APPENDIX A
VARIANT SMILES

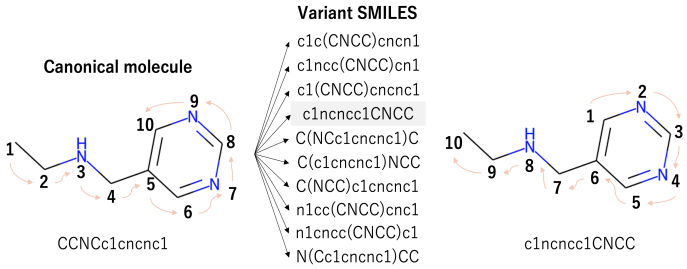

Fig. A.1. Example of producing variant SMILES strings.

Figure A.1 demonstrates an example of producing variant SMILES strings. Generally, a molecular graph has a canonical SMILES string (left figure). The numbers and arrows in the figure denote the traversal order of the atoms. Unlike natural language processing, the canonical molecule has various SMILES representations (middle figure) according to different traversal orders (right figure) called variant SMILES. However, the same molecular graph can be represented using these ten variant SMILES strings. Therefore, variant SMILES strings can be produced to improve the pretraining of the generator and prevent it from learning only a single semantic and syntactic feature.

APPENDIX B
DRUG-LIKENESS

Drug-likeness is evaluated by the QED scores. We generally assign different weights to eight molecular descriptors: the molecular weight (MW), octanol-water partition coefficient (ALOGP), number of hydrogen bond donors (HBDs), number of hydrogen bond acceptors (HBAs), molecular polar surface area (PSA), number of rotatable bonds (ROTBs), number of aromatic rings (AROMs), and number of structural alerts (ALERTS). The calculation is as follows:

$$\mathrm{QED} = \exp\left(\frac{\sum_{i=1}^{8} W_i \ln d_i}{\sum_{i=1}^{8} W_i}\right),$$

where d_i and W_i represent the desirability function and weight of the i-th descriptor, respectively. Usually, the weights of the eight molecular descriptors are obtained through chemical experiments. In practice, the QED score is calculated by a function in the RDKit tool. The larger the QED score, the more drug-like the molecule.

APPENDIX C
SMILESVAE VALIDATION

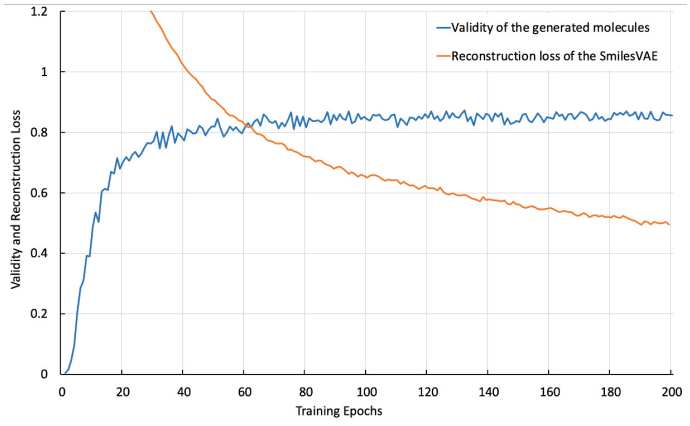

Fig. C.1. Change curves of the reconstruction loss of validation dataset and the ratio of valid molecules generated by the proposed GxVAEs.

Figure C.1 shows the training loss and the ratio of valid molecules generated by the proposed GxVAEs. The red curve indicates the training losses of SmilesVAE with the training epochs. The blue curve denotes the ratio of valid molecules generated by SmilesVAE with the training epochs. Note that the valid molecules are examined by the RDKit tool.