Exploring molecular heteroencoders with latent space arithmetic: atomic descriptors and molecular operators

Joao Aires-de-Sousa[1], Xinyue Gao[2], Natalia Baimacheva [3]

[1]LAQV and REQUIMTE, Chemistry Department, NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal. E-mail: jas@fct.unl.pt.
[2]Université Paris Cité, Faculty of Sciences, F-75013 Paris, France.
[3]University of Strasbourg, Faculty of Chemistry, 4, Blaise Pascal str., 67081, Strasbourg, France.

Optimization of chemical structures in the latent space became a major tool for *de novo* design but it is much based on random perturbations of a starting molecule. The establishment of arithmetic operations (rules) in the latent space associated with desired transformations opens the way to better guided algorithms for molecular optimization. We investigated the changes in the latent space vectors caused by modification of the input SMILES strings, to derive atomic descriptors and molecular operators.

A variational heteroencoder based on Recurrent Neural Networks [1] trained with SMILES linear notations was used to derive atomic descriptors: delta latent space vectors (DLSV) obtained from the original SMILES of the whole molecule and the SMILES of the same molecule with the target atom changed. Different perturbations of the target atom were explored, namely the change of the atomic element, the replacement by a character of the model vocabulary not used in the training set, or the removal of the target atom from the SMILES. Unsupervised mapping of the DLSV atomic descriptors with t-distributed stochastic neighbor embedding (t-SNE) shows a remarkable clustering according to the atom element, hybridization, atom type and aromaticity. Atomic DLSV descriptors were used to train QSPR models to predict ^{19}F NMR chemical shifts [2] with Random Forests and Gradient Boosting Regressor; predictions were obtained with R^2 up to 0.89 and mean absolute errors up to 5.5 ppm for an independent test set of 1046 molecules. Latent representations from a Transformer model [3] were similarly employed. DLSV could also be used as molecular operators in the latent space. We explored the DLSV of a halogenation (H→F substitution). It was summed to the LSV of 4135 new molecules with no fluorine atom and decoded into SMILES, yielding >99% of valid SMILES, a high percentage of which incorporating fluorine and most of them with no other structural change.

Bibliography :

[1] R. Winter; F. Montanari; F. Noé; D.-A. Clevert. Chem. Sci. 10 (2019) 1692–1701.
[2] P. Penner; A. Vulpetti. J. Comput. Aided. Mol. Des. 38 (2024) 4.
[3] Y. Yoshikai; T. Mizuno; S. Nemoto; H. Kusuhara. Nat. Commun. 15 (2024) 1197.

Acknowledgements

Support by national funds from Fundação para a Ciência e Tecnologia (FCT/MCTES), Portugal, under grants LA/P/0008/2020 DOI 10.54499/LA/P/0008/2020, UIDP/50006/2020 DOI 10.54499/UIDP/50006/2020 and UIDB/50006/2020 DOI 10.54499/UIDB/50006/2020. Co-funding by the European Union through scholarships awarded to N.B and X.G by the Erasmus Mundus Joint Masters ChEMoinformaticsplus project (Programme ERASMUS2027, ERASMUS-EDU-2021-PEX-EMJM-MOB, Project number 101050809).