Computing a collection of physico-chemical descriptors using a software package (e.g. formerly Dragon, now AlvaDesc) has been popular in olfaction research in the past (cite Haddad, Saito, etc). Those software packages are heavily used in computational drug discovery, and the selection of descriptors are well tuned to represent the space of drug-like molecules. Odorant molecules however are typically smaller than drug-like molecules and also differ significantly in other chemical features (Ruddigkeit et al 2014). Therefore, there exists the possibility that the descriptor sets provided by software packages like AlvaDesc are not ideally suited to describe the chemical space of odorants.

In pursuit of a numerical representation of odorant molecules which is better suited to odor space, we chose a bottom-up approach to defining features that is based on structural fingerprints. We defined a set of structural features in the SMARTS chemical pattern matching language (Daylight Chemical Information Systems Inc). Table 1 shows the SMARTS patterns and their descriptions we used to describe odorants. It has been designed to capture functional groups and other substructure features that are often present in odorants: e.g. carboxyl, aldehyde or ketone groups, ester bonds, aromatic rings, etc. It also encodes the presence of ortho-, meta- and para-substituted rings; a number of terpenoid scaffolds, and also single-bond aliphatic chains of lengths from 6 to 11 C atoms.

Compared to physicochemical features which are often abstract and difficult to interpret chemically, substructure fingerprints promise to provide better chemical insight into structure-activity relationships.

|  |  |  |
| --- | --- | --- |
| **Index** | **SMARTS** | **description** |
| 0 | \*-C(=O)-[OH1] | carboxylic acid |
| 1 | [CH1]=O | aldehyde |
| 2 | C-C(=O)-[O]-C | ester |
| 3 | C-C(=O)-[S]-C | thioester |
| 4 | [!O&!S]-C(=O)-[!O&!S] | ketone |
| 5 | [OX2H][CX4&!$(C([OX2H])[O,S,#7,#15]),c] | alcohol |
| 6 | c1ccccc1 | benzyl |
| 7 | C~C(~C)~C~C~C~C(~C)~C | monoterpene |
| 8 | [#8]1~[#6]~[#6]~[#6]~[#6]1 | furanoid |
| 9 | o1cccc1 | furan |
| 10 | [NH2][C] | primary amine |
| 11 | [NH](C)C | secondary amine |
| 12 | [NH0](C)(C)C | tertiary amine |
| 13 | [N,n]1~[C,c]~[C,c]~[C,c]~[C,c]~[C,c]1 | pyridine |
| 14 | [n,N]1~[C,c]~[C,c]~[C,c]~[C,c]1 | pyrrole |
| 15 | [N,n]1~[C,c]~[C,c]~[N,n]~[C,c]~[C,c]1 | pyrazine |
| 16 | [#16]1~[#6]~[#7]~[#6]~[#6]1 | thiazoline |
| 17 | [!#8]~C-S-C~[!#8] | thioether |
| 18 | [$(C-S-S-C),$(C-S-S-S-C)] | sulfide |
| 19 | [#6]-[SH] | thiol |
| 20 | [#6]=[#6] | Alkene |
| 21 | [#16] | sulfur |
| 22 | [#7] | nitrogen |
| 23 | [#8] | oxygen |
| 24 | [R] | ring |
| 25 | [CH3]-\*-[CH2]-\* | terminal 4-bond chain |
| 26 | \*!@\*@\*!@\* | ortho-substituted rings (1 bond) |
| 27 | \*!@\*@\*@\*!@\* | meta-substituted rings (2 bonds) |
| 28 | \*1(!@\*)@\*@\*@\*(!@\*)@\*@\*@1 | para substituted 6-ring but not fused ring |
| 29 | C~C(~C)~[R1]1~[R1]~[R1]~[R1](~C)~[R1]~[R1]~1 | menthane scaffold |
| 30 | C~C(~C)~2~[R2]1~[R2]~2~[R1]~[R1](~C)~[R1]~[R1]~1 | carene scaffold |
| 31 | C~C(~C)~[R2]12~[R1]~[R2]~2~[R1](~C)~[R1]~[R1]~1 | thujane scaffold |
| 32 | C~C2(~C)~[R]1~[R]~[R]~2~[R](~C)~[R]~[R]~1 | pinane scaffold |
| 33 | [!H]~[!H]2(~[!H])~[R]1~[R]~[R]~[R](~[!H])~2~[R]~[R]~1 | camphane scaffold |
| 34 | [!H]~[!H]2(~[!H])~[R]~[R](~[!H])1~[R]~[R]~2~[R]~[R]~1 | fenchane scaffold |
| 35 | C(-C)(-C)(-C)-C | quadra C, like in many terpenes |
| 36 | C-C-C-C-C-C | six carbon single bond |
| 37 | C-C-C-C-C-C-C | seven carbon single bond |
| 38 | C-C-C-C-C-C-C-C | eight carbon |
| 39 | C-C-C-C-C-C-C-C-C | nine carbon |
| 40 | C-C-C-C-C-C-C-C-C-C | ten carbon |
| 41 | C-C-C-C-C-C-C-C-C-C-C | eleven carbon |

Table 1: SMARTS V4, with explicit single bond aliphatic chains (keys 36 to 41).