KIAS Center for AI and Natural Sciences Fall Workshop



Autonomous design of products and processes: Generative Chemical Transformer to reinforcement learning-guided combinatorial chemistry

2024-11-05

Prof. Jonggeol Na

Department of Chemical Engineering and Materials Science

Ewha Womans University

jgna@ewha.ac.kr

EWHA WOMANS UNIVERSITY

Na Research Group

Advanced Manufacturing Systems Laboratory



"Space", as Douglas Adams famously said "is big. You just won't believe how vastly, hugely, mind-bogglingly big

1,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

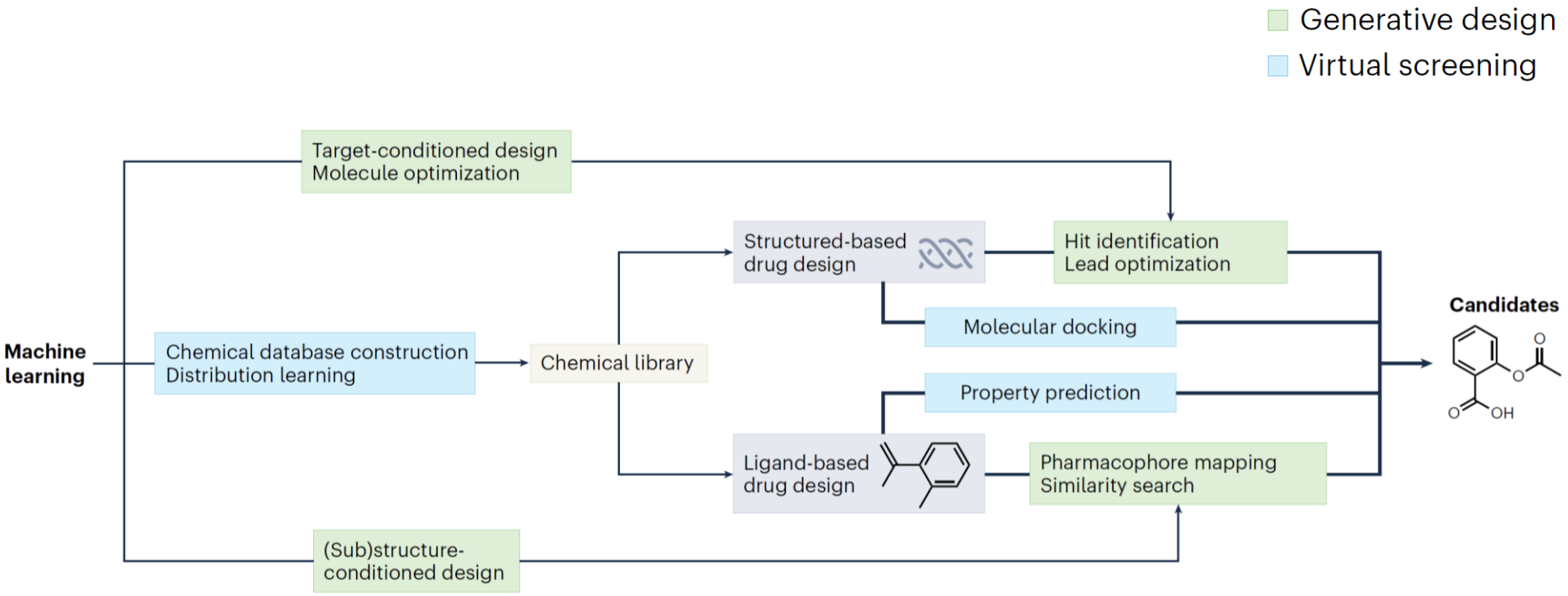
큰 수의 단위				
아승기(阿僧祇)	10,000 배 →	**나유타(那由他)**	10,000 배 →	불가사의(不可思議)

那由他 / 那由多 / Novemdecillion

1,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000 = 10^{60}

Kirkpatrick, P., & Ellis, C. (2004). Chemical space. *Nature*, *432*(7019), 823-824.

Generative ML-assisted molecular design pipeline



Machine learning-aided generative molecular design. ***Nature Machine Intelligence***, (2024), 1-16.

AI-Based Generative Molecular Design Methods

CVAE, GraphVAE, JT-VAE, DVAE, RationaleRL, ChemSpacE, QMO, GraphNVP, AAE, ORGAN, MolGAN, MolecularRNN, ChemTS, REINVENT, DeLinker, GraphAF, GraphEBM, DiGress, GCPN, GB-GA, GA+D, SynNet, AutoGrow 4.0, Reinforced GA, MIMOSA, GFlowNet, GraphDG, GeoMol, E-FM, EquiBind, Torsional Diff, DiffDock, liGAN, G-SchNet, 3D-SBDD, 3D-Scaffold, GraphBP, E-NF, EDM, EquiFM, DiffSBDD

+and a lot…

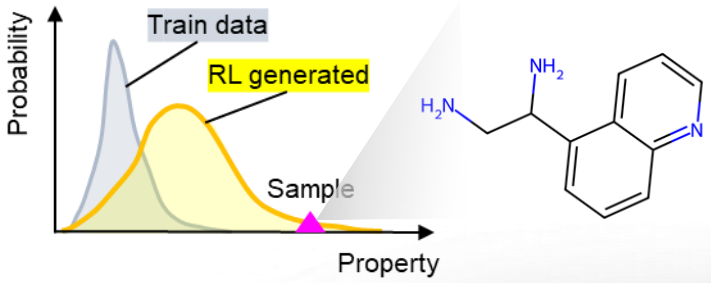
What is your “IDENTITY”?

Autonomous Process/Product Design

Development of Artificial Intelligence-based Design Framework
To Design Techno-economic and Environmental-sustainable
Products and Processes Simultaneously

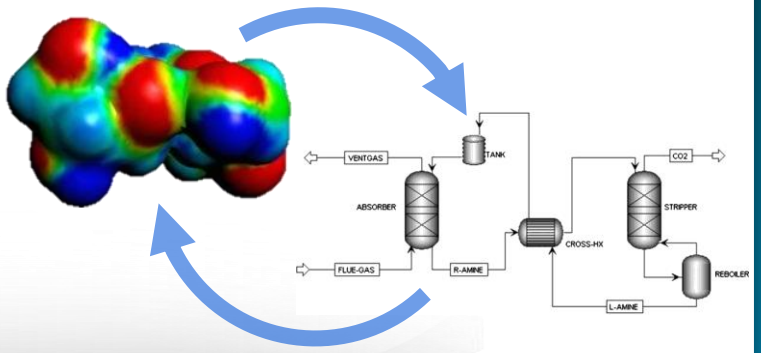
I. Material Inverse Design

- Reinforcement learning-guided combinatorial chemistry for extreme property
- Active-learning for DFT sampling



II. Optimal Process Design

- Closed-loop optimization through derivative-free optimization (or Bayesian optimization)
- Materials-Process multi-scale mapping tech.



Move traditional inefficient and non-intuitive sequential design procedure to
Simultaneous design space → "Design Paradigm Shift in Process Systems"

ChatGPT for
Molecule Product Design



Jonggeol Na, Won Bo Lee et al.
J. Chem. Inf. Model 2021 (Front Cover)

Inverse Design for
Unknown Chemical Space

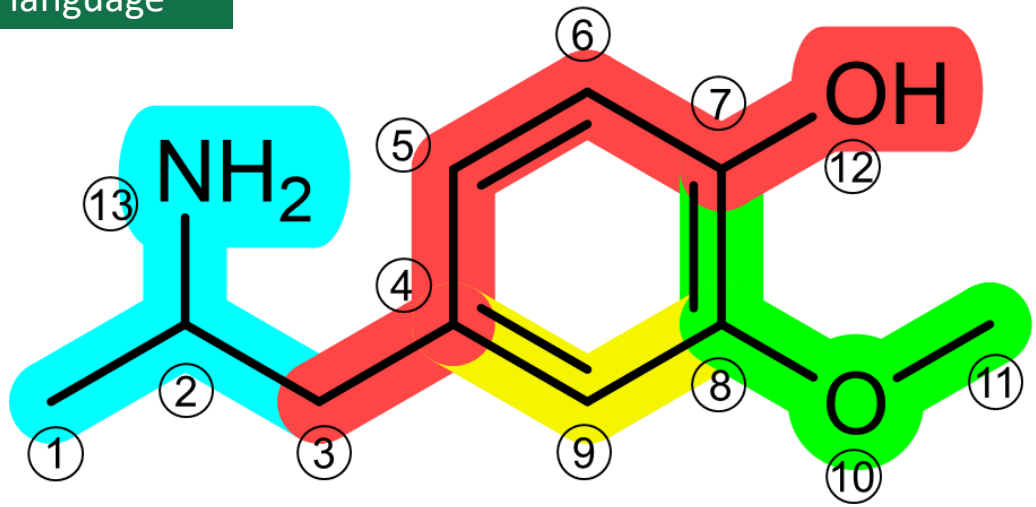


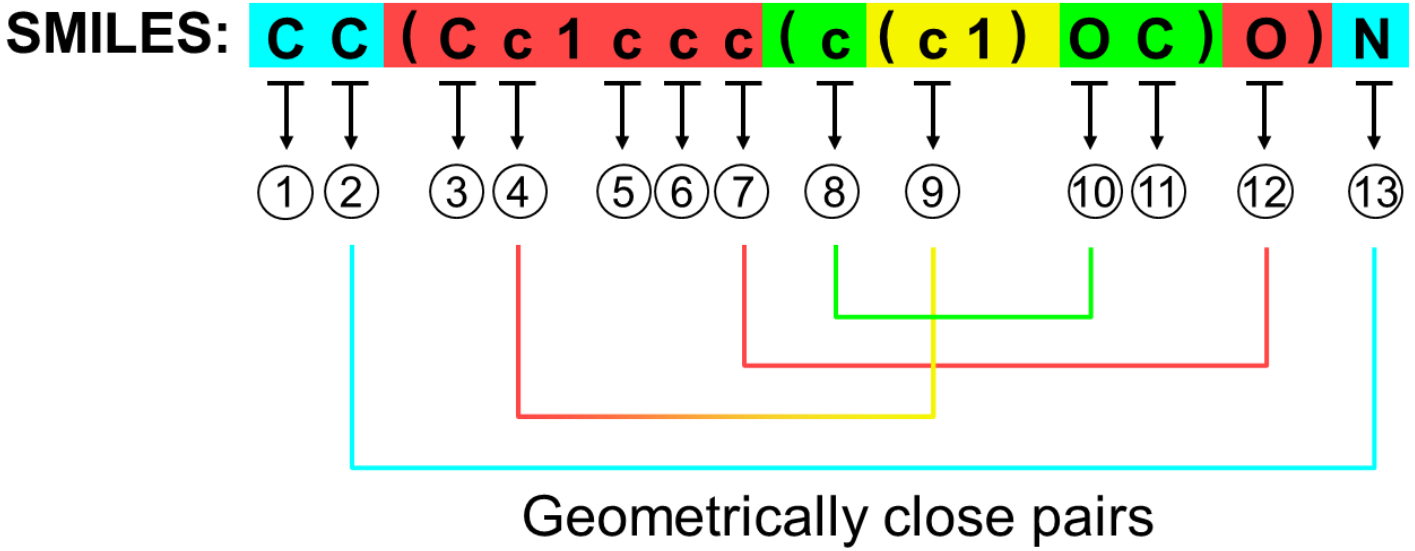
Won Bo Lee, Jonggeol Na et al.
Chemical Science 2024 (Front Cover)

Other Related Articles (Lead Author): *Journal of Materials Chemistry A 11(19), (2023), ACS Catalysis, 11(5), (2021), Physical Chemistry Chemical Physics, 22(35), (2020).*

Characteristics of Molecule Structure

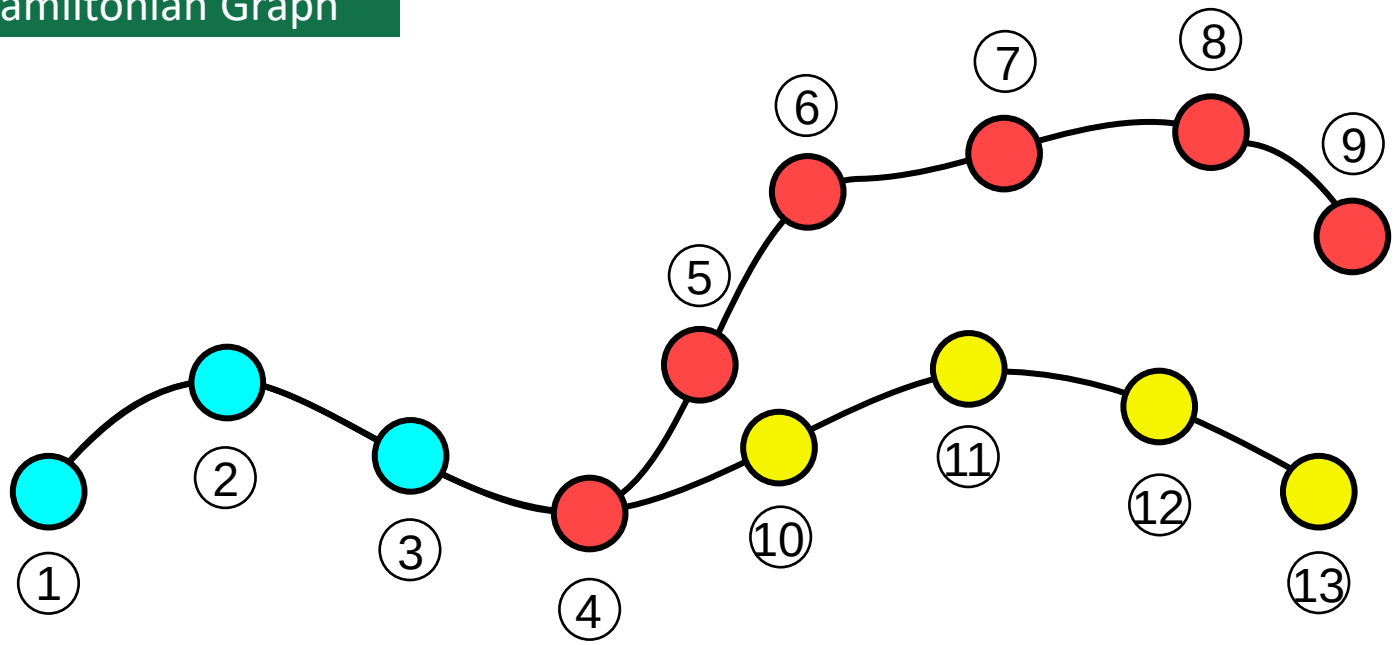
SMILES: artificial language





Characteristics of Molecule Structure

Non-Hamiltonian Graph

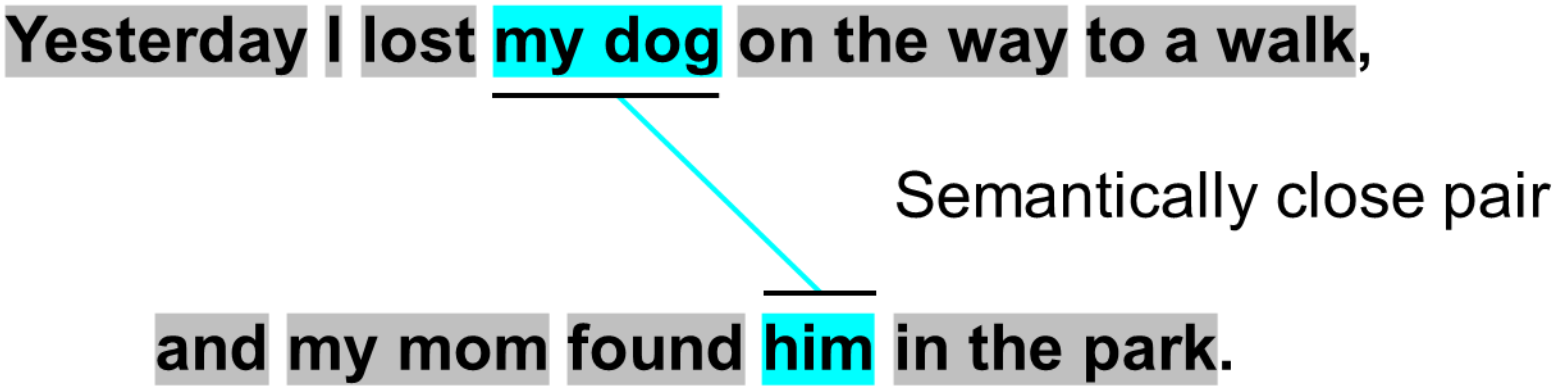


String: 1 2 3 — 4 5 6 7 8 9 — 10 11 12 13

Semantic discontinuity

Characteristics of Molecule Structure

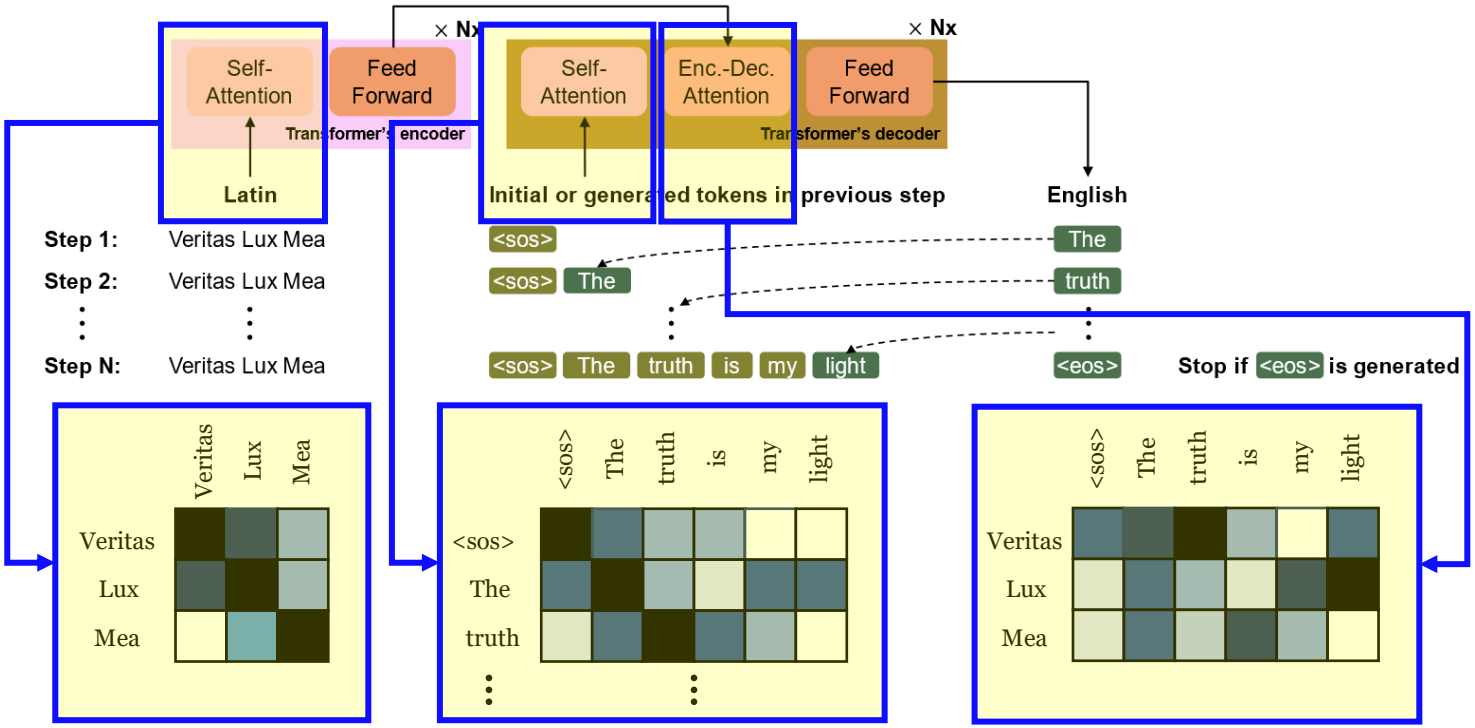
Natural Language



"Let's use Natural Language Processing to teach Chemical Grammar to AI"

Transformer: Attention is All You Need

Google

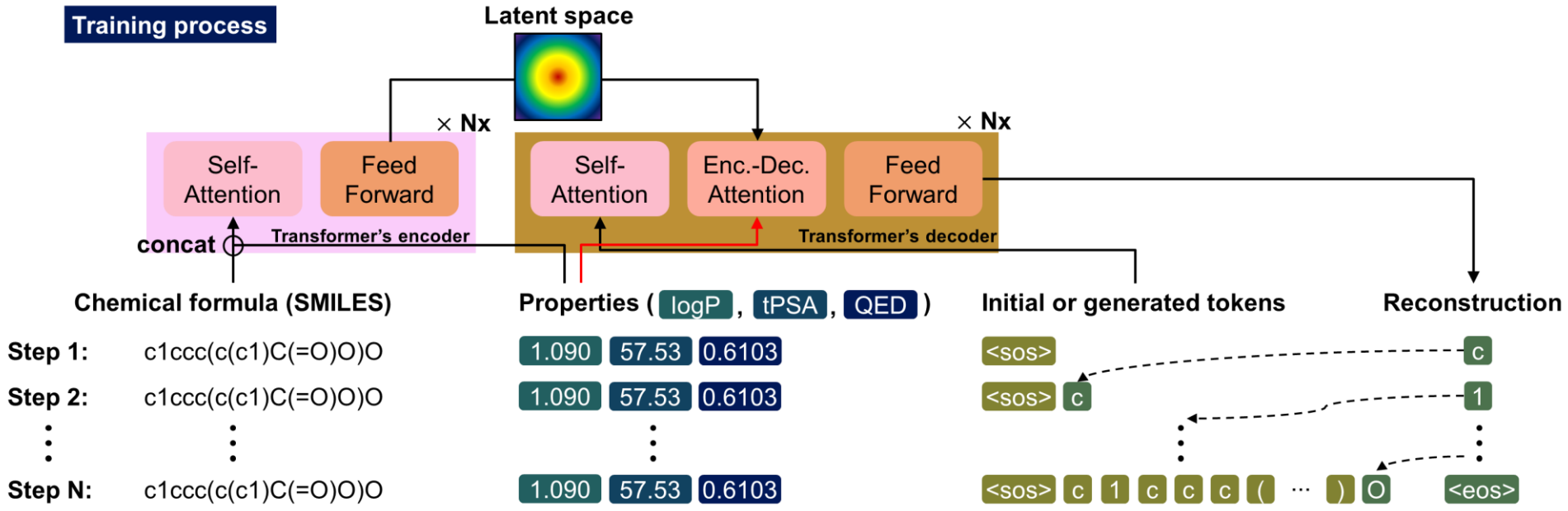


Transformer model for Neural Machine Translator (NMT): an example of translating Latin into English.

Vaswani, Ashish, et al. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017). **citation: ~139601 times**

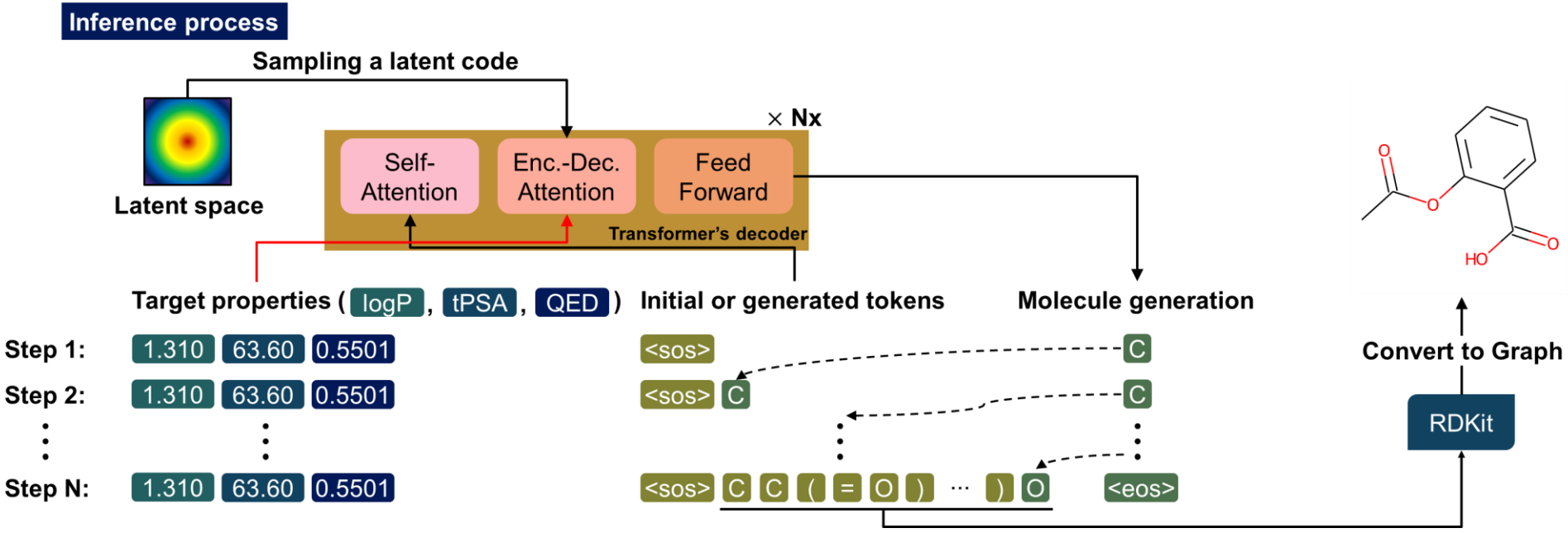
Generative Chemical Transformer

Na Research Group
Advanced Manufacturing Systems Laboratory



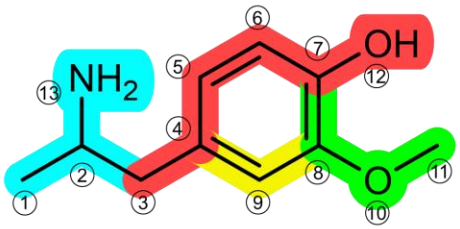
logP: Wildman-Crippen partition coefficient
tPSA: topological polar surface area
QED: drug-likeness

Generative Chemical Transformer

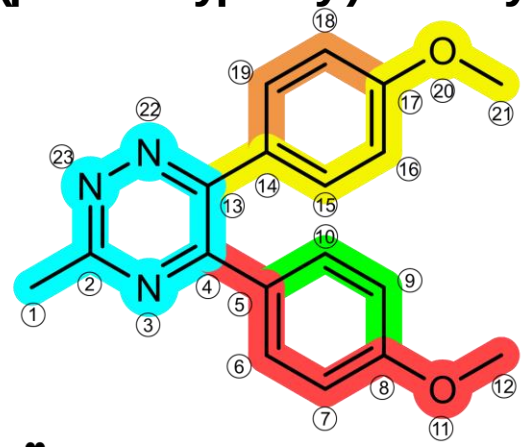


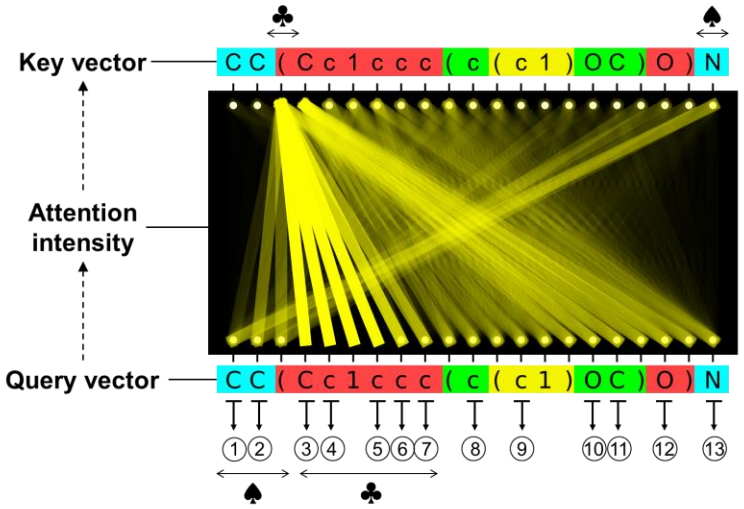
Visualization of self-attention scores

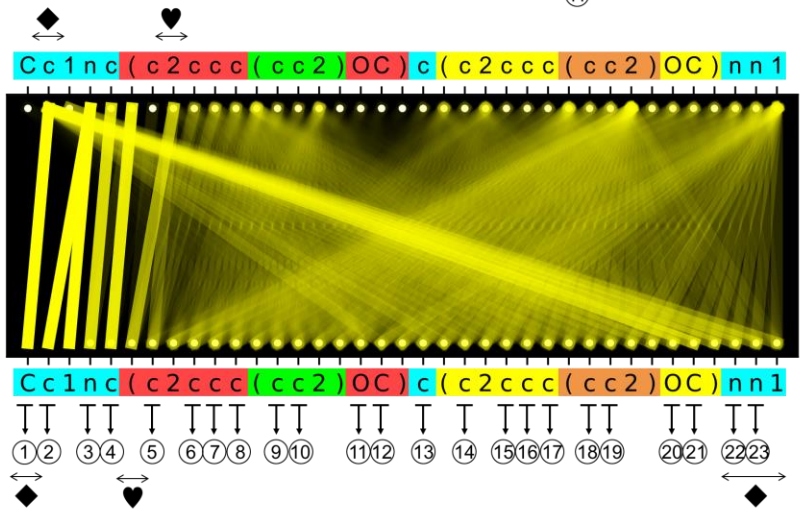
4-(2-aminopropyl)-2-methoxyphenol



5,6-bis(p-methoxyphenyl)-3-methyl-1,2,4-triazine

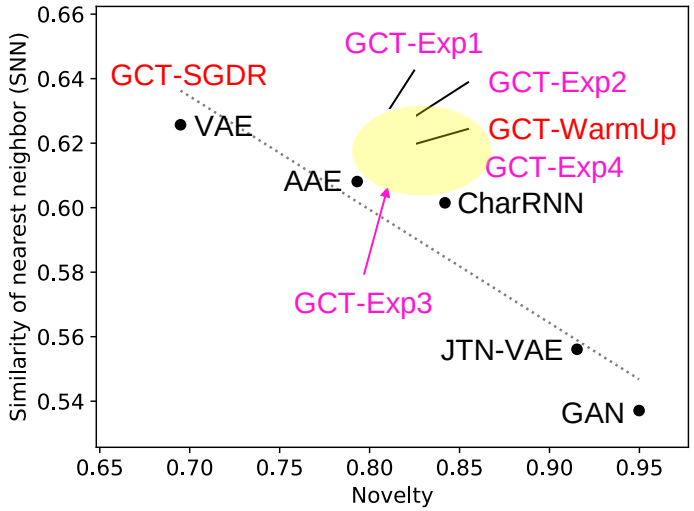


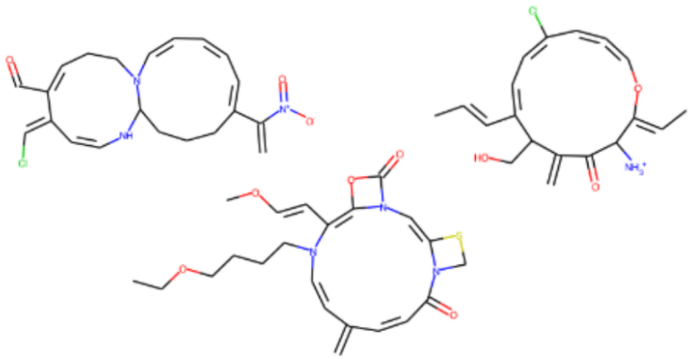




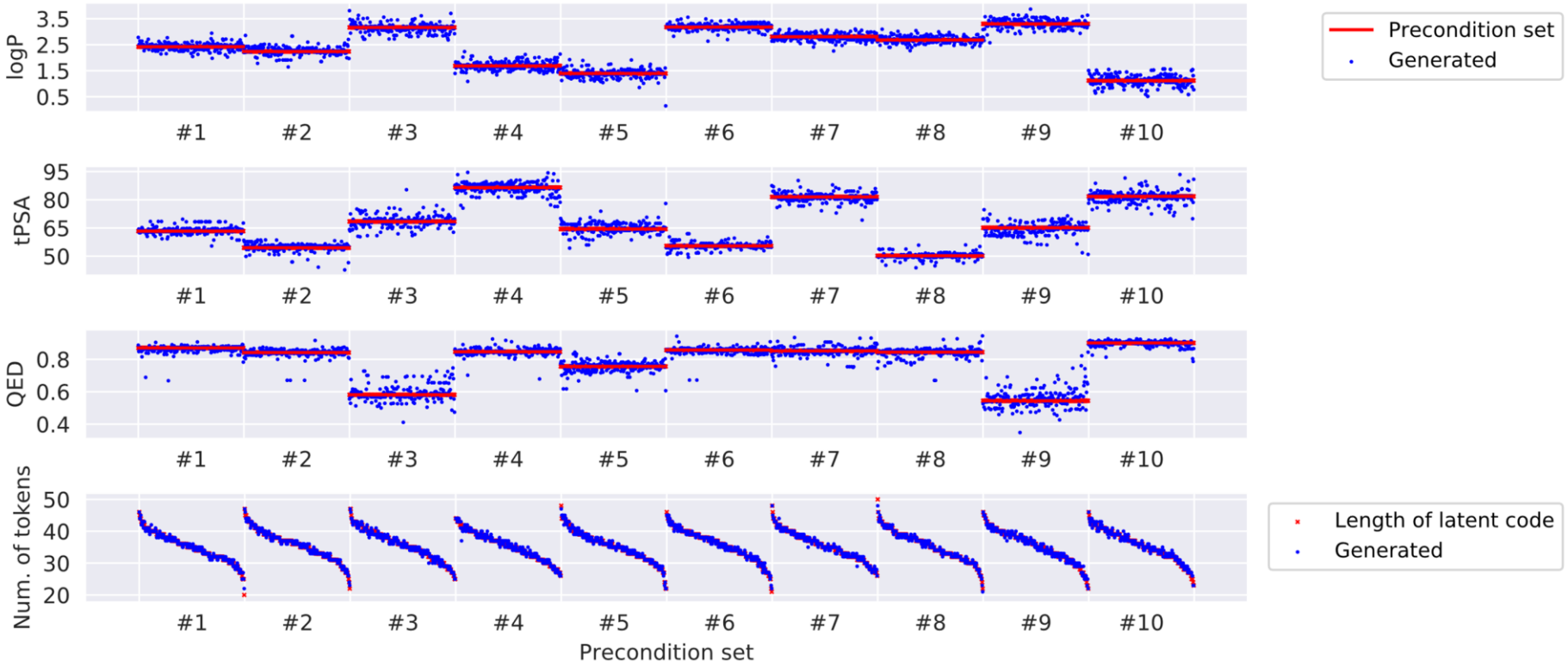
The stronger attention given through the important "Chemistry"

Results of Comparative Study





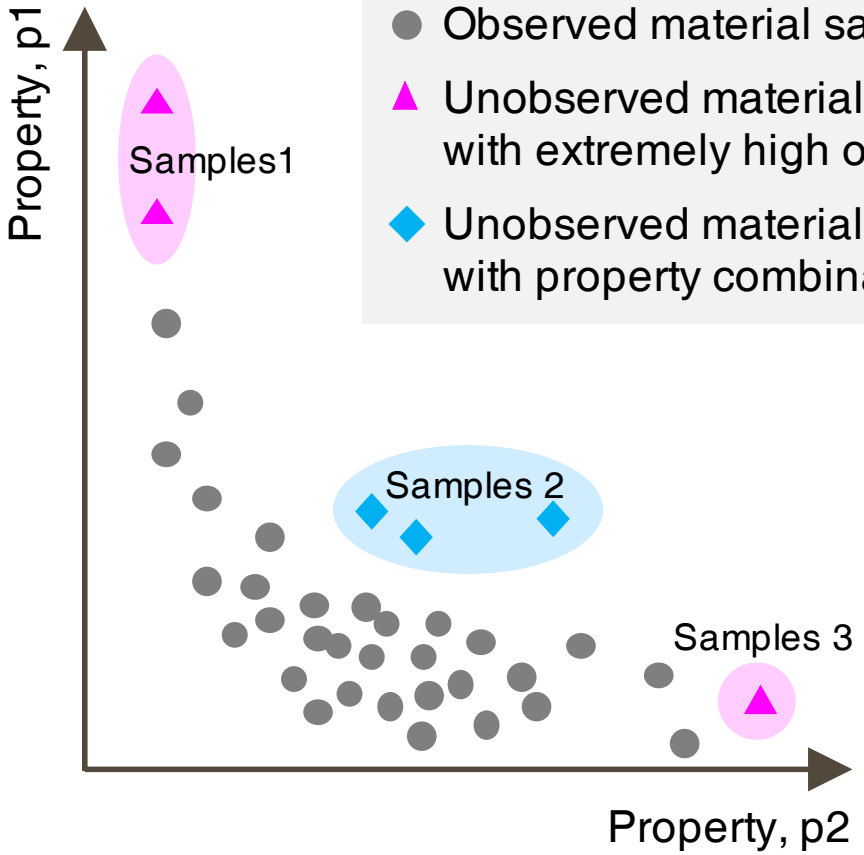
Multiple Molecules for Single Target Set



Possible to generate multiple de novo molecule designs!

Fragment-based RL for Molecular Discovery

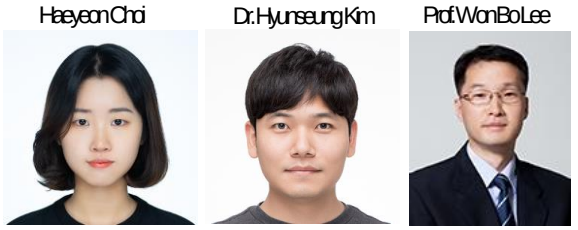
✓ *Supervised learning-based model cannot generate molecules which have unobserved properties*



- Observed material samples $\sim P_{data}$
- Unobserved materials with extremely high or low properties
- Unobserved materials with property combinations, hard to co-exist

$\sim true\ P$

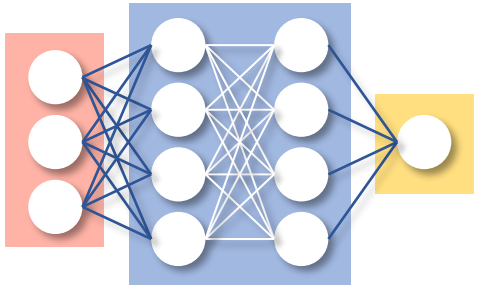
$$P_{data} \neq true\ P$$

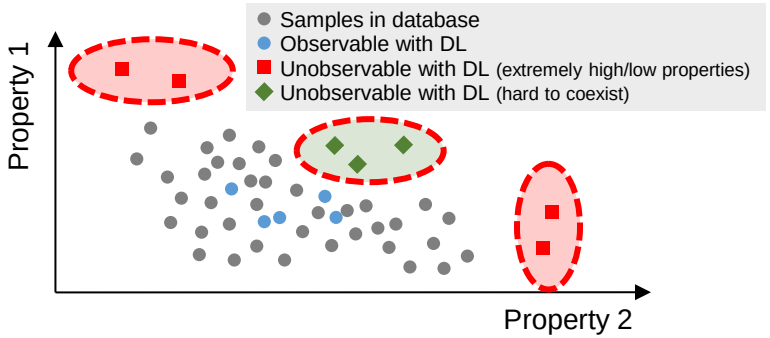


Inverse Design: Previous Research

Necessity of inverse design using Reinforcement Learning (RL)

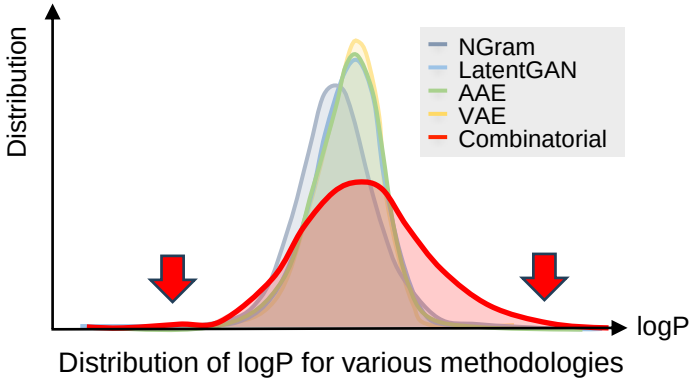
Deep Learning-based Models



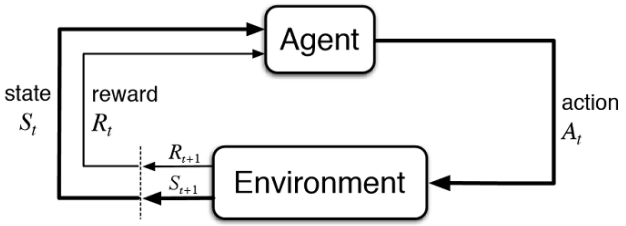


✓ **Impossible/Hard to generate molecules with extreme properties**

Combinatorial Chemistry + Reinforcement Learning



Distribution of logP for various methodologies

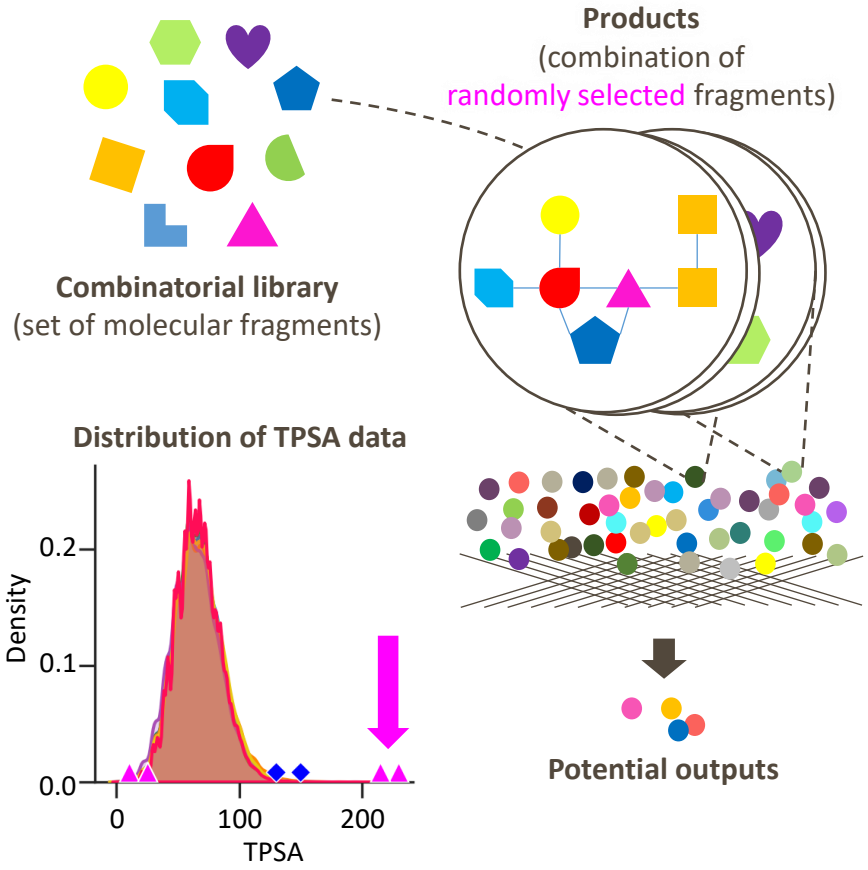


✓ **Molecules with extreme properties**
✓ **Learn policy for molecule generation**

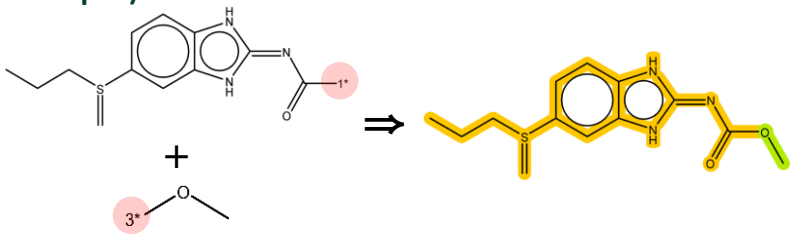
H. Kim, **H. Choi**, D. Kang, W. Lee, **J. Na**.*, *Chem. Sci.*, 2024,15, 7908-7925

Fragment-based RL for Molecular Discovery

Combinatorial Chemistry (notarized in 1982)



Example)



Rule-based ⇒ **100% valid**
Randomly select fragment ⇒ inefficient

Pros:

- able to generate any molecules physically possible, out-of-training-data-distribution

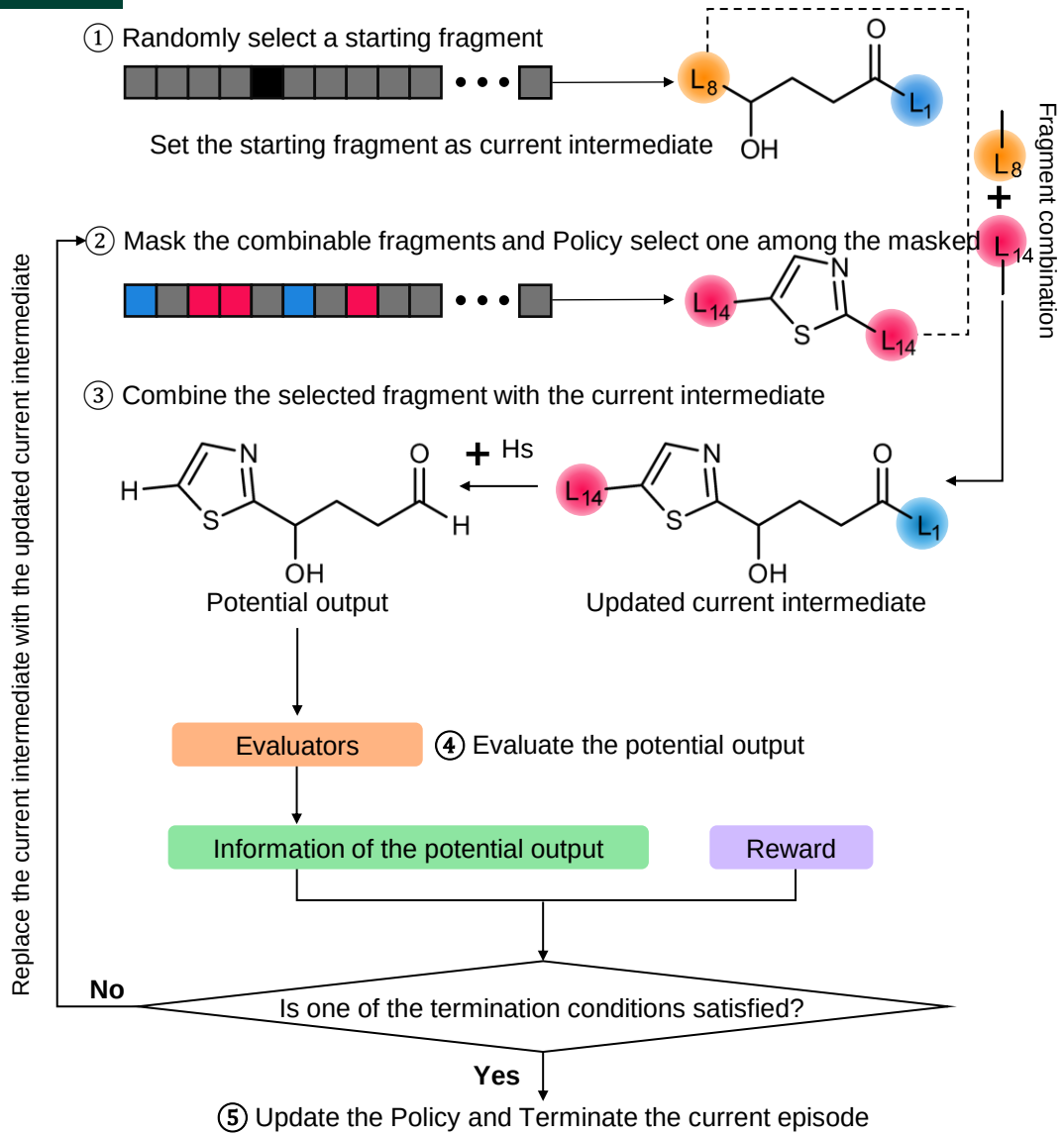
Cons:

- lack of policy toward the given targets

APPLY RL to Combinatorial Chemistry to get AI-Driven Fragment Selection POLICY!

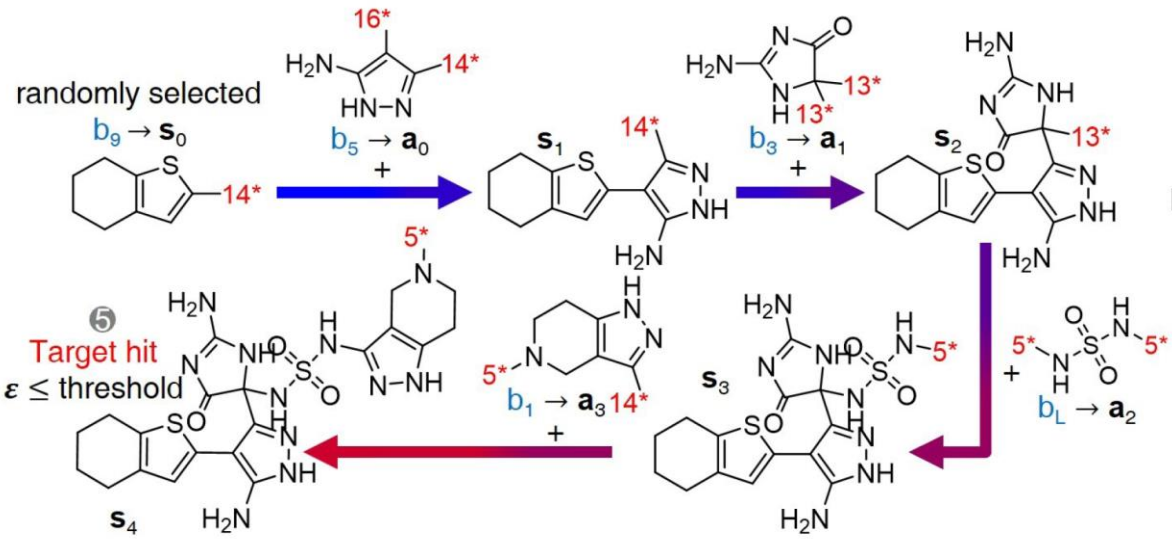
Fragment-based RL for Molecular Discovery

RL Training Loop

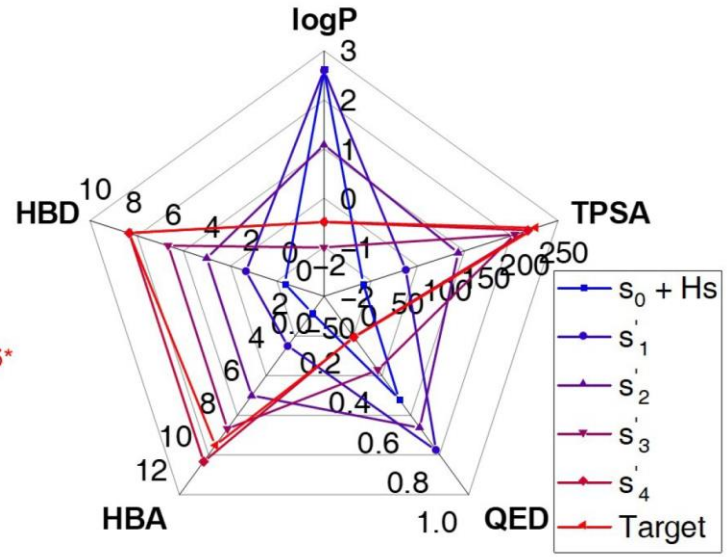


Fragment-based RL for Molecular Discovery

Target Molecule Generation



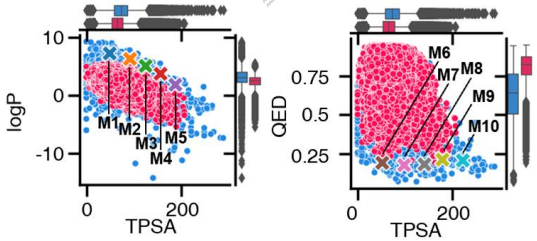
Target Tracing



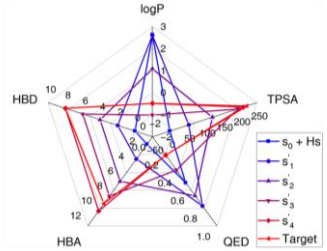
Computational Experiments

RL-CC: Reinforcement Learning-Driven Combinatorial Chemistry

Multi-target properties

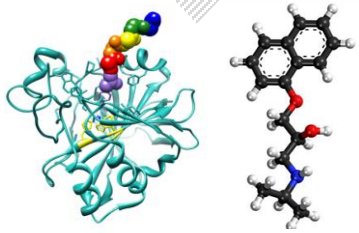


- Generate molecules with extreme properties, which deviate from dataset
- Considered logP, TPSA, QED, HBA, HBD, MW, DRD2; Properties related to drug

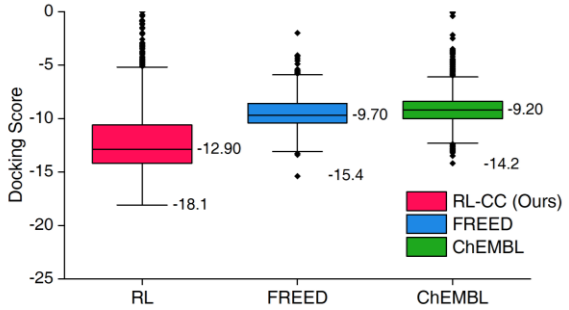


- Showed better performance than deep learning-based model on extreme points

Binding affinity towards 5-HT$_{1B}$

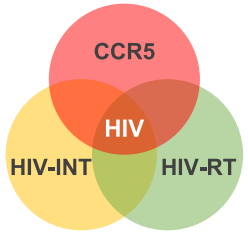


- Generate molecules with low binding affinity towards 5-HT$_{1B}$ receptor



- Showed better performance than existing reinforcement learning-based model

HIV-related target properties



- Generate molecules with low IC_{50} value for 3 HIV-related targets (CCR5, HIV-INT, HIV-RT)

Method	CCR5	HIV-INT	HIV-RT
ENAMINEBB	8.63	6.71	7.49
RS	8.79	7.25	7.65
GCPN	8.20	6.45	7.42
JT-VAE	8.15	7.25	7.58
MSO	8.68	7.28	7.76
PGFS	9.05	**7.5**	7.89
RL	**9.11**	7.17	**8.01**

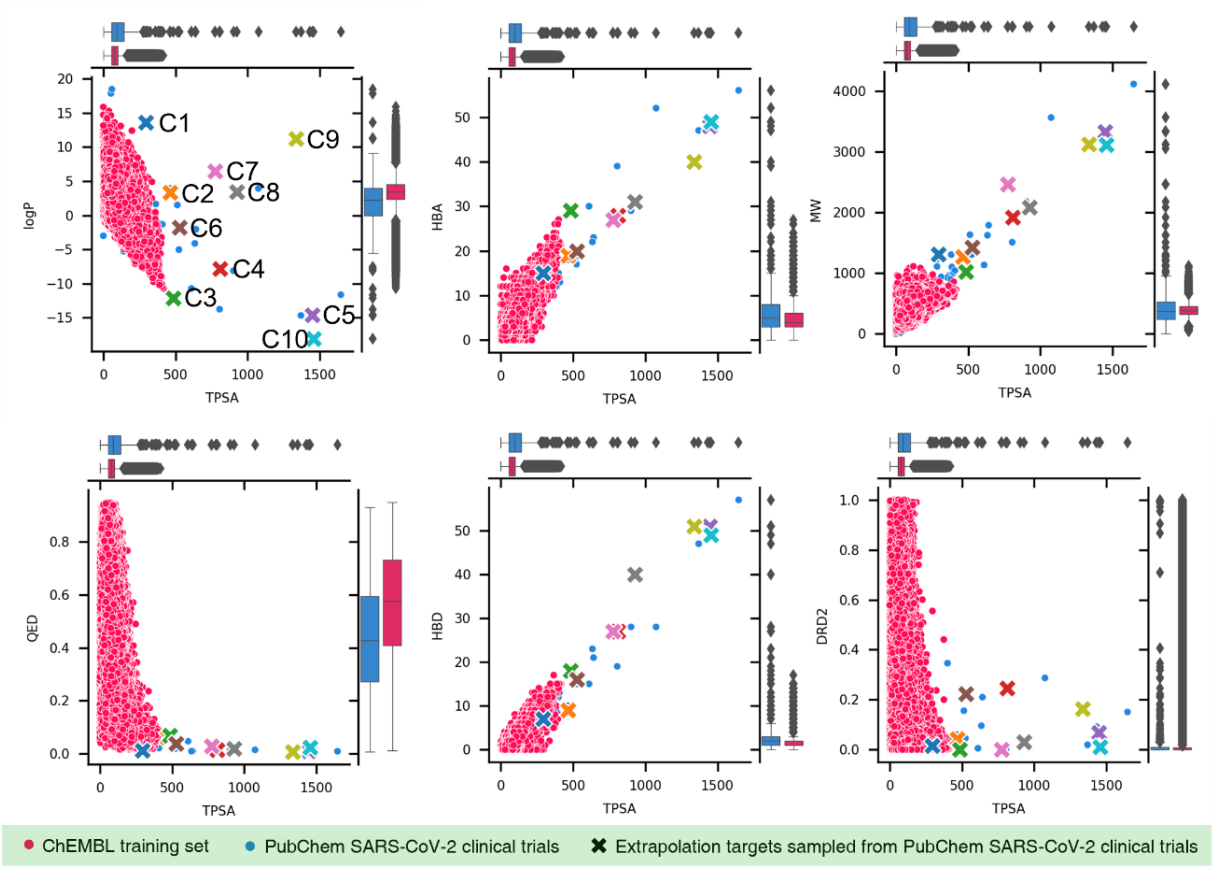
- Showed better performance than any other models in two properties(CCR5, HIV-INT), and similar performance in one property(HIV-RT)

Real-world Problem

Fundamental understanding

Multi-Target Properties: Drug Discovery

Reference Data and Targets with Extreme Properties (Seven-Auxiliary Indicators for Drug Discovery)



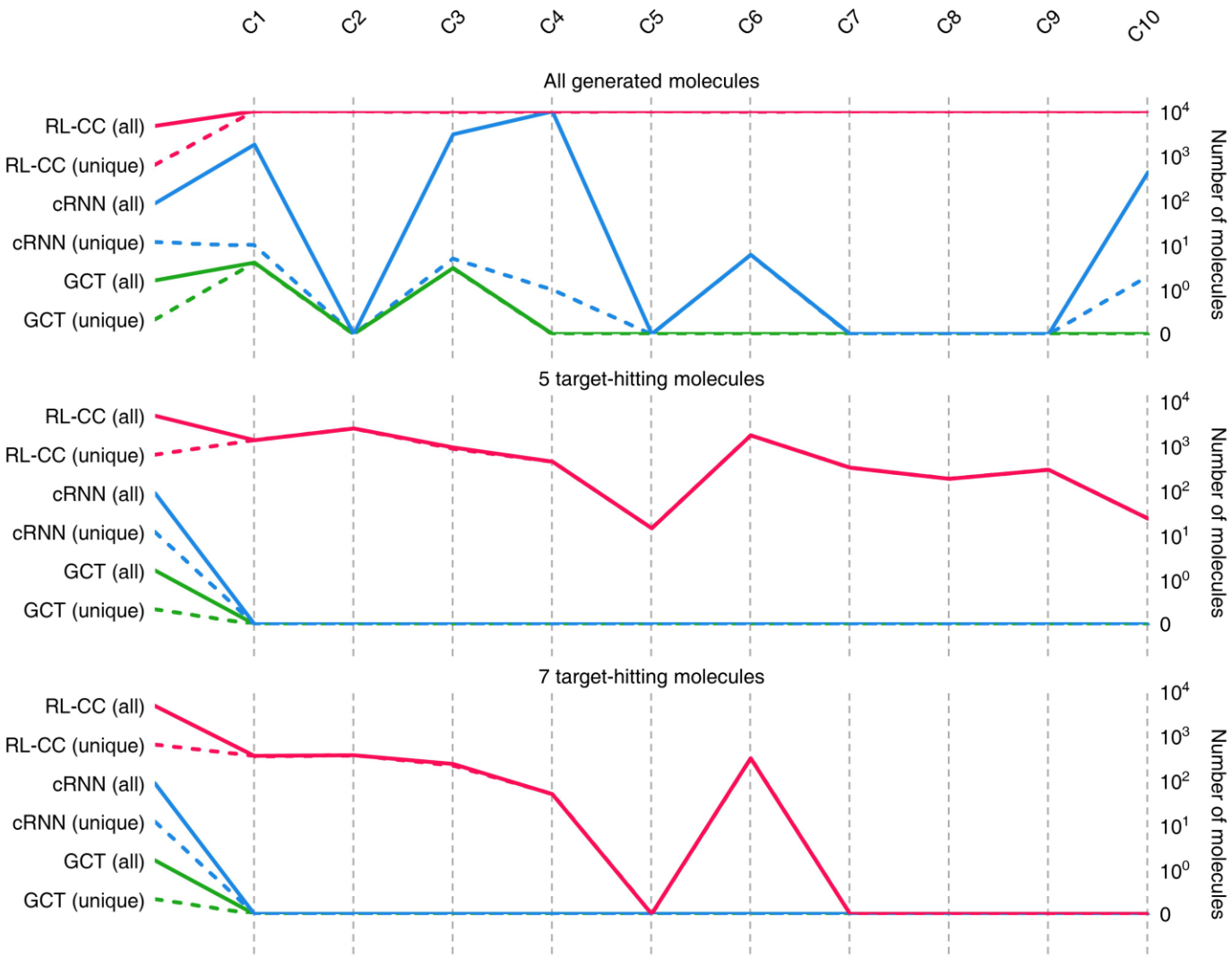
Performance evaluation for test set
(distributed in trained area)

	$RMSE_{i\in\{\text{logP,TPSA,QED,HBA,HBD,MW,DRD2}\}}$						
Model	logP	TPSA	QED	HBA	HBD	MW	DRD2
cRNN	0.379	5.476	0.081	0.932	0.223	5.954	0.113
GCT	0.368	5.109	0.075	1.204	0.247	8.272	0.098
Average	0.373	5.292	0.078	1.068	0.235	7.113	0.105

Averaged $RMSE_i$ are used for target bound of i

Results of Multi-Target Properties

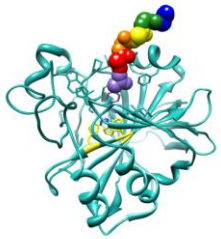
* cRNN: *Nature Machine Intelligence* 2020, 2, 254-265.
** GCT: *Journal of Chemical Information and Modeling* 2021, 61, 5804-5814.



"Only RL-CC can generate molecules which has unseen target properties"

Lead Molecules Discovery (Ligand-Protein Docking)

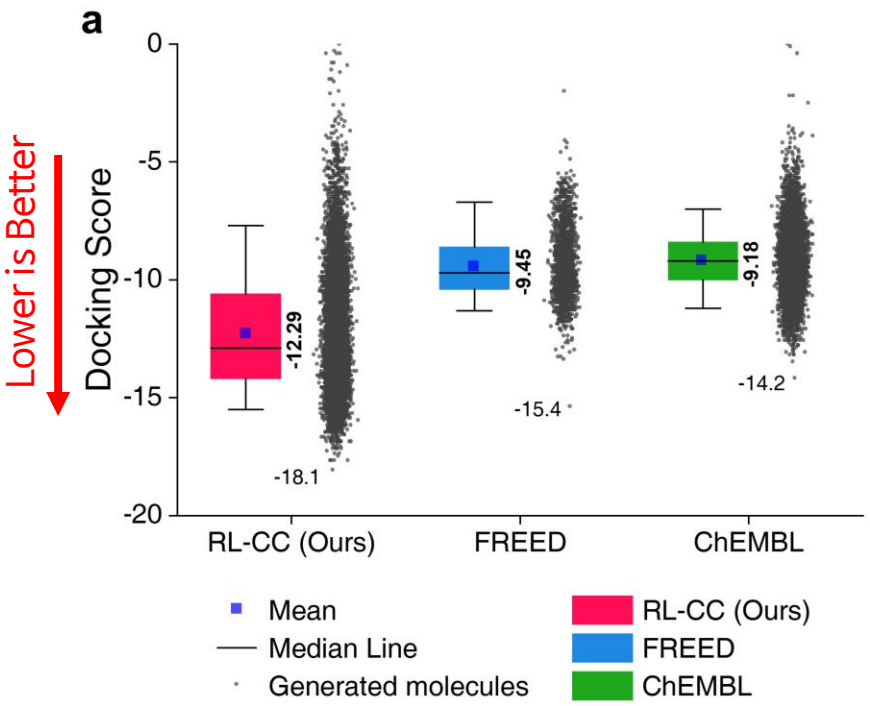
Application to the discovery of protein docking materials

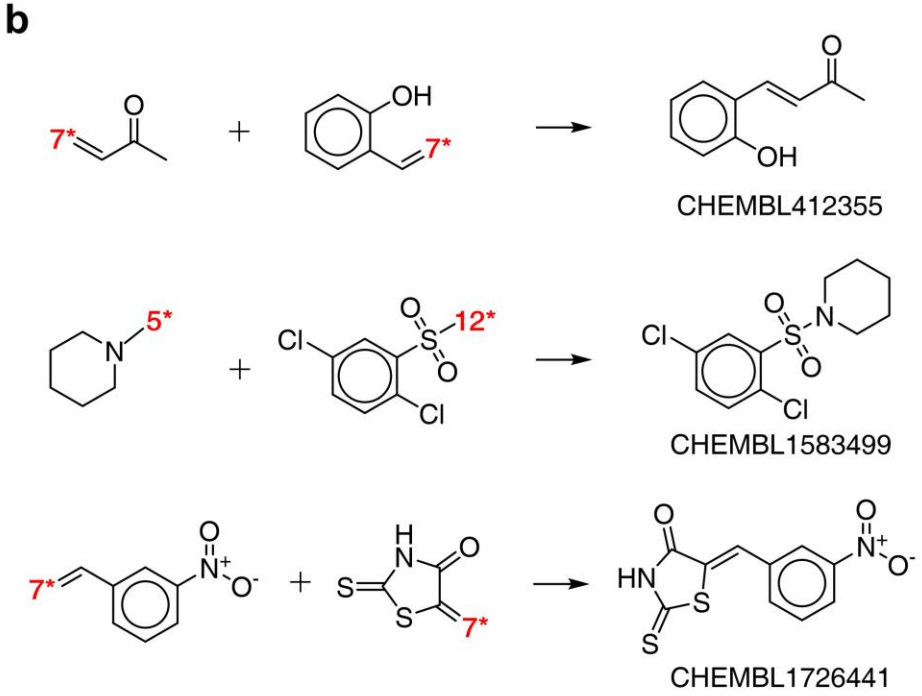


To find molecules with high binding affinity for 5-HT_{1B} receptor

QVina2 : docking score prediction
Docking score ∝ - (binding affinity)

Reward = - (docking score)



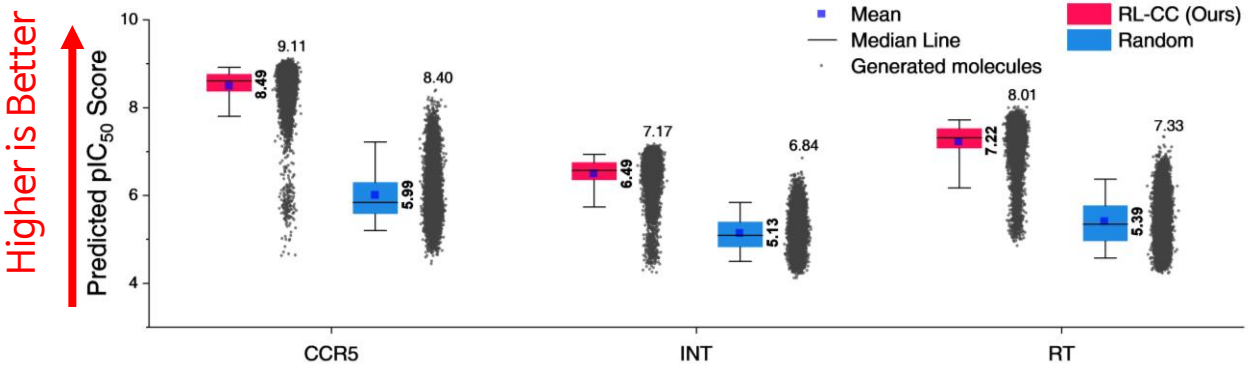


Lead Molecules Discovery (HIV-inhibitor)

Application to discovery of HIV-inhibitor

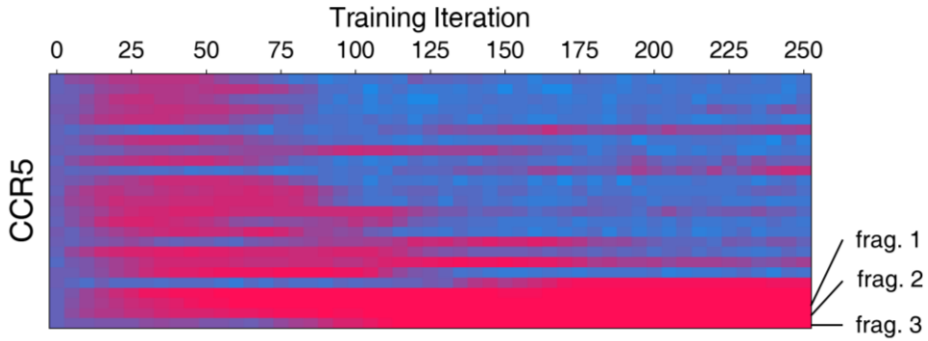
HIV-related target(CCR5, HIV-INT, HIV-RT)
pIC_{50}(=-logIC_{50}) prediction QSAR model:
* IC_{50}: half maximal inhibitory concentration

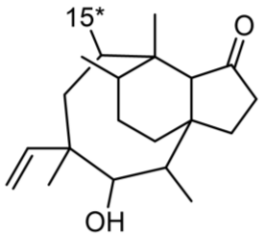
Reward = predicted pIC_{50}



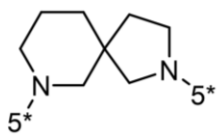
Method	HIV-RT	HIV-INT	HIV-CCR5
ENAMINEBB	7.49	6.71	8.63
RS	7.65	7.25	8.86
GCPN	7.45	6.45	8.62
JT-VAE	7.58	7.25	8.23
MSO	7.76	7.28	8.77
PGFS	7.89	**7.55**	**9.05**
This study	**8.01**	**7.17**	**9.11**

*Highest score *Second highest score

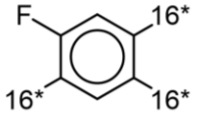




frag. 1

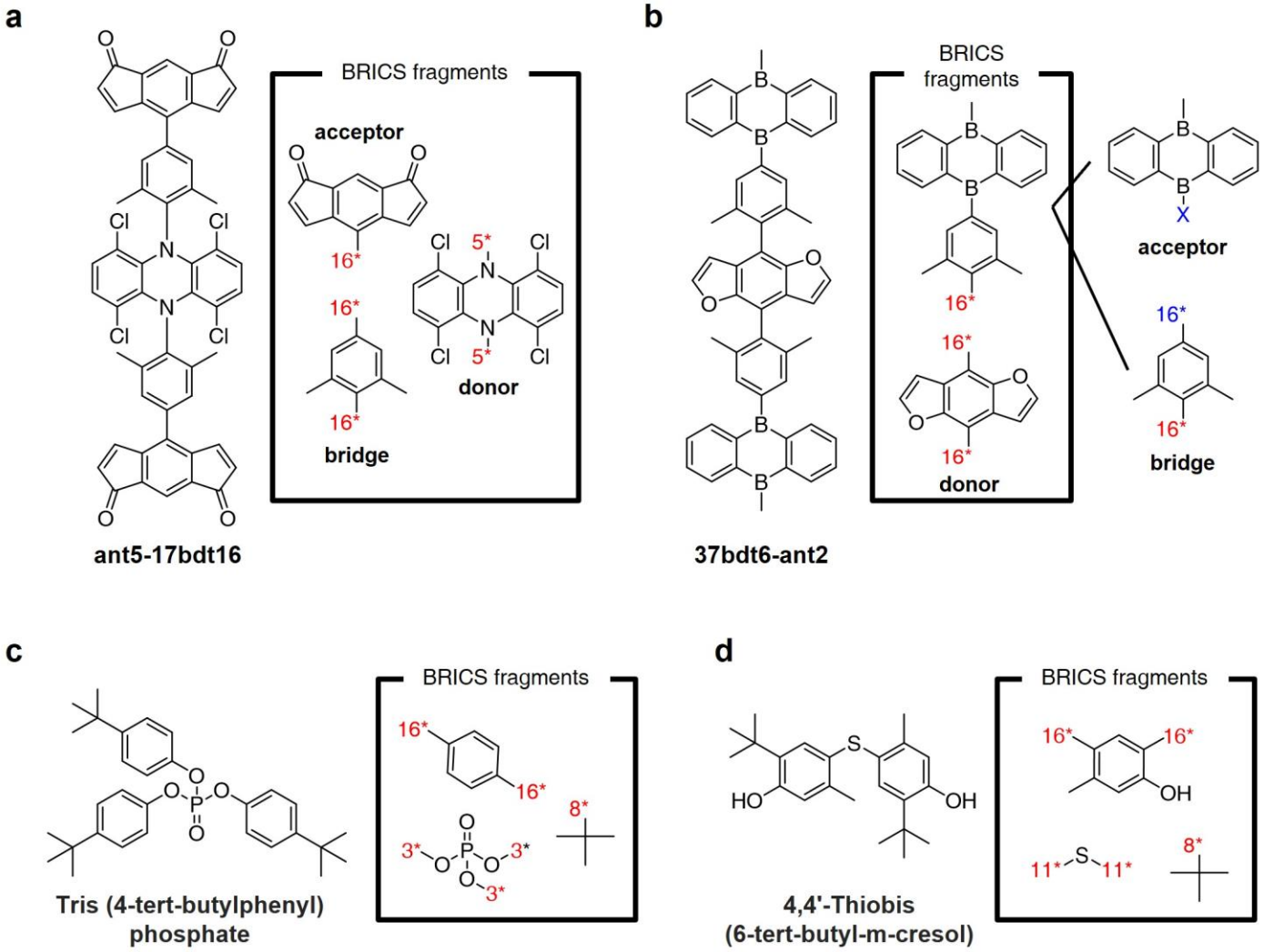


frag. 2



frag. 3

Can RL-CC Generate "Materials"?

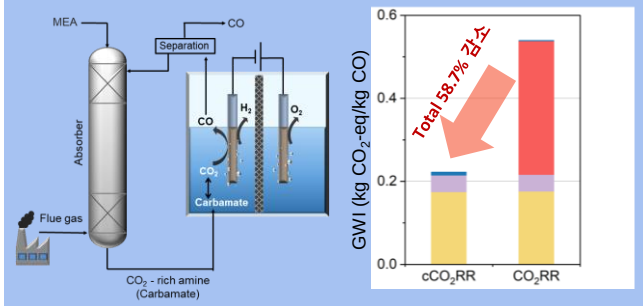


(a,b) Thermally-activated delayed fluorescence (TADF), (c,d) Flame retardant

On-Going Projects

IMPERIAL UCL

Optimal Amine Solvent Design
for Direct Reduction of Captured CO_2

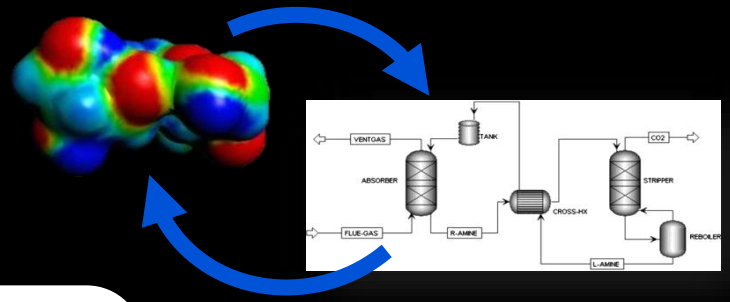




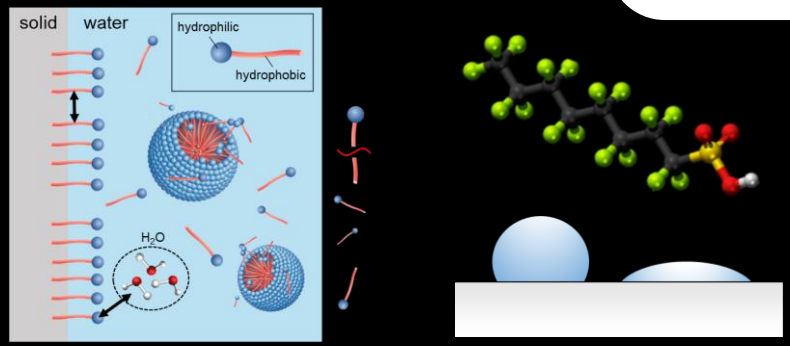
Collaboration with Prof. Yoon Jeong Hwang (SNU) & Dr. Ung Lee (KIST)

Green Chemistry 25 (24), 10398

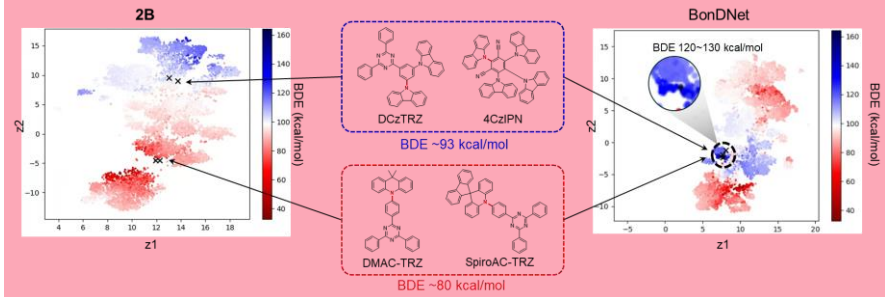
Material-Process Simultaneous Design
for Multi-Scale Design of "Systems"



Applied



Non-PFAS Surfactant Design
Sponsor: Solvay/Syensqo



Stability-guaranteed TADF OLED Design
AI understand OLED degradation

Conclusion

- We are developing technologies that enable **"product and process simultaneous design"** by fusing AI-based material design and process optimization technologies.

- We are the first to implement a successful **inverse design** system through a molecular/material design methodology with extreme properties using **reinforcement learning**.

- Unlike generative models based on supervised learning, reinforcement learning is specialized to create materials with the exact desired properties in a more efficient manner, as its purpose is "optimization".

- Currently, the methodology
  1. Amines for Reactive Capture & Conversion
  2. Multi-resonance TADFs with guaranteed lifetime
  3. Precursor reverse engineering for atomic layer deposition
  4. non-PFAS surfactant reverse engineering (with **Syensqo**).

Acknowledgements


이웅 박사님

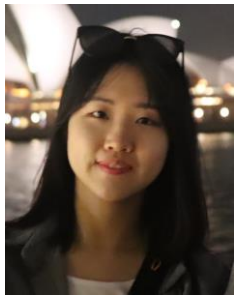

김경수 박사님


박담대 박사님


이원보 교수님


김현승 박사님


이화여자대학교
신다은 학생


이화여자대학교
최혜연 학생


이화여자대학교
김유현 학생


Prof. Lauren Lee

molGCT: https://github.com/Hyunseung-Kim/molGCT
RL-CC: https://github.com/Haeyeon-Choi/RL-CC

THANK YOU

감 사 합 니 다

Jonggeol Na, Ph. D.
Department of Chemical Engineering and Materials Science
Ewha Womans University
E-mail: jgna@ewha.ac.kr