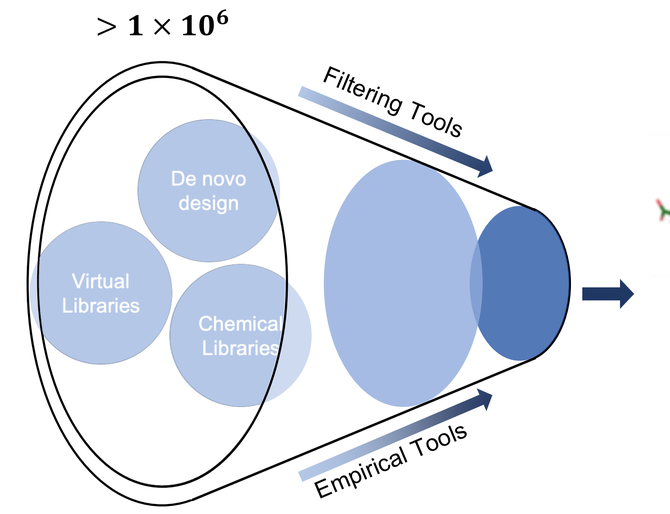
ENSS: Estimating the Number of Synthetic Steps by Graph-based Deep Learning for Virtual Screening

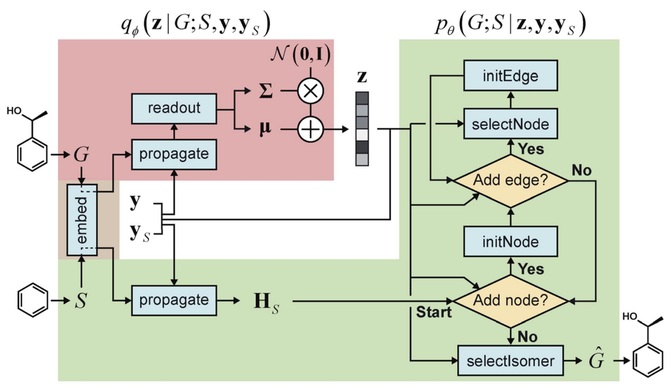
*Hyeongwoo Kim, Kyunghoon Lee, and Woo Youn Kim**
Department of Chemistry, 291, Daehak-ro, Yuseong-gu, KAIST, Daejeon 34141, Republic of Korea
**E-mail : wooyoun@kaist.ac.kr*

Intelligent Chemistry Lab

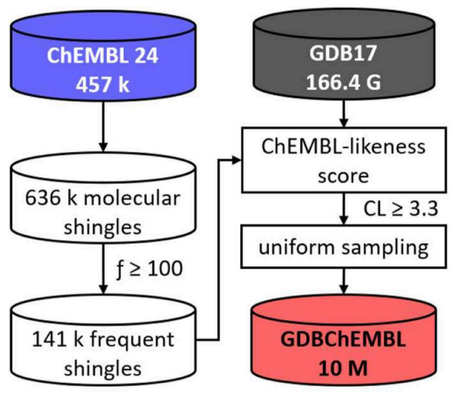
Introduction



Virtual Screening is a powerful pipeline for drug discovery



GGM

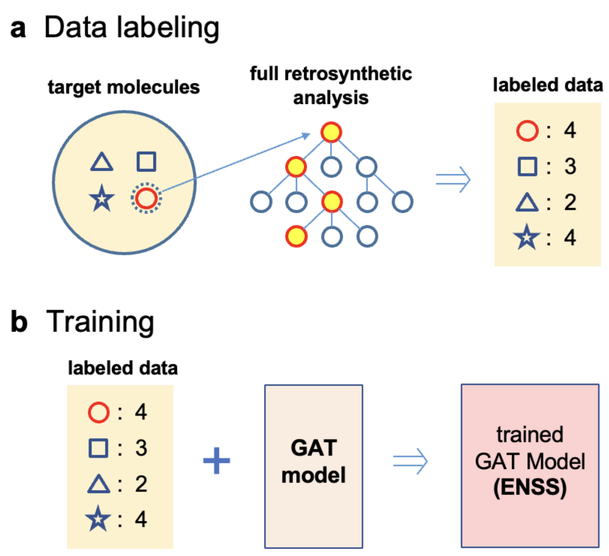


GDBChEMBL

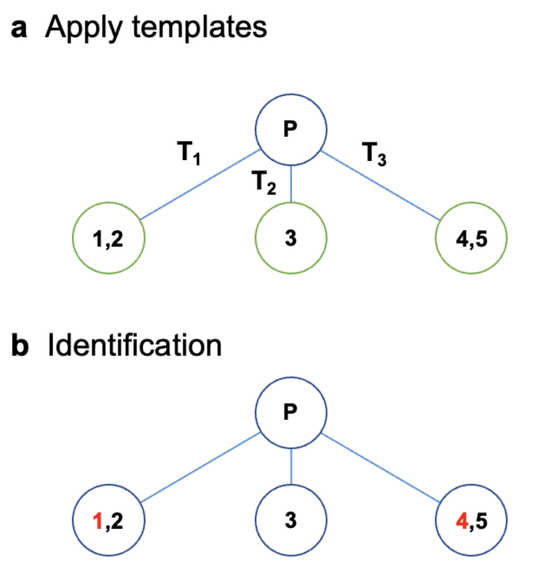
1. *De novo* design: including Generative models
2. Virtual Library: Enumeration based on algorithms

Method

- True labels for the number of synthetic steps are obtained using Full Retrosynthetic-Analysis (FRA).



Overview of obtaining ENSS



Algorithm of Full Retro-Analysis

- **Training Objective**

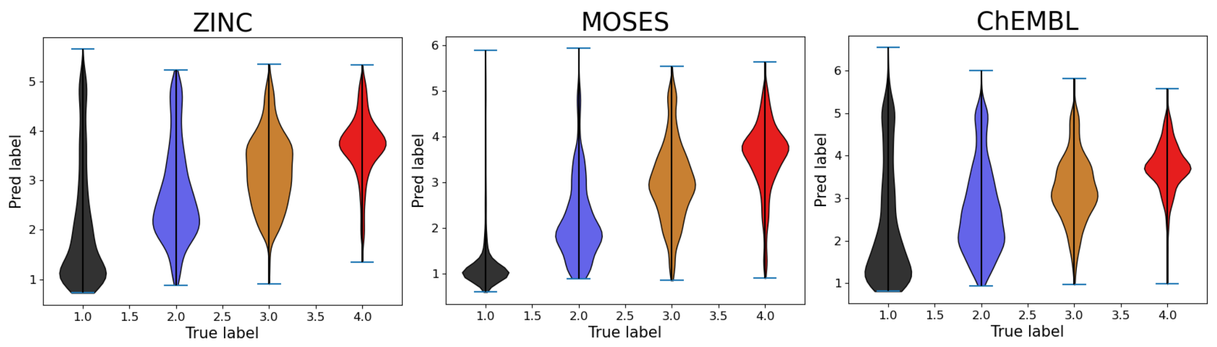
$$\mathcal{L}(\theta) = \sum_m \ell(m;\theta)$$

$$\ell(m;\theta) = \begin{cases} (N_{true,m} - ENSS(m;\theta))^2, & if\ N_{true,m} \le \text{max\_step} \\ \{\max(0, \text{max\_step} + 1 - ENSS(m;\theta))\}^2, & otherwise \end{cases}$$

Result

- **Full Retro-Analysis(FRA) setting**
  - Reaction Templates: 115 famous templates on drugs
  - Starting material: 9,289,950 eMolecules (M_w < 350Da)
- **Model Training setting**
  - 250,000 PubChem data labeled with FRA (≤ **4 steps**)

Exp 01. Ranking synthetic accessibility using ENSS



Exp 02. Evaluation Metrics (Regression & Classification)
10,000 molecules for each dataset (2,000 data for each class)

	ENSS				SA score	SC score
	Precision	Recall	Critical Err	AUROC	AUROC	AUROC
ZINC	0.915	0.922	0.131	0.863	0.641	0.489
MOSES	0.933	0.95	0.095	0.893	0.530	0.400
ChEMBL	0.939	0.905	0.123	0.896	0.678	0.467

Exp 03. Application on Virtual Screening Scheme
100,000 randomly selected molecules for each dataset

	Precision	Recall	Critical Err	AUROC	Filtered out True	Filtered out False
GGM	0.750	0.928	0.145	0.938	7.2%	81.1%
GDBChEMBL	0.282	0.698	0.152	0.843	30.2%	86.0%

Initial dataset distribution (True : False)
- GGM. 37916:62084 - GDBChEMBL. 7302:92698

Conclusion

- We developed a fast and generalizable scoring metric of synthetic accessibility, ENSS. (< 3 min for 100,000 mols)
- Using our model, we can shape a certain chemical library to be more synthetically tractable space.
- This approach can be easily applied to any other systems, by re-organizing reaction templates and starting materials.

Acknowledgement

This work was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF- 2017R1E1A1A01078109).

References

[1] Lim, Jaechang, et al. "Scaffold-based molecular design with a graph generative model." *Chemical science* 11.4 (2020): 1153-1164.
[2] Bühlmann, Sven, and Jean-Louis Reymond. "ChEMBL-likeness score and database GDBChEMBL." Frontiers in Chemistry (2020): 46.
[3] Ertl, Peter, and Ansgar Schuffenhauer. "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions." Journal of cheminformatics 1.1 (2009): 1-11.
[4] Coley, Connor W., et al. "SCScore: synthetic complexity learned from a reaction corpus." Journal of chemical information and modeling 58.2 (2018): 252-261.