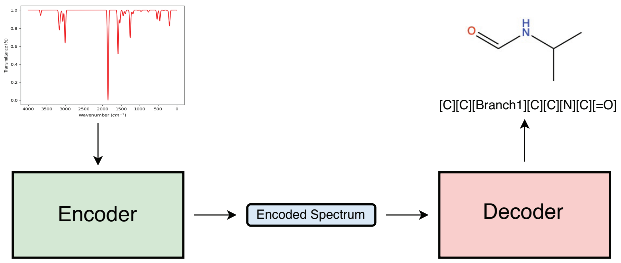
SEQUENCE-TO-SEQUENCE LEARNING FOR MOLECULAR STRUCTURE DERIVATION FROM INFRARED SPECTRA

ETHAN FRENCH, ZHOU LIN, *Department of Chemistry, University of Massachusetts, Amherst, MA, USA*.



Fully identifying unknown molecules via infrared spectroscopy can be a challenging task for even the most experienced researchers. Current data-driven computational methods usually identify unknown spectra by matching them against databases of known spectra. However, this method can be problematic for novel complex molecules given the relative lack of information. Deep learning provides a potential solution to this problem. Sequence-to-sequence learning has had great success in a wide range of areas such as language translation and speech recognition.[a] In this work, an unsupervised sequence-to-sequence model was extended to chemical systems and used to derive complete molecular structures from infrared spectra. The model was trained on the infrared spectra of small organic molecules containing C, H, O, N, and F atoms. These molecules were represented using SELFIES, an improved version of the SMILES string molecular fingerprint descriptor.[b] Our model is able to achieve state-of-the-art results in successfully identifying a wide variety of molecules from their infrared spectra.

[a] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, “Sequence to Sequence Learning with Neural Networks”, *NeurIPS*, 2014, **27**.

[b] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik, “Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation”, *Mach. Learn.: Sci. Technol.* 2020, **1**, 045024