Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

Data undersampling models for the efficient rule-based retrosynthetic planning

Min Sik Park,* Dongseon Lee, Youngchun Kwon, Eunji Kim, and Youn-Suk Choi

Autonomous Material Development Lab., Samsung Advanced Institute of Technology, Samsung Electronics, 130 Samsung-ro, Suwon, Gyeonggi-do 16678, Republic of Korea

* Correspondence: ms91.park@samsung.com (Min Sik Park)

S1. Effect of frequency cut-off on prediction accuracy

The prediction model in our study belongs to the classification model in the machine learning domain. Hence, the number of reaction templates in our study corresponds to the number of labels in the neural network model. In our study, as the frequency cut-off increases, the number of reaction templates decreases as shown in Table S1. Therefore, the large frequency cut-off induces the higher prediction accuracy, since the number of labels the neural network model has to predict is reduced. For reference, the effect of frequency cut-off on the accuracy of a prediction model based on a small dataset (total number of datasets is 161,574) is shown in the following table. Indeed, the top-1 prediction accuracy increases as the frequency cut-off (the number of reaction templates) increases (decreases).

Cut-off (≥)	# of templates	Top-1 accuracy (%)
1 (no-cut)	46,756	35.86
2	13,730	47.97
3	7,146	51.48
5	3,889	55.23
10	1,935	61.03

S2. Atomic mapping method

In this study, the atomic mapping method was used as an initial stage for extracting the reaction templates from the reaction SMILES dataset. The atomic mapping algorithm is based on an RDKit module [21]. In the module, the 'FindMCS' function finds a maximum common substructure between multiple molecules [21]. After that, the common substructure for each molecule can be numbered in common by using 'GetSubstructMatches' function [21]. Finally, the obtained results can be combined to generate atom-mapped SMARTS.

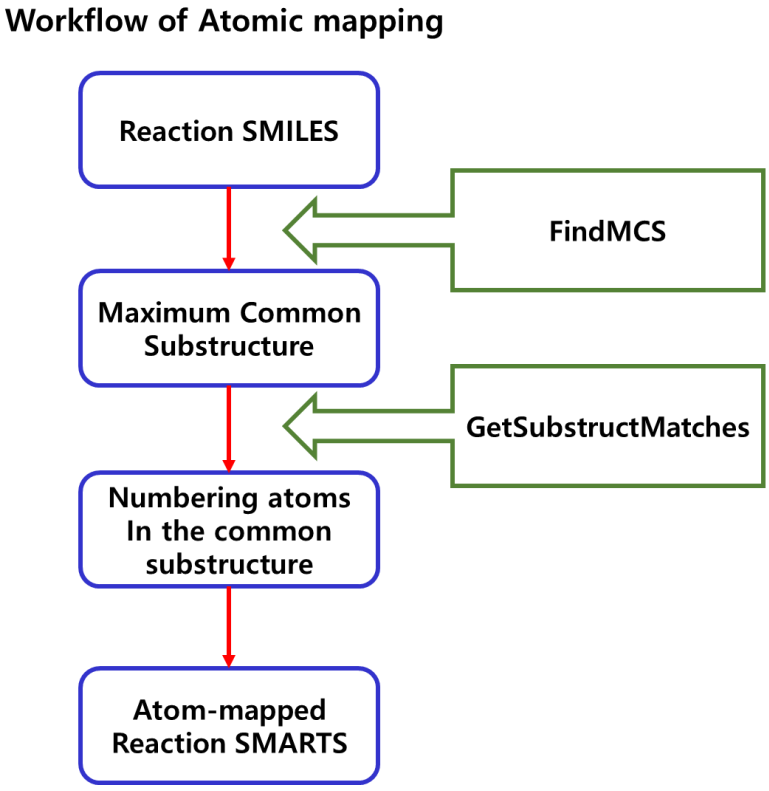


Table S1. Number of reaction templates and data samples depending on the frequency cut-off

cut-off (≥)	1	3	5	10	50	100
# of templates	9,672,940	386,901	165,778	61,234	8,031	3,899
# of data	15,930,914	4,797,540	4,062,238	3,395,642	2,454,530	2,171,771

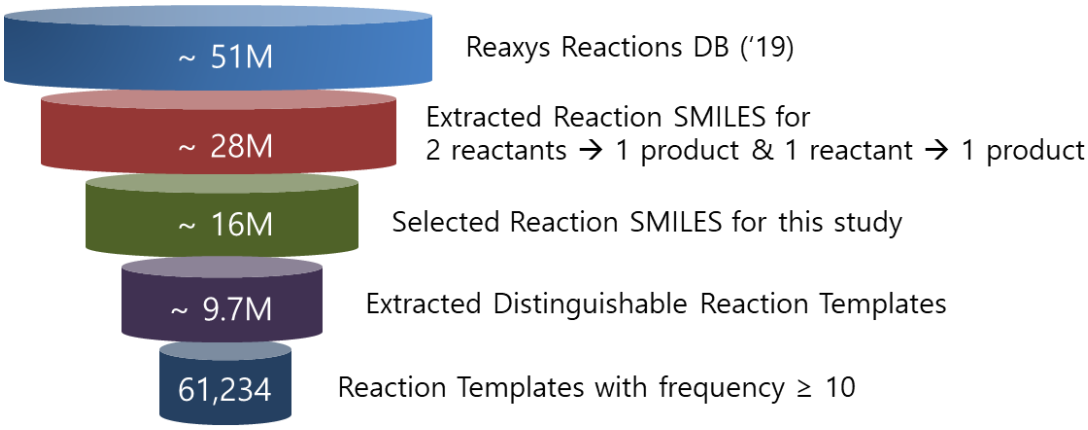


Fig. S1. Funnel-like depiction for dataset filtering steps

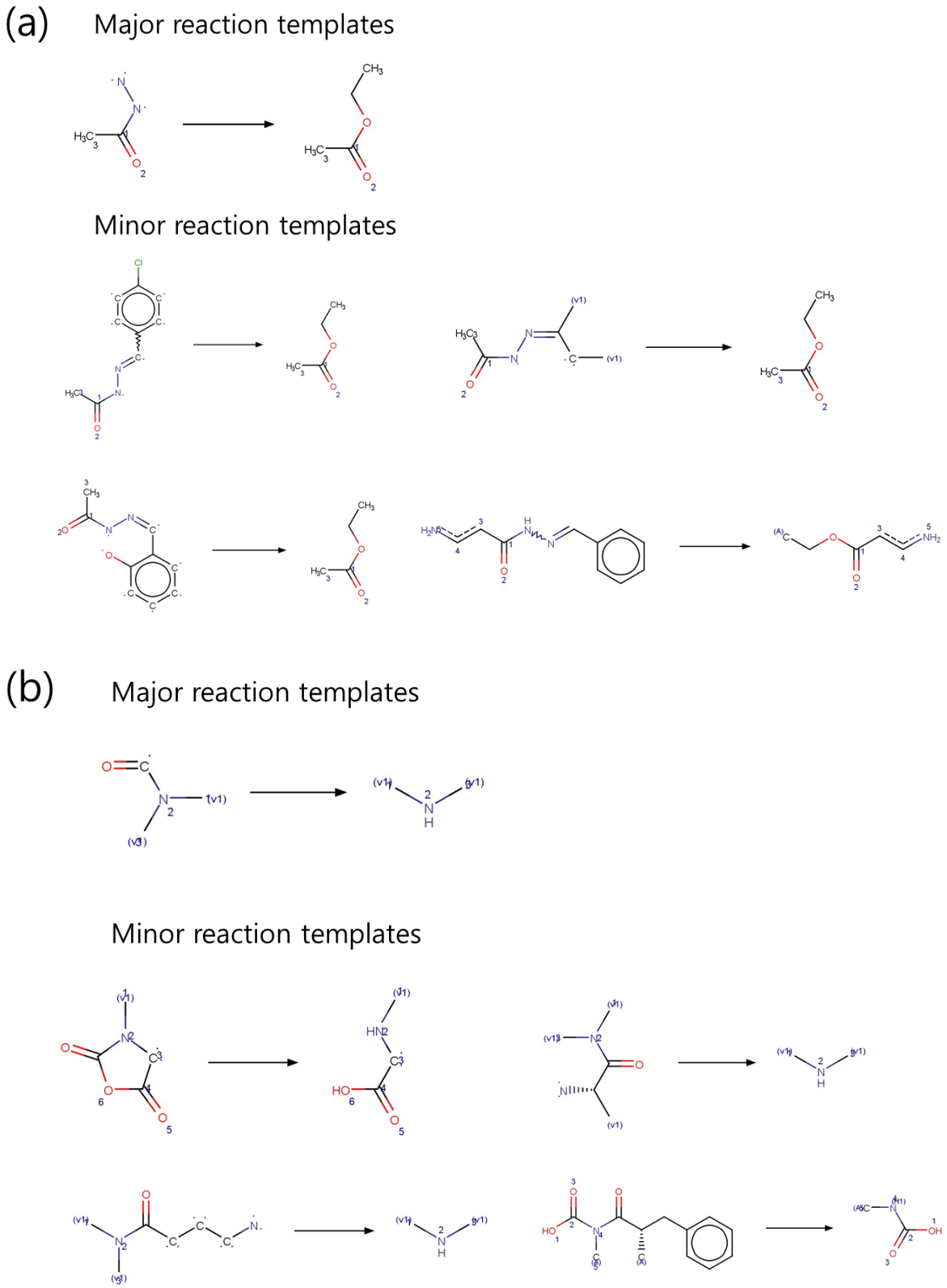
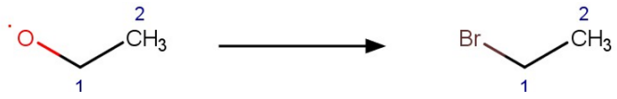


Fig. S2. Three examples represent that minor reaction templates (frequency $<$ 10) can be a subset of major reaction templates (frequency $\geq$ 10)

(C) Major reaction templates



Minor reaction templates

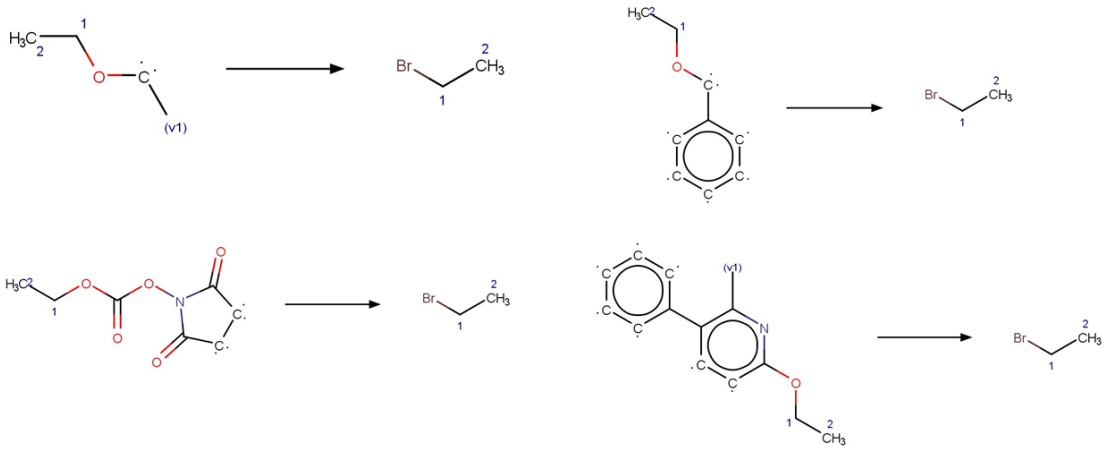


Fig. S2. (Cont.) Three examples represent that minor reaction templates (frequency $<$ 10) can be a subset of major reaction templates (frequency $\geq$ 10)

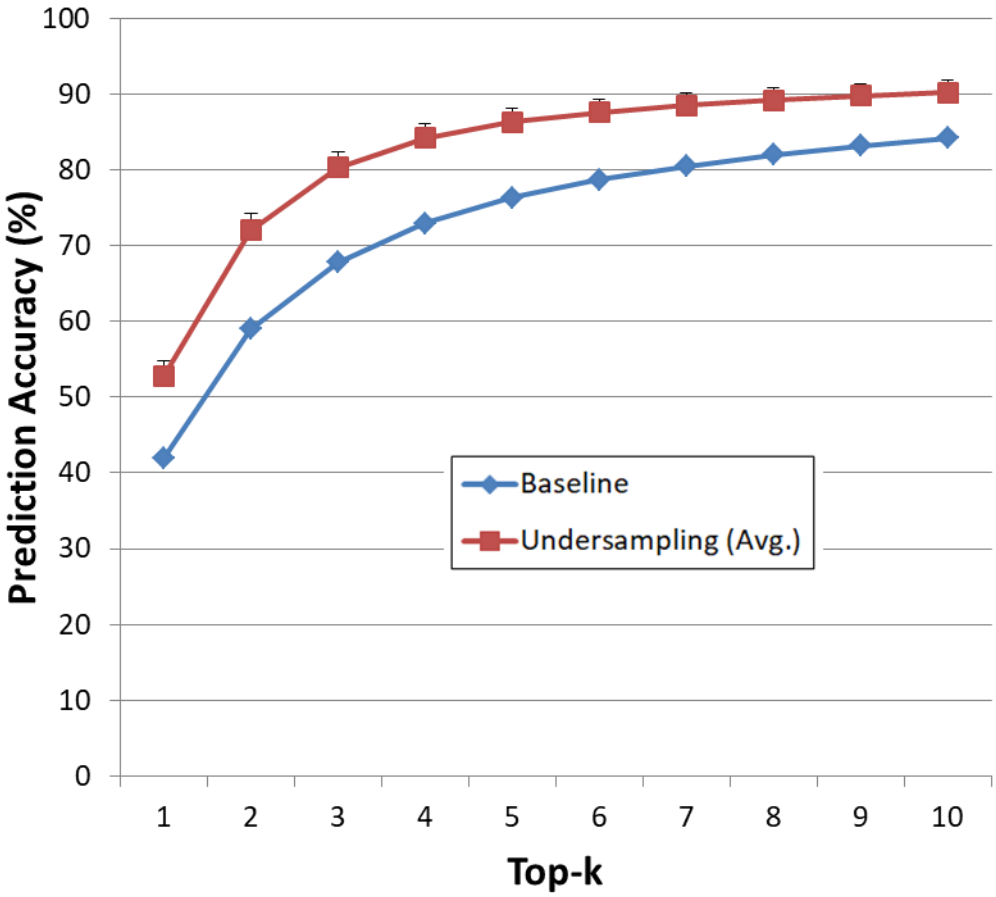


Fig. S3. Prediction accuracy for both the baseline model and the averaged undersampling models

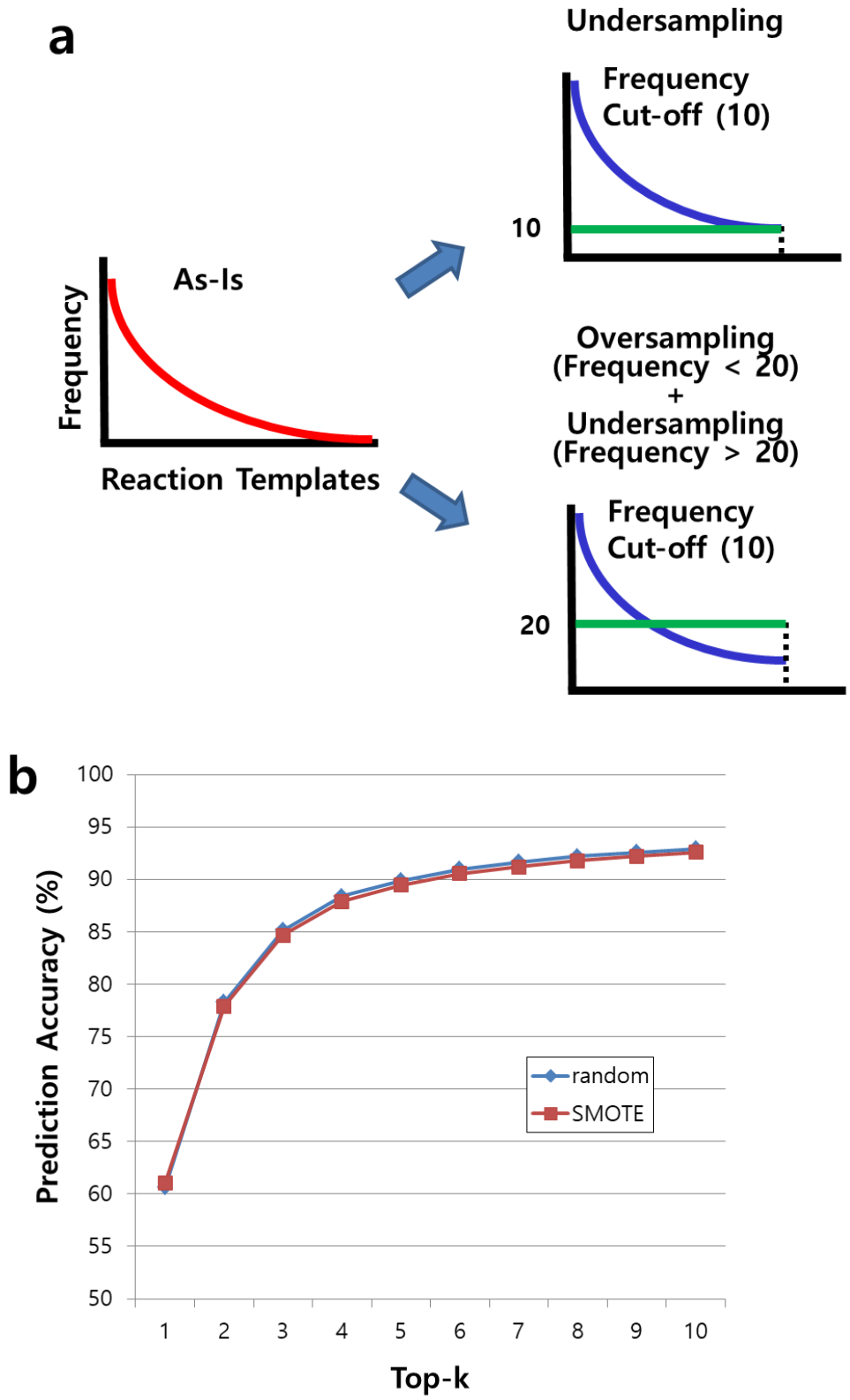


Fig. S4. (a) Schematic diagram of oversampling procedure; (b) Prediction accuracy for random and SMOTE oversampling models

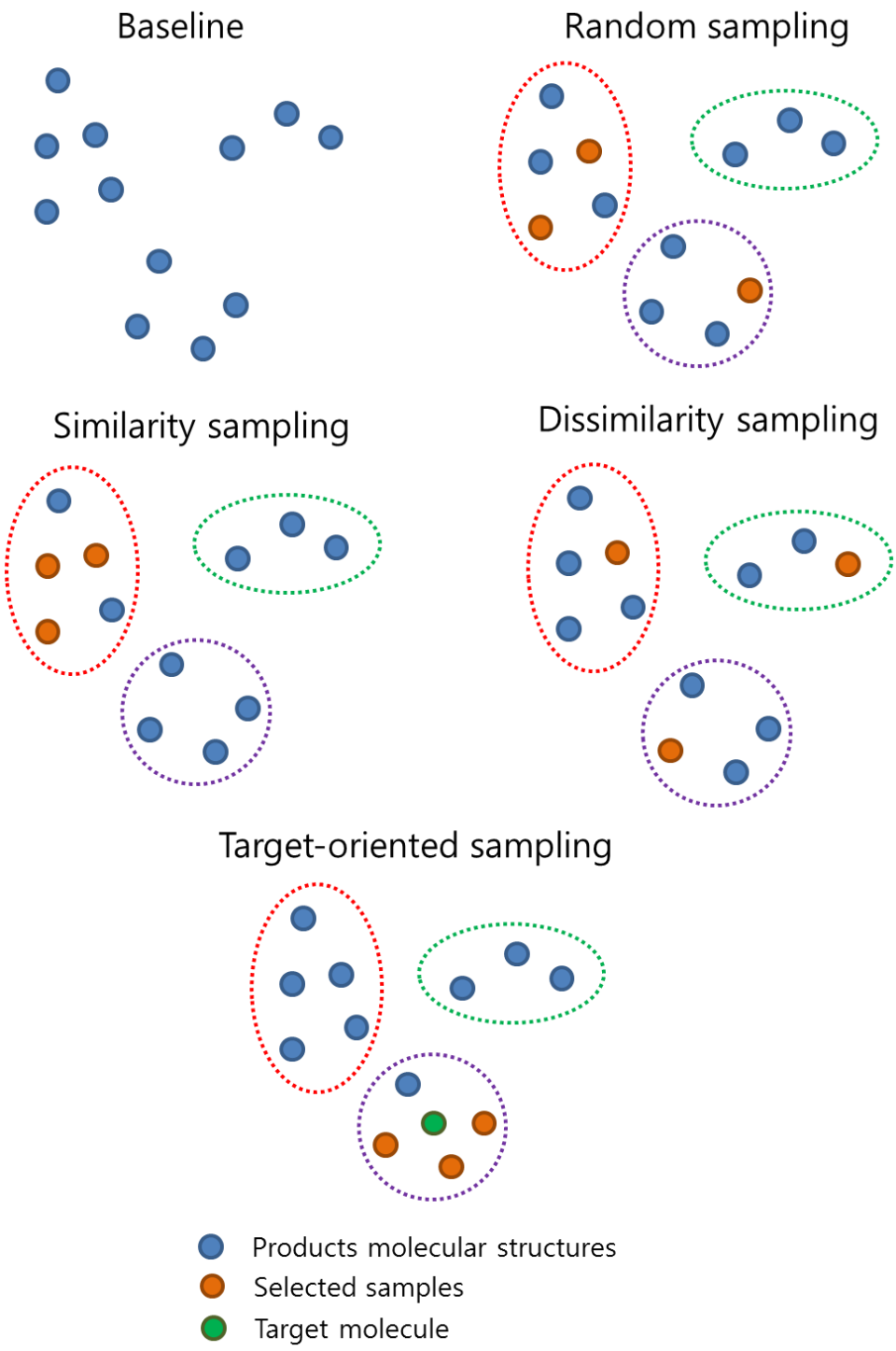


Fig. S5. Schematic diagram of the baseline and four undersampling methods.