

Large Language Models as Molecular Design Engines

D. Bhattacharya, H. J. Cassady, M. A. Hickner, W. F. Reinhart

Abstract: The design of small molecules is crucial for technological applications ranging from drug discovery to energy storage. Due to the vast design space available to modern synthetic chemistry, the community has increasingly sought to use data-driven and machine learning approaches to navigate this space. Although generative machine learning methods have recently shown potential for computational molecular design, their use is hindered by complex training procedures, and they often fail to generate valid and unique molecules. In this context, pretrained Large Language Models (LLMs) have emerged as potential tools for molecular design, as they appear to be capable of creating and modifying molecules based on simple instructions provided through natural language prompts. In this work, we show that the Claude 3 Opus LLM can read, write, and modify molecules according to prompts, with impressive 97% valid and unique molecules. By quantifying these modifications in a low-dimensional latent space, we systematically evaluate the model's behavior under different prompting conditions. Notably, the model is able to perform guided molecular generation when asked to manipulate the electronic structure of molecules using simple, natural-language prompts. Our findings highlight the potential of LLMs as powerful and versatile molecular design engines.