Prediction of Odour (PrOdour) – Data Exploration

Slavica Subic[1], Veniamin Morgenshtern[2], Thilo Bauer[3, 4]

[1]*Advanced Signal Processing and Communications Engineering, Elite Master's Study Programme, FAU Erlangen-Nürnberg*

[2]*Chair of Multimedia Communications and Signal Processing, Dept. Electrical, Electronic and Communication Engineering, FAU Erlangen-Nürnberg*

[3]*Computer Chemistry Centre (CCC), Department of Chemistry and Pharmacy, FAU Erlangen-Nürnberg*

[4]*Fraunhofer Institute for Process Engineering and Packaging IVV*

Where exactly does the odour come from? What causes a substance to produce a certain scent?

One could approach answering this question in many ways, and one way would be to look at the language used to represent the molecules. Structural formulae are exactly that, and that is where this data exploration task starts. Looking at the language of chemistry, and individual words in its dictionary, we search for the ones most responsible for particular odours.

Starting from molecular formulae, parsing them into millions of words (features, sub-molecular structures), and looking for largest overlapping structures, we collect a dictionary of molecular features. With data collected from experiments where experts evaluated the sensory properties of substances, it was possible to build a data set containing the applicability of a list of odour descriptors to the substances used in the experiment. Using LASSO (Least Absolute Shrinkage and Selection Operator) regression regularization for feature extraction, we aim to find the features most responsible for the applicability scores by which descriptors were rated in the experiment - in other words, features most responsible for the expression of a certain odour. By grouping descriptors with clustering techniques, we can explore similarities of features among the groups and search for chemically meaningful features that define a descriptor cluster.

[1] A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, and B. A. Grzybowski - Organic Chemistry as a Language and the Implications of Chemical Linguistics for Structural and Retrosynthetic Analyses, *Angew. Chem.* **2014**, 8246-8250, DOI: 10.1002/ange.201403708
[2] T. Hastie, R. Tibshirani, J. Friedman – The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, Second Edition, corrected 12th printing - 13. Jan. **2017**.