SUPPLEMENTARY METERIAL

The Structure of Human Olfactory Space

by Alexei A. Koulakov, Brian E. Kolterman, Armen G. Enikolopov, and Dmitry Rinberg

ODORANTS INCLUDED IN THE ANALYSIS

The following odorants were used from the Atlas of Odor Character profiles.

1	698-10-2	Abhexone
2	98-86-2	Acetophenone
3	1122-62-9	Acetyl Pyridine: ortho-Acetyl Pyridine
4	141-13-9	Adoxal
5	77-83-8	Aldehyde C-16 (So-Called) Lower Concentration
6	77-83-8	Aldehyde C-16 (So-Called) Higher Concentration
7	104-61-0	Aldehyde C-18 (So-Called)
8	123-68-2	Allyl Caproate
9	123-82-2	Amyl Acetate: iso-Amyl Acetate
10	540-18-1	Amyl Butyrate
11	60763-41-9	Amyl Cinnamic Aldehyde Diethyl Acetal
12	102-19-2	Amyl Phenyl Acetate
13	2173-56-0	Amyl Valerate
14	29597-36-2	Andrane
15	104-46-1	Anethole
16	100-66-3	Anisole
17	89-43-0	Auralva
18	100-52-7	Benzaldehyde
19	119-84-8	Benzo Dihydro Pyrone
20	5655-61-8	Bornyl Acetate: iso-Bornyl Acetate
21	107-92-6	Butanoic Acid
22	71-38-3	Butanol: 1-Butanol
23	544-40-1	Butyl Sulfide
24	67634-06-4	Butyl Quinoline: iso-Butyl Quinoline
25	78-22-2	Camphor: dl-Camphor
26	99-49-0	Carvone: l-Carvone
27	87-44-5	Caryophyllene (beta and gamma Isomers)
28	33704-61-9	Cashmeran
29	17369-59-4	Celeriax
30	89-68-9	Chlorothymol
31	104-55-2	Cinnamic Aldehyde
32	141-27-5	Citral
33	5585-39-7	Citralva
34	91-64-5	Coumarin
35	108-39-4	Cresol: m-Cresol
36	106-44-5	Cresol: p-Cresol
37	140-39-6	Cresyl Acetate: p-Cresyl Acetate
38	103-93-5	Cresyl Butyrate: p-Cresyl-iso-Butyrate
39	104-93-8	Cresyl Methyl Ether: p-Cresyl Methyl Ether
40	122-03-2	Cuminic Aldehyde

41	1423-46-7	Cyclocitral: iso-Cyclocitral
42	55704-78-4	Cyclodithalfarol
43	765-87-7	Cyclohexanedione: 1,2-Cyclohexanedione
44	108-93-0	Cyclohexanol
45	80-71-7	Cyclotene
46	67634-23-5	Cyclotropal
47	25152-84-5	Decadienal: 2,4-trans-trans-Decadienal
48	91-17-8	Decahydro Naphthalene
49	111-92-2	Dibutyl Amine
50	352-93-2	Diethyl Sulfide
51	10094-34-5	Dimethyl Benzyl Carbinyl Butyrate
52	103-05-9	Dimethyl Phenyl Ethyl Carbinol
53	5910-89-4	Dimethyl Pyrazine: 2,3-Dimethyl Pyrazine
54	123-32-0	Dimethyl Pyrazine: 2,5-Dimethyl Pyrazine
55	625-84-3	Dimethyl Pyrrole: 2,5-Dimethyl Pyrrole
56	3658-80-8	Dimethyl Trisulfide
57	4747-07-3	Diola
58	101-84-8	Diphenyl Oxide
59	105-54-4	Ethyl Butyrate
60	105-37-3	Ethyl Propionate
61	13925-00-3	2-Ethyl Pyrazine (Lower Concentration)
62	13925-00-3	2-Ethyl Pyrazine (Higher Concentration)
63	470-82-6	Eucalyptol
64	97-53-0	Eugenol
65	67634-15-5	Floralozone
66	6413-10-1	Fructone
67	98-01-1	Furfural
68	98-02-2	Furfuryl Mercaptan
69	88683-93-6	Grisalva
70	90-05-1	Guaiacol
71	111-71-7	Heptanal
72	111-70-6	Heptanol: 1-Heptanol
73	68-25-1	Hexanal
74	142-62-1	Hexanoic acid
75	111-27-3	Hexanol: 1-Hexanol
76	623-37-0	Hexanol: 3-Hexanol
77	6728-26-3	Hexenal: trans-1-Hexenal
78	111-26-2	Hexyl Amine (Lower Concentration)
79	111-26-2	Hexyl Amine (Higher Concentration)
80	101-86-0	Hexyl Cinnamic Aldehyde
81	90-87-9	Hydratropic Aldehyde Dimethyl Acetal
82	107-75-5	Hydroxy Citronellal
83	120-72-9	Indole
84	67801-36-9	Indolene
85	75-47-8	Iodoform
86	14901-07-6	Ionone: beta-Ionone (Lower Concentration)
87	14901-07-6	Ionone: beta-Ionone (Higher Concentration)
88	79-69-6	Irone: alpha-Irone
89	126-91-0	Linalool
90	138-86-3	Limonene: d-Limonene

91	31906-04-4	Lyral
92	67258-87-1	Maritima
93	106-72-9	Melonal
94	2216-51-5	Menthol: l-Menthol
95	93-04-9	Methoxy-Naphthalene: 2-Methoxy Naphthalene
96	134-20-3	Methyl Anthranilate
97	462-95-3	Methyl Acetaldehyde Dimethyl Acetal
98	1334-76-5	Methyl Furoate
99	2271-428	Methyl-iso-Borneol: 2-Methyl-iso-Borneol
100	491-35-0	Methyl Quinoline: para-Methyl Quinoline
101	2459-09-8	Methyl iso-Nicotinate
102	119-36-8	Methyl Salicylate
103	2432-51-1	Methyl Thiobutyrate
104	1222-05-5	Musk Galaxolide
105	1508-02-1	Musk Tonalid
106	37677-14-8	Myracaldehyde
107	143-13-5	Nonyl Acetate
108	4674-50-4	Nootkatone
109	111-87-5	Octanol: 1-Octanol
110	3391-86-4	Octenol: 1-Octen-3-OL
111	109-52-4	Pentanoic Acid
112	591-80-0	Pentenoic Acid: 4-Pentenoic Acid
113	103-82-2	Phenyl Acetic Acid
114	536-74-3	Phenyl Acetylene
115	60-12-8	Phenyl Ethanol (Lower Concentration)
116	60-12-8	Phenyl Ethanol (Higher Concentration)
117	78-59-1	Phorone: iso-Phorone
118	80-56-8	Pinene: alpha-Pinene
119	105-66-8	Propyl Butyrate
120	135-79-5	Propyl Quinoline: iso-Propyl Quinoline
121	111-47-7	Propyl Sulfide
122	110-86-1	Pyridine
123	94-59-7	Safrole
124	69460-08-8	Sandiff
125	115-71-9	Santalol
126	83-34-1	Skatole
127	10482-56-1	Terpineol, mostly alpha-Terpineol
128	110-01-0	Tetrahydro Thiophene
129	91-61-2	Tetraquinone
130	36267-71-7	Thienopyrimidine
131	123-93-3	Thioglycolic Acid
132	110-02-1	Thiophene
133	89-83-8	Thymol
134	529-20-4	Tolualdehyde: ortho-Tolualdehyde
135	108-88-3	Toluene (Lower Concentration)
136	108-88-3	Toluene (Higher Concentration)
137	75-50-3	Trimethyl Amine
138	104-67-6	Undecalactone: gamma-Undecalactone
139	112-38-9	Undecylenic Acid
140	590-86-3	Valeraldehyde: iso-Valeraldehyde

141	503-74-2	Valeric Acid: iso-Valeric Acid
142	108-29-2	Valerolactone: gamma-Valerolactone
143	121-33-5	Vanillin
144	122-48-5	Zingerone

PERCEPTUAL DESCRIPTORS

1	FRUITY, CITRUS
2	LEMON
3	GRAPEFRUIT
4	ORANGE
5	FRUITY, OTHER THAN CITRUS
6	PINEAPPLE
7	GRAPE JUICE
8	STRAWBERRY
9	APPLE (FRUIT)
10	PEAR
11	CANTALOUPE, HONEY DEW MELON
12	PEACH (FRUIT)
13	BANANA
14	FLORAL
15	ROSE
16	VIOLETS
17	LAVENDER
18	COLOGNE
19	MUSK
20	PERFUMERY
21	FRAGRANT
22	AROMATIC
23	HONEY
24	CHERRY (BERRY)
25	ALMOND
26	NAIL POLISH REMOVER
27	NUTTY (WALNUT ETC)
28	SPICY
29	CLOVE
30	CINNAMON
31	LAUREL LEAVES
32	TEA LEAVES
33	SEASONING (FOR MEAT)
34	BLACK PEPPER
35	GREEN PEPPER
36	DILL
37	CARAWAY
38	OAK WOOD, COGNAC
39	WOODY, RESINOUS
40	CEDARWOOD
41	MOTHBALLS
42	MINTY, PEPPERMINT
43	CAMPHOR
44	EUCALIPTUS
45	CHOCOLATE
46	VANILLA
47	SWEET
48	MAPLE SYRUP

CARAMEL
MALTY
RAISINS
MOLASSES
COCONUT
ANISE (LICORICE)
ALCOHOLIC
ETHERISH, ANAESTHETIC
CLEANING FLUID
GASOLINE, SOLVENT
TURPENTINE (PINE OIL)
GERANIUM LEAVES
CELERY
FRESH GREEN VEGETABLES
CRUSHED WEEDS
CRUSHED GRASS
HERBAL, GREEN, CUT GRASS
RAW CUCUMBER
HAY
GRAINY (AS GRAIN)
YEASTY
BAKERY (FRESH BREAD)
SOUR MILK
FERMENTED (ROTTEN) FRUIT
BEERY
SOAPY
LEATHER
CARDBOARD
ROPE
WET PAPER
WET WOOL, WET DOG
DIRTY LINEN
STALE
MUSTY, EARTHY, MOLDY
RAW POTATO
MOUSE
MUSHROOM
PEANUT BUTTER
BEANY
EGGY (FRESH EGGS)
BARK, BIRCH BARK
CORK

BURNT, SMOKY
FRESH TOBACCO SMOKE
INCENSE
COFFEE
STALE TOBACCO SMOKE
BURNT PAPER
BURNT MILK
BURNT RUBBER

99 TAR
100 CREOSOTE
101 DISINFECTANT, CARBOLIC
102 MEDICINAL
103 CHEMICAL
104 BITTER
105 SHARP, PUNGENT, ACID
106 SOUR, VINEGAR
107 SAUERKRAUT
108 AMMONIA
109 URINE
110 CAT URINE
111 FISHY
112 KIPPERY (SMOKED FISH)
113 SEMINAL, SPERM
114 NEW RUBBER
115 SOOTY
116 BURNT CANDLE
117 KEROSENE
118 OILY, FATTY
119 BUTTERY, FRESH BUTTER
120 PAINT
121 VARNISH
122 POPCORN
123 FRIED CHICKEN
124 MEATY (COOKED, GOOD)
125 SOUPY
126 COOKED VEGETABLES
127 RANCID
128 SWEATY
129 CHEESY
130 HOUSEHOLD GAS
131 SULFIDIC
132 GARLIC, ONION
133 METALLIC
134 BLOOD, RAW MEAT
135 ANIMAL
136 SEWER
137 PUTRID, FOUL, DECAYED
138 FECAL (LIKE MANURE)
139 CADAVEROUS (DEAD ANIMAL)
140 SICKENING
141 DRY, POWDERY
142 CHALKY
143 LIGHT
144 HEAVY
145 COOL, COOLING
146 WARM

LIST OF PHYSICO-CHEMICAL PARAMETERS USED

A more detailed description of the parameters is given at the bottom of the list.

1	C
2	H
3	O
4	N
5	S
6	I
7	L
8	molecular_weight
9	molecular_volume
10	molecular_length
11	molecular_width
12	molecular_depth
13	density
14	surface_area
15	Log_Kow_fragments
16	HLB
17	solubility_parameter
18	dispersion_3D
19	polarity_3D
20	hydrogen_bond_3D
21	hydrogen_bond_acceptor
22	hydrogen_bond_donor
23	dipole_moment_debye
24	hydrophilic_surface_area
25	water_of_hydration
26	boiling_point_C
27	vapor_pressure_torr
28	MR
29	parachor
30	connectivity_0
31	connectivity_1
32	connectivity_2
33	connectivity_3
34	connectivity_4
35	valence_0
36	valence_1
37	valence_2
38	valence_3
39	valence_4
40	kappa_2
41	log_water_solubility
42	Log_P__atom_based
43	Z_chain_length
44	glass_transition_temperature
45	melt_transition_temperature
46	water_content_30_RH

47 water_content_50_RH
48 water_content_70_RH
49 water_content_90_RH
50 water_content_100_RH
51 molar_volume
52 Surface_tension
53 Viscosity_cp_at_25C
54 Surface_tension_in_water
55 Critical_Temperature_K
56 Critical_Pressure_bar
57 Normal_Boiling_Point_K
58 Normal_Freezing_Point_K
59 Enthalpy_of_formation
60 Gibbs_energy_of_formation
61 enthalpy_of_vaporization
62 enthalpy_of_fusion
63 liquid_viscosity
64 heat_capacity_25C
65 Effective_number_of_torsional_bonds
66 hydrogen_bond_number
67 Entropy_of_boiling_JKmol
68 Heat_capacity_change_on_boiling_JKmol
69 CIM_1
70 CIM_2
71 CIM_3
72 CIM_4
73 CIM_5
74 CIM_6
75 CIM_7
76 CIM_8
77 CIM_9
78 CIM_10
79 Polar_surface_area
80 C1C
81 C1H
82 C1O
83 C1N
84 C1S
85 C1I
86 C1L
87 H1O
88 H1N
89 H1S
90 S1S
91 C2C
92 C2O
93 C2N
94 C3C
95 C3N
96 C1C1C

97 C2C1C
98 C1C1H
99 C2C1H
100 C3C1H
101 C1C1O
102 C1C2O
103 C2C1O
104 C1C1N
105 C1C2N
106 C1C3N
107 C2C1N
108 C1C1S
109 C2C1S
110 C2C1L
111 C1O1C
112 C1O1H
113 C1N1C
114 C2N1C
115 C1N1H
116 C1S1C
117 C1S1S
118 H1C1O
119 H1C2O
120 H1C1N
121 H1C2N
122 H1C1S
123 O1C1O
124 O2C1O
125 O1C1S
126 S1S1S

Parameters 1-7: These parameters represent atom counts per molecule [C (carbon) through L (chlorine)].

Parameters 8-79: These parameters were calculated by the Molecular Modeling Pro software (ChemSW, Fairfield, CA, USA). The algorithms for calculating these parameters are described below:

Calculations made by Molecular Modeling Pro:

Mass, size

- Molecular weight
- Van der Waals volume (calculated with geometry)
- Molar volume (van Krevelen type method)
- Surface area (calculated with geometry)
- Length, width, depth (current, maximum and minimum calculated by geometry)
- Density (proprietary method for small molecules)
- Mass Percent

Partition coefficients, hydrophobicity, solubility etc.

- Log water octanol partition coefficient (4 methods, Fragment addition generally following the methods of Hansch and Leo, atom based generally following Ghose and Crippen, charge and atom based, and Q Log P after N. Bodor and P. Buchwald, J. Phys. Chem. B, 1997, 101: 3404-3412)
- HLB (hydrophilic lipophilic balance, proprietary method)
- Hydrophilic surface area (proprietary method)
- Percent hydrophilic surface area (proprietary method)
- Polar surface area (J. Med. Chem. 43: 3714-3717)
- Hydration number
- Water solubility (after Klopman et.al. J. Chem. Inf. Comput. Sci. 32:474 and S. Yalkowsky, J. Pharm Sci., 70:971)
- Olive oil gas partition coefficient (after Klopman et.al. J. Med. Chem. 43: 3714-3717)

Properties used in QSAR

- Sterimol properties (L1, B1, B2, B3, B4, B5 and 3 more)
- Hammett Sigma (sigma para, meta, sigma induction (SIND), sigma star)(proprietary method)
- MR (molar refractivity after Ghose and Crippen)

Dipole moment and other charge related properties

- Dipole moment (Modified methods based on Del Re method: G. Del Re, J. Chem. Soc. 4031 (1958); D. Poland and H.A. Scheraga, Biochemistry 6: 3791 (1967); Coefficients modified in MAP 4.0 to take into account pi contributions ; PEOE method: J. Gasteiger and M. Marsili, Tetrahedron 36:3219 (1980); MPEOE (DQP) method: K.T. No, J.A. Grant and H.A. Scheraga, J. Phys. Chem. 94:4732 (1990) and K.T. No, J.A. Grant, M.S. Jhou and H.A. Scheraga, J. Phys. Chem. 94: 4740 (1990); J.M. Park, K.T. No, M.S. Jhou and H.A. Scheraga, J. Comp. Chem. 14:1482 (1993). Semi-empirical Quantum Mechanics methods in CNDO and MOPAC are alternative methods used by MMP to calculate dipole moment.
- Partial charge (many methods - see Dipole moment)
- HOMO/LUMO (via CNDO or MOPAC)
- Hydrogen bond acceptor and donor from charge calculations

Connectivity indices

- Randic, Hall, Kier type connectivity indices 0-4
- Randic, Hall, Kier type valence indices 0-4
- Kier type Kappa shape index 2
- Wiener index
- Chemically Intuitive Molecular Index (F. Burden, Quant. Struct.-Act.Relat. 16:309-314 (1997))

Thermodynamics

- Critical temperature, pressure and volume (after Joback and Reid)
- Normal boiling and freezing point (after Joback and Reid)
- Enthalpy of formation, ideal gas at 298 K (after Joback and Reid)
- Gibbs energy of formation, ideal gas, unit fugacity at 298 K
- Enthalpy of vaporization at the boiling point (after Joback and Reid)
- Enthalpy of vaporization at the boiling point (after Joback and Reid)
- Enthalpy of fusion (after Joback and Reid)
- Liquid viscosity (after Joback and Reid)
- Heat capacity, ideal gas (after Joback and Reid)
- Effective number of torsional bonds (tau) (after S. Yalkowsky et.al.)
- Hydrogen Bond Number (after S. Yalkowsky et.al.)
- Entropy of boiling (after S. Yalkowsky et.al.)

- Effective number of torsional bonds (tau) (after S. Yalkowsky et.al.)
- Heat capacity change on boiling (after S. Yalkowsky et.al.)
- Vapor pressure (after S. Yalkowsky et.al.)
- Vapor pressure (after The Handbook of Chemical Property Estimation Methods)
- Boiling point (after The Handbook of Chemical Property Estimation Methods)
- Parachor (after The Handbook of Chemical Property Estimation Methods)

More properties are available through the MOPAC program included., such as heat of formation, ionization potential and many more.

Polymer and Surfactant properties

- Solubility parameter
- 3-D solubility parameters (dispersion, polarity and hydrogen bonding)
- Water content of polymers at different relative humidities
- Melt transition temperature
- Glass transition temperature
- Chain length (van Krevelen Z)
- Surface tension of liquids
- Surface tension in water
- Molecular weight, molar volume, van der Waals volume, surface area (listed above)
- HLB, hydrophilic surface area, % hydrophilic surface area (listed above)

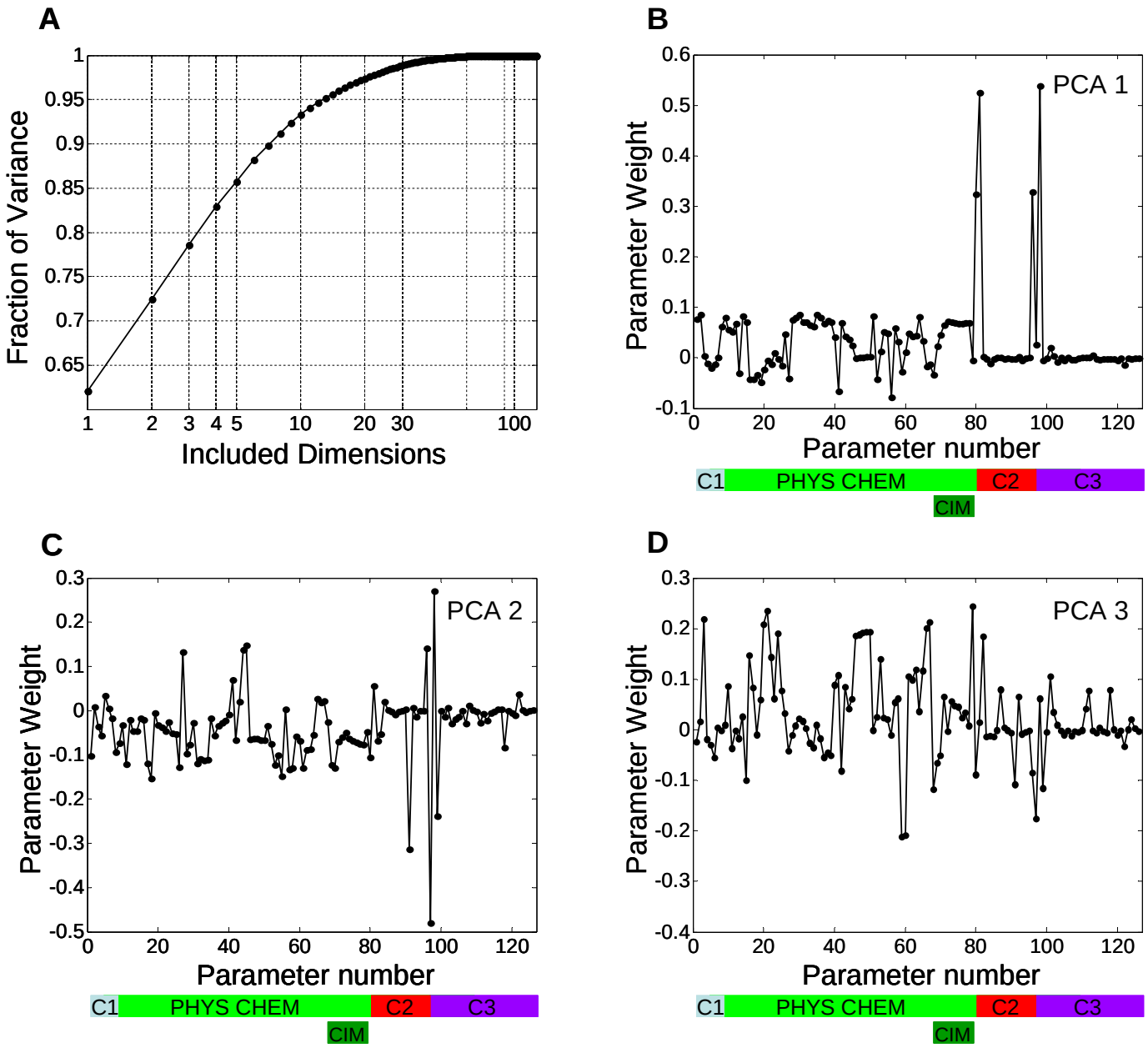
Parameters 80-95: Number of pairs per molecule (C1C or C-C through C3N or C≡N)

Parameters 96-126: Number of triples per molecule (C1C1C stands for C-C-C, while S2S1S represents S=S-S). All triples observed had linear topology. No loops were observed in triples.

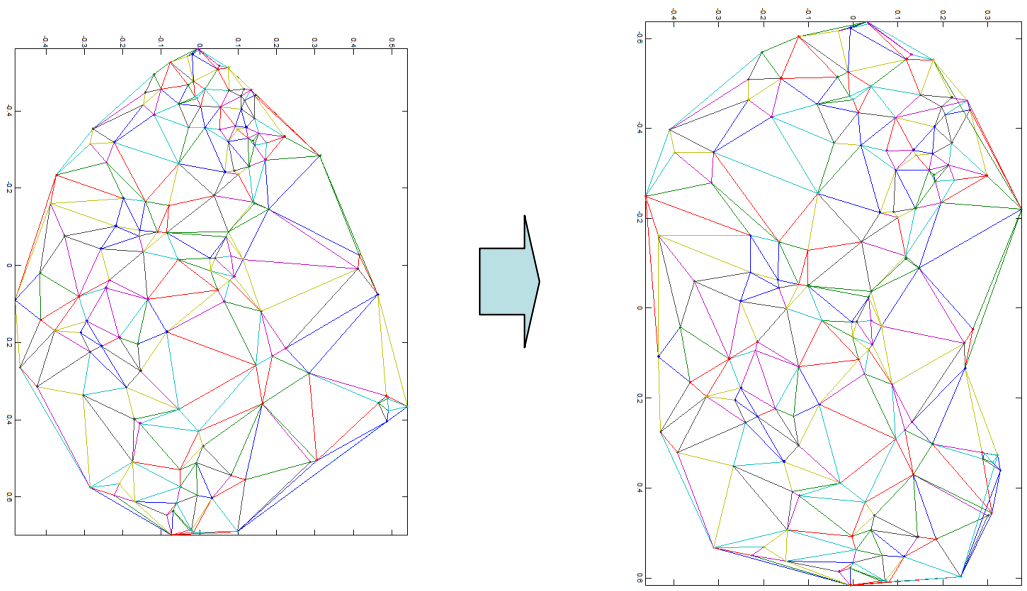
PCA ANALYSIS OF THE PHYSICO-CHEMICAL PARAMETER SPACE

We have conducted a basic PCA analysis of the previously enumerated physical parameters for the odorants used in the AOCP database (Dravnieks dataset). This analysis was performed without using any of the perceptual information contained in the AOCP database. As such, it reflects the structure of the physico-chemical space alone. The results are shown in Supplementary Figure 1. Figure 1A shows that 3 PCA dimensions cover nearly 80% of the variance. Figures 1B-1D show the weight with which each parameter contributes to each PCA dimension. The parameters contributing the most to the first 2 PCA dimensions are those which count the number of carbon atom pairs and triples (including double bonds). That these contributions dominate is to be expected given these values are directly related to the chemical formula of each molecule and as such, act as good discriminators. The third PCA dimension involves a much more complicated combination of parameters which do not suggest any immediate interpretation. PCA dimensions of molecules' properties are expected to be strongly dependent by the choice of these properties and may reflect the redundancies in this choice. Redundant properties or their combinations are expected to contribute strongly to the principal components. That these results are different from those in found in the main text (Table 1) is due to the fact that in the main analysis we sought correlations between the parameters and the perceptual dimensions on the basis of multiple linear regression, which is a different form of analysis from PCA. The interpretation of this difference is that the olfactory system discriminates molecules based on features which are more subtle than the dominant ones shown by PCA. We also show below that

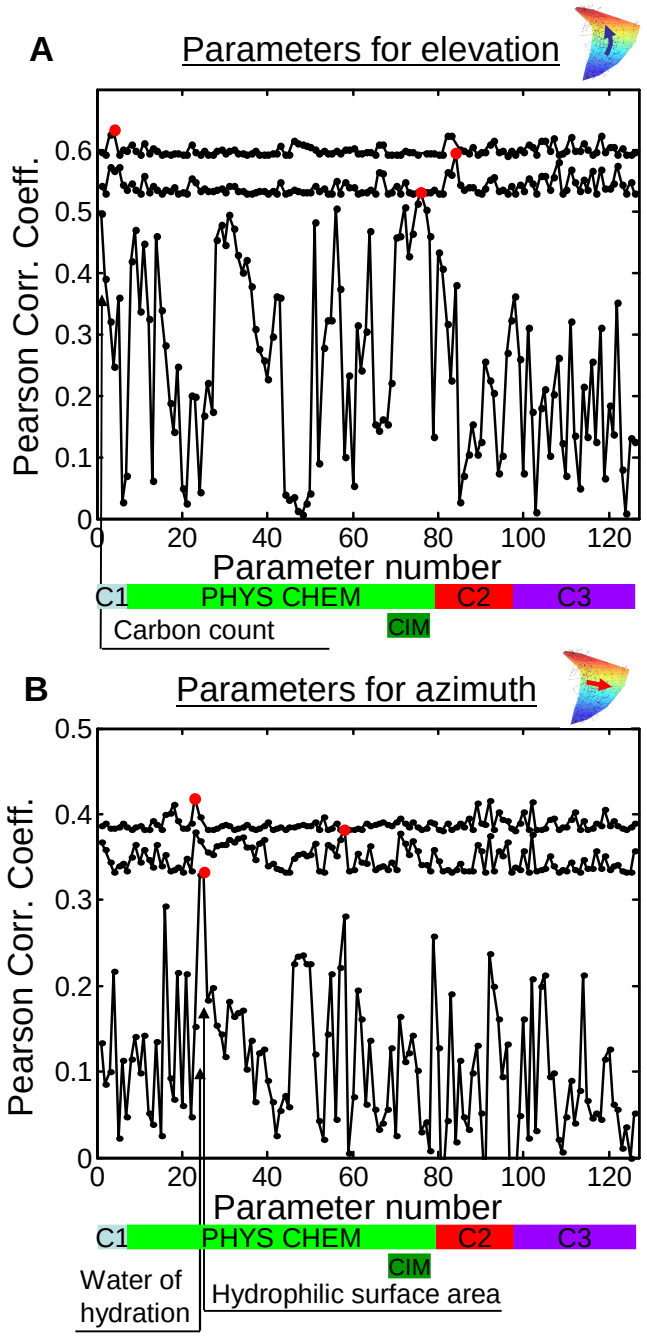
PCA dimensionality of the semantic space used in AOCP database is sufficiently high (~60D), which implies that semantic descriptors used by database are not overly redundant.



Supplementary Figure 1. PCA of odorant physico-chemical parameters. (A) Fraction of variance explained vs. number of included PCA dimensions. (B), (C) & (D) Weights of individual physico-chemical parameters contributing to the first 3 PCA dimensions. The horizontal axis also contains markings indicating the corresponding block of molecular parameters: element counts (C1), Molecular modeling Pro physico-chemical parameters (PHYS CHEM), pairs counts (C2), and triples counts (C3). CIM is the block of ten Burden chemical intuitive indexes.



Supplementary Figure 2. Equilibrating the density of the odorants in two dimensions. Left: the original set of odorants projected onto a flat 2D space and Delaunay triangulated. Right: the same set of odorants after relaxing the elastic energy of edges that are assumed to be springs with unit equilibrium length and the same elastic coefficient. The transformation (arrow) was constrained to be of second order as in equation (1) of the main text. The final two coordinates were studied for correlations with the structural and physico-chemical parameters (Table 1, Supplementary figure 3).



Supplementary Figure 3. The results of greedy algorithm for elevation (A) and azimuth (B) variables on the 2D fit to psychophysical data. Pearson correlation coefficient is shown as a function of the number of physico-chemical/structural parameter (see above). Three iterations are shown for each parameter by three lines with dots. The parameters yielding maximal correlation on each iteration are shown by the red dots. Some parameters are highlighted, such as Carbon count (R=0.50), hydrophilic surface area (R=0.33), and water of hydration (R=0.33). Horizontal axis also contains markings indicating the corresponding block of parameters included: element counts (C1), Molecular modeling Pro physico-chemical parameters (PHYS CHEM), pairs counts (C2), and triples counts (C3). CIM is the block of ten Burden chemical intuitive indexes.

THE PERCEPTUAL SPACE OF MIXTURES

List of 15 mixtures from the AOCP database used in the analysis

01 Cedartone	MIXTURE OF HYDROCARBONS FROM AMERICAN CEDARWOOD ACETYLATED
02 CedroneS	MIXTURE OF OXYGENATED CEDARWOOD HYDROCARBONS
03 Cinnamon Bark Oil (Ceylon)	MAJOR COMPONENTS: CINNAMALDEHYDE EUGENOL ACETEUGENOL
04 Cinnamon Leaf Oil (Ceylon)	MAJOR COMPONENT: EUGENOL
05 Clove Bud Oil	MAJOR COMPONENTS: EUGENOL
06 Eucalyptus Oil	MAJOR COMPONENT: CITRONELLAL
07 Garlic Oil	MAJOR COMPONENTS: ALLICIN
08 Oenantic Ether	MIXTURE OF ETHYL ESTERS OF THE FATTY ACIDS ISOLATED FROM COCONUT OIL
09 Onion Oil	MAJOR COMPONENTS: ORGANIC SULFIDES
10 Patchouli Oil	MAJOR COMPONENT: PATCHOULI ALCOHOL
11 Perfume "Charlie"	COMMERCIAL PERFUME
12 Phenoxaflor	A FRAGRANCE COMPOUND WITH ROSE CHARACTER
13 Pyrroline + Pyrrolidone (mixture)	3-Pyrroline
14 Rosemarel	MIXTURE OF COMPOUNDS A AND B COMPOUND A: BETA-PINENE EPOXIDE
15 Spearmint Oil	MAJOR COMPONENT: L-CARVONE

ANALYSIS OF THE SEMANTIC SPACE

As stated in the main article, the dimensionality of the olfactory space was determined using the results of odorant profiling in which a set of 146 semantic perceptual descriptors were used. The methodology of this study is outlined in (Dravnieks, 1982). They create odor profiles by presenting an odor to a participant and have them rate each of the 146 semantic descriptors by applicability. Using a large number of participants allows for the calculation of the percentage of applicability for a given descriptor to a particular odorant. They find this method generates odor profiles that give high correlations ($p < 0.001$) when confronted by the results of an earlier study using a nearly independent set of participants. This stability of the odor profiles in the Dravnieks catalog makes their results an excellent basis for the study we present in the main article. However, there is an important question that needs to be answered about how independent these semantic descriptors are.

If the descriptors were essentially synonymous, the low dimensionality of the olfactory space could be attributed to the dependencies between descriptors. The redundancy between descriptors could render the semantic space spanned by them low dimensional, imposing low dimensionality on the odorant space. On the face of things, the descriptor space in the Dravnieks catalog has 146 dimensions. However, if the perceptual descriptors used in this psychophysical study are related to one another (in a semantic sense), the dimensionality of the semantic space may be considerably lower. Therefore, here we will study the dimensionality of the semantic space of the descriptors used in the AOCP catalog. We will argue that the dimensionality of the semantic space spanned by these descriptors is substantially larger than the dimensionality of the olfactory space based on the study of texts found by web searches, as described below.

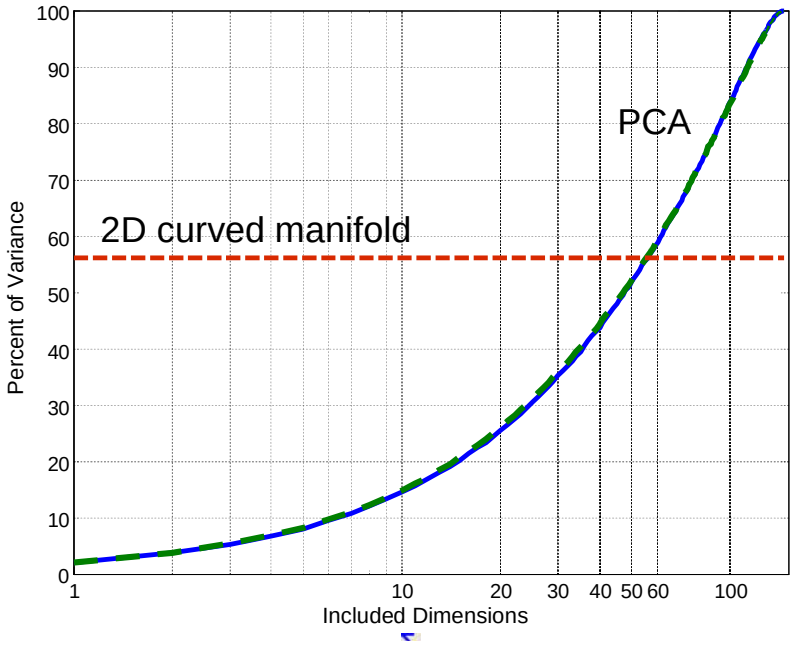
To evaluate the lower bound on the dimensionality of semantic space we employed a “bag-of-words” technique in this analysis (Manning and Schütze, 1999). For each perceptual descriptor (PD), a Google search was performed. Links provided by the search were collected for a given number of result pages. Each link was then followed and all formatting strings were removed (scripts, html tags, etc.). What is left over was the main text of the website. Of course, some of the search results link to pages that cannot be treated in this manner (e.g. flash sites, Word docs, pdfs etc.). These links were excluded from this analysis. The “bag-of-words” for the PD was constructed by saving a specified number of words surrounding each instance of the PD in the website text (contextual window). As we are only interested in comparing substantive words between different PDs, all closed-class words (i.e.

articles, pronouns, prepositions etc.) were then removed. Performing this process over many search results allowed us to calculate the probability of finding a given word within the specified contextual window for the PD of interest. After completing the procedure for the entire list of PDs we created a probability matrix where columns corresponded to one PD and the rows to the entire set of words found for all 146 PDs. Performing principle component analysis on this matrix allowed us to calculate the variance covered for the number of Euclidean “dimensions” included.

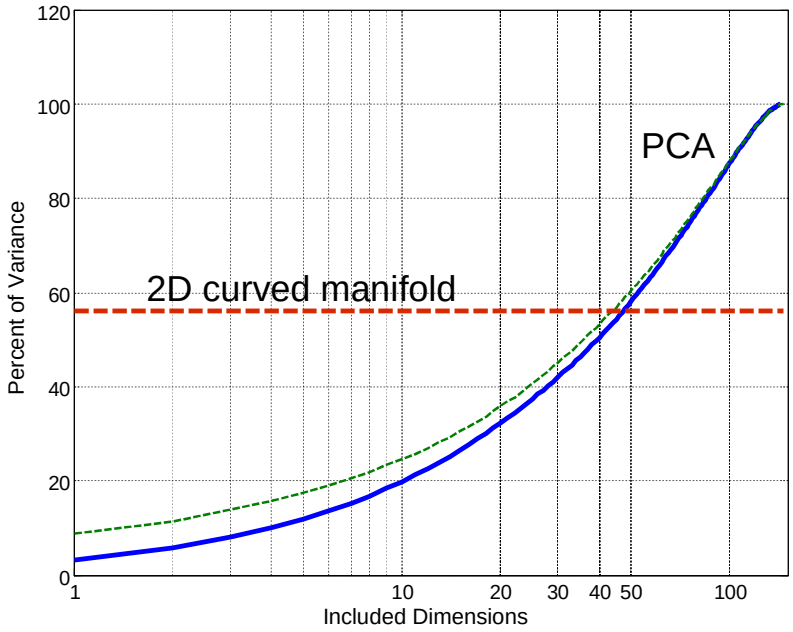
We carried out this analysis using a mean of 50 different websites for each of the 146 PDs with a contextual window of 20 words. The mean number of words (not unique) for each PD is found to be ~13000. The number of unique words is about a factor of 10 less. The results of the analysis can be seen in Supplementary Figure 4. The dimensionality of the PD semantic space found using this method is approximately 28 times larger than that found for the olfactory space. For example, the (non-jackknife) variance covered by the curved 2-D manifold is approximately 56% (Figure 1E, main text). The same amount of variance can be explained by including 56 dimensions of the semantic space (Supplementary Figure 4). We conclude that the semantic space spanned by the descriptors used in (Dravnieks, 1985) is substantially higher dimensional than the space formed by the odorants.

To verify that this result is statistically stable we reduced the number of websites included by a factor of two (from 50 to 25 for each PD). After conducting the same PCA analysis, we find that the amount of variance covered as a function of the number of included dimensions is very similar to the full analysis (Supplementary Figure 4, green dashed line). The number of dimensions corresponding to 56% variance is 55 (vs. 56 for the analysis with full dataset, as described above). We conclude that the estimate for the dimensionality of the semantic space of PD used by the AOCP is stable with respect to the statistical variability in the search data. The dimensionality of the semantic space is therefore substantially larger than the dimensionality of olfactory space. Although 56% of variance of olfactory data can be explained by 2 curved dimensions, the same amount of variance in the semantic space can be accounted for by 56 dimensions.

As an additional check we have run the same analysis, however, this time including the word 'olfaction' in the search for each PD. This method is expected to bias the searches toward the texts that have relevance to olfaction and analyze the descriptors in the olfactory context. This method of sampling is expected therefore to lower the dimensionality of the semantic space. It is not clear if this method of sampling can isolate the influence of context from the effects of olfactory percepts that, as we know from the main text, can be viewed as low-dimensional. Despite these limitations, we expect that the searches related to olfaction can provide the lower bound for the dimensionality of the semantic space relevant to olfaction. Supplementary Figure 5 shows the result of this analysis. The effect of requiring each PD search to have a bias toward olfaction has the expected result of reducing the overall dimensionality of the PD space. However, this dimensionality is still approximately 22-24 times larger than that of the olfactory space found in the main article. More precisely, at the level of 56% of the variance, the olfactory space can be accounted for by a 2D curved manifold (non-jackknifed data is used here, as PCA of the semantic space is not jackknifed). The same amount of variance is captured by 47 dimensions of the semantic data. When only 50% of the websites are used, the dimensionality at 56% variance level can be estimated to be 44. We conclude that the semantic space is 22-24 times higher dimensional than the olfactory space. The correlations in the olfactory space reported in the main text are therefore not caused by a poverty of diversity in the semantic space.



Supplementary Figure 4. The dimensionality of semantic space. Percent variance explained vs. number of included PCA dimensions for the semantic space of the 146 perceptual descriptors used in the psychophysical study (blue line). The red line shows the variance captured by the 2-D curved surface found in the main analysis. The intersection is at 56 dimensions. The dimensionality of the semantic space spanned by descriptors is therefore 28 times larger than the dimensionality of the olfactory space. The green dashed line shows the results of PCA with 50% of data used (25 vs. 50 web pages per PD).



Supplementary Figure 5. Percent variance vs. number of included PCA dimensions for the semantic space of the 146 perceptual descriptors plus the term “olfaction” (blue line). The red line shows the variance captured by the 2-D curved surface found in the main analysis. The intersection is at 47 dimensions. The green dashed line shows the results of PCA with 50% of data used (25 vs. 50 web pages per PD). The intersection with the red line is at 44 dimensions.

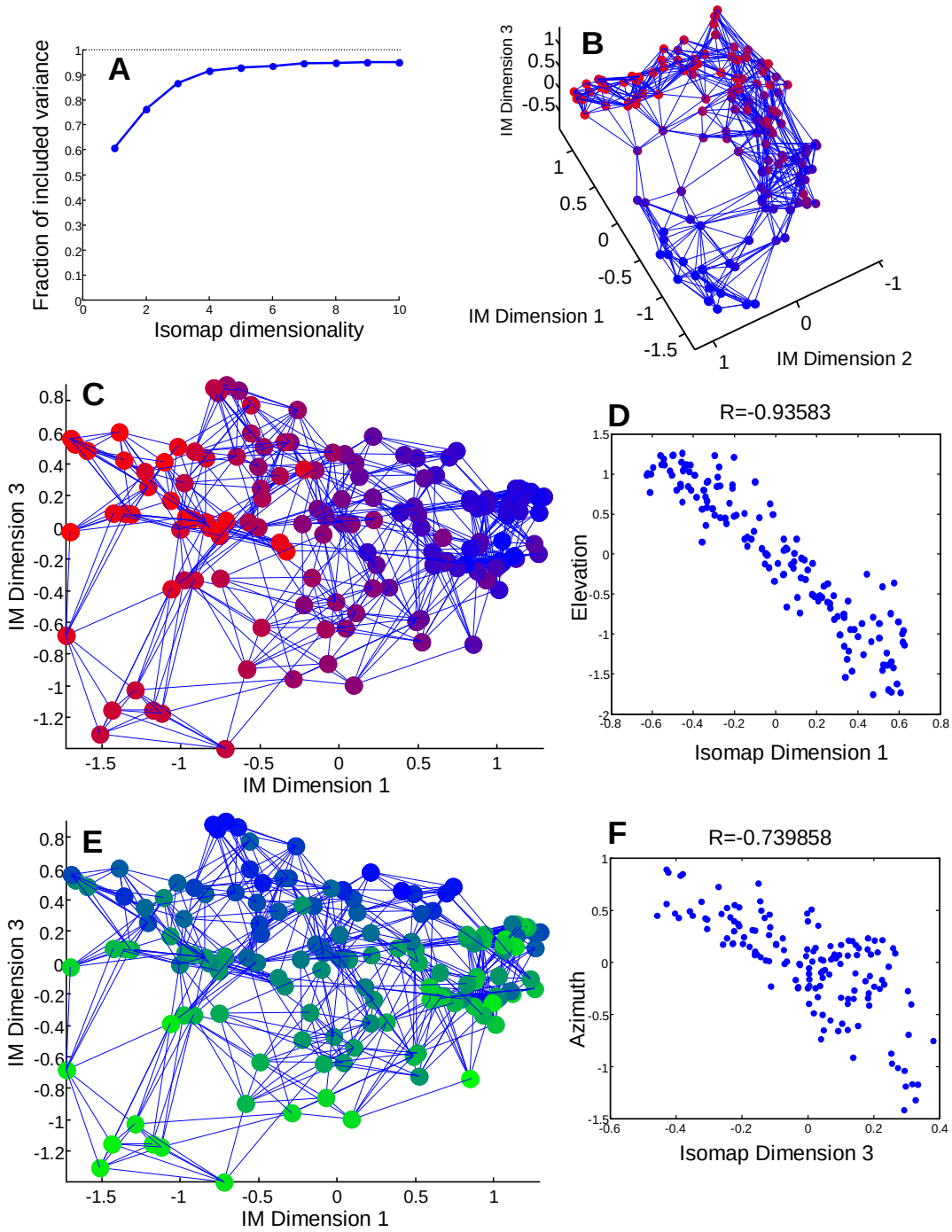
OTHER METHODS OF ANALYSIS

Here we applied other published methods to embed the perceptual data into a low-dimensional subspaces. These methods include Isomap (Tenenbaum et al., 2000) and locally linear embedding (LLE) (Roweis and Saul, 2000). We argue here that these methods give similar results to the method used by us. Because several elements needed for our purposes are not developed in these two methods, we employed the method of non-linear regression described in the main article. Here we report the results of applying both Isomap and LLE.

Isomap (IM)

In this method the shortest path between every two points is first calculated on the basis of the graph that includes K nearest neighbors. Because only proximal points are used in evaluating the distances, this "geodesic" distance is expected to perform better for the data in which large distances are inaccurate. The classical MDS algorithm is then performed with this set of pair-wise distances to establish the coordinates of points in the embedded space. We downloaded the algorithm from the authors' website (http://waldron.stanford.edu/~isomap/). The results of applying this algorithm to the AOCP database are shown in the Supplementary Figure 6. As shown in Supplementary Figure 6A, the 3D embedding can account for 87% of variance in the data (note here that this number does not pertain to the original perceptual data but to the variability of "geodesic" distances calculated on the graph of nearest neighbors, see below). The 3D embedding is shown in Supplementary Figure 6B. It is clear that the odorant in the plane defined IM dimensions 1 versus 2 show a configuration similar to the letter "C" reported in the main text (cf. Figure 1B). The IM dimension number 1 can therefore be mapped upon perceptual dimension number one (pleasantness, elevation on the 2D manifold) reported in the main text and found by our method. To confirm this, we show in Supplementary Figure 6D the plot of elevation coordinate versus IM dimension 1 that displays the high level of correlation (R=0.94). We therefore interpret the first Isomap dimension as elevation in our method on embedding. The IM dimension number 3 can be mapped upon the second perceptual dimension (azimuth or hydrophobicity). Indeed, as we show in Supplementary Figure 6F, the two variables show a high degree of correlation (R=0.74). These identifications are further shown by the color coded elevation and azimuth in Supplementary Figures 7C and E respectively. We argue that our method of analysis gives results similar to Isomap (Tenenbaum et al., 2000).

Although the results of Isomap embedding are quite impressive (Supplementary Figure 6A), several features of this algorithm may require further development. First, the included variances are calculated on the basis of geodesic distances, and, as such, do not reflect directly the included variance in the original data. Because the geodesics are evaluated on the subspace of restricted dimensionality, variability within the subspace is expected to be lower than that in the original data. Second, it is not clear how to establish a mapping between the direct space and the embedded space. For the forward mapping (direct space -> embedded space) one could just rerun the algorithm for every new point added. The reverse mapping is not clear. This reverse mapping is used in our analysis to validate the results of embedding. Indeed, by removing the odorants from the database, calculating the embedding, and finding the smallest distance between the embedded space and the removed point (jackknife analysis) we were able to determine whether our embedding generalizes on odorants not included in the database. The calculations require a model for the position of the embedded space in the original data space, i.e. the reverse mapping. These limitations restrict embedding quality control within the Isomap method.



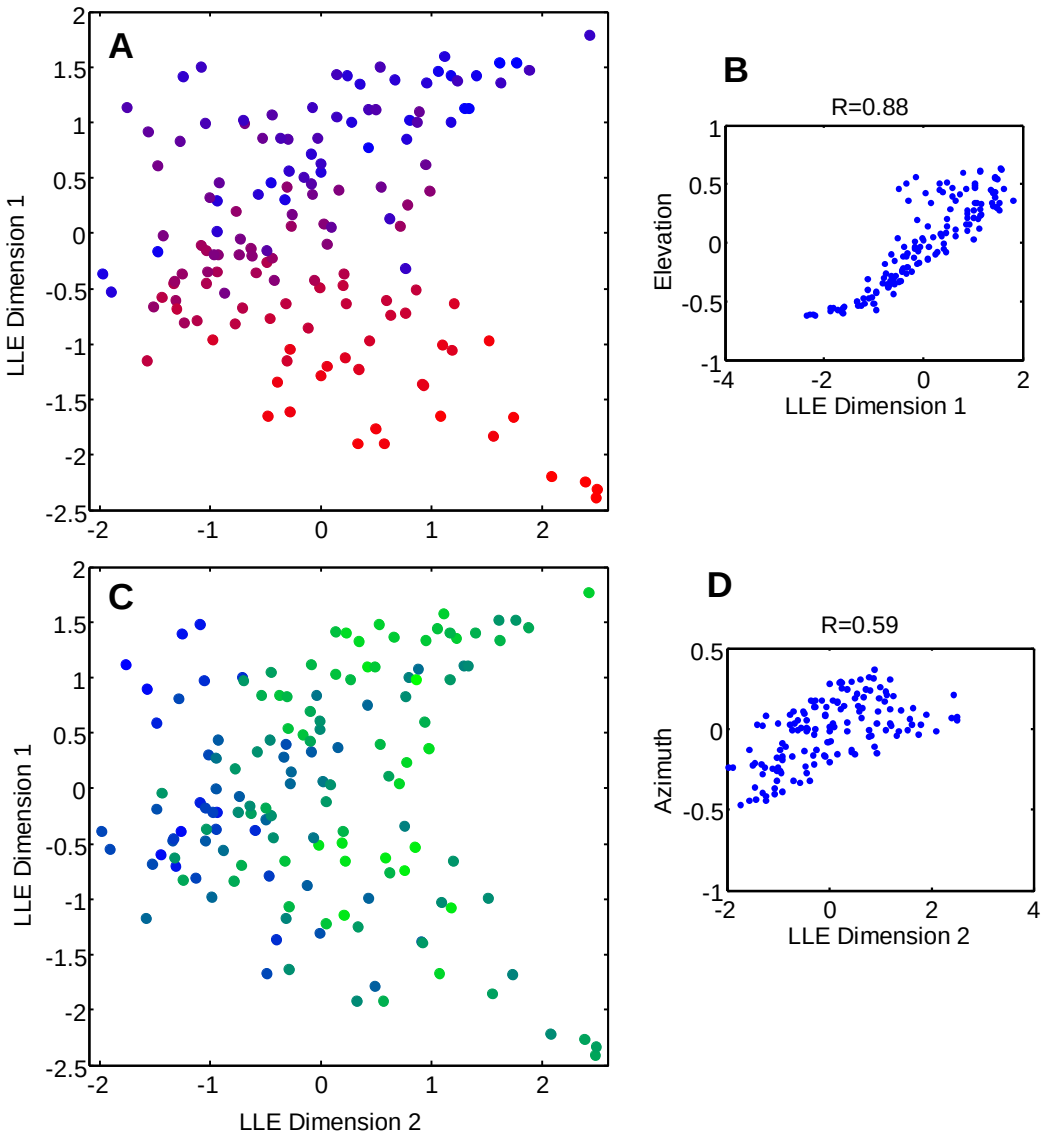
Supplementary Figure 6. The results of analysis of the same dataset using the Isomap algorithm. (A) The fraction of variance explained by Isomap as a function of the number of included dimensions. The fraction is calculated on the basis of geodesic pairwise distances. (B) The embedding in 3D can that can explain 87% of data. The red color represents the first perceptual dimension identified in the main paper. The data points reside near a letter "C" as in the main paper (Figure 1B). The links between points represent connections to the nearest neighbors (K=7) used to calculated the shortest pairwise distances (geodesics). (C) The view on the same set of points from the direction of Isomap (IM) dimension 2. IM dimension 1 clearly correlates with the first perceptual dimension discussed in the main paper (redness of points). (D) This is confirmed by high degree of correlation between two variables. (E) Similarly, IM dimension 3 correlates with the second perceptual dimension (azimuth). (F) The correlation between these two variables obtained from different analyses is high. The correlation between IM dimension 2 and azimuth is lower (R=0.20, not shown). We conclude that IM yields similar results to our analysis with IM dimensions 1 and 3 identified with elevation and azimuth coordinates respectively.

Local Linear Embedding (LLE)

In this algorithm a set of weights is found that relates each point to K of its nearest neighbors. The weights are found by minimizing the squared error between the position of the point and the linear weighted sum of the neighbors. The weights are expected to carry information about local neighborhoods in the dataset. The low dimensional embedding is then found that minimizes the same error for points of lower dimensionality by optimizing their positions with the weights fixed (Roweis and Saul, 2000). Because the sets of weights are the same for high and low dimensional data, the local neighborhood relationships are expected to be preserved approximately by the algorithm.

We applied the LLE algorithm to the AOCP dataset (Supplementary Figure 7). The algorithm was downloaded from the author's website (http://www.cs.nyu.edu/~roweis/lle/code.html). We used K=20 of nearest neighbors, as recommended by (Roweis and Saul, 2000). The results of these calculations agree to a large degree with the results presented in the main article. Thus the LLE dimension 1 correlates strongly with the first perceptual dimension identified in the main article (elevation or pleasantness, Supplementary Figure 7 A and B). The second LLE dimension is correlated strongly with the second perceptual dimension identified in the main article (azimuth or hydrophobicity). We conclude therefore that the non-linear embedding method used in the main paper yields similar results to a related embedding method (LLE).

There are several ways how the LLE algorithm could be further extended to facilitate application to olfactory data. First, the embedding provides no quality control; the variance in the data accounted for is not available using this method. Second, similarly to Isomap, no mapping from the embedded space to the original data space is provided. It is therefore not possible to evaluate the embedding quality for novel odorants that were not included in the calculation of the embedding space. The validation of the results of this embedding is not readily possible.



Supplementary Figure 7. The results of analysis of the AOCP dataset using Locally Linear Embedding (LLE) algorithm. (A) The embedding of the AOCP dataset into 2D space. The dots represent individual odorants positioned in the 2D embedded space by the LLE algorithm. The degree of red color represents the first perceptual dimension identified in the main paper. K=20 neighbors were used in LLE algorithm. No data about the fraction of variance of data covered by the embedding is provided by the algorithm. The first perceptual dimension appears to correlate with LLE dimension 1. (B) The high correlation between the first perceptual dimension (elevation) and LLE dimension 1 is confirmed by the high Pearson correlation coefficient (R=0.88). (C) The same embedding with the degree of green color representing the second perceptual dimension identified in the main study. (D) The correlation between the second perceptual dimension (azimuth) and the second dimension identified by LLE is high (R=0.59). On the basis of these findings we conclude that 2D LLE yields similar results to our analysis with LLE dimensions 1 and 2 identified with elevation and azimuth coordinates respectively.

DETAILED METHODS

Preparation of responses for analysis. Responses to 144 odorants were obtained from Ref. [(Dravnieks, 1985)] and represented in a set of 146D vectors $\vec{r}_i$ (i=1…144). We used percent used (PU) set of responses from Ref [(Dravnieks, 1985)]. PU describes the fraction of about 150 observers that thought that a given descriptor applies to an odorant. We verified that our conclusions do not change substantially if other parameters are used instead of PU, such as PA.

All computations were performed using MATLAB (Mathworks, Inc.) Before applying PCA we normalized response vectors to have unit length in terms of the L_2 measure. This implies that the vectors resided on a unit sphere in 146D. This reduced somewhat the dimensionality of the dataset to 145D. The normalization step was intended to equalize the odorants in their perceived intensity or concentration. We verified that our conclusions do not change qualitatively if other measures (L_2 through L_9) are used for normalization. We noticed some deterioration of the fits beyond this range. For further analysis the data were centered so that the mean response to each semantic descriptor is zero. This step resulted in the response matrix $\hat{R}$ that contained responses to individual odorants in its columns. It was therefore 146 (number of descriptors, height) by 144 (number of odorants, width). The elements in the rows are centered i.e. have zero mean.

Principal component analysis (PCA). The matrix of responses was represented as $R^T = USV^T$ using SVD algorithm. Here U and V are 144 by 144 and 146 by 146 orthogonal matrices ($V^TV = U^TU = I$) and S is the 144 by 146 diagonal matrix. The principal components are contained in the columns of 144 by 146 matrix $Y = R^TV$. Thus, the first three columns of Y were used to visualize data in Figures 1A and B. The variance explained by each PC is equal to the diagonal elements of diagonal matrix S. In Figure 1E the cumulative variance is shown as a fraction of total variance ($\sum_{ij} R_{ij}^2$). The first n columns of the orthogonal matrix V represent a projection operator P_n onto the n-D PCA space. The PC loadings can be found as the coefficients in the columns of matrix V.

The inverse participation ratio. This measure is commonly used to evaluate how many parameters contribute to data in a threshold-free manner (Eriksen et al., 2003). Thus, here we wanted to evaluate how many perceptual descriptors contribute to PC1 and PC2 of the data. To this end we calculated the participation ratio for the loading of each PC

$$P_n = \sum_i V_{in}^4 . \qquad (1.1)$$

We then calculated the inverse participation ratios as $iP_n = 1/P_n$. These variables describe how many semantic descriptors contribute to each PC. Indeed, assume that d descriptors contribute to a PC uniformly. Assume that other descriptors do not contribute. This implies that the value of each descriptor loading is $1/\sqrt{d}$ due to normalization ($\sum_i V_{in}^2 = 1$). The value of the participation ratio is then $P_n = 1/d$, while the inverse participation ratio is $iP_n = 1/P_n = d$, i.e. describes accurately the number of non-zero loadings. For the olfactory data the number of contributing loadings was found to be $iP_1 \simeq 17$, $iP_2 \simeq 23$, and $iP_3 \simeq 26$. This means, for example, that 17 semantic descriptors contributed substantially to the first principal components (pleasantness).

Approximating odorant response with curved spaces. Each odorant vector $\vec{r}_i$ was approximated with the 'projected' vector $\vec{p}_i$. Here index i enumerates the odorants while each vector contains 146 components corresponding to semantic descriptors. The projected vectors were sought in the form

$$\vec{p}_i = \vec{A} + \sum_{\alpha=1}^{D} \vec{B}_\alpha x_{\alpha i} + \sum_{\alpha=1}^{D}\sum_{\beta=1}^{D} \vec{C}_{\alpha\beta} x_{\alpha i} x_{\beta i} \,. \qquad (1.2)$$

Here $\vec{A}$, $\vec{B}_\alpha$, and $\vec{C}_{\alpha\beta}$ are odorant-independent parameters of the surface. Parameters $\vec{C}_{\alpha\beta}$ allowed the surface to be curved. Parameters $x_{\alpha i}$ define positions of odorants on the surface. D is the number of parameters per odorant which is the dimensionality of the surface. The manifold defined by this equation is D-dimensional. In Figure 2 we used $D = 2$, while in Figure 4 the dimensionality was varied. To find $\vec{A}$, $\vec{B}_\alpha$, $\vec{C}_{\alpha\beta}$, and $x_{\alpha i}$ we minimized $\sum_i \left\| \vec{r}_i - \vec{p}_i \right\|^2$ using the conjugate gradient algorithm (CGA). The set of parameters $x_{\alpha i}$ was determined therefore as the nearest points on the curved manifold. The nearest points define 'projections' onto the curved manifold. The remaining variance for approximation is estimated as $\sum_i \left\| \vec{r}_i - \vec{p}_i \right\|^2$.

To remove possible ambiguity in the data, the initial set of nearest points on the surface was determined from the PCA projection. The initial nearest points for the elevation coordinate were chosen to match PC1 of given odorant. The initial azimuth coordinate was chosen to be a linear combination of PC2 and PC3 with coefficients 0.96 and 0.29 respectively. When looking from this rotated direction, the projection of the surface did not have folds and the parameterization (1.2) was expected to yield accurate results. In the n-D surface case the initial coordinates were chosen to be the remaining PCs. Before running CGA to optimize the surface, the nearest points on the surface were found for each odorant individually using CGA. The optimal surface was then found using 20 iteration of CGA with both parameters of the surface and the positions of nearest points subject to optimization. We verified that 50 iterations of CGA did not improve the result by more than 1% of variance. Finally, the nearest points for each odorant individually were fine tuned by running CGA on the positions of these points. The positions of projected onto the curved surface odorant responses $\vec{p}_i$ were the results of this step.

Jackknife procedure. Approximating human sensory responses with higher dimensional curved manifolds is confounded by a dramatic increase in the number of parameters of fit. Because the number of parameters increases as a second power of the number of dimensions in our quadratic regression, for a moderately low-dimensional manifold we find that we can perfectly fit all of the experimental data (Figure 4A, dashed line). To avoid this overfitting problem we employed the jackknife technique, in which we remove a single odorant from the perceptual database, obtain a high-dimensional fit with the curved surface (1.2) for the responses to the remaining compounds, and calculate the distance between the fitted manifold and the removed odorant. By applying this procedure for all odorants in the database sequentially we evaluated a variance of the approximation with curved manifolds. The remaining variance does not vanish for spaces of high dimensionality due to overfitting (Figure 4A, solid line).

The natural system of coordinates of the 2D surface was used to equilibrate the density of odorants (grid in Figure 3). The odorants were projected onto the 2D plane and the Delaunay triangulation was calculated. The edges of triangulation were replaced with elastic strings of unit equilibrium length and a coordinate transformation was found that minimizes the elastic energy of the strings. The coordinate transformation was constrained to the form used above [equation (1.2)] with the mapping of 2D to 2D space. The results are shown in the Supplementary Figure 2. These natural coordinates (Supplementary Figure 2B) were used to evaluate the correlations between perceptual dimensions and the physico-chemical parameters.

Estimating the variability due to a finite number of observers. The perceptual variable used here (percent used, PU) is convenient for estimating the experimental variability. We resampled the data for every entry in the database independently using 149 observers as specified in (Dravnieks, 1985). We estimated the variance of the resulting ensemble to be equal to 7% of the experimental variance present in (Dravnieks, 1985).

Structural and Physico-chemical parameters (SPCP). The values of 72 physico-chemical properties were calculated using the program Molecular Modeling Pro™ (ChemSW, Failfield, CA) as described in the section

titled "The list of physico-chemical parameters used" above. We verified that the use of 1999 parameters generated by E-Dragon (VCCLAB.org) did not improve the result suggesting a redundancy in the data. To evaluate the properties of the odorants we used their 3D structure provided by the chemical database maintained by the National Cancer Institute (CADD group) located at http://cactus.nci.nih.gov (release 3, September 2003). The structures were identified through the CAS numbers provided by the AOCP database. For small number of compounds the 3D structure was not found in the database. For such compounds the mol files were downloaded from several sources on the internet, the consensus of structural formulas was found, and the 3D geometry was optimized by the MM2 algorithm provided by Molecular Modeling Pro. All 3D structures were examined visually and in case of clear deficiencies MM2 algorithm was applied. The discrepancy between MM2 algorithm and the 3D structure provided by NCI database was found to be small.

The set of 72 physico-chemical properties was then calculated in the batch format, using Molecular Modeling Pro with the algorithms described in the section titled "The list of physico-chemical parameters used" above. Calculation of CIM indexes that are important for the vertical perceptual dimension was checked independently using the known procedure (Burden, 1997).

To normalize the properties we used the following procedure. For the properties that took negative or zero values we used the z-score [$z = \left(x - \overline{x}\right) / \sigma(x)$], as was done by (Khan et al., 2007). The properties that took positive values for all odorants often had log-normal distributions, i.e. the logarithm of these quantities had a Gaussian distribution. The use of z-score in this case was not appropriate. We therefore evaluated the standard deviation of the logarithm of positive properties. If this (unitless) standard deviation was found to be larger than one, indicating closeness to lognormal distribution, we used the z-score of the logarithm of such quantity for fitting. For quantities with smaller than unit standard deviation of the logarithm we used the direct z-score, as for the properties with some negative values.

The greedy algorithm. To approximate the perceptual variables (elevation and azimuth) we used the set of 126 structural/physical/chemical parameters (SPCPs) that are described above in the section titled "The list of physico-chemical parameters used". We used the algorithm that essentially reproduced the method described in (Saito et al., 2009) and was pioneered by Sobel's group (Haddad et al., 2008). We will describe the algorithm briefly for the elevation coordinate (called here "y"). The usage for the azimuth coordinate is identical.

On the first step of the algorithm, we found the SPCP that yields the largest Pearson correlation coefficient with the given perceptual coordinate "y". The results are shown by the lowest black curve of Supplementary Figure 3A. The best correlated SPCP turned out to be the Burden CIM index 8 as indicated in the figure by the red dot (see also Table 1). We then calculated the Pearson correlation coefficient for the approximation of the elevation coordinate with two SPCPs: The best property found in the first iteration (Burden CIM 8) and each of the remaining 125 properties. The approximation was found using multiple linear regression that employed pseudoinverse. The results are shown by the second lowest black curve in Supplementary Figure 3A. The best SPCP obtained on the second iteration was the number of C-S pairs (Carbon and Sulfur connected by a single bond) as indicated in Table 1 and Supplementary Figure 3A by the second red dot. The set of two best SPCPs from the first two iterations (CIM8 and C-S pairs) was used along with 124 additional SPCP in the third iteration. The results of the Person correlation coefficient obtained during the third iteration are shown in Supplementary Figure 3A, top black curve. The best correlation can only increase during these iterations. For this reason the procedure is called "greedy".

The best five SPCPs for coordinate "y" are shown in Table 1, left. These names mean that, for example, a multiple linear regression of "y" with five parameters listed yields a correlation coefficient of r=0.68. Similarly, the use of four parameters (CIM8 through molecular width) yields the correlation coefficient of r=0.66 after multiple linear regression.

References

Burden, F.R. (1997). A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. Quant Struct-Act Rel *16*, 309-314.

Dravnieks, A. (1982). Odor quality: semantically generated multidimensional profiles are stable. Science *218*, 799-801.

Dravnieks, A. (1985). Atlas of odor character profiles (Philadelphia, PA: ASTM).

Eriksen, K.A., Simonsen, I., Maslov, S., and Sneppen, K. (2003). Modularity and extreme edges of the internet. Phys Rev Lett *90*, 148701.

Haddad, R., Khan, R., Takahashi, Y.K., Mori, K., Harel, D., and Sobel, N. (2008). A metric for odorant comparison. Nat Methods *5*, 425-429.

Khan, R.M., Luk, C.H., Flinker, A., Aggarwal, A., Lapid, H., Haddad, R., and Sobel, N. (2007). Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. J Neurosci *27*, 10015-10023.

Manning, C.D., and Schütze, H. (1999). Foundations of statistical natural language processing (Cambridge, Mass.: MIT Press).

Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science *290*, 2323-2326.

Saito, H., Chi, Q., Zhuang, H., Matsunami, H., and Mainland, J.D. (2009). Odor coding by a Mammalian receptor repertoire. Sci Signal *2*, ra9.

Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science *290*, 2319-2323.