Chirality Encoding for Machine Learning: NLM Latent Space Arithmetic and Atom-Centered Triple Product Descriptors

Joao Aires-de-Sousa[1], Natalia Baimacheva[2], Xinyue Gao[3], Alexandre V. Costa[1], Johander Azuaje[1], Abdulwahab Hussein[1]

[1]LAQV REQUIMTE, Chemistry Department, NOVA School of Science and Technology | NOVA - FCT, 2829-516 Caparica, Portugal. E-mail: jas@fct.unl.pt.

[2]University of Strasbourg, Faculty of Chemistry, 4, Blaise Pascal str., 67081, Strasbourg, France.

[3]Université Paris Cité, Faculty of Sciences, F-75013 Paris, France.

The assignment of stereochemical configuration is an essential component of structure elucidation with major implications in biological and analytical chemistry and in materials design. Direct determination of absolute configuration is rarely feasible, and assignments typically rely on assumed reaction mechanisms, comparison of experimental and theoretical VCD spectra, or correlations of (optical or NMR) properties. Machine learning (ML) models that predict chiral observable properties from molecular structure can assist in configuration assignment, curation of experimental databases, and reporting of chiral properties. However, among the molecular descriptors available in accessible software packages, only a few encode chirality and are suitable for chiral ML models.

We explored two approaches to chirality encoding: (i) descriptors derived from the latent space of SMILES heteroencoders through latent space arithmetic [1]; and (ii) atom-centered ESEC-like descriptors based on scalar triple products calculated from three first-order moment vectors [2] centered on each chiral atom, restricted to spheres of neighboring atoms to reduce conformation dependence. ML models were built using the Random Forest algorithm. The first approach was evaluated on a dataset of 3858 molecules (1929 enantiomeric pairs) for the prediction of elution order in chiral chromatography and for the assignment of intrinsic chirality labels (R/S configuration and canonical SMILES @/@@ notation). The second approach was assessed on elution order prediction across four HPLC columns and on the prediction of compatibility between a chiral molecule and a theoretical VCD spectrum.

Both descriptor strategies demonstrated the feasibility of encoding molecular chirality for ML-based prediction of chiral properties, offering complementary tools for stereochemical assignment and analytical method development.

Bibliography :

[1] N. Baimacheva; X. Gao; J. Aires-de-Sousa. J. Cheminform. 17 (2025) 137.
[2] J. Peeters; P. De Gauquier; F. Ameli; Y.V. Heyden; D. Mangelings; K. Vanommeslaeghe. PLoS One 20 (2025) e0322580.

Acknowledgments :

S
u
p
p
o
r
t

b
y

n