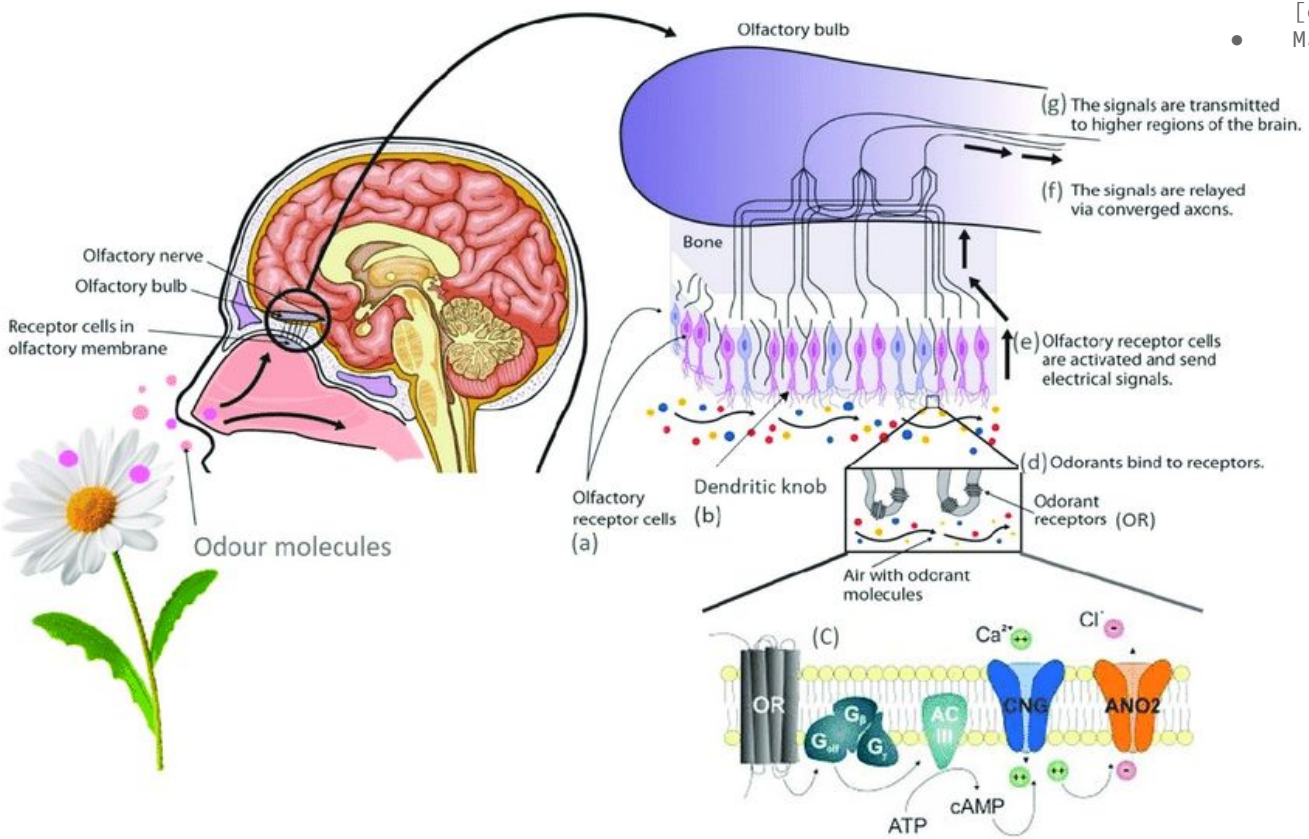
AI Bachelor Projects

Predicting human olfactory perception from the molecular structure of odorants

Human olfactory system

- Odorants:organic volatile molecules
- Millions of different odorants
- ~10 million olfactory receptor cells [dogs: ~300M]
- ~400 odorant receptor types [dogs: ~1000]
- Many-to-many rel. odorants & receptors



Note: Adopted from “everage and Food Fragrance Biotechnology, Novel Applications, Sensory and Sensor Techniques: An Overview” by Alice Vilela et al.

Odorants & odors

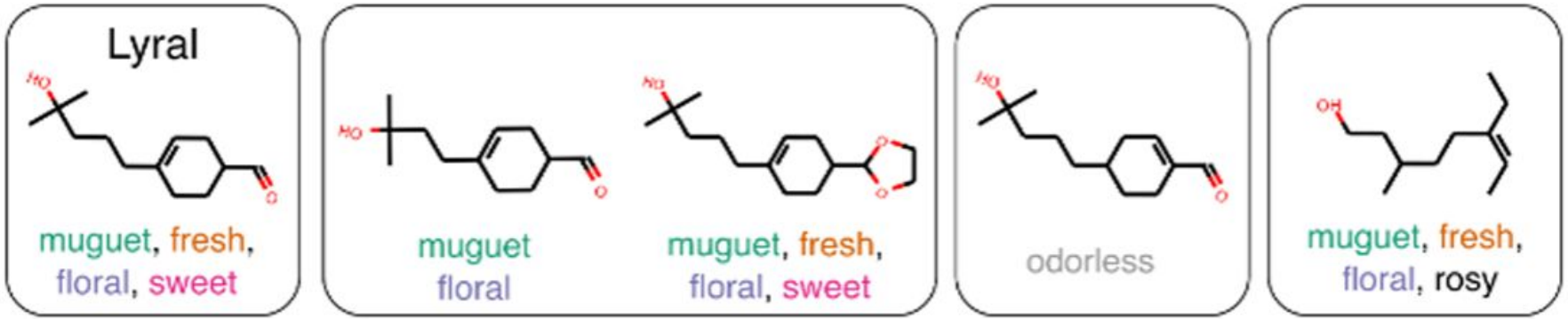


Figure 2: The molecule called Lyral as a reference on the left, followed by two structurally and perceptually similar molecules. The two molecules on the right show that a molecule with a very small structural change can have a completely different odor, while a molecule with a completely different structure can have a very similar odor. Example taken from Ohloff *et al.* [18].

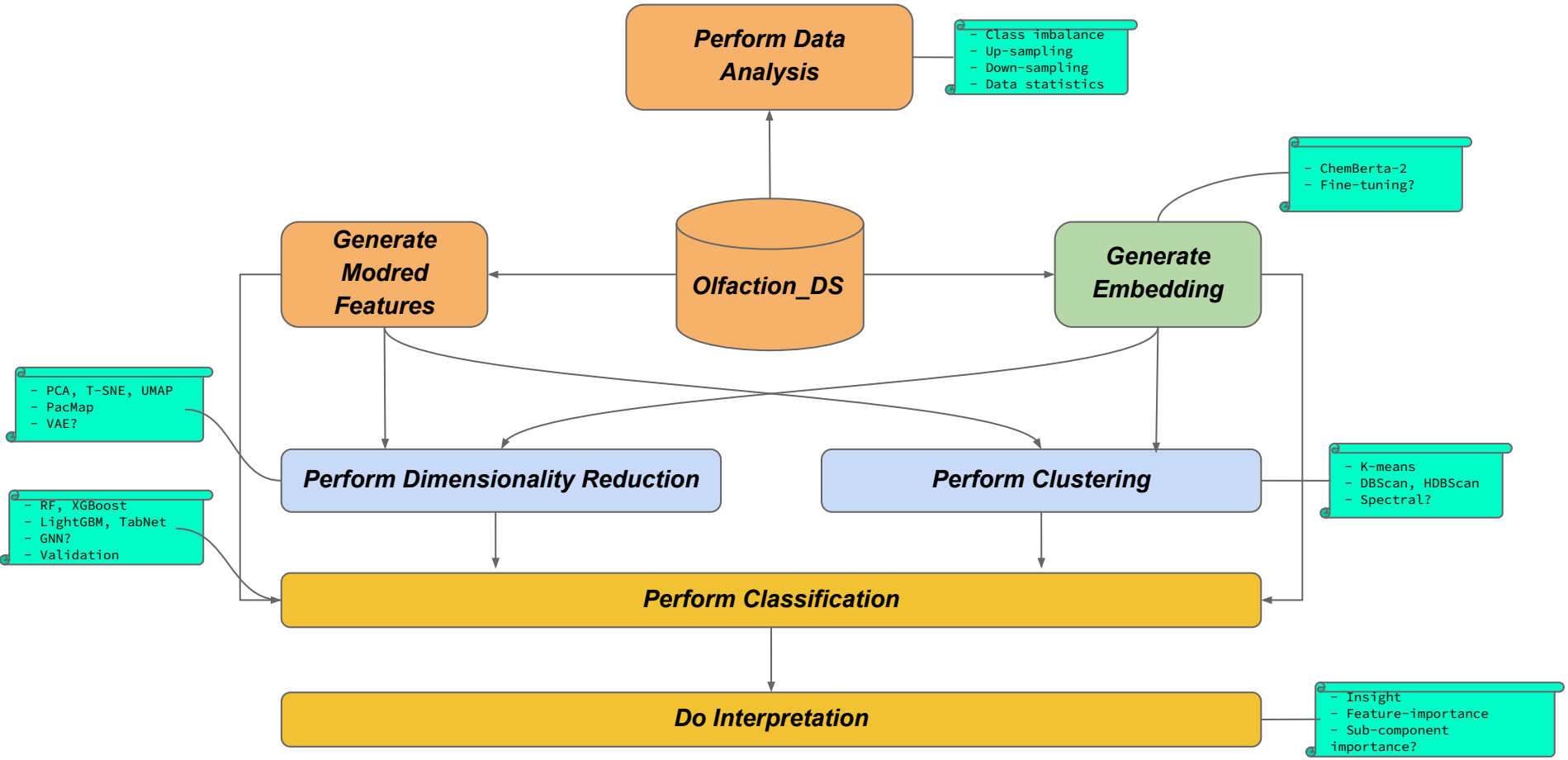
Note: Adopted from the master thesis (Stijn Sluis)

Data Set (Columns)

Table 1: Example of structure-odor data. The odorants are identified by their compound ID, the structure is stored as an isomeric SMILES string and each compound can have multiple descriptors, in no particular order.

Compound ID	Isomeric SMILES	Descriptors	Molecular Weight	Descriptor Count
4	CC(CN)O	fish	75.068	1
49	CC(C)C(=O)C(=O)O	fruit	116.047	1
58	CCC(=O)C(=O)O	lactone, sweet, cream, caramel, brown, fat	102.032	6
70	CC(C)CC(=O)C(=O)O	fruit	130.063	1
72	C1=CC(=C(C=C1C(=O)O)O)O	phenol, balsamic	154.026	2

Note: Adopted from the master thesis (Stijn Sluis)


Perform Data Analysis
- Class imbalance
- Up-sampling
- Down-sampling
- Data statistics
- ChemBerta-2
- Fine-tuning?
Generate Modred Features
Olfaction_DS
Generate Embedding
- PCA, T-SNE, UMAP
- PacMap
- VAE?
Perform Dimensionality Reduction
Perform Clustering
- K-means
- DBScan, HDBScan
- Spectral?
- RF, XGBoost
- LightGBM, TabNet
- GNN?
- Validation
Perform Classification
Do Interpretation
- Insight
- Feature-importance
- Sub-component importance?

References

1. Predicting odor with a SMILES; Master thesis - Stijn Sluis - already shared with you.
2. A principal odor map unifies diverse tasks in olfactory perception; BRIAN K. LEE et al.
3. Identifying the Primary Odor Perception Descriptors by Multi-Output Linear Regression Models; Xin Li et al.
4. Simplified molecular-input line-entry system (SMILES)
5. Mordred: a molecular descriptor calculator; Hirotomo Moriwaki et al.
6. ChemBERTa-2: Towards Chemical Foundation Models; Walid Ahmad et al.
7. ChemBERTa Github
8. ChemBert2a Smiles embeddings for beginners (Kaggle)
9. ChemBert2a model download (ChemBERTa-77M-MTR)
10. This project Github
11. Auto-ML for choosing number of DR components??