Embedded molecular representations for more efficient machine learning in molecular discovery and chemical property prediction

Francisco J. Martin-Martinez[1,2], Emilio Nuñez-Andrade[2], Isaac Vidal-Daza[2], Rafael Gómez-Bombarelli[3], James W. Ryan[2]

[1]Department of Chemistry, Faculty of Natural, Mathematical & Engineering Sciences, King's College London, London, UK.

[2]Department of Chemistry, Swansea University, Singleton Park, Sketty, SA28PP, Swansea.

[3]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

francisco.martin-martinez@kcl.ac.uk

The practical deployment of machine learning (ML) and deep learning (DL) methods for molecular discovery and property prediction relies heavily on encoding chemical structures into machine-readable numerical formats. Standard encoding schemes such as One Hot Encoding (OHE) and Morgan Fingerprints (MFP) produce high-dimensional, sparse representations that impose significant computational overhead, including excessive memory usage, overfitting risks, and prolonged training times, particularly when scaling to large and diverse molecular datasets. In this work, we present a unified embedding framework that addresses these challenges through two complementary methods: embedded One Hot Encoding (eOHE) and embedded Morgan Fingerprints (eMFP).

eOHE compresses the sparse OHE matrices used to encode string-based molecular representations (SMILES, DeepSMILES, and SELFIES) into compact numerical arrays by reducing the dictionary dimensionality by a tuneable factor *q*. Benchmarked across Variational Autoencoders (VAEs) and Recurrent Neural Networks (RNNs) on the ZINC, QM9, and GDB-13 databases, eOHE reduces vRAM usage by up to 50% and achieves disk memory reduction efficiencies of approximately 80%, while maintaining comparable model accuracy, molecular validity, diversity, and uniqueness [1].

eMFP extends this embedding philosophy to Morgan Fingerprints by reshaping binary fingerprint vectors into compressed decimal representations normalized by the compression factor. Evaluated across five regression models (Random Forest, Multi-layer Perceptron, K-Neighbors Regressor, Gradient Booster Regressor, and a Deep Neural Network) on the RedDB, NFA, and QM9 databases for HOMO-LUMO energy gap prediction, eMFP consistently outperforms standard MFP, with optimal compression factors of q = 16, 32, and 64. The reduced input dimensionality enables more extensive hyperparameter optimization within fixed computational budgets, mitigates overfitting, and accelerates training without compromising predictive performance [2].

Together, eOHE and eMFP demonstrate that embedding-based dimensionality reduction of molecular representations constitutes a general, interpretable, and resource-efficient strategy for both generative and predictive ML tasks in molecular and materials science. These methods promote scalable, energy-efficient computing, and are broadly applicable beyond chemistry to any domain relying on high-dimensional categorical data encoding.

References

[1] Emilio Nuñez-Andrade, Isaac Vidal-Daza, James W. Ryan, Rafael Gómez-Bombarelli, Francisco J. Martin-Martinez, Digital Discovery, 4 (2025) 776.

[2] Emilio Nuñez-Andrade, Isaac Vidal-Daza, Rafael Gómez-Bombarelli, James W. Ryan, Francisco J. Martin-Martinez, ChemRxiv. (2025).

Figures

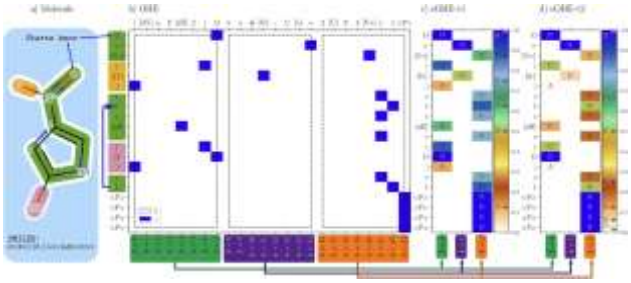

Figure 1. Comparison of OHE representation and two embedded methods, eOHE-v1 and eOHE-v2 for a SMILES representation of a 4-nitro-1H-pyrrol-2-ol sample molecule.