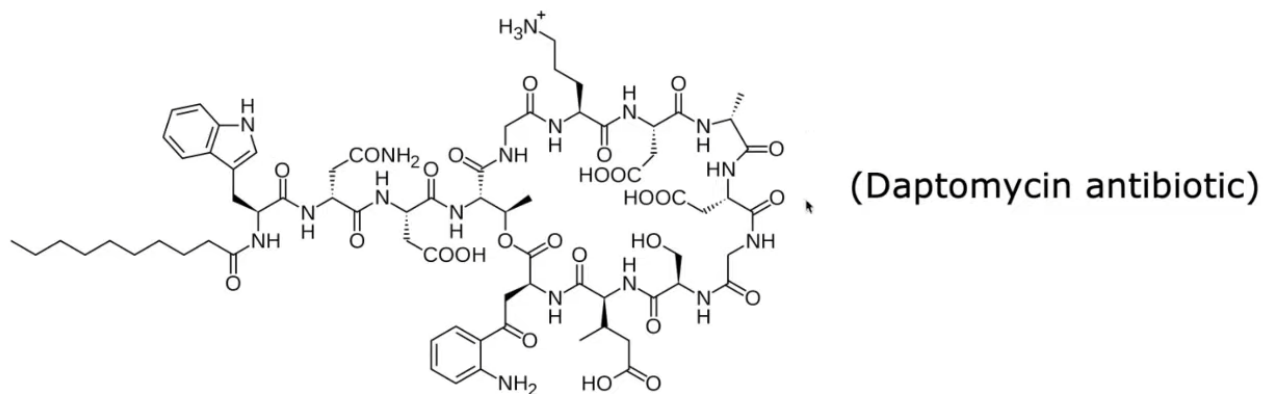
Department of Artificial Intelligence Korea University

Younghan Son

Apr 2022

Large scale molecules

- 기존 small scale molecules는 생각보다 성능이 높게 잡히지만 large molecules에서 성능이 급격하게 하락



Previous methods

- Atom based (CG-VAE 18', GCPN 18')



Atom based

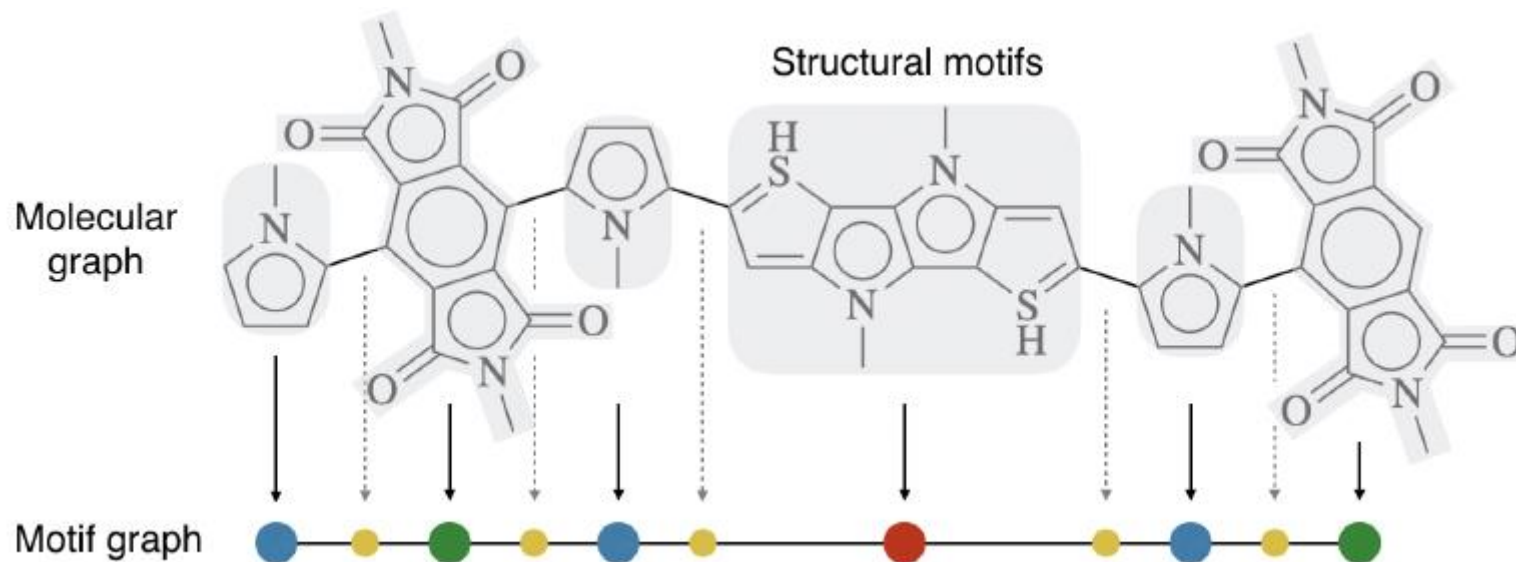
- Substructure based (JT-VAE, graph to graph translation, ICLR 18', 19')



Substructure based

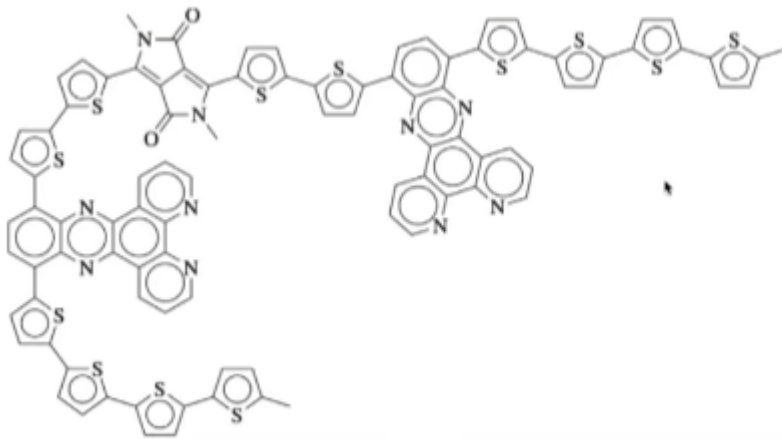
Structure motifs

- 기존 atom based 220 steps → 11 steps



Structure motifs

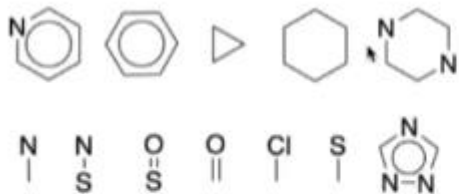
- *“Large molecules such as polymers exhibit clear hierarchical structure, being built from repeated structural motifs”*
- Incorporating such motifs as building blocks in the generation process



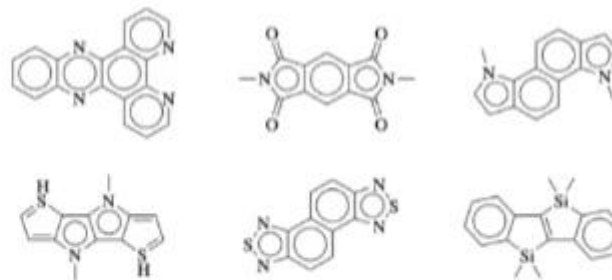
(Polymer)

Substructure vs Motif

- Motif는 어떠한 성질을 가지는 최소한의 단위
- 그렇다면 Motif를 어떻게 정의하는가



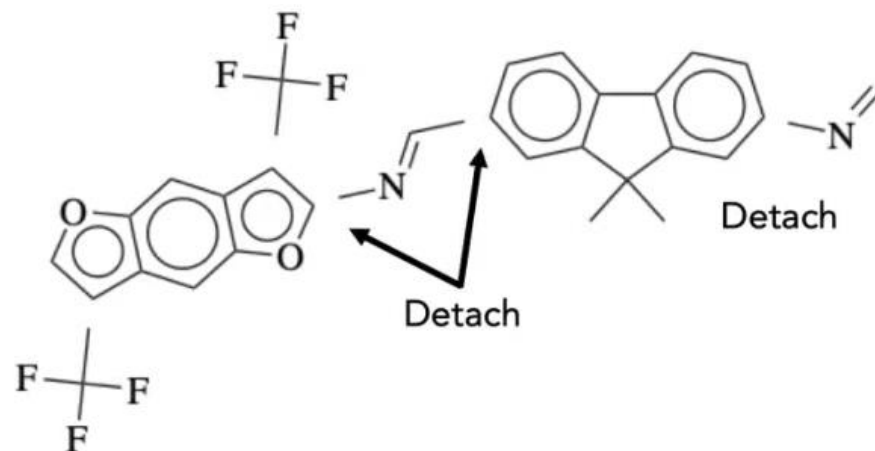
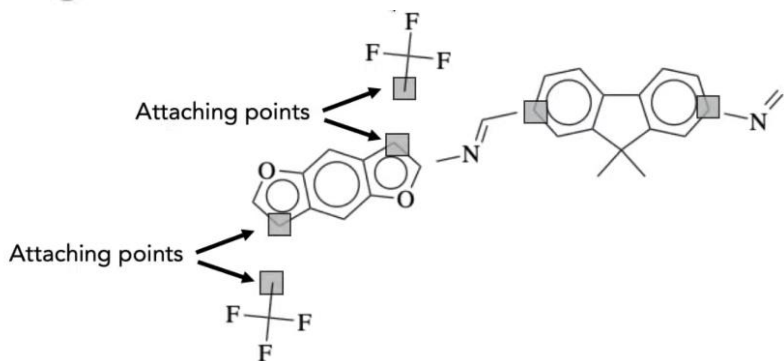
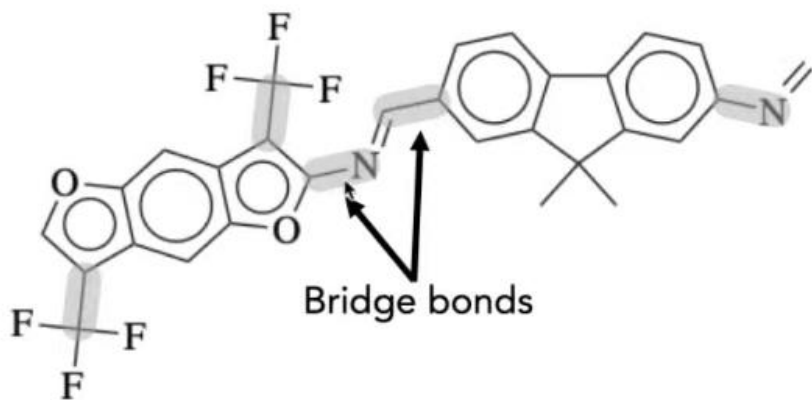
Substructures
(ring and bond only)



Motifs
(structures can be flexible)

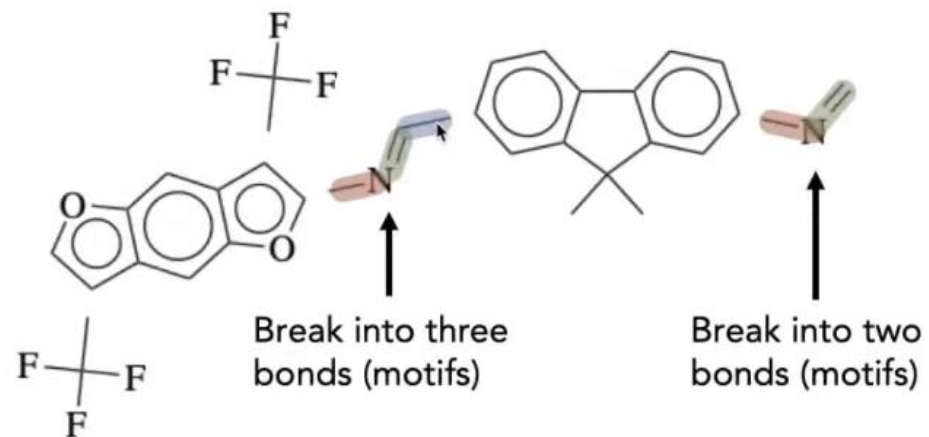
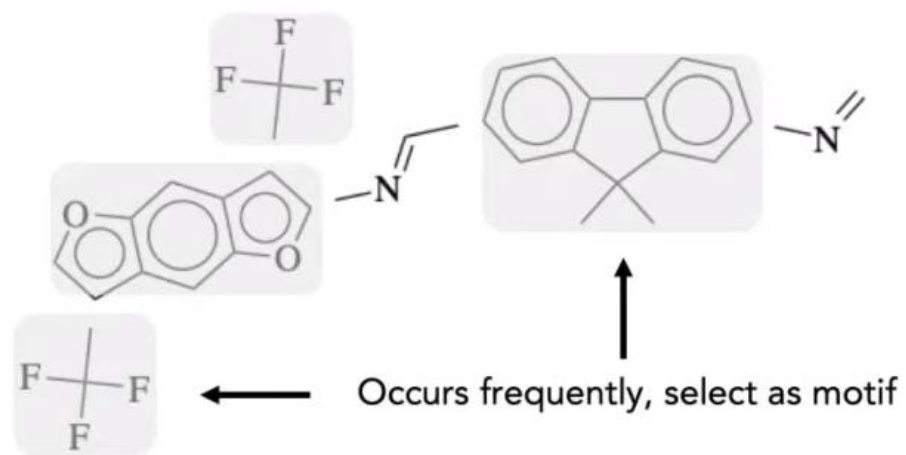
Motif extraction

- 모티프를 정의하는 것은 하기 나름
- 1) 모티프를 잇는 bridge bonds를 정의한다.
 - Bridge bonds: chemical validity를 헤치지 않는 bonds
- 2) Bridge bonds를 끊는다
 - 끊긴 Bridge bonds에 다른 motif의 Attaching points가 됨



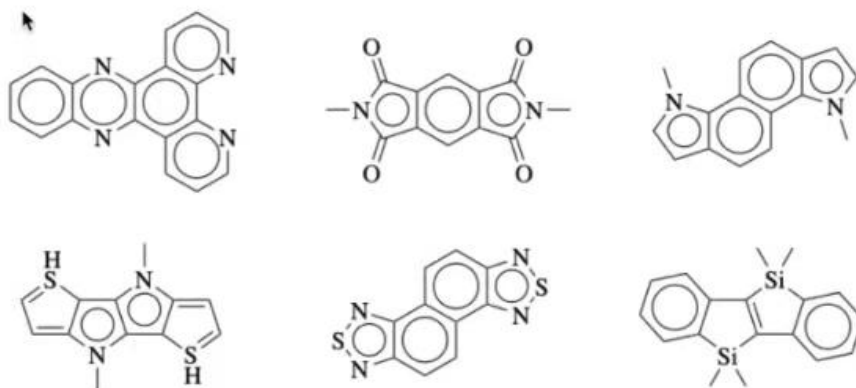
Motif extraction

- 3) 100번 이상 등장하는 motif를 저장
- 4) 선택되지 못한 motif는 분해해서 하위 motif로 저장
- 모티프를 찾는 것은 substructure counting이라 비싼거 아닌가? 그냥 Lookup table로 다 만들어놔서 괜찮

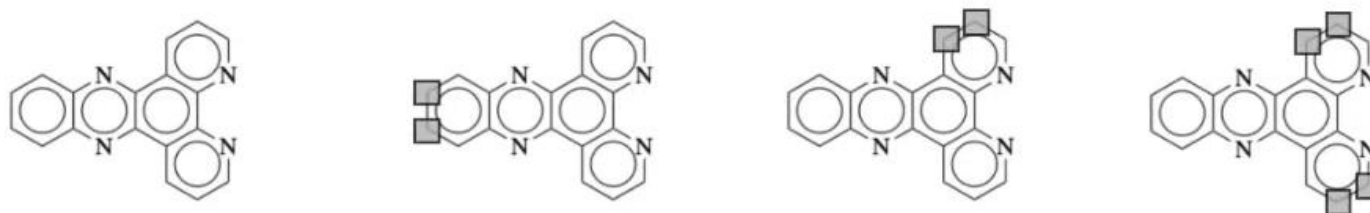


— Motif Vocabulary

- We can construct a motif vocabulary given a training set (usually <500)



- Each motif also has a vocabulary of possible attaching points (usually < 10).



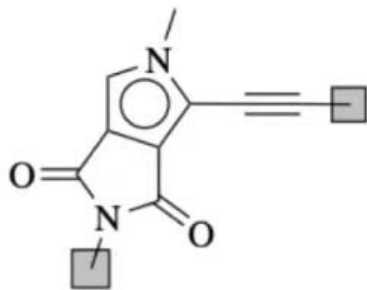
Summary

- 1) motif extract(vocab) and attachment vocab (one hot encoding)
- 2) Hierarchical encoder atom to motif
- 3) Autoregressive hierarchical decoder(one by one)

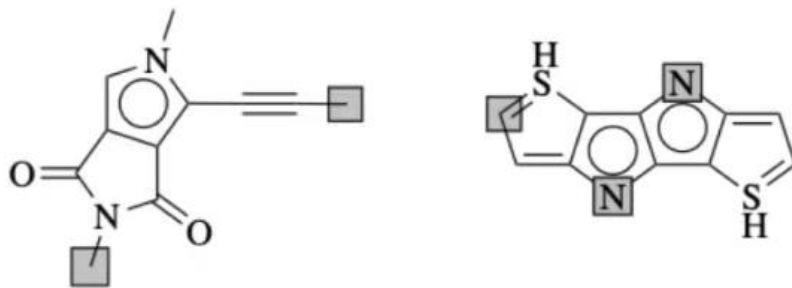
Generation

- JT-VAE는 한번에 다 붙였는데, 이번엔 한개씩 붙임(one go)

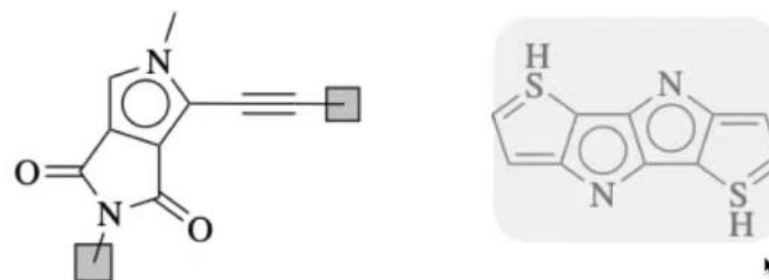
Current state



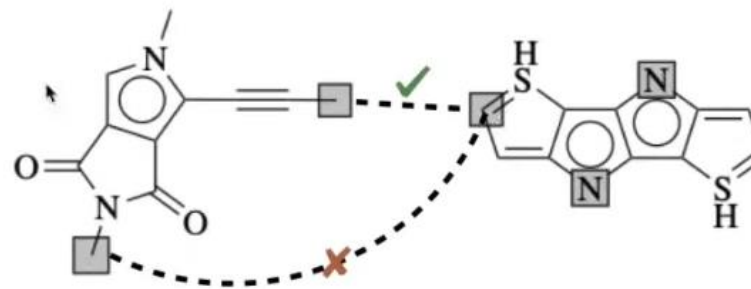
Step 2: Attachment Prediction




Step 1: Motif Prediction



Step 3: Graph Prediction



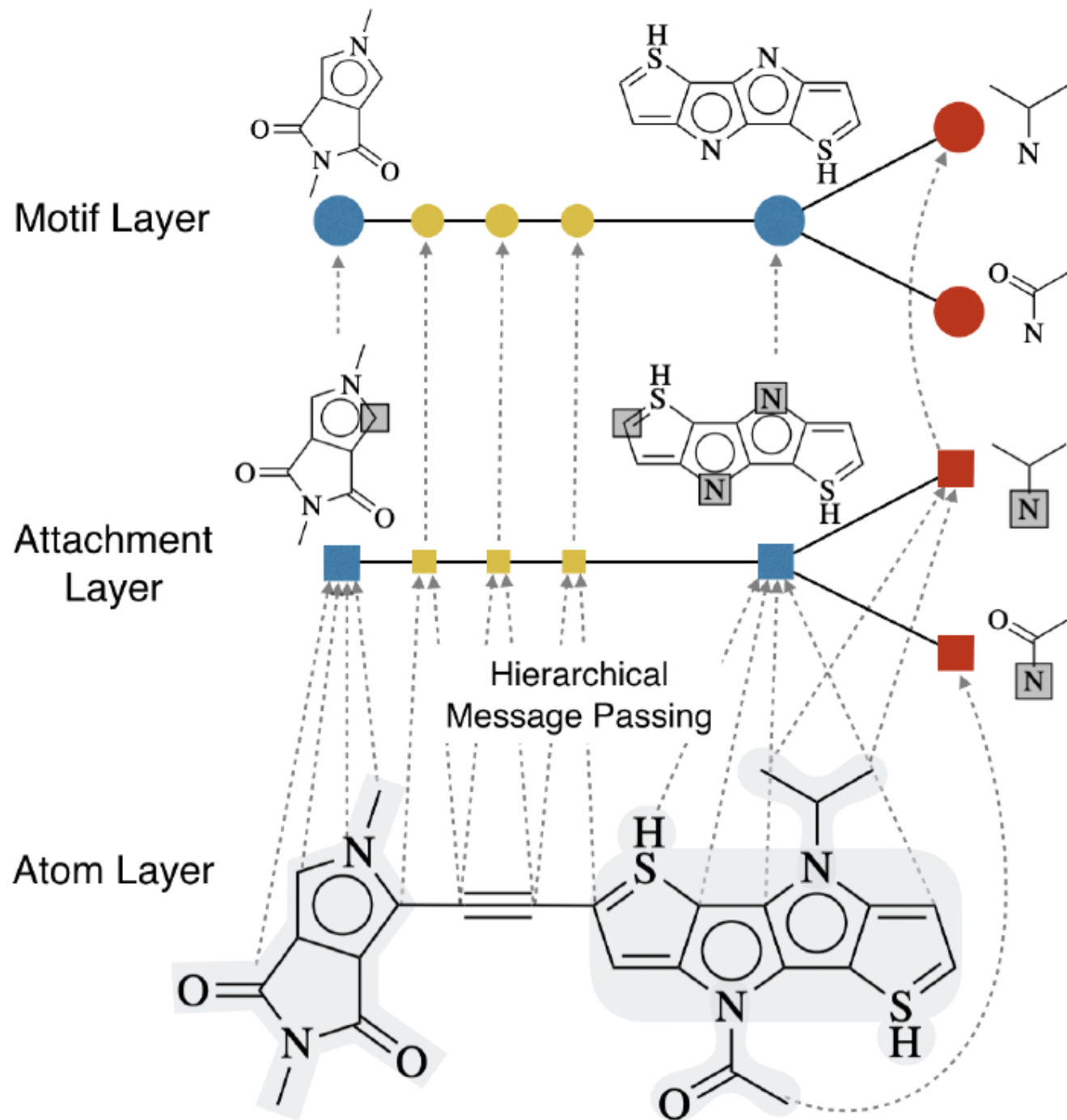
- 
- S는 motif, A는 attach, z는 latent variables VAE방식임

$$P(\mathcal{G}) = \int_{\mathbf{z}} P(\mathbf{z}) \prod_k P(\mathcal{S}_k, \mathcal{A}_k \mid \mathcal{S}_{<k}, \mathcal{A}_{<k}, \mathbf{z}) d\mathbf{z}$$

Hierarchical graph generation

- Pooling 과정과도 유사
- 1) Motif layer
 - DFS 형식으로 motif를 구하기 때문에 tree 구조임
 - Next motif prediction step
- 2) Attachment layer $\mathcal{A}_i = (\mathcal{S}_i, \{\overset{\text{Intersect node}}{v_j}\})$
 - Encode connectivity between motif
 - Attachment prediction step
- 3) Atom layer
 - Graph prediction step
 - Node: atom type and charge
 - Edge: bond type

3개의 층을 directed edge로 연결하여 정보를 전달



Hierarchical graph encoder

- Atom layer MPN

hierarchical message

$$c_g^g = \{h_v\} = \text{MPN}_{\psi_1}(\mathcal{H}_g^g, \{e(a_u)\}, \{e(b_{uv})\})$$

node edge

- Attachment layer MPN

attachment message

$$f_{\mathcal{A}_i} = \text{MLP}\left(e(\mathcal{A}_i), \sum_{v \in \mathcal{S}_i} h_v\right)$$

adjacent motif atom embed

hierarchical message

$$c_g^a = \{h_{\mathcal{A}_i}\} = \text{MPN}_{\psi_2}(\mathcal{H}_g^a, \{f_{\mathcal{A}_i}\}, \{e(d_{ij})\})$$

directed edge

- Motif layer MPN

$$f_{\mathcal{S}_i} = \text{MLP}(e(\mathcal{S}_i), h_{\mathcal{A}_i})$$

$$c_g^s = \{h_{\mathcal{S}_i}\} = \text{MPN}_{\psi_3}(\mathcal{H}_g^s, \{f_{\mathcal{S}_i}\}, \{e(d_{ij})\})$$

graph embedding directed edge

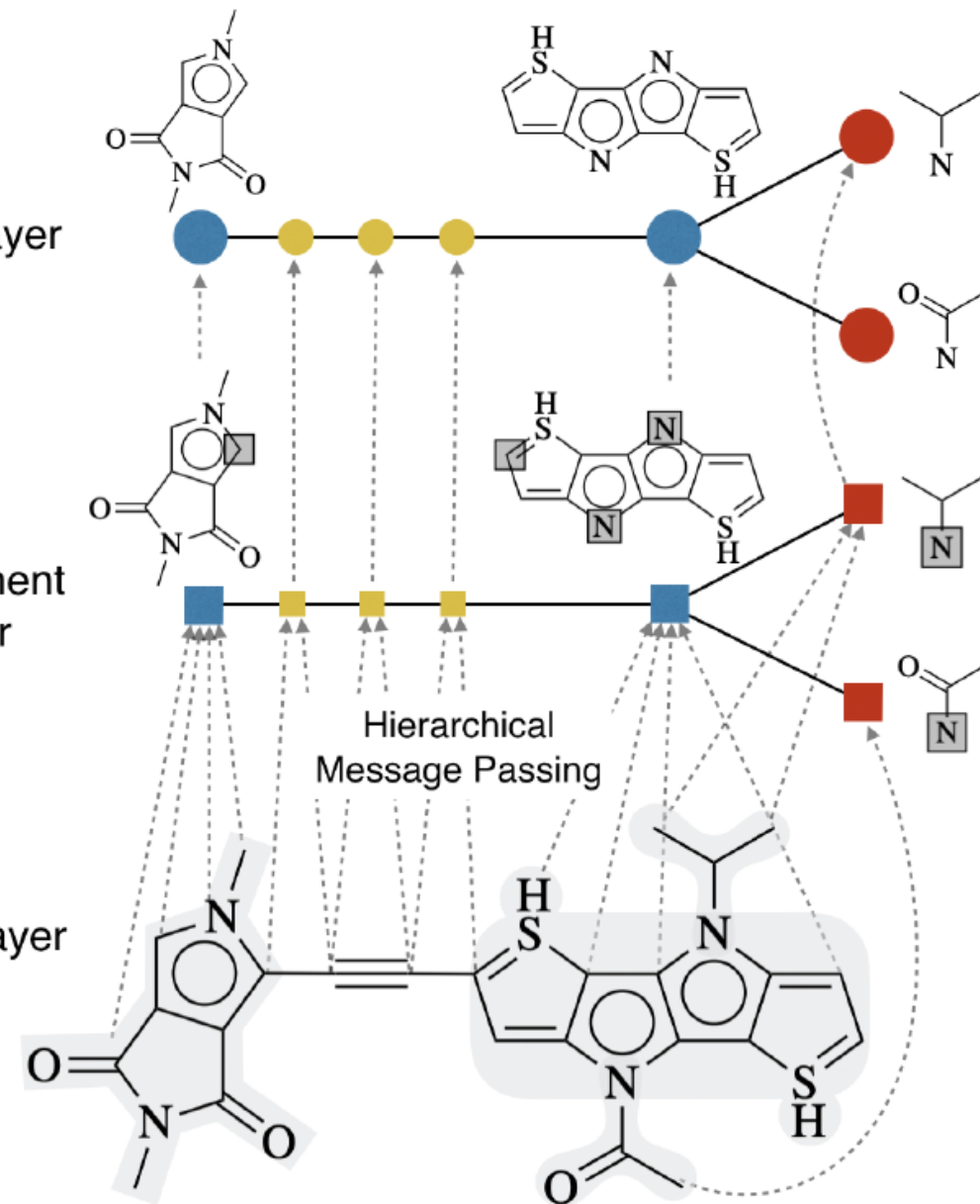
- Molecular Graph latent vector

$$z_g = \mu(h_{\mathcal{S}_1}) + \exp(\Sigma(h_{\mathcal{S}_1})) \cdot \epsilon; \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Motif Layer

Attachment Layer

Atom Layer



Hierarchical graph decoder

- Next Motif Prediction - 어떤 모티프를 붙일지

$$p_{S_t} = \text{softmax}(\text{MLP}(h_{S_k}, z_G))$$

Graph embedding latent vector

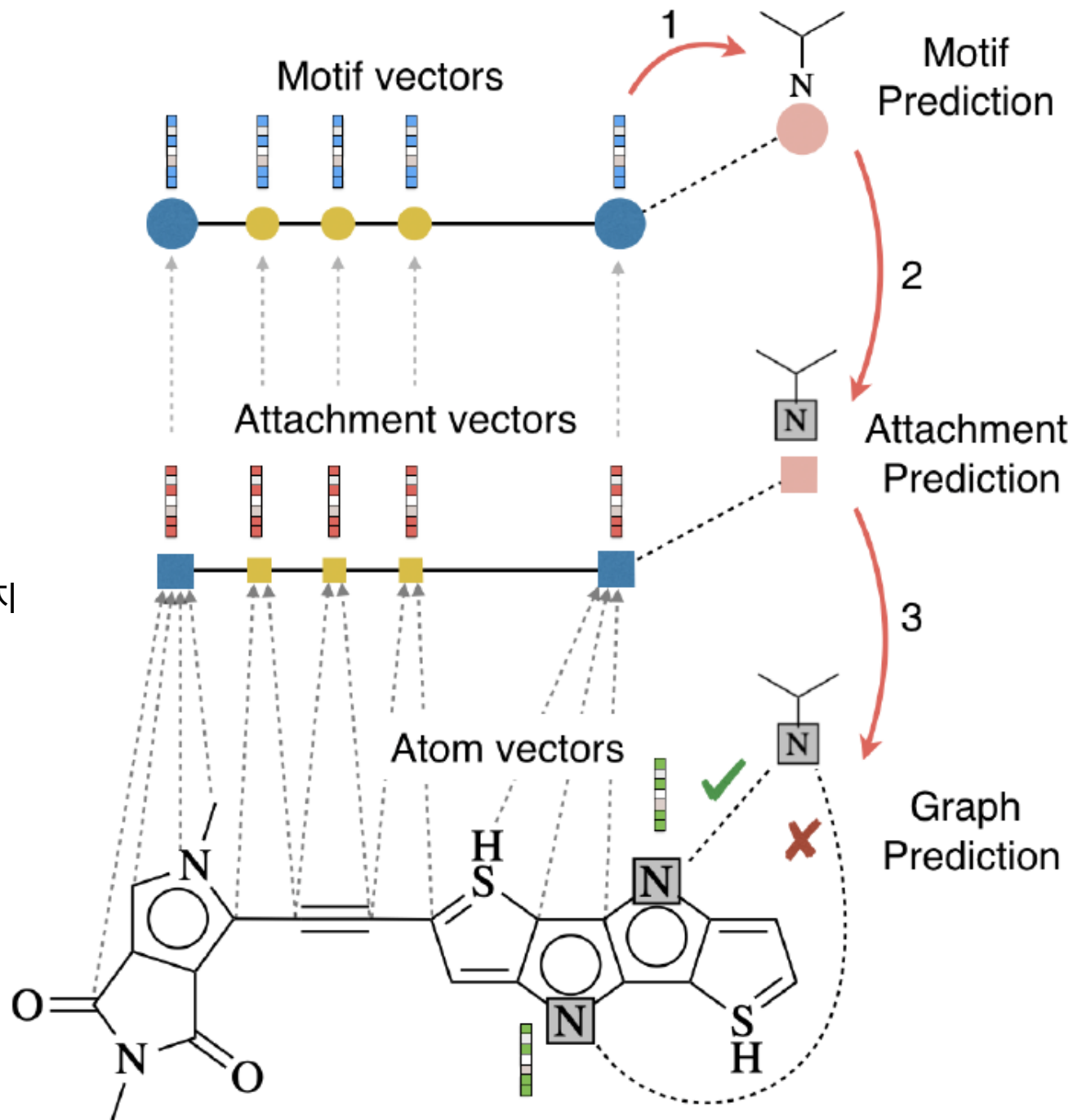
- Attachment Prediction - 어느 모티프에 붙일지

$$p_{A_t} = \text{softmax}(\text{MLP}(h_{S_k}, z_G))$$

- Graph Prediction - 어느 attachment point(atom)에 붙일지

$$p_M = \text{softmax}(h_M \cdot z_G)$$

$$h_M = \sum_i \text{MLP}(h_{u_j}, h_{v_j})$$



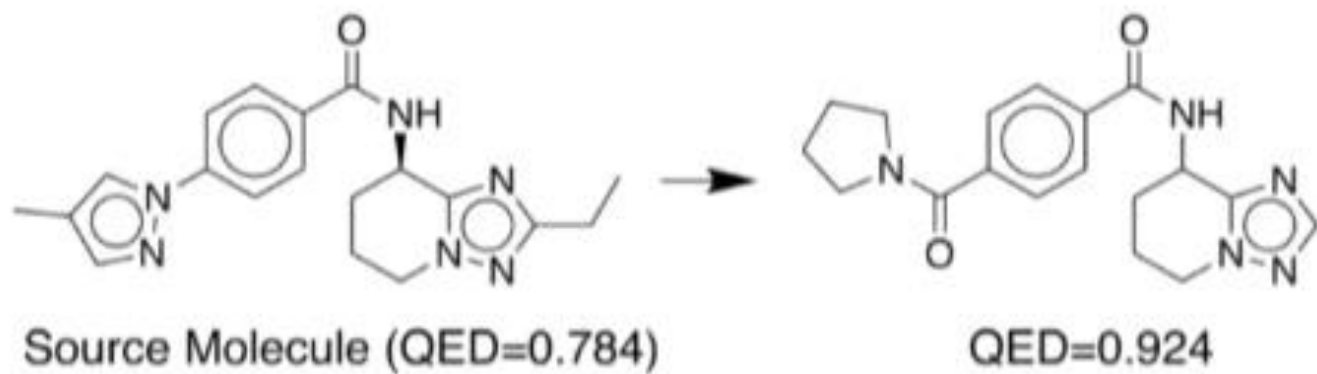
Training

- Teacher Forcing
- 실제 molecule을 가지고 RNN계열 처럼 teacher forcing 방법으로 학습
- Optimize negative ELBO

$$- \mathbb{E}_{z \sim Q} [\log P(\mathcal{G}|z)] + \lambda_{\text{KL}} \mathcal{D}_{\text{KL}}[Q(z|\mathcal{G}) || P(z)]$$

Extension to Graph-to-Graph translation(Molecular optimize)

- molecular optimization, which seeks to modify compounds in order to improve their biochemical properties
- 기존 molecule을 수정하여 더 새로운 molecule 생성
- Attention module추가



Training

- Given X molecule(source), get diverse Y molecules(target)

$$P(Y|X) = \int_z P(Y|X, z)P(z)dz$$

$$Q(z|X, Y) = \mathcal{N}(\mu_{X,Y}, \sigma_{X,Y})$$

- difference - motif level(Cs), atom level(Cg)

$$\delta_{X,Y}^s = \sum c_Y^s - \sum c_X^s \quad \delta_{X,Y}^g = \sum c_Y^g - \sum c_X^g$$

- reparameterization trick

$$[\mu_{X,Y}, \sigma_{X,Y}] = \text{MLP}(\delta_{X,Y}^s, \delta_{X,Y}^g)$$

Training

- Input

$$\mathbf{c}_X = \mathbf{c}_X^s \cup \mathbf{c}_X^a \cup \mathbf{c}_X^g$$

- Motif prediction & Attention(Attachment prediction)

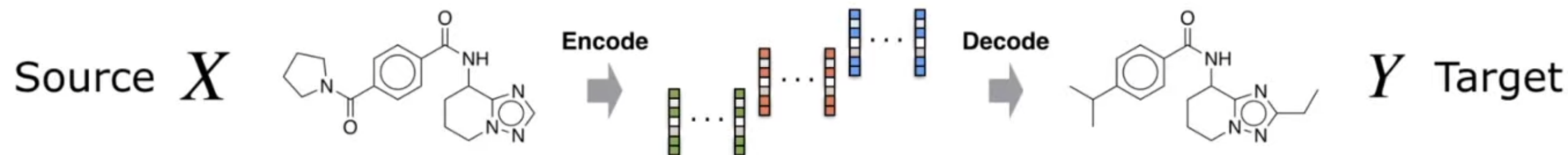
$$p_{\mathcal{S}_t} = \text{softmax}(\text{MLP}(\mathbf{h}_{\mathcal{S}_k}, \alpha_k^s, \mathbf{z}))$$

$$\alpha_k^s = \text{attention}(\mathbf{h}_{\mathcal{S}_k}, \mathbf{c}_X^s)$$

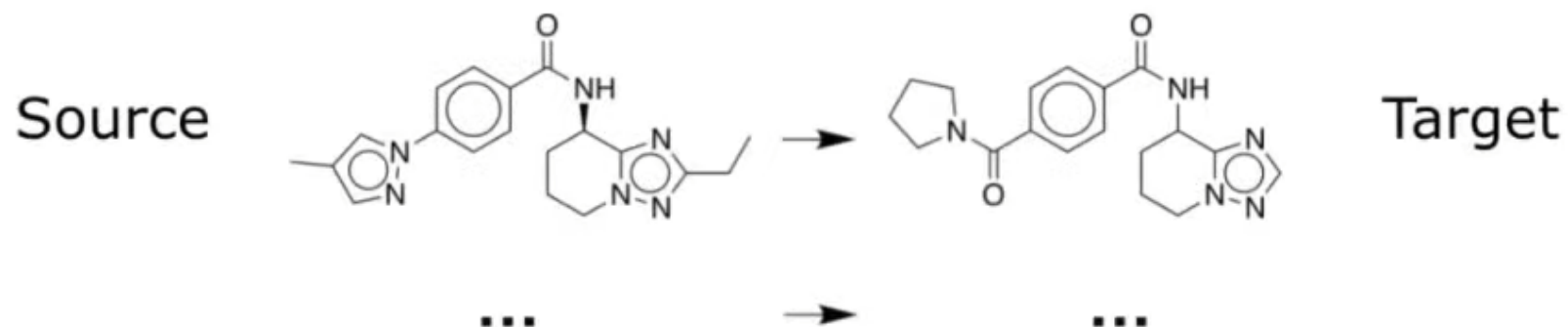
- Graph prediction

$$p_M = \text{softmax}(\mathbf{h}_M \cdot \text{attention}(\mathbf{h}_M, \mathbf{c}_X^g))$$

$$\mathbf{h}_M = \sum_j \text{MLP}(\mathbf{h}_{u_j}, \mathbf{h}_{v_j}, \mathbf{z})$$



- The training set consists of (source, target) molecular pairs, e.g.,



Experiments

- 1) Polymer generation
- 2) Graph-to-graph translation for small molecules

Polymer generation

- 436 motifs, avg 5.24 attachment for each motif
- 속도도 빠르고 molecular size에 따른 accuracy 변화도 적었다.
- Reconstruction이란 test molecule을 encoder - decoder 방법으로 reconstruction

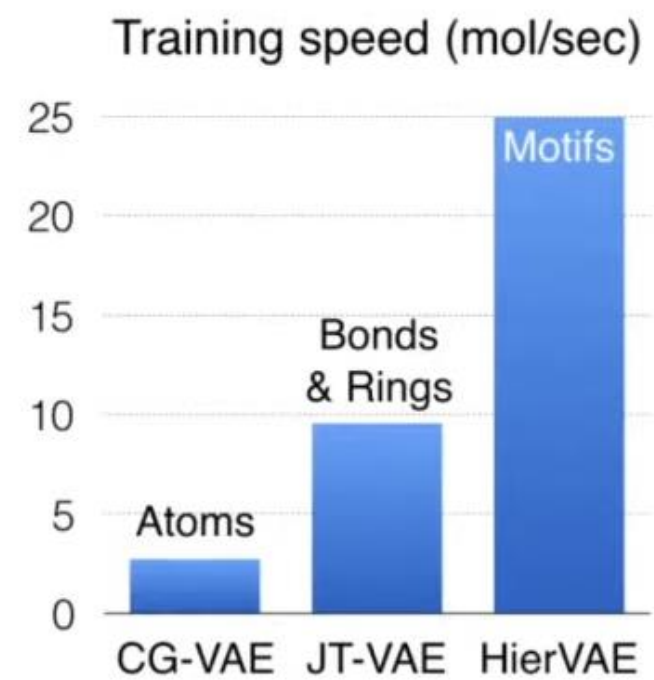
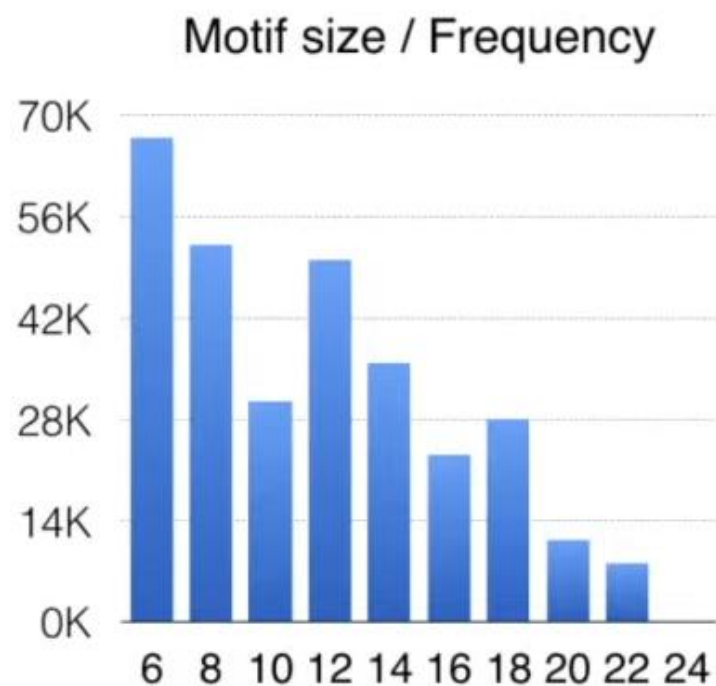
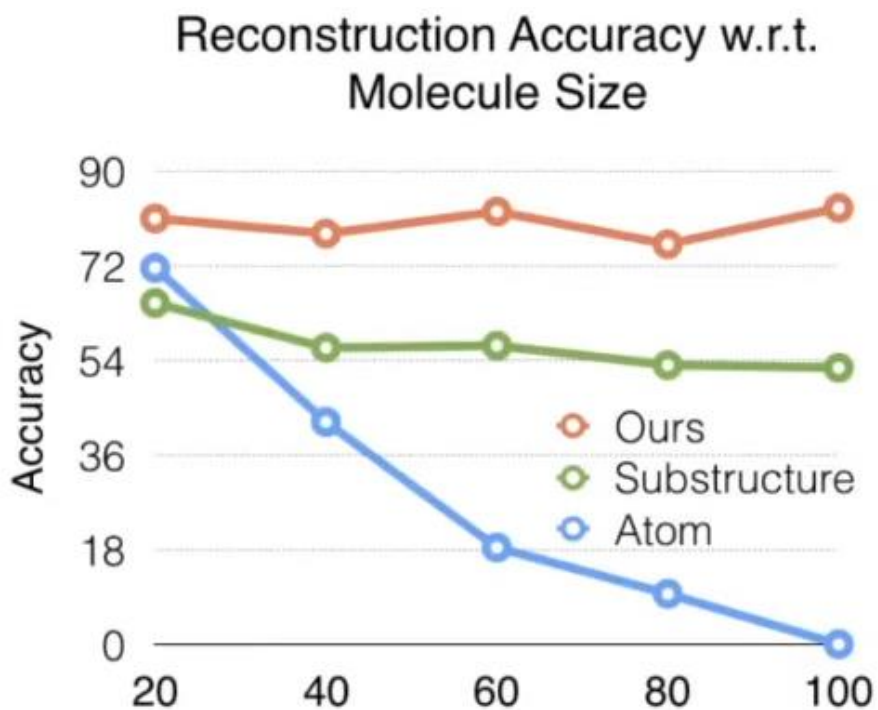
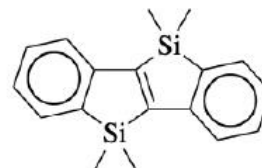
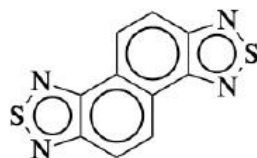
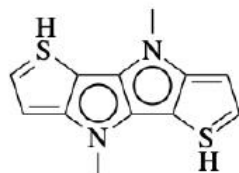
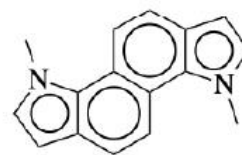
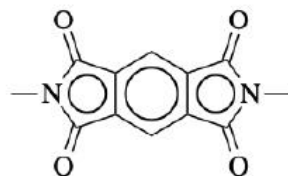
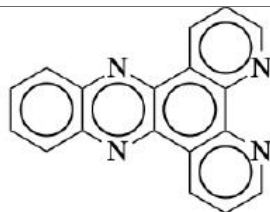


Table 1. Results on polymer generative modeling. The first row reports the oracle performance using real data as generated samples. The last row (small motif) is a variant of our model where we restrict the motif vocabulary to contain only single rings and bonds (similar to JT-VAE). “Recon.” means reconstruction accuracy; “Div.” means diversity; SNN means nearest neighbor similarity; “Frag / Scaf” means fragment and scaffold similarity. Except property statistics, all metrics are the higher the better.

How similar to real molecules

Method	Reconstruction / Sample Quality (\uparrow)				Property Statistics (\downarrow)				Structural Statistics (\uparrow)		
	Recon.	Valid	Unique	Div.	logP	SA	QED	MW	SNN	Frag.	Scaf.
Real data	-	100%	100%	0.823	0.094	6.7e-5	1.7e-5	82.3	0.706	0.995	0.462
SMILES	21.5%	93.1%	97.3%	0.821	1.471	0.011	5.4e-4	4963	0.704	0.981	0.385
CG-VAE	42.4%	100%	96.2%	0.879	3.958	2.600	0.0030	3944	0.204	0.372	0.001
JT-VAE	58.5%	100%	94.1%	0.864	2.645	0.157	0.0075	10867	0.522	0.925	0.297
HierVAE	79.9%	100%	97.0%	0.817	0.525	0.007	5.7e-4	1928	0.708	0.984	0.390
• Small motif	71.0%	100%	97.2%	0.835	0.872	0.042	0.0019	5320	0.575	0.949	0.191



Graph-to-Graph Translation

- Faster, SOTA accuracy, improved diversity

Table 2. Results on graph translation tasks from Jin et al. (2019). We report average improvement for continuous properties (logP), and success rate for binary properties (e.g., DRD2).

Method	solubility, synthetic accessibility				biological activity against dopamine type 2 receptor			
	logP (sim ≥ 0.6)		logP (sim ≥ 0.4)		Drug likeness		DRD2	
	Improvement	Diversity	Improvement	Diversity	Success	Diversity	Success	Diversity
JT-VAE	0.28 ± 0.79	-	1.03 ± 1.39	-	8.8%	-	3.4%	-
CG-VAE	0.25 ± 0.74	-	0.61 ± 1.09	-	4.8%	-	2.3%	-
GCPN	0.79 ± 0.63	-	2.49 ± 1.30	-	9.4%	0.216	4.4%	0.152
MMPA	1.65 ± 1.44	0.329	3.29 ± 1.12	0.496	32.9%	0.236	46.4%	0.275
Seq2Seq	2.33 ± 1.17	0.331	3.37 ± 1.75	0.471	58.5%	0.331	75.9%	0.176
JTNN	2.33 ± 1.24	0.333	3.55 ± 1.67	0.480	59.9%	0.373	77.8%	0.156
AtomG2G	2.41 ± 1.19	0.379	3.98 ± 1.54	0.563	73.6%	0.421	75.8%	0.128
HierG2G	2.49 ± 1.09	0.381	3.98 ± 1.46	0.564	76.9%	0.477	85.9%	0.192

Ablation

- Decoder input에서 motif를 뺌 $c_X = \text{ⓧ} \cup c_X^a \cup c_X^g$
- Motif layer 자체를 없앴
- Attachment layer까지 없앴

Table 3. Ablation study: the importance of hierarchical graph encoding, LSTM MPN architecture and structure-based decoding.

Method	QED	DRD2
HierG2G	76.9%	85.9%
• atom-based decoder	76.1%	75.0%
• two-layer encoder	75.8%	83.5%
• one-layer encoder	67.8%	74.1%

Graph encoders

- 각 레이어에서 Hierarchical 한 정보를 뽑는다는 점에서 ChebyGNN, Unet, SAGPool과 유사하다고 (그들은 생각)
- Cheby 는 K로 hierarchical 조절, topk, SAGpool은 pooling과정이 들어가므로 자연스럽게 hierarchical



Thank you
