

Chapter 2

The Chemical Space of Flavours

Lars Ruddigkeit and Jean-Louis Reymond

2.1 Introduction

In the complex array of molecules composing foods, flavourant molecules, although present in relatively small amounts, play a central role in determining the food flavour in terms of taste and smell. Taste molecules, which have very diverse chemical structures and properties, interact directly with receptors in the mouth to trigger taste perceptions of bitter, sweet, sour, acidic, salty and umami [1]. Fragrances are generally small, apolar and volatile compounds, which must reach olfactory receptor neurons in the upper part of the nose to trigger the complex perception of smell through interactions with approximately 900 genetically distinct G-protein-coupled olfactory receptors [2–6]. Fragrances are also used as ingredients in perfumes, soaps, shampoos or lotions. Classifications of fragrances, according to their perceived smell, produce tens to hundreds of fragrance families, although a general characterization system of smell is still difficult due to perceptual qualities [7]. The relationship between structural types and odour types is very diverse. Herein, we discuss flavourant molecules collected from the open-access databases, SuperScent [8], Flavornet [9], BitterDB [10] and SuperSweet [11], in an overall perspective of the chemical space classification of molecules to convey a global understanding of this molecular class independent of detailed structure–activity relationships [12]. This global view provides a conceptual framework to understand the chemical structural diversity of taste and smell and suggests approaches to discover new flavours through chemical space exploration.

J.-L. Reymond (✉) · L. Ruddigkeit
Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3,
3012 Bern, Switzerland
e-mail: jean-louis.reymond@dcb.unibe.ch

2.2 Flavour Molecules

2.2.1 Databases of Organic Molecules

Organic molecules consist of a few tens of atoms of various types (carbon, hydrogen, nitrogen, oxygen, sulphur, halogens and a few others) linked together via kinetically stable covalent single or multiple bonds. The atoms and their connectivity pattern including their three-dimensional relative positions define the molecule's identity, its molecular shape and its physicochemical and biological properties. Since the discovery of organic molecules, as the elementary building blocks of living matter, many millions of different organic molecules have been reported in the literature either as naturally occurring compounds or as the products of chemical syntheses.

Most efforts have been devoted to the area of medicinal chemistry where molecules are investigated for their drug properties. The cumulated knowledge acquired there has been placed, in part, in the public domain thanks to open-access initiative, such as the US National Institute of Health PubChem database, in which the structure and possible biological evaluation of more than 30 million of organic molecules are freely accessible [13]. The Royal Society of Chemistry runs a similar but broader open-access archive in the form of ChemSpider, a repository in which authors are encouraged to deposit their structures [14]. Additional public databases of molecules of medicinal interest are listed in Table 2.1, including collections of commercially available compounds in ZINC [15], annotated database of bioactive molecules such as ChEMBL [16] and DrugBank [17], and very large databases of theoretically possible molecules covering the entire range of what is feasible with organic chemistry, such as the chemical universe databases GDB-11 [18], GDB-13 [19] and GDB-17 [20], which list all organic molecules possible up to 11, 13 and 17 atoms obeying simple rules for chemical stability and synthetic feasibility [21].

When considering flavourants, hundreds of thousands of molecules have been investigated for their fragrant properties by various fragrance companies worldwide. However, there has been only very limited effort to establish a broad repository of flavour molecules. Nevertheless, several relatively small databases have been made accessible online in the last few years: SuperScent [8] and Flavornet [1], which list almost 2000 documented fragrances and their properties; BitterDB [10], which lists 606 molecules with documented bitter taste, containing many alkaloids; and SuperSweet [11], which list 342 molecules with proven or likely sweet taste, containing, in particular, a broad range of glycosides. When combined together, SuperScent and Flavornet assemble to a collection of 1760 different fragrance molecules, here named FragranceDB. BitterDB and SuperSweet similarly combine to 806 taste molecules, here named TasteDB.

2.2.2 Property Profiles

The properties of drug-like molecules have been extensively discussed in the literature focussing on the characteristics necessary for oral bioavailability in the form of

Table 2.1 Databases of organic molecules as of December 2013

Database	Description	Size	Web address
PubChem	Database of known molecules from various public sources	38.8 M	http://pubchem.ncbi.nlm.nih.gov
ChemSpider	Integrated resource of Royal Society of Chemistry	28.0 M	http://www.chemspider.com/
ZINC	Commercial small molecules	13.5 M	http://zinc.docking.org
ChEMBL	Bioactive drug-like small molecules annotated with experimental data	1.5 M	https://www.ebi.ac.uk/chembl/db
DrugBank	Experimental and approved small-molecule drugs	6825 M	http://www.drugbank.ca
SuperScent	Database of scents from literature	1591 M	http://bioinf-applied.charite.de/superscent/
Flavornet	Volatile compounds from the literature based on GC–MS	738 M	http://flavornet.org
FragranceDB	SuperScent + Flavornet	1760 M	–
SuperSweet	Database of carbohydrates and artificial sweeteners	342 M	http://bioinf-applied.charite.de/sweet/index.php?site=home
BitterDB	Database of bitter Cpd from literature and Merck index	606 M	http://bitterdb.agri.huji.ac.il/bitterdb/
TasteDB	SuperSweet + BitterDB	806 M	–
GDB-11	Possible small molecules up to 11 atoms of C, N, O, F	26.4 M	http://www.gdb.unibe.ch
GDB-13	Possible small molecules up to 13 atoms of C, N, O, S, Cl	980 M	http://www.gdb.unibe.ch
GDB-17	Possible small molecules up to 17 atoms of C, N, O, S, halogen	166.4 G	http://www.gdb.unibe.ch

Lipinski's "rule of five", which sets boundaries to molecular weight ($MW \leq 500$ Da), the octanol–water partition coefficient P ($\log P \leq 5$), and the number of hydrogen-bond-donor atoms ($HBD \leq 5$) and hydrogen-bond-acceptor atoms ($HBA \leq 10$) [9]. A narrower definition with tighter boundaries on molecular weight ($MW \leq 300$ Da), polarity ($\log P \leq 3$) and flexibility in terms of rotatable bonds ($RBC \leq 3$) have also been defined to select molecules suitable as "fragments", which are generally smaller molecules showing weak activities, but which can be optimized by adding substituents [22].

A similar set of boundaries has not been proposed for flavours. While the property ranges necessary for taste molecules is a priori rather large, one can guess that for fragrant molecules, upper values in terms of molecular weight and polarity are necessary to enable a minimum amount of volatility, which is the key feature necessary for fragrances to reach their site of action. To understand which boundaries are suitable, we present herein the property profiles of the flavour collections, FragranceDB and TasteDB, and compare them with those of drug-like molecules in ChEMBL (bioactive molecules) [16], ZINC (commercial compounds for bioactivity screening) [15] and GDB-13 (possible molecules up to 13 atoms) [19].

The heavy-atom count (HAC, heavy atoms = all non-hydrogen atoms) profile shows that FragranceDB contains predominantly very small molecules with an upper boundary at approximately 21 atoms (Fig. 2.1a). A frequency peak appears at

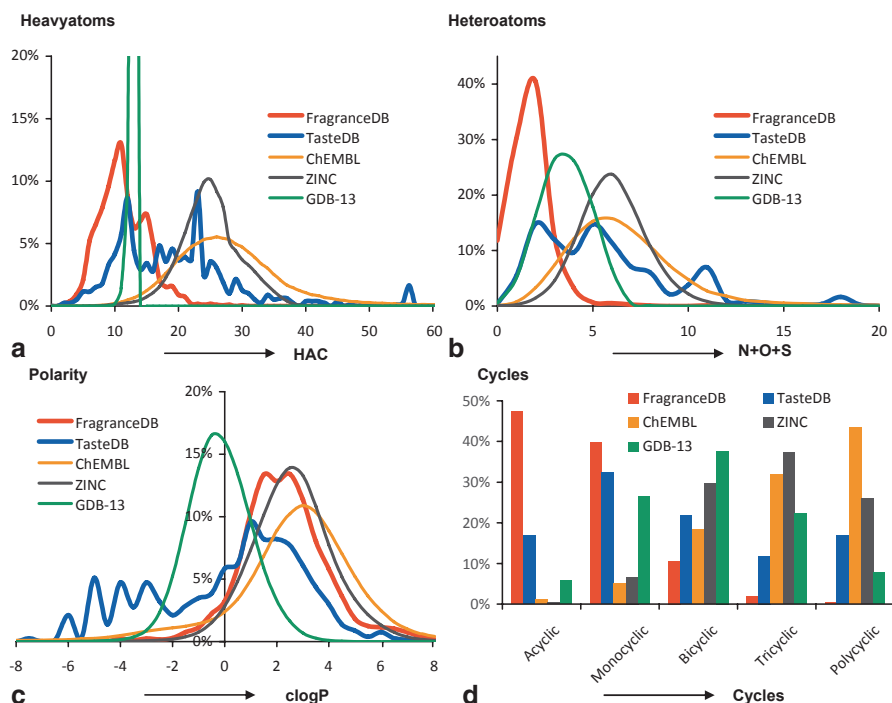


Fig. 2.1 Property histograms of fragrance and taste databases in comparison to ChEMBL, ZINC and GDB-13

9–11 heavy atoms corresponding to a diverse constellation comprising aliphatic linear and branched alkenes, aldehydes, alcohols, ketones and esters, various simple benzene, phenol and benzaldehyde analogues, furanones and monoterpenes. FragranceDB shows only very limited size overlap with drugs (ChEMBL) and commercial drug-like compounds (ZINC), which peak at the size of 20–30 heavy atoms. The chemical universe database GDB-13 falls within the size boundary of FragranceDB and offers a very large diversity of potential fragrances, including, in particular, analogues of monoterpenes with 10–11 atoms. TasteDB, on the other hand, covers a much broader size range, in agreement with the fact that flavours do not require volatility to reach their site of action. An abundance peak is nevertheless visible at 10–12 atoms and corresponds to various hexoses and their reduced hexitols, together with monoterpenes (menthone, camphor, citronellol), coumarins, anisols and some amino acids. Taste molecules in the size range of drugs (20–30 atoms) correspond to simple di-glycosides as well as various alkaloids and aromatic compounds and peptides. The frequency peak at HAC = 56 corresponds to steviol glycosides listed in the database SuperSweet [23].

The heteroatom composition of flavours versus drugs is best compared by considering the sum of oxygen, nitrogen and sulphur atoms (Fig. 2.1b). Halogens are

rather rare in flavours, although organochlorine compounds such as sucralose have a sweet taste. FragranceDB stands out with a very low number of heteroatoms peaking at just two heteroatoms, which are mostly oxygen atoms as found in volatile aldehydes and ketones, alcohols, carboxylic esters and acids. As for the HAC profile, the overlap with drug molecules in ChEMBL and drug-like compounds in ZINC, in terms of heteroatom numbers, is small because drug molecules generally have a larger number of functional groups due to their larger size. Note that drug molecules very often contain multiple nitrogen atoms as well as amide bonds which are almost entirely absent in fragrances. The GDB-13 database displays relatively more heteroatoms despite of the small molecular size due to a combinatorial enumeration favouring highly functionalized molecules. The heteroatom profile of TasteDB is much broader in line with the broader range of molecular weights, again a consequence of the abundance of sweet-tasting oligosaccharides, including the steviol glycosides with a high density of hydroxyl groups.

A further insight into global properties can be gained by considering the logarithm of the calculated octanol/water partition coefficient $\text{clog}P$ as a measure of polarity (Fig. 2.1c). $\text{Clog}P$ indicates lipophilic molecules at high positive values, water-soluble molecules at strongly negative values and amphiphilic molecules around zero. Here, FragranceDB overlaps nicely with the drug and drug-like molecules in ChEMBL and ZINC by covering the range $0 < \text{clog}P < 5$, which is a polarity range well suitable for rapid diffusion in biological media. This probably reflects the necessity of fragrances to diffuse from the gas phase to the olfactory neurons to reach their receptors, which requires properties similar to those necessary for drugs to reach their site of action. This property is also shared by the majority of TasteDB; however, in this case a significant fraction of the database extends into negative $\text{clog}P$ values, comprising monosaccharides, disaccharides and related polyols, steviol glycosides, and amino acids and peptides such as aspartame. It should be noted that GDB-13, which reflects the combinatorial enumeration of the entire chemical space, peaks at $\text{clog}P=0$ due to the large fraction of cationic polyamines in the database which extend into negative $\text{clog}P$ values. Due to the large size of GDB-13, however (almost one billion molecules), the database still contains an extremely large number of molecules in the polarity range of fragrances compared to the other databases.

Structural rigidity is a defining molecular property in drugs because conformational entropy strongly reduces binding affinity. Generally, molecules with large number of cycles are more rigid and have a better chance to bind strongly and selectively to their target. Remarkably, FragranceDB is predominantly a collection of acyclic compounds, with an abundance of acyclic aliphatic alcohols, aldehydes, acids and esters, such as butter and fruit aroma (Fig. 2.1d). Monocyclic molecules are also abundant, in particular cyclic terpenes, such as limonene or menthol; and monocyclic aromatic molecules, such as cinnamaldehyde. The abundance of acyclic and monocyclic compounds in FragranceDB contrasts with the typical drug molecules in ChEMBL and ZINC, which tend to be polycyclic, also as a consequence of their size. The combinatorial enumeration of molecules in GDB-13 correspond-

ing to the size range of fragrances favours bicyclic molecules as the most abundant topology. TasteDB contains mostly monocyclic molecules, many of which are monosaccharides, but also extends into polycyclic molecules due to the presence of oligosaccharides and steroids in the collection.

2.3 Visualizing the Chemical Space of Flavours

2.3.1 The Chemical Space

In the context of organic chemistry, the term “chemical space” describes the ensemble of all known and/or possible molecules, but also the various multidimensional “property spaces” that can be defined by assigning dimensions to numerical descriptors of molecular structures [24, 25]. Such property spaces provide a general organization principle, which helps understand the molecular diversity available in large databases often containing many millions of molecules (Table 2.1). To obtain visual representations of property spaces, one usually performs principal component analysis (PCA) and representation of the (PC1, PC2)-plane containing the largest variance. This mathematical procedure is equivalent to taking a picture of the multidimensional space from the angle showing the largest diversity (Fig. 2.2) [26–32].

Thousands of numerical descriptors of molecular structure are known, and the number of possible property spaces is therefore unlimited. Recently, we showed that the chemical space of molecular quantum numbers (MQN), a set of 42 simple integer value descriptors counting atoms, bonds, polar groups and topological fea-

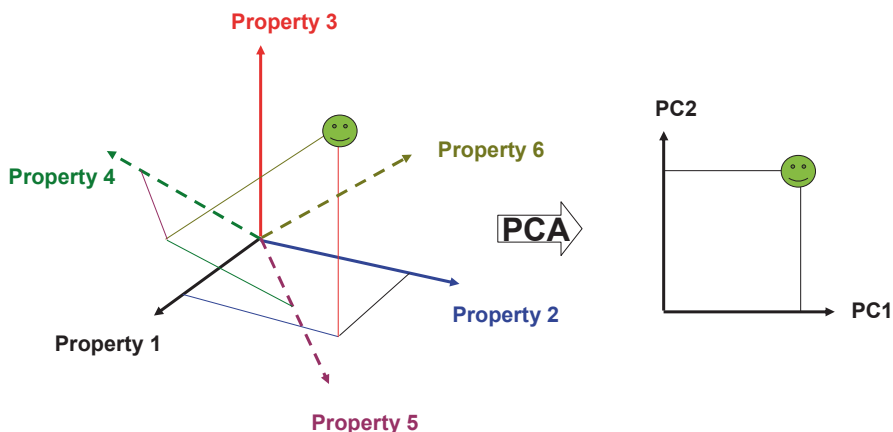


Fig. 2.2 Principal component analysis (PCA) projects a multidimensional property space into the plane of the largest variance

tures, such as cycles, provides a simple classification system of large databases and produces insightful (PC1, PC2)-maps for a variety of databases [33]. These PC-maps separate molecules by their mass, the number of cycles and rotatable bonds and their polarity, as can be illustrated by colour coding with property values. We have used such MQN-space maps to design interactive searchable maps of various public databases including zoom-in function and visualization of the molecules with links to their source database in the form of a “Google-map”-type application freely available from www.gdb.unibe.ch [34]. A related classification system and interactive visualization system were also realized using a simplified molecular-input line-entry system (SMILES) fingerprint (SMIfp), counting the occurrences of characters occurring in the SMILES representation of molecules [35]. One of the most striking features of these classification systems is that they group molecules by their pharmacophoric features and biological activities, and thus enable virtual screening in prospective searches [36].

2.3.2 Maps of the Flavours—Chemical Space

To gain an overview of the chemical space of flavours, we have performed a PCA visualization of the merged database containing FragranceDB and TasteDB, totaling 2517 compounds. These databases are represented in their (PC1, PC2)-plane which can be considered as a general 2-D map of their chemical space.

For the case of the MQN-space representation shown in Fig. 2.3a–d, the molecules spread by increasing size in the horizontal PC1-axis covering 67.97% of data variability. The vertical PC2-axis separates molecules by structural rigidity covering 15.54% of data variability. The total data variability represented by the (PC1, PC2)-plane amounts to 83.51%, which is typical for the projection of large databases from MQN-space. The molecules are grouped in descending diagonal stripes grouping molecules with an increasing number of cycles and ring atoms. Acyclic and monocyclic compounds are the most abundant category in FragranceDB, respectively, TasteDB. The category map in Fig. 2.3d shows that FragranceDB is essentially an acyclic/monocyclic compound database of small molecules, while TasteDB extends in large and polycyclic molecules.

In the maps of the SMIfp-space shown in Fig. 2.3e–h, the PC1-axis covers 66.9% of data variability and the PC2-axis covers 18.97%, totalling to 85.87% of data variability visible in the (PC1, PC2)-plane. Molecules spread by increasing size along the descending diagonal (Fig. 2.3e). The horizontal PC1-axis separates molecules according to the number of nonaromatic carbons (Fig. 2.3g), and the vertical axis according to the number of aromatic carbons (Fig. 2.3f). When comparing the category map in Fig. 2.3h with the property values in Fig. 2.3e–h, one can appreciate that FragranceDB contains mostly nonaromatic molecules, which correspond, in large part, to the acyclic molecules seen in the MQN-map of Fig. 2.3b. On the

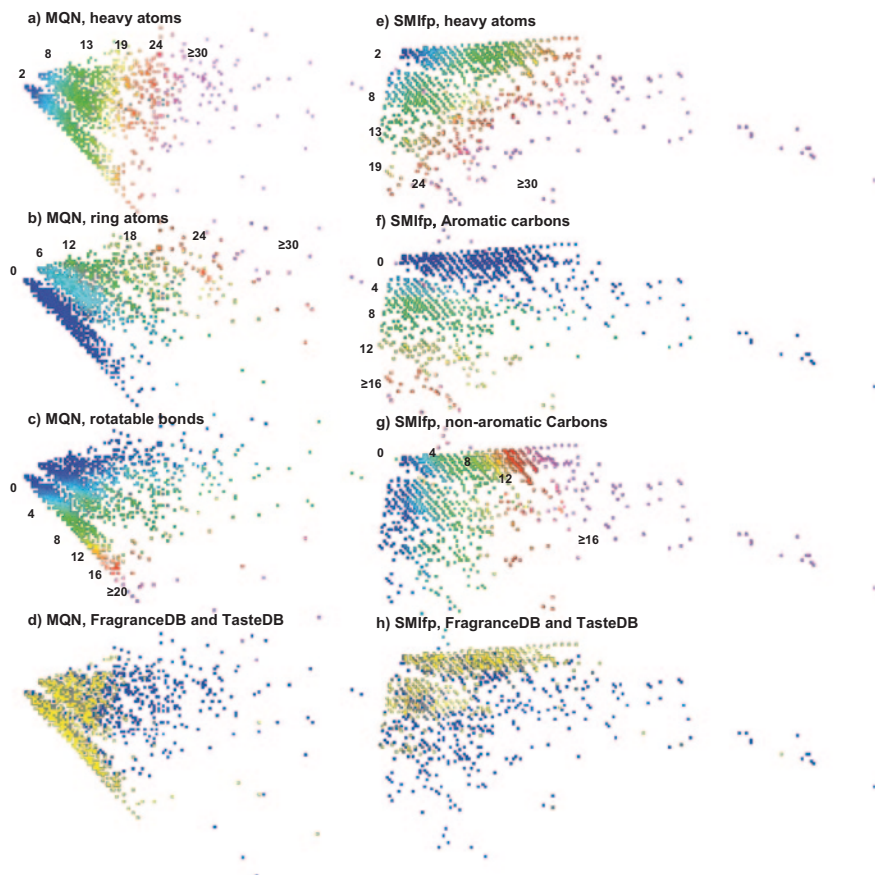


Fig. 2.3 Colour-coded maps of the flavours and taste chemical space. (PC1, PC2)-maps for PCA of the 42-dimensional MQN-space (**a–d**) and 34-dimensional SMIfp-space (**e–h**) are colour-coded by increasing value of the indicated property in the scale *blue–cyan–green–yellow–orange–red–magenta* with the corresponding value indicated on the map, for (**d, h**) *yellow* = flavour, *blue* = taste, and *grey* = pixel with mixed categories

other hand, TasteDB spans a broader range of SMIfp values, in particular, many taste molecules contain a large number of aromatic carbon atoms.

Overall, the MQN- and SMIfp-maps of the combined FragranceDB and TasteDB illustrate the broad range of structural types encountered in flavours. Note that the (PC1, PC2)-plane does not reflect any distribution of polarity properties. These are generally to be found in the PC3-dimension which requires additional representations not discussed here.

2.4 Fragrance Analogues in Chemical Space

2.4.1 Similarity Searching by City-Block Distance

The MQN- and SMIfp-spaces discussed in the previous section allow not only simple PCA-mapping of chemical space but also an extremely fast search for analogues using dedicated online browsers, which are freely accessible for use at www.gdb.unibe.ch. The browsers search for analogues of any query molecule as drawn in the query window using the principle of nearest neighbours in the multidimensional property space by measuring the city-block distance (CBD) between molecules. The CBD separating two molecules is the sum of the absolute differences between descriptor pairs across the 42 MQN and the 34 SMIfp descriptors. By pre-organizing databases according to file systems named X-MQN and X-SMIfp, databases of many millions of compounds can be searched within seconds for CBD_{MQN} and CBD_{SMIfp} neighbours, respectively, of any query molecule [37].

We have performed extensive comparisons between CBD and the more common Tanimoto coefficient as pairwise similarity measured between molecules and found the performance of both methods to be largely comparable, in particular, for the high-similarity pairs, i.e. both similarity measures will indicate the same molecules as the most similar, but differ substantially when considering very dissimilar compounds. On the other hand, searching according to the Tanimoto similarity is much slower than searching by CBD. The X-MQN and X-SMIfp systems incorporate additional options to direct any analogue search by restricting certain parameters in the analogues shown to certain subclasses (charges, HBD, HBA, elemental formula or compliance with drug-likeness rules), as visible in the search-window interface for the database ZINC using MQN-similarity searching (Fig. 2.4).

2.4.2 Fragrance Analogues from MQN-Space

The chemical space neighbourhood search gives particularly interesting results when considering fragrances. In the context of an analogue search within databases of commercially available compounds such as ZINC, one can identify interesting analogues by MQN- or SMIfp-similarity searching by preserving the number of HBD and HBA atoms, the electrostatic charges and optionally the elemental formula to avoid the selection of analogues with multiple heteroatoms, in particular nitrogen-rich heterocycles which are particularly abundant due to their importance in drug-discovery applications. Only the MQN-similarity search is exemplified here, but the SMIfp-similarity gives comparable results.

In Fig. 2.5, the MQN neighbours of the peppermint fragrance component, menthone, are shown. There are 27 commercially available compounds within $CBD_{MQN} \leq 12$, which is a useful distance boundary in the MQN-space [37]. These commercial analogues not only contain menthone itself (hit no. 1), a regioisomer (hit no. 2), but also various other cyclohexanones with the same number of acyclic

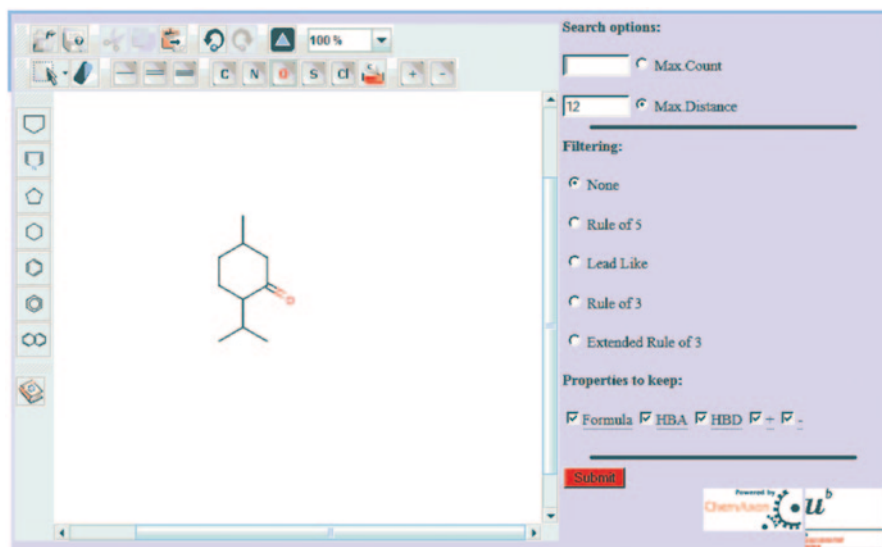


Fig. 2.4 The search-window option to identify the nearest neighbours of menthone in the MQN-space of the database ZINC (see also Sect. 1.4.2)

carbon atom substituents (hit nos. 3–9). Cycloheptanones (hit nos. 13–15) and cyclopentanones (hit nos. 26–27) are also proposed by the MQN-similarity search.

One can also extend the search to other databases containing a larger diversity of molecules. The chemical universe database GDB-13, which lists 977 million molecules of up to 13 atoms of C, N, O, S and Cl possible following simple rules of chemical stability and synthetic feasibility, is the largest database of small molecules to date [19]. GDB-13 is particularly relevant for fragrance analogue searches since it contains molecules in the size range most populated by fragrances; in particular, the majority of monoterpenes have less than 13 atoms. When applying the MQN-similarity search to typical fragrances, one can appreciate the very large number of high-similarity fragrance analogues that are possible, including isomers (Table 2.2). The vast majority of these molecules are presently unknown, and many do not pose any particular synthetic challenge, suggesting that large numbers of fragrant molecules remain to be explored.

2.5 Conclusion and Outlook

The general properties of flavour molecules, comprising fragrances which are relatively small organic compounds with few polar functional groups, such as to be volatile, and the more polar and diverse taste molecules, define a subset of the chemical space that is clearly separated from the well-known drug-like molecules. A global understanding of chemical space aided by representations such as the

Foodinformatics

Applications of Chemical Information to Food Chemistry

Martinez-Mayorga, K.; Medina-Franco, J.L. (Eds.)

2014, XII, 251 p. 51 illus., 39 illus. in color., Hardcover

ISBN: 978-3-319-10225-2