CLAMP: A Contrastive Language and Molecule Pre-Training Network for Scaling Artificial Photosynthesis Candidate Identification

Redkar, Neel (School: Dougherty Valley High School)

Artificial photosynthesis is the process of turning sunlight and CO2 into chemical products. This is especially desirable due to the ability for carbon neutral fuel and scalable carbon dioxide reduction. Metal Organic Frameworks (MOFs) are porous crystals that can be synthesized to have various active properties—carbon capture, water treatment, or even artificial photosynthesis. The largest obstacle to current research and development of crystals is finding new sites & reactions to synthesize crystals. Machine learning is applied in this space due to its ability to find new permutations that can be tested. For artificial photosynthesis there are no candidates for visible light artificial photosynthesis, which makes machine learning with prior data almost impossible. Similar problems are faced in image generation where due to low performance small amounts of specific data, large amounts of unsupervised generic text and image data are used to generate images with high accuracy. Therefore the engineering goal is to build an unsupervised contrastive language and molecule pre-training framework which could predict possible MOF photo-catalysts. Building a web scraper that gathered 222k crystal text pairs from research papers, an unsupervised zero shot classification model was trained taking advantage of linguistic structure. Without any specific training data, an ~82% accuracy was achieved and ~75% accuracy for photocatalyst prediction. With no prior specific data, this could be a breakthrough technology. This novel network can be cross applied to any task that can be described via text, opening completely new methods to think about chemical generation.