A metric for odorant comparison

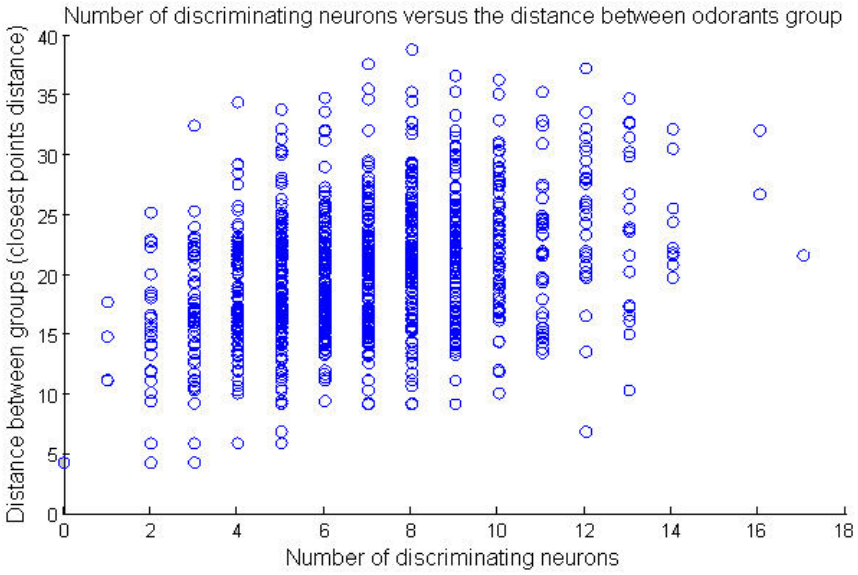
Rafi Haddad, Rehan Khan, Yuji K Takahashi, Kensaku Mori, David Harel & Noam Sobel

Supplementary figures and text:

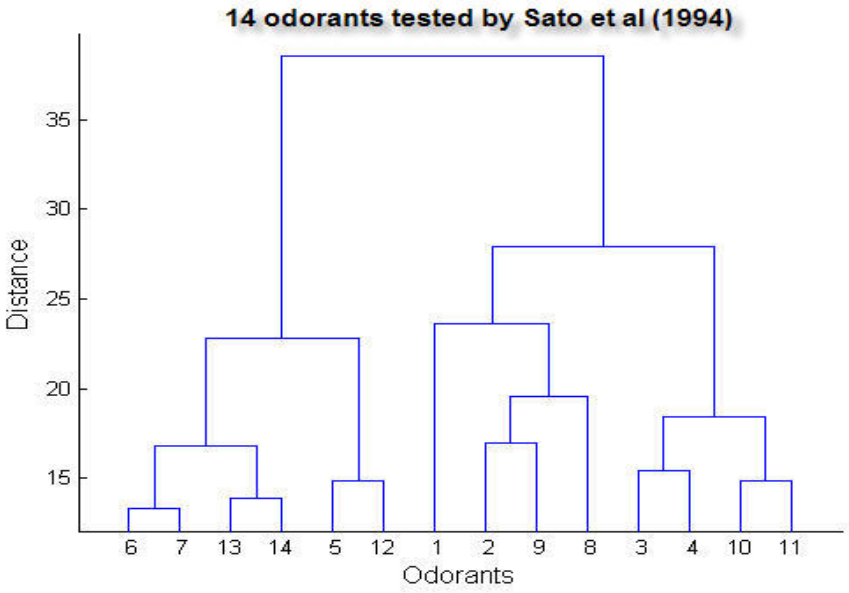
Note: Supplementary Data 1–3 are available on the Nature Methods website.

Supplementary Figure 1. Two examples of the power of the multidimensional metric. **(a)** The number of discriminating neurons between two groups of odorants versus the distance between these groups. Each point in the graph was calculated by generating two groups containing five randomly selected odorants from the set of 110 odorants. This process was repeated 1000 times. We found that the further away the groups were, the higher the number of discriminating receptors we could find (r = 0.38, P = 2e-36). The distance between groups was defined to be the minimum distance between all pairs of odorants in the two groups. Changing the group size (from 3 to 10) did not change the correlation significantly. The data is taken from.[1] **(b)** In a study[2] reporting the optical recordings of [Ca+] of 30 olfactory neurons to 14 odorants (seven straight chained acids and alcohols) suggested that 10 neurons were discriminating neurons – neurons that responded to odorants from one functional group and not to the others – and 20 neurons responded to both. Based on this result, the authors suggested that neurons are tuned to functional groups. However, grouping the odorants using the multidimensional metric revealed 14 discriminating neurons. This number of discriminating neurons is significantly higher than any random selection of odorants with the same group size ($P < 0.01$, 1000 trials). Thus, the multidimensional metric provided a better account of neural response than did functional group, even though the odorants were initially selected based on differences in functional groups.

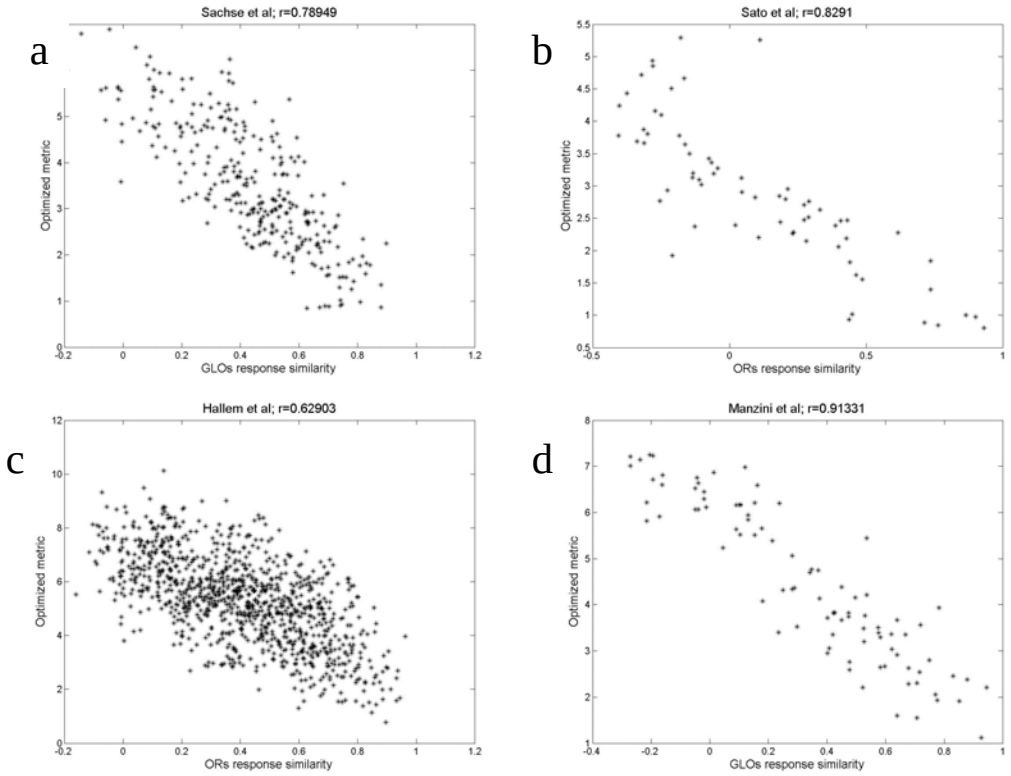
a


Number of discriminating neurons versus the distance between odorants group
Distance between groups (closest points distance)
Number of discriminating neurons
0
5
10
15
20
25
30
35
40
0
2
4
6
8
10
12
14
16
18

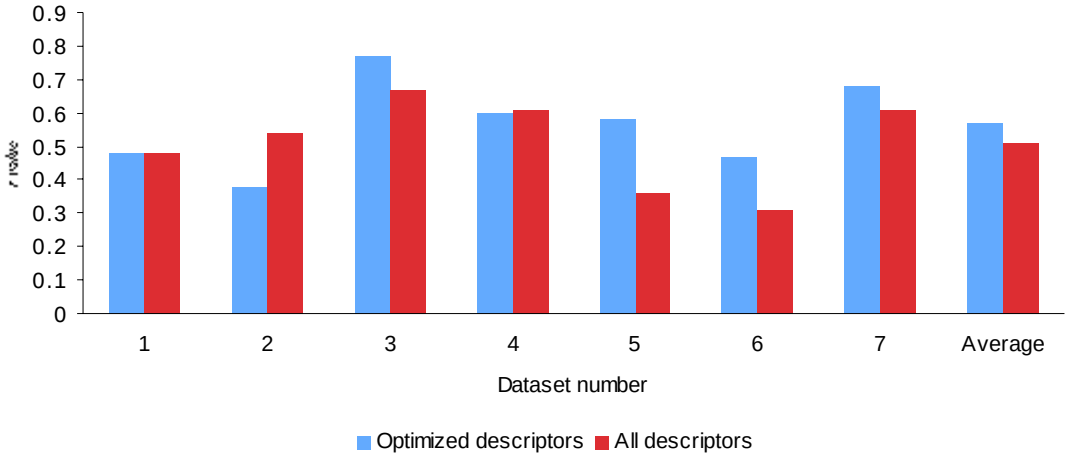
b

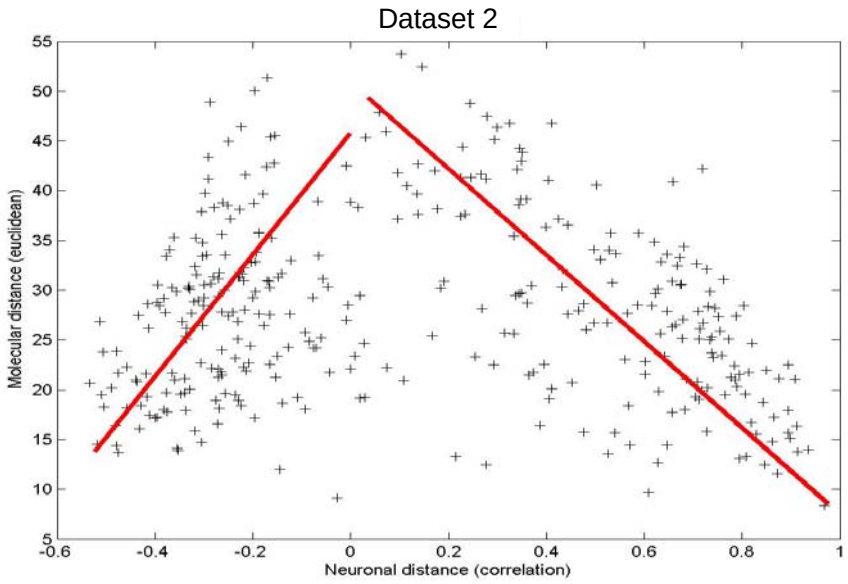

14 odorants tested by Sato et al (1994)
Distance
15
20
25
30
35
6
7
13
14
5
12
1
2
9
8
3
4
10
11
Odorants

Supplementary Figure 2: Correlation plots of four unrelated datasets. Each point in the graphs represents the distances between two odorants in both the neural space and the multidimensional physicochemical space.

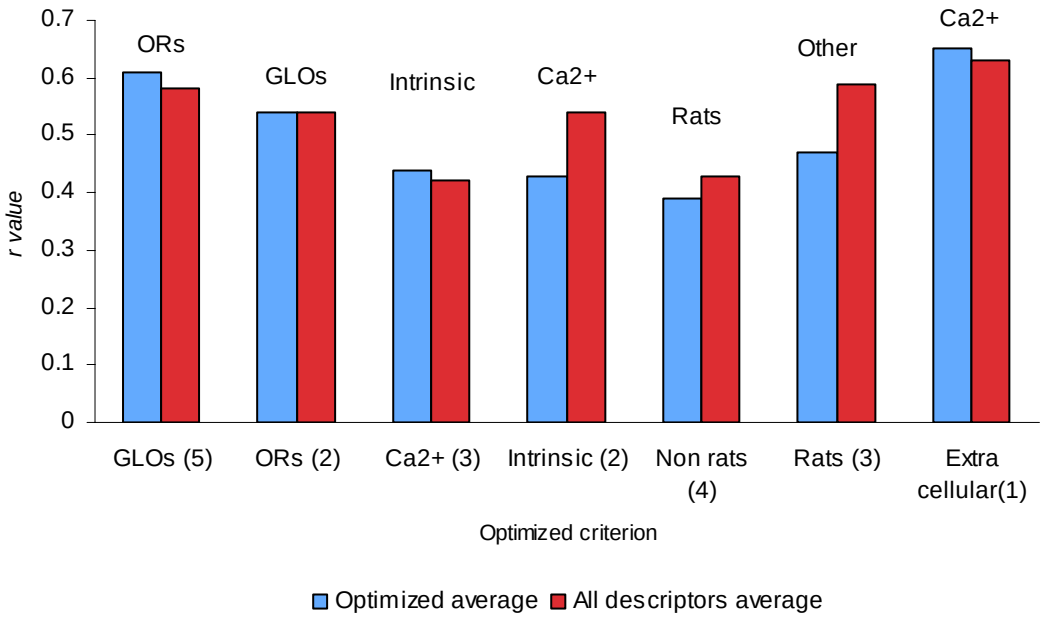


Supplementary Figure 3. Leave-one-out learning scheme. (**a**) The r value of each dataset when using the set of descriptors learned from using the other six datasets. (**b**) In dataset 2 only, the optimized metric was significantly lower than the all descriptors metric. However, the correlation when restricting to positive (0 to 1) range was 0.55, and -0.31 when in the negative range (0 to -1) ($P < 0.0001$, in both cases). It is if similar odorant can elicit positively or negatively correlated response patterns.





Supplementary Figure 4: Optimizing while using datasets that share a specific attribute. The labels above each bar state the attribute used for the tested datasets and the labels below each bar state the attribute used for grouping dataset for the optimizing process. For example, when optimizing the metric using datasets reporting GLOs response, the average r value for the datasets reporting olfactory receptors (ORs) response was 0.61 (leftmost bar). The number in brackets is the number of datasets used for the learning process.



Supplementary Table 1: The list of odorant's CAS number used to compare the carbon atom metric to the multidimensional metric. The blue odorants are the ones used to calculate the blue bars in Figure 1a and the red odorants are the additional odorants used to calculate the red bars in Figure 1a. To calculate the right most red bar we used 233 odorants that are commonly used in olfaction research.

Odorant functional group	List of odorants used (CAS)
Aldehydes	50-00-0 75-07-0 123-38-6 123-72-8 110-62-3 66-25-1 111-71-7 124-13-0 124-19-6 112-31-2 112-44-7 100-52-7 6728-26-3
Alcohols	67-56-1 64-17-5 71-23-8 71-36-3 71-41-0 111-27-3 111-70-6 111-87-5 143-08-8 112-30-1 112-42-5 106-24-1 108-93-0 626-93-7 928-96-1 67-63-0 78-92-2 6032-29-7 626-93-7 543-49-7 123-96-6 628-99-9 623-37-0 100-51-6 2216-51-5 99-48-9
Acids	64-18-6 64-19-7 79-09-4 107-92-6 109-52-4 142-62-1 111-14-8 124-07-2 112-05-0 334-48-5 112-37-8 503-74-2 65-85-0 103-82-2 79-31-2 600-07-7 140-10-3 107-93-7
Alkanes	74-82-8 74-84-0 74-98-6 106-97-8 109-66-0 110-54-3 142-82-5 111-65-9 111-84-2 124-18-5 1120-21-4 75-28-5 78-78-4 463-82-1 589-34-4 563-16-6 19398-77-7

Supplementary Table 2: The list of 32 descriptors composing the optimized multidimensional metric. Note that some descriptors have a weight higher than one.

	Descriptor Index	Weight	Description	Group
1	959		3D-MoRSE - signal 04 / weighted by atomic van der Waals volumes	3D-MoRSE descriptors
2	96		maximal electrotopological negative variation	topological descriptors
3	946		3D-MoRSE - signal 23 / weighted by atomic masses	
4	1558	2	phenol / enol / carboxyl OH	atom-centred fragments
5	688		aromaticity index	geometrical descriptors
6	948	2	3D-MoRSE - signal 25 / weighted by atomic masses	3D-MoRSE descriptors
7	592		Eigenvalue sum from electronegativity weighted distance matrix	eigenvalue-based indices
8	1528		R--CH—X	atom-centred fragments
9	690		d COMMA2 value / weighted by atomic masses	geometrical descriptors
10	963	3	3D-MoRSE - signal 08 / weighted by atomic van der Waals volumes	3D-MoRSE descriptors
11	1321		R maximal autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities	GETAWAY descriptors
12	1012		3D-MoRSE - signal 25 / weighted by atomic	Sanderson electronegativities
13	1191		leverage-weighted autocorrelation of lag 6 / weighted by atomic masses	GETAWAY descriptors
14	433		Eigenvalue 11 from edge adj. matrix weighted by resonance integrals	edge adjacency indices
15	513		lowest eigenvalue n. 8 of Burden matrix / weighted by atomic masses	Burden eigenvalues
16	698		d COMMA2 value / weighted by atomic Sanderson electronegativities	geometrical descriptors
17	1110		1st component shape directional WHIM index / weighted by atomic electrotopological states	WHIM descriptors
18	1541	2	R-C(=X)-X / R-C#X / X=C=X	atom-centered fragments
19	48		number of benzene-like rings	constitutional descriptors
20	359		Geary autocorrelation - lag 1 / weighted by atomic masses	2D autocorrelations
21	582		global topological charge index	topological charge indices
22	1430		number of aromatic hydroxyls	functional group counts

23	404		Eigenvalue 12 from edge adj. matrix weighted by edge degrees	edge adjacency indices
24	1576		R--N--R / R--N--X	atom-centred fragments
25	1373	2	number of carboxylic acids (aliphatic)	functional group counts
26	1348		number of terminal primary C(sp3)	functional group counts
27	1331	3	R autocorrelation of lag 2 / weighted by atomic polarizabilities	GETAWAY descriptors
28	92		Balaban-type index from mass weighted distance matrix	topological descriptors
29	1069		2st component symmetry directional WHIM index / weighted by atomic masses	WHIM descriptors
30	678		average span R	geometrical descriptors
31	22		number of double bonds	constitutional descriptors
32	1286		R maximal autocorrelation of lag 2 / weighted by atomic masses	GETAWAY descriptors

Supplementary Table 3: The list of odorants used for blind testing the metrics. The spreadsheet titled "Supplementary Data 1" contains the predicted distance between all pair wise odorants and the observed distances (odorants marked in red were removed from the analysis due to sparseness of responding neurons).

	CAS		CAS		CAS		CAS
1	79-09-4	19	623-37-0	37	96-22-0	55	2244-16-8
2	107-92-6	20	100-51-6	38	107-87-9	56	6485-40-1
3	109-52-4	21	2216-51-5	39	589-38-8	57	10458-14-7
4	53896-26-7	22	99-48-9	40	123-19-3	58	76-22-2
5	111-14-8	23	108-95-2	41	591-78-6	59	464-48-2
6	124-07-2	24	95-48-7	42	106-35-4	60	488-10-8
7	123-38-6	25	108-39-4	43	589-63-9	61	89-81-6
8	110-62-3	26	106-44-5	44	502-56-7	62	89-82-7
9	111-71-7	27	90-00-6	45	110-43-0	63	1128-08-1
10	100-52-7	28	620-17-7	46	106-68-3	64	108-94-1
11	71-36-3	29	123-07-9	47	111-13-7	65	628-28-4
12	71-41-0	30	90-05-1	48	624-16-8	66	111-43-3
13	111-27-3	31	93-51-6	49	821-55-6	67	470-82-6
14	111-70-6	32	97-53-0	50	928-80-3	68	71-43-2
15	111-87-5	33	100-66-3	51	14476-37-0	69	108-88-3
16	928-96-1	34	103-73-1	52	98-86-2	70	7785-70-8
17	67-63-0	35	4437-51-8	53	122-57-6	71	79-92-5
18	626-93-7	36	3848-24-6	54	127-41-3	72	99-86-5

Supplementary Methods

Forward greedy algorithm for finding the subset of best descriptors:

1. Set DescriptorSet to be the empty set
2. For each of the 1664 descriptors Di:
3. Calculate the total r value using the descriptors: DescriptorSet U Di
4. Select the descriptor Di that had the maximum r value and add it to DescriptorSet
5. Repeat stages 3 and 4 until the difference in the increase in the r value diminish to less than 0.004 in three consecutive runs.

Note that the forward greedy algorithm may select the same descriptor more than once. This is the same as adding weights to the Euclidean metric (integer weights). Dataset[3] gave better correlation values for positively correlated response pattern (see **Supplementary Fig. 3b**) and a significant negative value for the negatively correlated response pattern. We reported the positive range only for this dataset.

Selecting odorants for use in an experiment

We provide a spreadsheet titled "Supplementary Data 2" where one can probe the distance between any of more than 400 commonly used odorants with a simple push of a button.
To further facilitate the selection of odorants we suggest different odorant sets of different sizes (10,20,30,40,50,75,100 and 150 odorants) that guarantee to span the physicochemical space according to the multidimensional metric. To generate these groups we used the 'clusterdata' function from Matlab using the 'average' algorithm for the 'linkage' method. A spreadsheet named 'Supplementary Data 3' contains the list of odorants and their clustering according to the number of clusters requested. The suggested groups of odorants are calculated using all descriptors metric and the optimized metric (a separate workbook for each metric).

Reference

1. Hallem, E.A. & Carlson, J.R. Coding of odors by a receptor repertoire. *Cell* **125**, 143-160 (2006).
2. Sato, T., Hirono, J., Tonoike, M. & Takebayashi, M. Tuning specificities to aliphatic odorants in mouse olfactory receptor neurons and their local distribution. *J Neurophysiol* **72**, 2980-2989 (1994).
3. Uchida, N., Takahashi, Y.K., Tanifuji, M. & Mori, K. Odor maps in the mammalian olfactory bulb: domain organization and odorant structural features. *Nat Neurosci* **3**, 1035-1043 (2000).