STRING-BASED MOLECULE GENERATION VIA MULTI-DECODER VAE

Kisoo Kwon

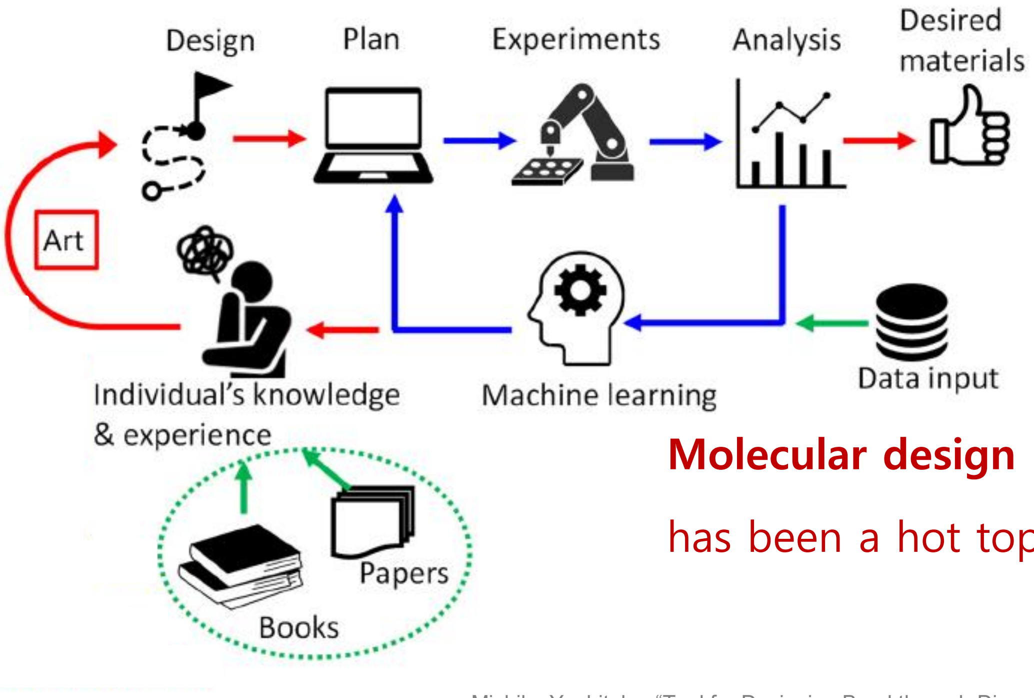
Samsung Advanced Institute of Technology

Novel Molecular Generation

- **Material (molecular) discovery?**
  - Find a **novel structure** which has **desired** physical and chemical **properties**

- **Domain knowledge based Molecular generation (design)**
  - High dependency to human knowledge (experience)

  → Bad Bias leads to a wrong structure and low diversity. + high time-consuming



Molecular design using ML

has been a hot topic recently

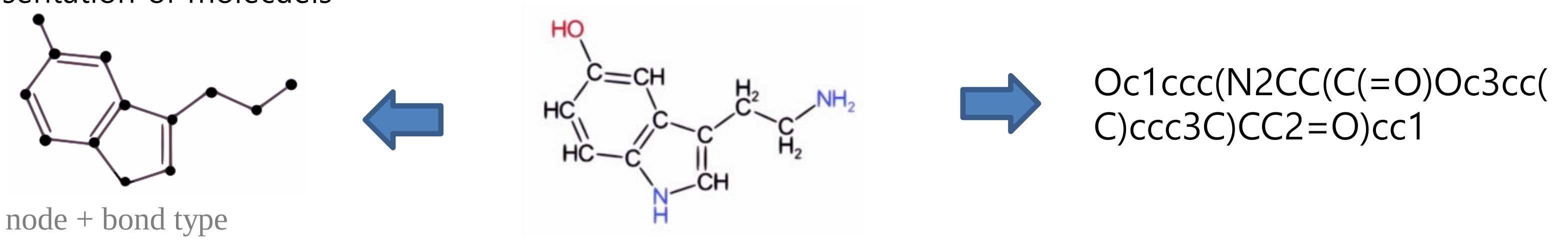
Michiko Yoshitake, "Tool for Designing Breakthrough Discovery in Materials Science", Materials, 2021

Novel Molecule Generation

- **Machine learning (deep learning) for molecular generation**
  - **Molecule representations** for ML: Graphs, **SMILES** (string type), Images…
  - **Algorithms** : GANs, **VAEs**, flow-based, score-based, **diffusion**-based approaches, …
  - Need generation models with a high percentage of **vaildity, novelty, uniqness**

★ The ability to generate **out-of-distribution (OOD) domain's structure** is essential.

Representation of molecuels



node + bond type

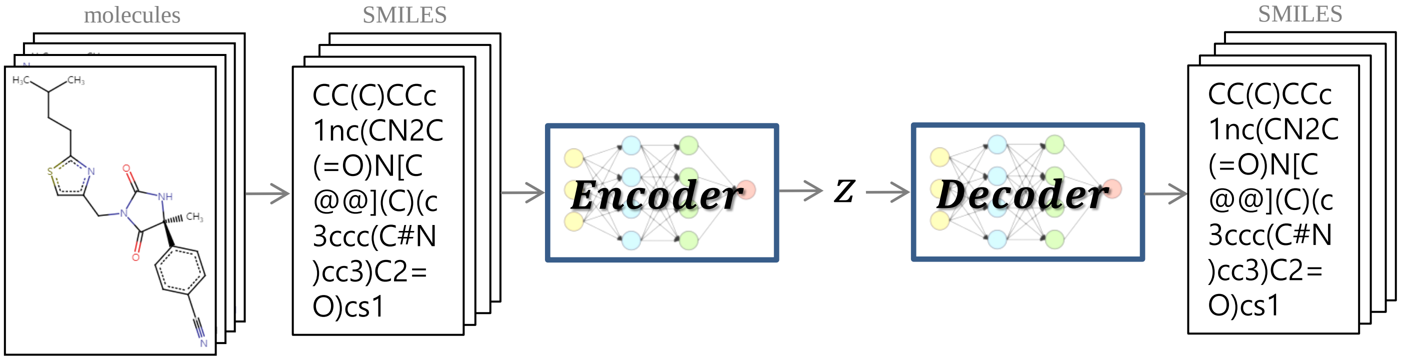
<Graph>

<molecule>

Oc1ccc(N2CC(C(=O)Oc3cc(
C)ccc3C)CC2=O)cc1

<SMILES>

(simple) VAEs for moleucles



molecules

SMILES

CC(C)CCc
1nc(CN2C
(=O)N[C
@@](C)(c
3ccc(C#N
)cc3)C2=
O)cs1

Encoder

z

Decoder

SMILES

CC(C)CCc
1nc(CN2C
(=O)N[C
@@](C)(c
3ccc(C#N
)cc3)C2=
O)cs1

Ensemble Method

- **Ensemble learning**

  "**Ensemble methods** use **multiple learning algorithms** to obtain better predictive performance than could be obtained from any algorithms alone." – Wiki.

  - Usually, **average predictions** or **bottleneck features** from multiple models

  - However, there is less research on ensembles in generative models

- **We propose appropriate ensemble techniques for generative models**

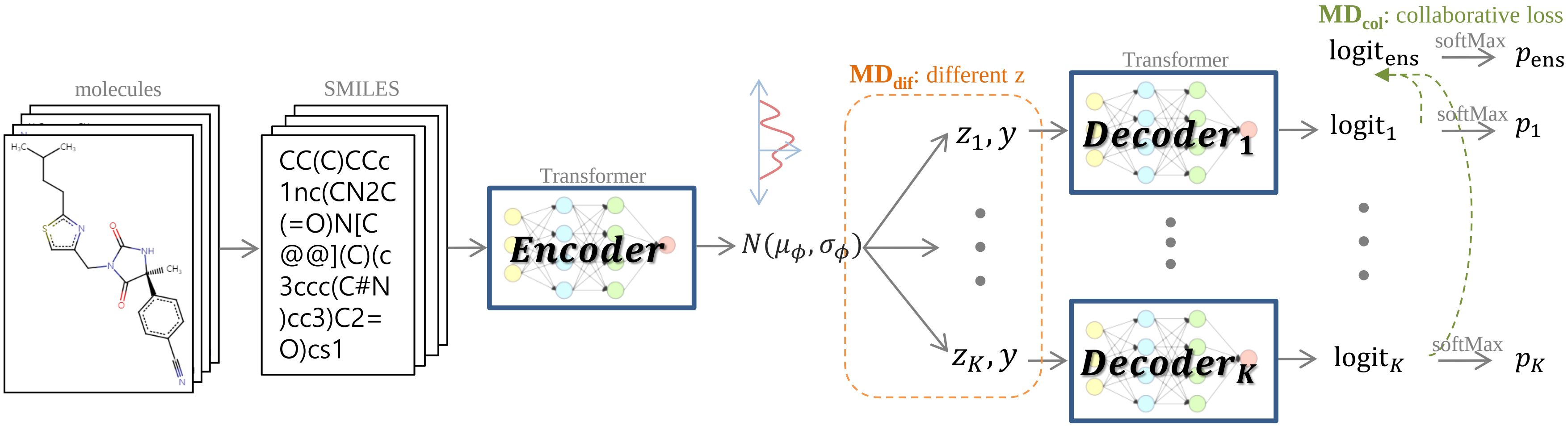
  - **Multi-decoder** structure with auto-regressive decoders

    + A **collaborative loss**

    + A simple way to **differentiate between decoders**

→ *It can be applied to other domains.*

Multi-decoder based VAE (1/2)



- **Multi-decoder (MD) Variational Autoencoder**
  - Multiple decoders are **trained simultaneously** with **shared single encoder**
  - **Ensemble logits*** of each decoder and generate **each string with auto-regression**

* In our experiments, the **ensemble on logits** showed better performance than those on **softmax(logits)**.

Multi-decoder based VAE (2/2)

- **Add two approaches for MD-VAE**
  - **Collaborative loss** ($\mathcal{L}_{col}$)
    - Cross-entropy loss of the ensembled logits
    - Apply with the previous reconstruction loss ($\mathcal{L}_{ind}$*)

$$\mathcal{L}_{col} = log \frac{1}{K} \sum_k p_{\theta_k}(x|y, z_k) \qquad \mathcal{L}_{ind} = \frac{1}{K} \sum_k log p_{\theta_k}(x|y, z_k)$$

* Cross-entropy of the each decoder's logits

  - **Different latent variables** for each decoder
    - Each decoder has different z_k → **strengthen each decoder's specialty**

$$z_k \sim \mathcal{N}(z_k | \mu_\phi(x, y), diag(\sigma_\phi(x, y)))$$

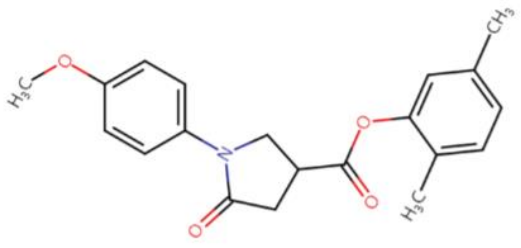
- **Training loss of the proposed method**

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{col} + \alpha \mathcal{L}_{ind} + KLD(q_\phi(\cdot\,|x, y)||p(\cdot))$$

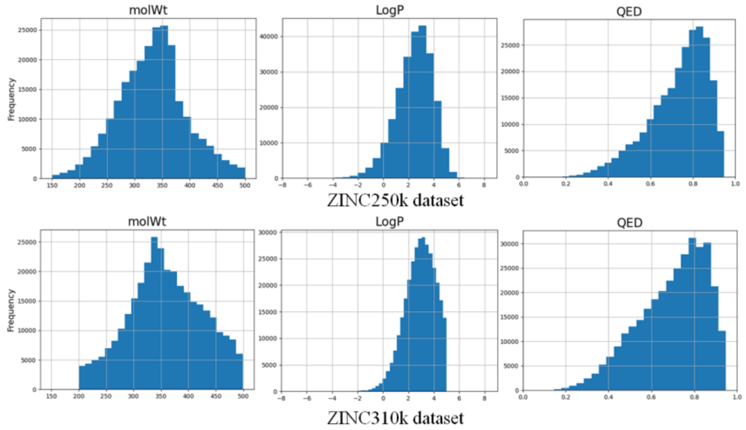
Experimental Result (Reconstruction)

- **Dataset: ZINC-250k DB (training), ZINC-310k DB (evaluation)**
    - Organic molecules, drug-like molecules
    - Input: SMILES, Ouput: 3-properties (continuous values)
    - Properties
        - molWt (molecular weight), LogP (partition coefficient), QED (quantitative estimation of drug-likeness)

- **Back-bone**: transfomer based conditional VAE + controlVAE*



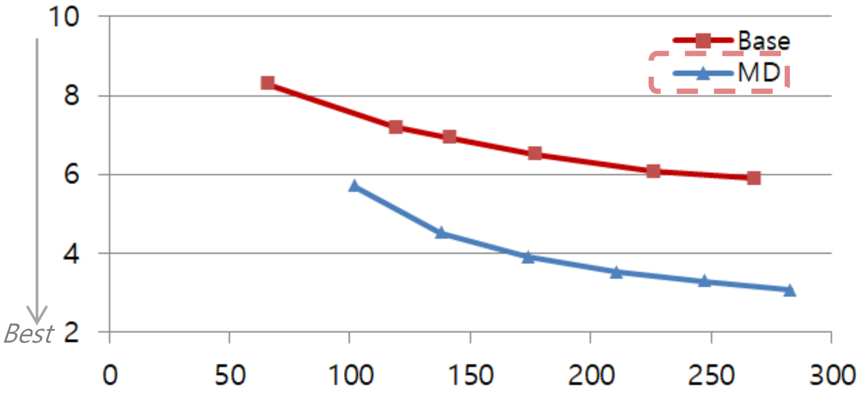
(a) An example of SMILES in ZINC DB:
COc1ccc(N2CC(C(=O)Oc3cc(C)ccc3C)CC2=O)cc1



(b) Distributions of 3-property

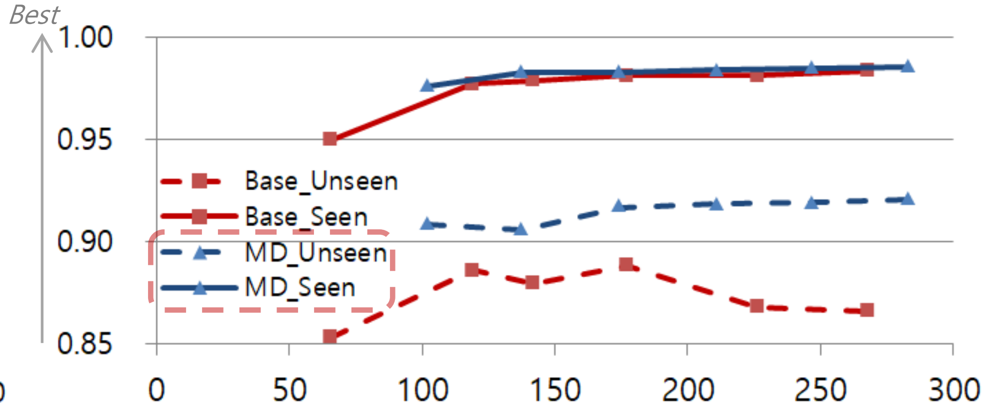
*Huajie Shao, et al., "Controlvae: Controllable variational autoencoder. International Conference on Machine", ICML, 2020

Experimental Result (Reconstruction)



(c) **Reconstruction loss**: x-axis=model size (mb).
In case of MD, #decoder has increased (3~7)

*140mb: relative reduction 36.2%



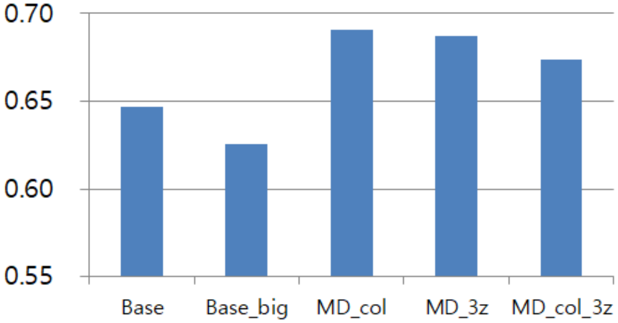
(d) **Reconstruction success rate**: x-axis=model size (mb)
Seen DB=ZINC250k, **Unseen DB=ZINC310k**

*Unseen case: relative improvement 4.8%

Model	model size	Recon. Loss	KL Loss	Reconstruction success rate (Unseen)
Vanilla VAE	142MB	17.276	0.000	0.783
Control VAE (Base)	142MB	6.851	15.168	0.880
3-Decoder	138MB	7.001	15.207	0.898
3-Decoder+collaborative	138MB	5.508	14.937	0.891
3-Decoder+different z	138MB	6.555	15.145	0.902
3-Decoder+collaborative +differenct z	138MB	**4.482**	15.068	**0.909**

Experimental Result (Generation)

- **Generative efficiency** **(validity, novelty, uniqueness)**
  - Target: **out-ouf-distribution (OOD) conditions**
  - Using out of property-range of training DB as a generative condtions of cVAE
  - 10k generative tries per property



(a) Molecular Generative Efficiency (%)

*relative improvment 9.3%

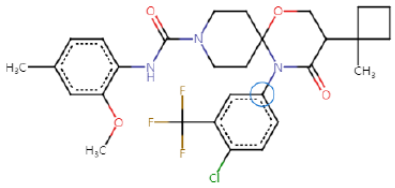
*RDkit calculations

- **Conditions satisfaction** **(Top1 molecule, absolute error)**
  - Difference between ground-truth* and generative conditions

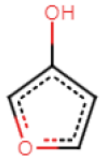
	In-domain Condition			Out-of-distribution Condition		
	molWt	**logP**	**QED**	**molWt**	**logP**	**QED**
Control VAE (Base)	0.1520	0.0008	0.0041	0.0800	1.3598	0.0008
MD	**0.0940**	**0.0003**	**0.0040**	0.1740	**0.0204**	0.0015
MD_{col}	**0.0497**	0.0013	0.0042	**0.0760**	**0.0069**	**0.0002**
MD_{dif}	**0.0797**	**0.0007**	**0.0041**	**0.0470**	**0.0003**	**0.0002**
$MD_{dif,col}$	**0.0513**	**0.0004**	**0.0041**	**0.0620**	**0.0013**	**0.0006**

Experimental Result (Generation)

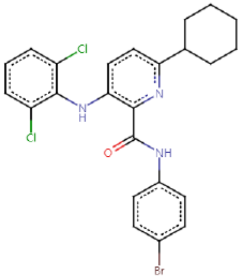
- **Examples of generated molecules**
  - Each **condition value** is $\boldsymbol{\mu \pm 3\sigma}$ of the properties of **ZINC-250k DB**



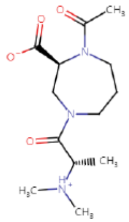
molwt: 580 → 580.047



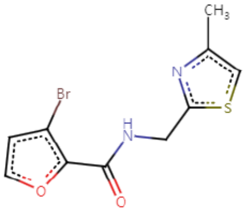
molwt: 84 → 84.074



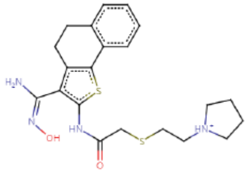
LogP: 8.194 → 8.1945



LogP: −3.281 → −3.2813



QED: 0.9598 → 0.948094



QED: 0.1775 → 0.177306

*property **name**: condition **value** → generated **molecule's property** (by RDkit)*

- Property value range of ZINC-250k DB

Property	Value	
molwt	Max	**500.00**
	Min	**150.12**
LogP	Max	8.252
	Min	-6.876
QED	Max	0.9484
	Min	0.1166