Appendix A Chemical Properties Calculation

Drug-likeness (QED) scores are calculated based on eight molecular descriptors: 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 mathematical description is as follows:

$$QED = e^{\frac{\sum_{n=1}^{8} W_n ln d_n}{\sum_{n=1}^{8} W_n}}, \tag{1}$$

where d_n and W_n are the desirability function and the weight of n-th molecular descriptor, respectively.

Solubility (LogP) is determined from the partition coefficient. This coefficient represents the equilibrium concentration ratio of a molecule between two nonmixing solvents. It is calculated as follows:

$$logP = log(\frac{c_o}{c_w}), \tag{2}$$

where c_o represents the substance activity in the organic phase and c_w indicates the one in water phase.

Synthesizability (SA) is calculated as follows:

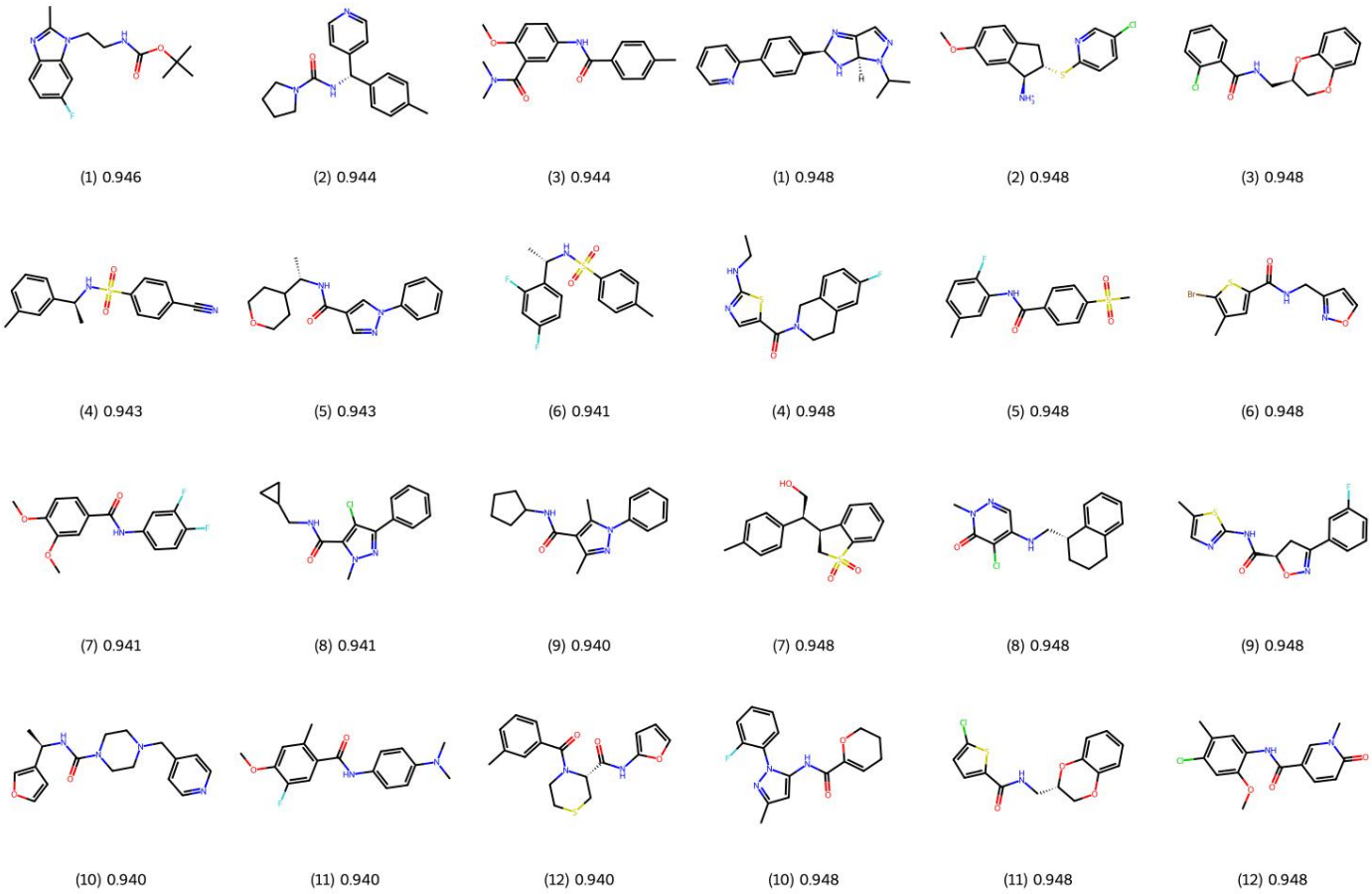
$$SA = r_s - \sum_{n=1}^{5} p_n, \tag{3}$$

where r_s is the ratio of the total contributions of all segments to the number of segments in the molecule. This ratio is obtained from the analysis of synthetic molecules. We obtained r_s from the SpotGAN [33]. p_n $(n \in 1, \cdot, 5)$ represents five features: ring complexity, stereo complexity, macrocycle penalty, size penalty, and bridge penalty.

For implementation, the QED scores are computed directly by RDKit.chem.QED.qed in the RDKit tool. The QED scores range from 0 to 1. The Logp and SA scores also use the RDKit tool, but we have normalized and clipped them from 0.1 to 1. The higher these scores are, the more desirable these properties are.

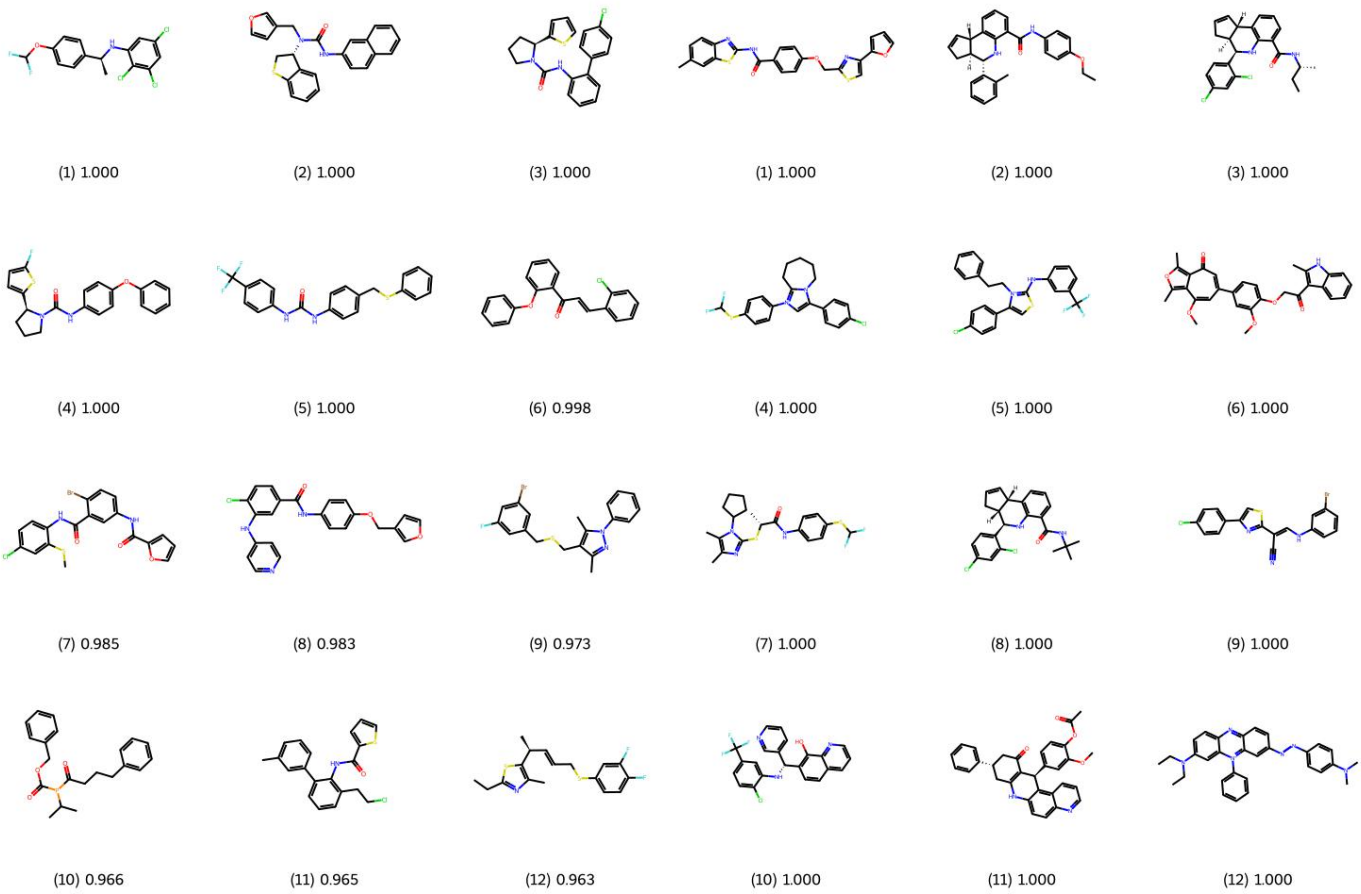
Appendix B Experimental Details

Since the novelty of the baselines on the QM9 dataset is based on the proportion of valid SMILES strings not present in the training dataset, we also include the same novelty (65.37%) of EarlGAN. To facilitate future research in this field, we also provide the novelty (99.73%) based on valid SMILES of EarlGAN on the ZINC dataset.



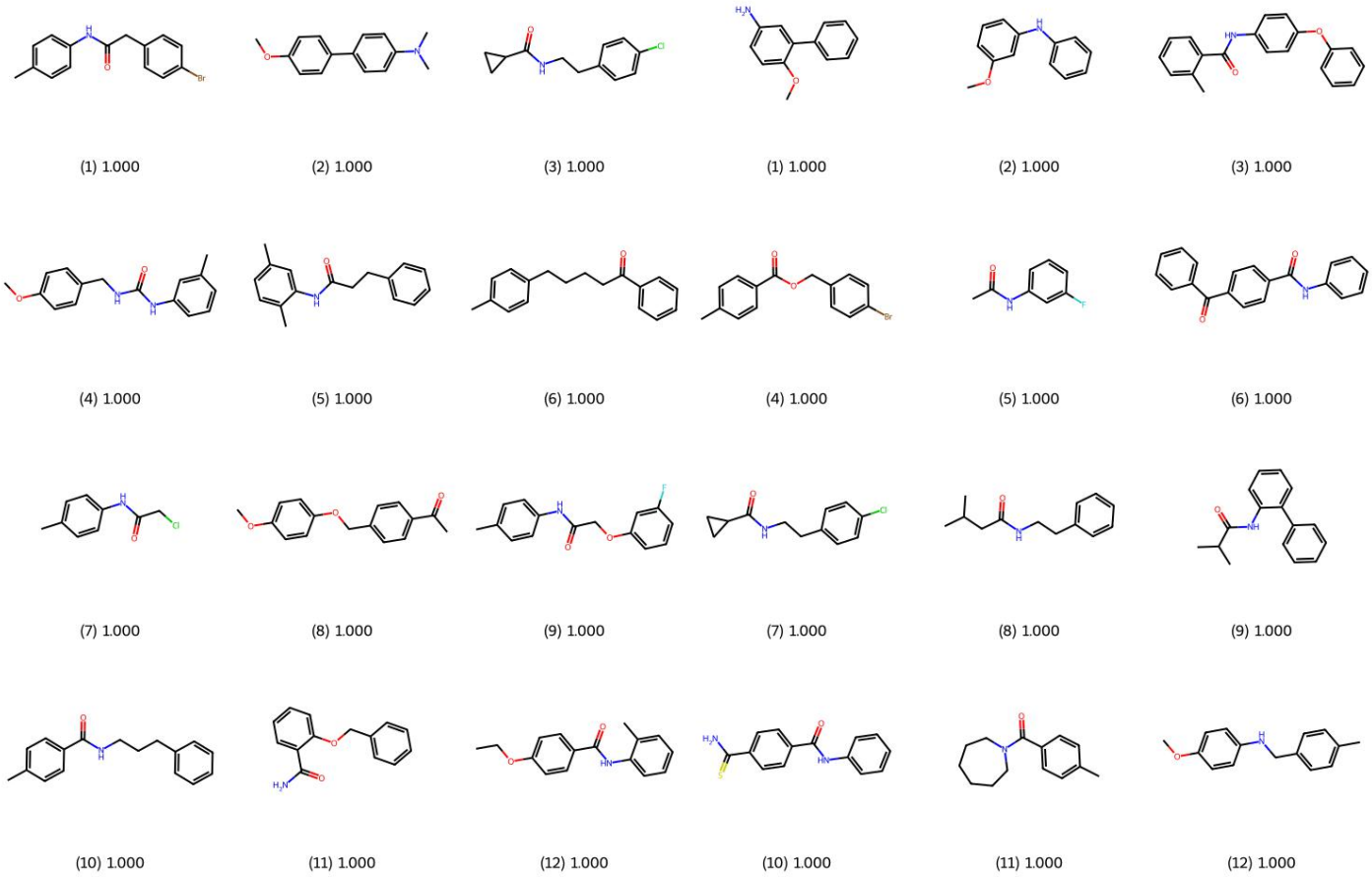
(a) Generated molecules. (b) Training molecules.

Figure B.1: Top 12 Drug-likeness (QED) molecules.



(a) Generated molecules. (b) Training molecules.

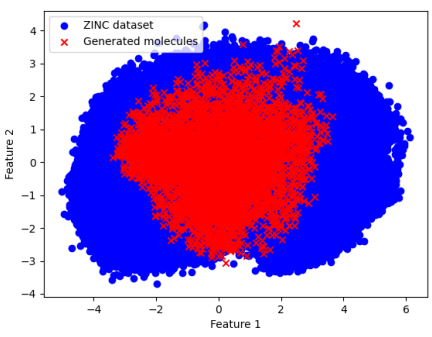
Figure B.2: Top 12 Solubility (Logp) molecules.



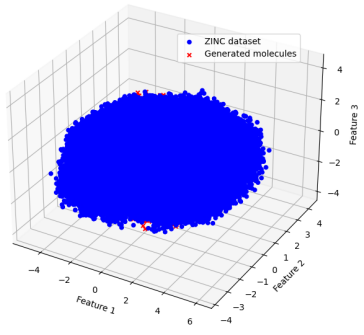
(a) Generated molecules.

(b) Training molecules.

Figure B.3: Top 12 Synthesizability molecules.



(a) Visualization of the two-dimensional scatter plot.



(b) Visualization of the three-dimensional scatter plot.

Figure B.4: Visualization of trained and generated molecules in two and three-dimensional space.