Patterns, Volume *5*

Supplemental information

Enhancing molecular design efficiency: Uniting language models and generative networks with genetic algorithms

Debsindhu Bhowmik, Pei Zhang, Zachary Fox, Stephan Irle, and John Gounley

Supplemental information

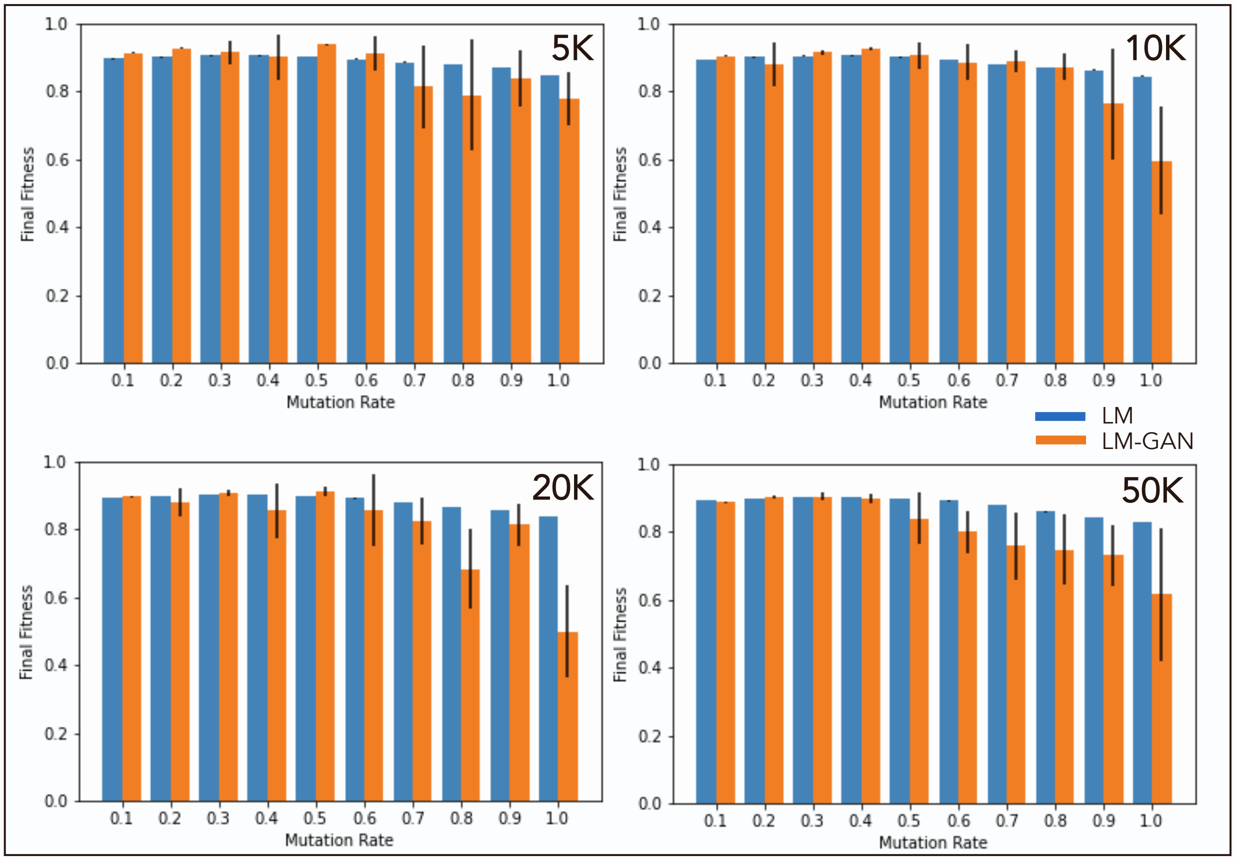


Fig. S1: Plot showing the fitness scores concerning the population size of the initial molecular dataset and mutation rate. The LM and LM-GAN methods exhibit similar performance, with the LM consistently outperforming the LM-GAN, albeit by a small margin. However, this performance gap becomes more significant, especially when dealing with larger populations (above 20,000) or higher mutation rates (above 0.5).

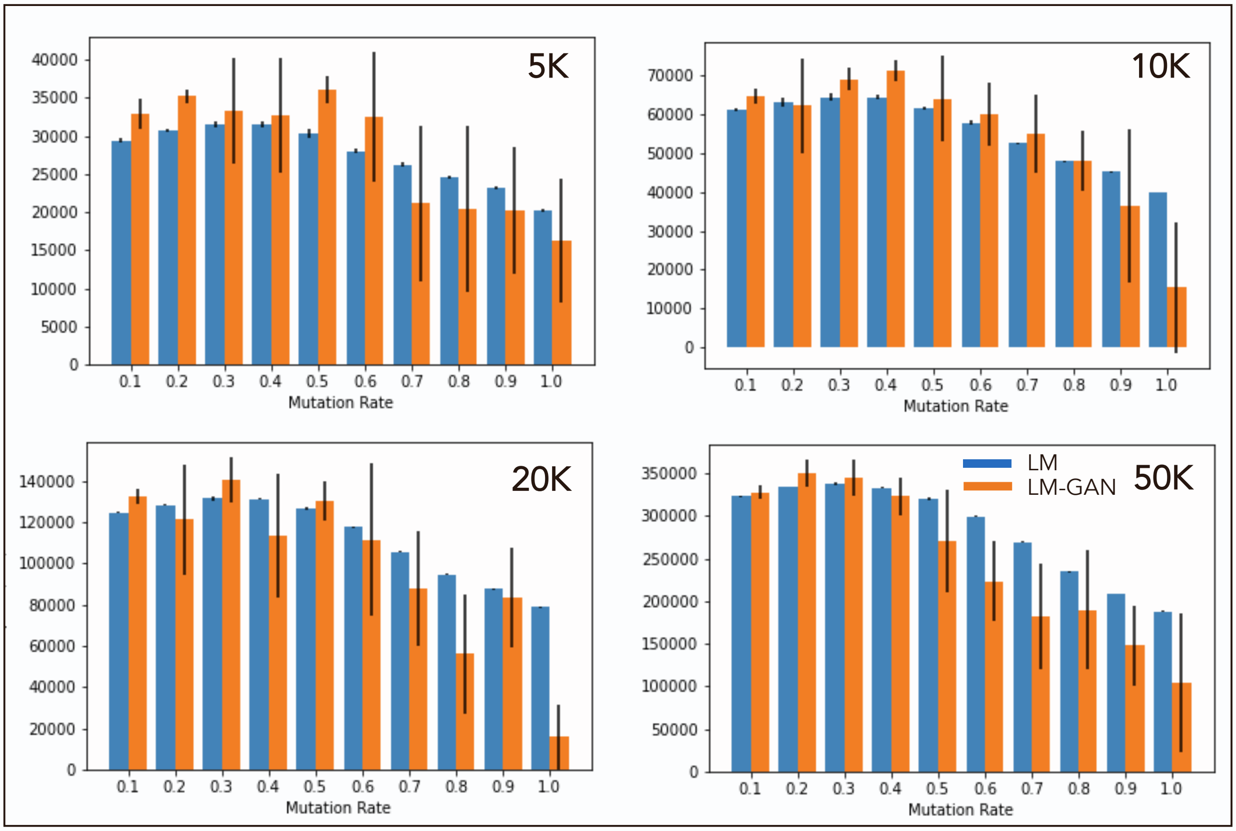


Fig. S2: Plot showcasing the capability of two methods to generate novel molecules not found in the original dataset. The LM method outperforms the LM-GAN method, producing a greater quantity of novel molecules. For the LM method, there is an initial rise in the number of novel molecules until a mutation rate of 50\%, after which there is a gradual decline. Conversely, the LM-GAN method generates significantly fewer novel molecules, and this count decreases gradually as the mutation rate increases.

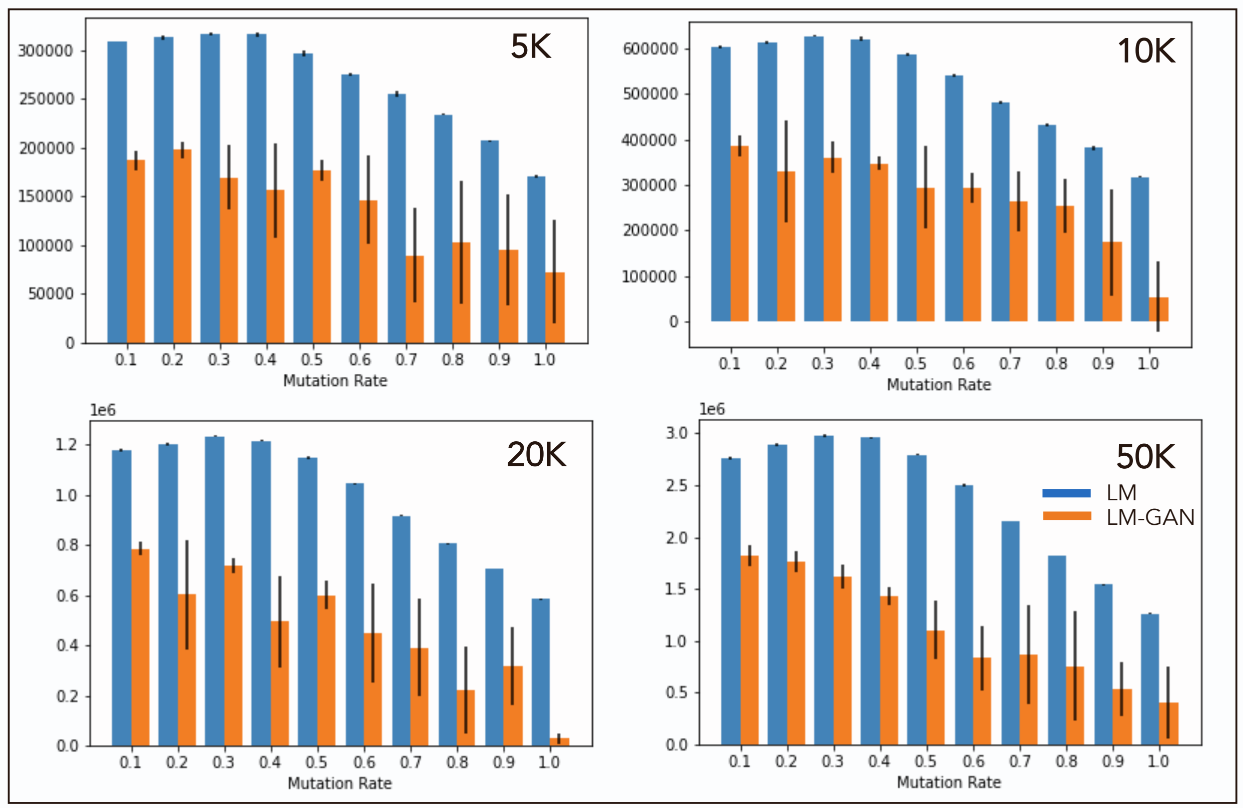


Fig. S3: Plot demonstrating the number of accepted molecules included in the generated dataset. The findings demonstrate that the LM-GAN method performs slightly better than the LM method when the population sizes are smaller and the mutation rates are lower. However, as the population size and mutation rate increase, the performance of the LM method gradually improves and eventually surpasses that of the LM-GAN method.

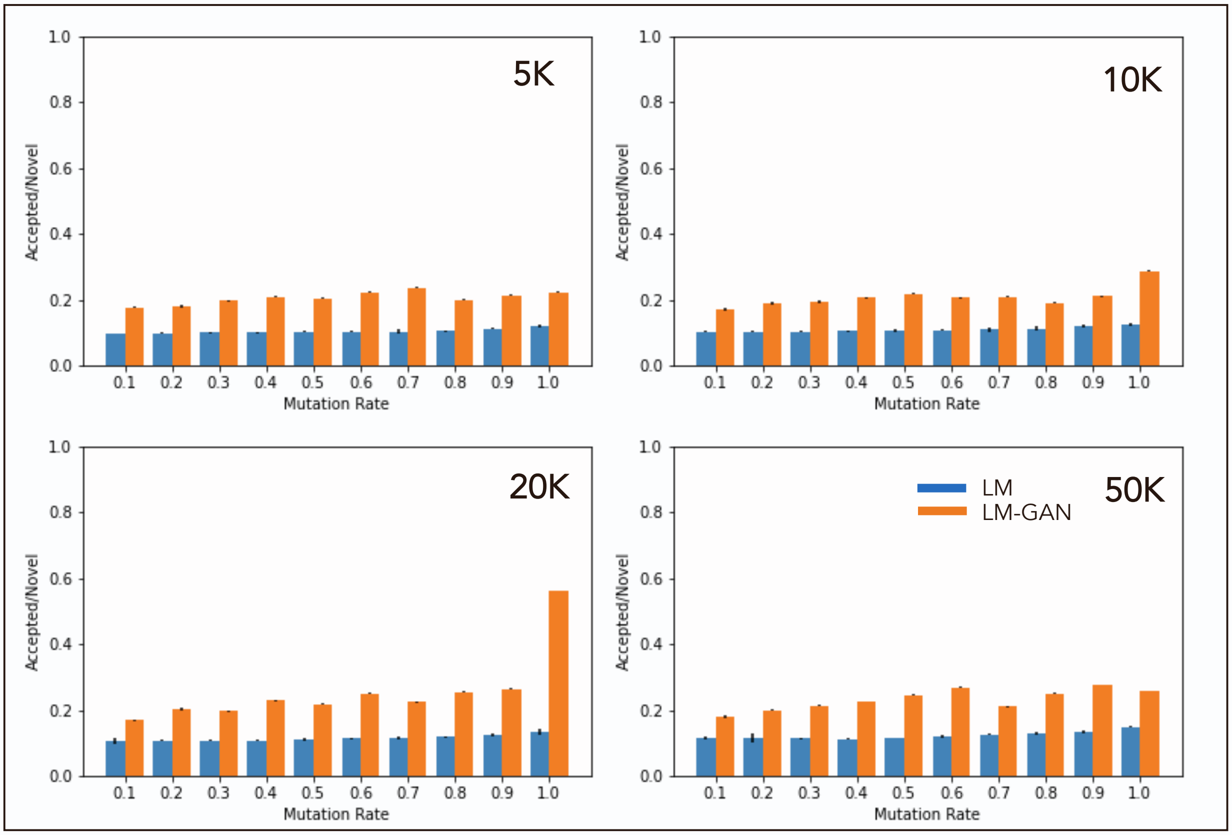


Fig. S4: Plot comparing the efficiency values of two different techniques across different population sizes and mutation rates. The LM-GAN technique consistently outperforms the LM-only technique, with an efficiency value that is nearly double. The effect of population size on efficiency values is not significant. However, there is a slight trend suggesting that LM-GAN performs slightly better at higher masking rates.

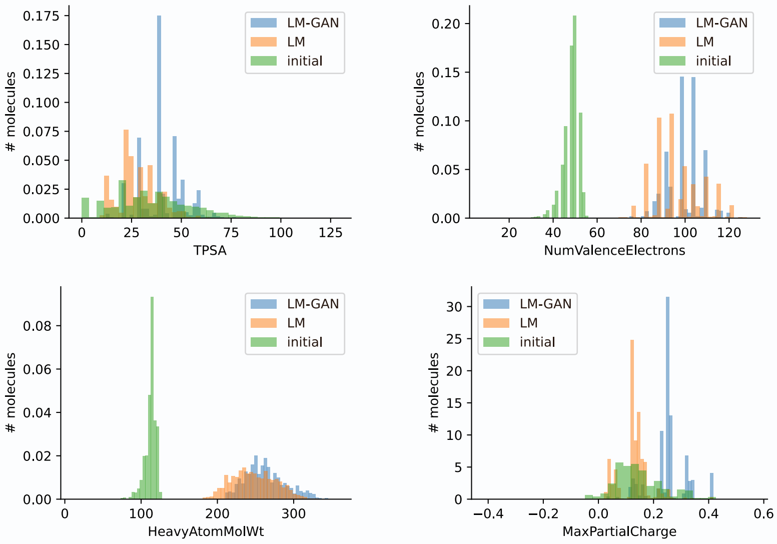


Fig. S5: Distributions of TPSA, Number of Valence electrons, Heavy atom molecular weight, and max partial charge across the initial population (green) and the final populations for the LM-GAN and LM based approaches.

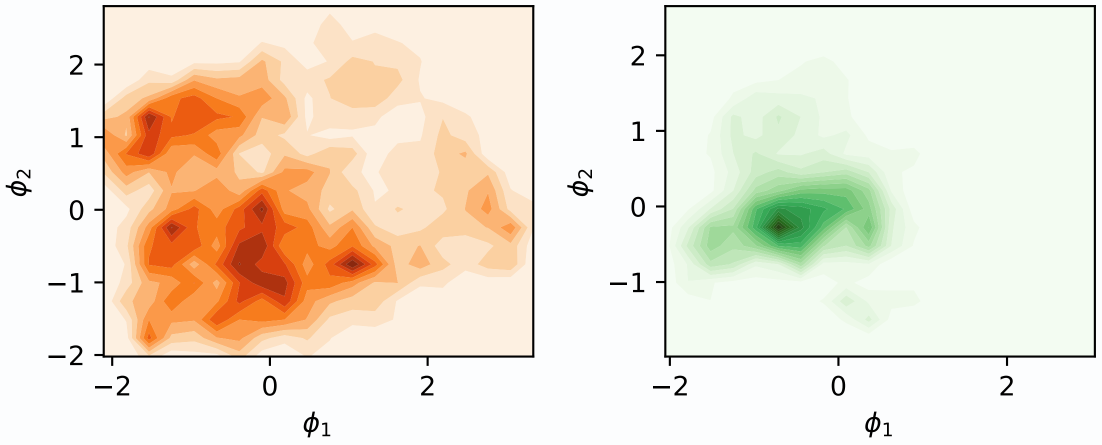
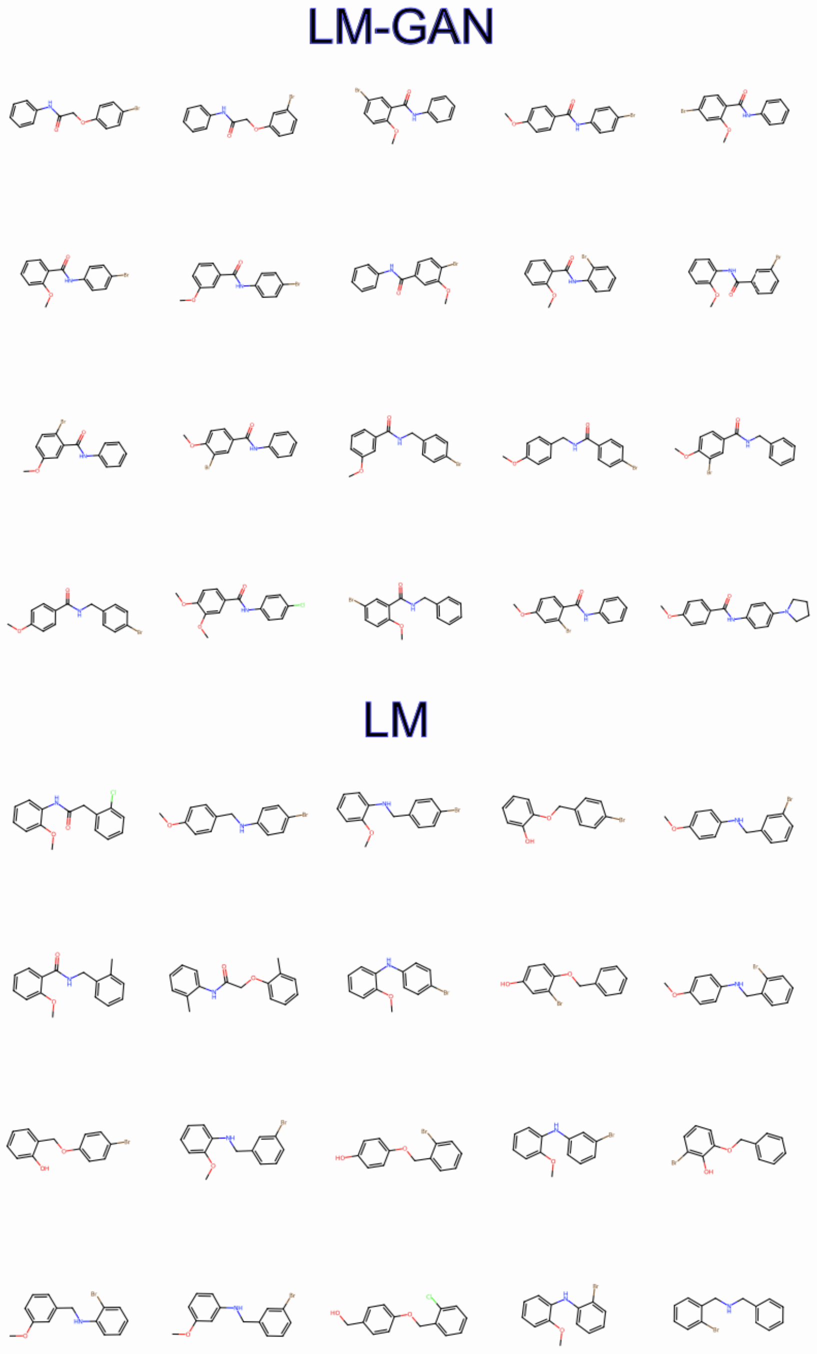


Fig. S6: Density maps showing the difference between the LM (left) and LM-GAN (right) molecular distributions. The ECFP's for each molecule in each population were projected onto the first two principal components of the LM-only population of molecules, which are denoted φ1 and φ2.



LM-GAN

LM

Fig S7: Molecules sampled from the population after 50 generations for the LM-GAN and LM generated molecules.