Contribution ID : **20** Type : **Oral Presentation**

Embedded machine-readable molecular representation for resource-efficient deep learning applications

The practical implementation of deep learning (DL) methods for chemistry applications relies on encoding chemical structures into machine-readable formats that can be efficiently processed by computational tools. One Hot Encoding (OHE) and Morgan fingerprints (MF) are established representations of alphanumeric categorical data in expanded numerical matrices or vectors. We have developed embedded alternatives to OHE and MP that encode discrete alphanumeric tokens of an N-sized alphabet into a few real numbers that constitute a simpler matrix representation of chemical structures. The implementation of this embedded representations in training machine learning models achieves comparable results to traditional representations in model accuracy and robustness while significantly reducing the use of computational resources. Our benchmarks across molecular representations (SMILES, DeepSMILES, and SELFIES) and different molecular databases for Variational Autoencoders (VAEs), Recurrent Neural Networks (RNNs) and other DL models show a reduction in vRAM memory usage by up to 50% while increasing disk Memory Reduction Efficiency to 80% on average, in some cases. These encoding methods open new avenues for data representation in embedded formats that promote energy efficiency and scalable computing in resource-constrained devices, or in scenarios with limited computing resources. The application of these embeddings impacts other disciplines that rely on the use of OHE and MF.

Primary author(s) : MARTIN-MARTINEZ, Francisco (King's College London); Mr. NUÑEZ-ANDRADE, Emilio (Swansea University); Dr. VIDAL-DAZA, Isaac (University of Granada); Dr. RYAN, James W. (Swansea University); Dr. GÓMEZ-BOMBARELLI, Rafael (Massachusetts Institute of Technology)

Presenter(s) : MARTIN-MARTINEZ, Francisco (King's College London)