OPEN ACCESS Freely available online

PLOS ONE

Categorical Dimensions of Human Odor Descriptor Space Revealed by Non-Negative Matrix Factorization

Jason B. Castro[1,2]*, Arvind Ramanathan[3], Chakra S. Chennubhotla[4]*

1 Department of Psychology, Bates College, Lewiston, Maine, United States of America, **2** Program in Neuroscience, Bates College, Lewiston, Maine, United States of America, **3** Computational Data Analytics Group, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States of America, **4** Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

Abstract

In contrast to most other sensory modalities, the basic perceptual dimensions of olfaction remain unclear. Here, we use non-negative matrix factorization (NMF) – a dimensionality reduction technique – to uncover structure in a panel of odor profiles, with each odor defined as a point in multi-dimensional descriptor space. The properties of NMF are favorable for the analysis of such lexical and perceptual data, and lead to a high-dimensional account of odor space. We further provide evidence that odor dimensions apply categorically. That is, odor space is not occupied homogenously, but rather in a discrete and intrinsically clustered manner. We discuss the potential implications of these results for the neural coding of odors, as well as for developing classifiers on larger datasets that may be useful for predicting perceptual qualities from chemical structures.

Citation: Castro JB, Ramanathan A, Chennubhotla CS (2013) Categorical Dimensions of Human Odor Descriptor Space Revealed by Non-Negative Matrix Factorization. PLoS ONE 8(9): e73289. doi:10.1371/journal.pone.0073289

Editor: Andreas Schaefer, MPI f. med. Research, Germany

Received April 26, 2013; **Accepted** July 18, 2013; **Published** September 18, 2013

Funding: CSC was partially supported by NIH GM086238. No additional external funding was received for this study. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jcastro@bates.edu (JBC); chakracs@pitt.edu (CSC)

Introduction

Our understanding of a sensory modality is marked, in part, by our ability to explain its characteristic perceptual qualities [1,2]. To take the familiar example of vision, we know that the experience of color depends on the wavelength of light, and we have principled ways of referring to distances between percepts such as 'red', 'yellow' and 'blue' [2,3]. In olfaction, by contrast, we lack a complete understanding of how odor perceptual space is organized. Indeed, it is still unclear whether olfaction even *has* fundamental perceptual axes that correspond to basic stimulus features.

Early efforts to systematically characterize odor space focused on identifying small numbers of perceptual primaries, which, when taken as a set, were hypothesized to span the full range of possible olfactory experiences [4–6]. Parallel work applied multidimensional scaling to odor discrimination data to derive a two-dimensional representation of odor space [7,8], and recent studies using dimensionality reduction techniques such as Principal Components Analysis (PCA) on odor profiling data have affirmed these low-dimensional models of human olfactory perception [9–11]. A consistent finding of these latter studies is that odor percepts smoothly occupy a low dimensional manifold whose principal axis corresponds to hedonic valence, or "pleasantness". Indeed, the primacy of pleasantness in olfactory experience may be reflected in the receptor topography of the olfactory epithelium [12] as well as in early central brain representations [13].

Here, we were interested in explicitly retaining additional degrees of freedom to describe olfactory percepts. Motivated by studies suggesting the existence of discrete perceptual clusters in olfaction [14,15] we asked whether odor space is amenable to a description in terms of sparse perceptual dimensions that apply categorically. To do so, we applied non-negative matrix factorization (NMF) [16–19] to the odor profile database compiled by Dravnieks [20] and analyzed in a number of recent studies [9–11]. NMF and PCA are similar in that both methods attempt to capture the potentially low-dimensional structure of a data set; they differ, however, in the conditions that drive dimensionality reduction. Whereas basis vectors obtained from PCA are chosen to maximize variance, those obtained from NMF are constrained to be non-negative. This constraint has proven especially useful in the analysis of documents and other semantic data where data are intrinsically non-negative [19,21] – a condition that is met by the Dravnieks database.

Applying NMF, we derive a 10-dimensional representation of odor perceptual space, with each dimension characterized by only a handful of positive valued semantic descriptors. Odor profiles tended to be categorically defined by their membership in a single one of these dimensions, which readily allowed co-clustering of odor features and odors. While the analysis of larger odor profile databases will be needed to generalize these results, the techniques described herein provide a conceptual and quantitative framework for investigating the potential mapping between chemicals and their corresponding odor percepts.

Materials and Methods

Non-Negative Matrix Factorization (NMF)

Non-negative matrix factorization (NMF) is a technique proposed for deriving low-rank approximations of the kind [16–18]:

$$\mathbf{A} \approx \mathbf{WH}, \quad (1)$$

where $\mathbf{A}$ is a matrix of size $m \times n$ with non-negative entries, and $\mathbf{W}$ and $\mathbf{H}$ are low-dimensional, non-negative matrices of sizes $m \times s$ and $s \times n$ respectively, with $s < \min(m,n)$. The matrices $\mathbf{W}$ and $\mathbf{H}$ represent feature vectors and their weightings. NMF has been widely used for its ability to extract perceptually meaningful features, from high dimensional datasets, that are highly relevant to recognition and classification tasks in several different application domains.

To derive $\mathbf{W}$ and $\mathbf{H}$ we used the alternate least squares algorithm originally proposed by Paatero [17]. Realizing that the optimization problem is convex in either $\mathbf{W}$ and $\mathbf{H}$, but not both, the algorithm iterates over the following steps:

1. assume $\mathbf{W}$ is known and solve the least squares problem for $\mathbf{H}$ using:

$$(\mathbf{W}^T\mathbf{W})\mathbf{H} = \mathbf{W}^T\mathbf{A}$$

2. set negative elements of $\mathbf{H} \rightarrow 0$
3. assume $\mathbf{H}$ is known and solve the least squares problem for $\mathbf{W}$ using

$$(\mathbf{H}\mathbf{H}^T)\mathbf{W}^T = \mathbf{H}\mathbf{A}^T$$

4. set negative elements of $\mathbf{W} \rightarrow 0$.

We used the standard implementation of non-negative factorization algorithm (nnmf.m) in Matlab (Mathworks, Inc.). Given the size of the odor profile matrix (146×140), the speed of convergence was not an issue. As a stopping criterion, we chose a value of 1000 for the maximum number of iterations. Given the iterative nature of the algorithm and small size of the dataset, we expect the algorithm to reach a global minimum for small s and a fixed point for large s.

Note that a minimum solution obtained by matrices $\mathbf{W}$ and $\mathbf{H}$ can also be satisifed by the pairs such as $\mathbf{WD}$ and $\mathbf{D}^{-1}\mathbf{H}$ for any nonnegative $\mathbf{D}$ and $\mathbf{D}^{-1}$. Thus, scaling and permutation can cause uniqueness problems, and hence the optimization algorithm typically enforces either row or column normalization in each iteration of the procedure outlined above.

Cross-validation procedure with training and testing sets

The choice of sub-space dimension s is problem dependent. Our strategy was to iterate over the sub-space dimension from $s = 1$ to 50, dividing the data matrix $\mathbf{A}$ each time into random but equal-sized training and testing halves. We kept track of the residual error in the form of the Frobenius error norm: $||\mathbf{A} - \mathbf{WH}||_F^2$ for both training and testing sets. For each choice of s we repeated this division 250 times, with a stopping criterion of 1000 iterations, to report the statistics on residual errors. In addition, once an optimal sub-space dimension is chosen, we report the most stable version of the basis matrix, by computing KL-divergence between every pair of the 250 instances of $\mathbf{W}$ from the training set and picking $\mathbf{W}$ with the lowest mean KL-divergence value.

Scrambling odor profiles

We applied NMF to scrambled perceptual data, that is elements of A are scrambled (randomly reorganized) before analyzing with NMF. Three different scrambling procedure were implemented. First was odorant shuffling where the column values of A are randomly permuted in each row. The second was descriptor shuffling where the row values of matrix A are randomly permuted in each column. Finally, we scrambled the elements of the entire matrix, that is indiscriminate shuffling of both descriptors and odorants entries.

Consensus matrix

We tested the stability of the NMF results on the original and scrambled versions of the perceptual data using a consensus clustering algorithm proposed in [22,23]. Because NMF is an iterative optimization algorithm, it may not converge to the same solution each time it is run (with random initial conditions). For a sub-space of dimension s, NMF algorithm groups descriptors and odorants into s different clusters. If the clustering into s classes is strong, we expect the assignment of descriptors or odorants to their respective clusters will change only slightly from one run to another. We quantified this with a consensus matrix. For illustration, we will work with cluster assignments made to the descriptors. In particular, each descriptor i is assigned to a meta-descriptor s', where $\mathbf{W}(i,s')$ is the highest among all the values of $\mathbf{W}(i,k)$ with $1 <= k <= s$.

We first initiated a zero-valued connectivity matrix $\tilde{\mathbf{C}}$ of size $m \times m$. For each run of NMF, we updated the entries of the connectivity matrix by 1, that is $\tilde{\mathbf{C}}_{ij} = \tilde{\mathbf{C}}_{ij} + 1$ if descriptors i and j belong to the same cluster, or 0 if they belong to different clusters. Averaging the connectivity matrix over all the runs of NMF gives the consensus matrix $\mathbf{C}$, where the maximum value of 1 indicates that descriptors i and j are always assigned to the same cluster. We ran NMF for 250 runs to ensure stability of the consensus matrix. If the clustering is stable, we expect the values in $\mathbf{C}$ to be close to either 0 or 1. To see the cluster boundaries, we can use off-diagonal elements of $\mathbf{C}$ as a measure of similarity among descriptors, and invoke an agglomerative clustering method where one starts by assigning each descriptor to its own cluster and then recursively merges two or more most similar clusters until a stopping criterion is fulfilled. The output from the agglomerative clustering method can be used to reorder the rows and columns of $\mathbf{C}$ and make the cluster boundaries explicit.

Cophenetic correlation coefficient

We then evaluated the stability of the clustering induced by a given sub-space dimension s. While visual inspection of the reordered $\mathbf{C}$ can provide qualitative insights into the stablity of cluster boundaries, we seek a quantitative measure by using the cophenetic correlation coefficient approach suggested in [23]. Note that there are two *distance* matrices to work with. The first distance matrix is induced by the consensus matrix generated by s-dim NMF decomposition. In particular, the distance between two descriptors is taken to be $1 - \mathbf{C}_{ij}$. The second distance matrix is one induced by a agglomerative clustering method, such as the average linkage hierarchical clustering (HC). In particular the off-diagonal elements of the consensus matrix can be used as distance values to generate hierarchical clustering (HC) of the data (in Matlab, invoke: linkage.m with average linkage option). HC imposes a tree structure on the data, even if the data does not have

a tree-like dependencies and is also sensitive to the distance metric in use. HC generates a dendrogram and the height h_{ij} of the tree at which two elements are merged provide for the elements of the second distance matrix. The cophenetic correlation coefficient ρ_s is defined to be the Pearson correlation value between the two distance matrices. If the consensus matrix is perfect, with elements being either 0 or 1, then ρ_s is 1. When the consensus matrix elements are between 0 and 1, then $\rho_s < 1$.

We plot ρ_s vs s for increasing values of s. The results of such analyses are in some cases helpful for choosing an optimal subspace size. If a given clustering (say, for subspace size of $s-1$) is highly reliable across repeated factorizations (that is, the same sets of descriptors and the same sets of odors tend to co-cluster), and hence ρ_{s-1} is very high, then one is motivated to retain at least $(s-1)$ dimensions. If increasing this subspace size (to s, $s+1$, etc) leads to systematically less reliable clustering, $\rho_{s-1} > (\rho_s, \rho_{s+1}, \cdots)$, one is motivated to retain the more conservative estimate of dimensionality $(s-1)$. That said, we note that cophenetic correlation analyses can often provide better grounds for excluding certain choices of subspace size that lead to unreliable clustering, rather than privileging a specific number as 'the' dimensionality of the data. Note we seek solutions where $s > 1$ because for $s = 1$ the correlation coefficient $\rho_1 = 1$. We also performed a similar consensus clustering and cophenetic coefficient analysis in the odorant space using the entries in **H**.

Odor space visualization

We use a variant of stochastic neighbor embedding method [24,25] to visualize the high-dimensional odor space organized by NMF. In particular, we first generated the consensus matrices for clustering descriptors and odorants, and used them separately as similarity matrices in the stochastic neighbor embedding algorithm. We used the code from http://homepage.tudelft.nl/19j49/t-SNE.html and ran it with default parameters.

Results

Dimensionality of odor space

We analyzed the published data set of Dravnieks [20], which catalogs perceptual characteristics of 144 monomolecular odors. Each odor in this data set is represented as a 146 dimensional vector (an odor profile), with each dimension corresponding to the rated applicability of a given semantic label, such as 'sweet', 'floral', or 'heavy'. Because these are strictly non-negative quantities (i.e. a given semantic label either applies, or does not), we reasoned this could be meaningfully exploited when reducing the dimensionality of profiling data. Thus, we applied NMF to the profiling data in an effort to obtain a perceptual basis set corresponding to 'parts' or 'features', as has been observed in the analysis of images [18] and text [18,21].

NMF seeks a low-rank approximation of a matrix **A** (146 descriptors × 144 odors in the present case) as the product **WH**, where the s columns of **W** are non-negative basis vectors (146-D vectors of odor descriptors in the present case), and the columns of **H** are the new s-dimensional representations of the original odors (144 columns, in the present case) (Fig. 1A). Figure 1B shows the root-mean-squared (RMS) residual (see Methods) between **A** and its approximation **WH** for subspaces ranging from 1 to 50 (100 equal divisions of **A** into training and testing subsets, for each choice of subspace). The residual attained a minimum for a subspace choice of 25, and increased for larger subspaces. In addition, the width of the error bars increased on the training and testing residuals after subspace 25. Increasing the number of iterations used for training the NMF model only marginally reduced the size of the error bars. We speculate that the energy landscape is becoming increasingly rugged, with the existence of many more local minima to potentially trap the learning of NMF model parameters. In particular, NMF employs a non-linear optimization method, and hence it is possible that the each time the method is run, it finds a local minimum that is different and far away from a global minimum. Hence, the error bars on the residuals are large and continue to increase with increasing subspace dimensionality s because of the ruggedness in the landscape and the limited size of odor profile data used for training the model.

Notably, for subspaces 1–25 – a regime in which training error decreases continuously – the testing error decreases, attains a minimum, and then begins to increase. Thus, while a 25 dimensional representation of the original perceptual data is evidently the most accurate achievable with NMF, it is not necessarily the most parsimonious. Inspecting low-order basis vectors, we observed that descriptors with largest-amplitudes were consistent across repetitions of the factorization, and corresponded to broadly applicable labels such as 'fragrant', and 'sickening' (see Figure 2 for examples). By contrast, higher order basis vectors (>10) had peak-value descriptors that were highly specific ('anise', 'cinammon', etc), and somewhat variable between NMF repetitions.

To more quantitatively motivate the choice of subspace size, we applied two techniques commonly used in problems of NMF model selection [23,26]. First, we plotted reconstruction error (that is, the fraction of unexplained variance) vs subspace size for 250 different repetitions of NMF (Fig. 1C), and compared this to the reconstruction error obtained with PCA performed on the original data ($\mathbf{PCA}_{\mathrm{orig}}$) as well as on scrambled data ($\mathbf{PCA}_{\mathrm{scram}}$) (Fig. 1D) [26]. The slope of $\mathbf{PCA}_{\mathrm{scram}}$ is small and relatively constant for increasing subspace sizes (Fig. 1D), and provides a means for estimating the point after which a given model is explaining noise rather than correlations in data . To visualize this cutoff point, Figure 1D plots the change in variance for each added dimension (differences between successive points in Figure 1C). The reconstruction error rates of both $\mathbf{PCA}_{\mathrm{orig}}$ and NMF intersect with $\mathbf{PCA}_{\mathrm{scram}}$ at subspace size 10 (Fig. 1D), indicating that there is no gain in retaining dimensions >10 for either dimensionality reduction method. This is consistent with a recently published estimate of the intrinsic dimensionality of this same dataset [11], using PCA. For a further comparison of NMF with PCA, we show cumulative variance plots of PCA and several runs of NMF in Fig. S1.

As a second means for quantifying the intrinsic dimensionality of the Dravnieks data set, we calculated the cophenetic correlation coefficient [23] for several choices of subspace size. Briefly, this method exploits the stochasticity inherent in NMF to determine how reproducible the derived basis set and odor weights are across repetitions of the factorization. Cophenetic correlations ≈ 1 indicate highly reproducible basis sets (see Methods for further explanation). We note that cophenetic correlation analyses can often provide better grounds for excluding certain choices of subspace size that lead to unreliable clustering, rather than privileging a specific number as 'the' dimensionality of the data.

The results of our cophenetic correlation analysis are shown in supplementary Fig. S2. Two features are readily apparent: First, there are some notably poor choices of subspace size (such as $s = 4$ or $s = 5$). We speculate that the sharp drop at these values is because at these subspace choices, the classification scheme has lost the advantage of being simple and dichotomous, but has yet to support enough categories for accurate and reliable classification. Second, unlike with the reconstruction error criterion (above),

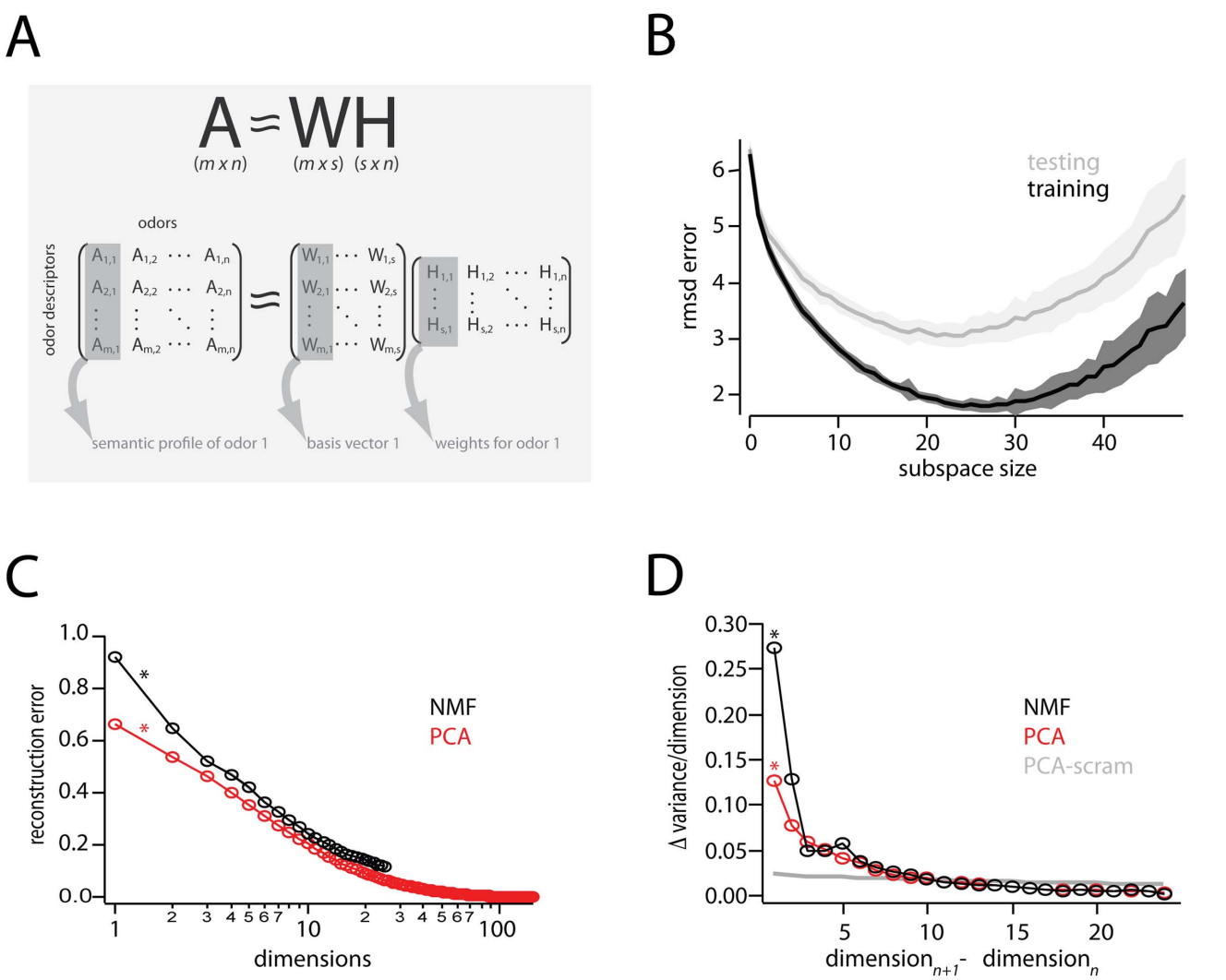


Figure 1. Summary of non-negative matrix factorization (NMF) applied to odor profiling data. (A) Schematic Overview: NMF seeks a lower, *s*-dimensional approximation of a matrix **A** as the product of matrices **W** and **H**. **A** is $m \times n$, consisting in the present study of 146 odor descriptors $\times$ 144 odors. A given column of **A** is the semantic profile of one odor, with each entry providing the percent-used value (see methods) of a given descriptor. Columns of **W** are basis vectors of the reduced, s-dimensional odor descriptor space. Columns of **H** are *s*-dimensional representations (weights) of the odors in the new basis. (B) Plot of residual error between perceptual data, **A**, and different NMF-derived approximations. **WH**. For each choice of subspace, data were divided into random training and testing halves, and residual error between **A** and **WH** computed. One-hundred such divisions into training and testing were used to compute the standard errors shown (shaded areas). (C) Reconstruction error (fraction of *unexplained* variance) for PCA and NMF vs. number of dimensions. The change in reconstruction error for the first interval is indicated by asterisks(*), and corresponds to the first point in the next panel. (D) Change in reconstruction error for PCA and NMF, compared to the change in reconstruction error for PCA performed on a scrambled matrix (PCA_{scram}). PCA_{scram} is used to estimate the cutoff number of dimensions for which a given dimensionality reduction method is explaining only noise in a dataset. Note that each point, *n*, is actually the difference in reconstruction error between dimensions *n* and $n+1$ (by way of illustration, points with an asterisk in this panel denote corresponding intervals in the previous panel C).
doi:10.1371/journal.pone.0073289.g001

there is no monotone decreasing relationship between cophenetic correlation and dimension size that provides an obvious stopping criterion. Our interpretation of this is that there are many good, reduced-dimensionality representations of the Dravnieks data that exhibit sparse structure.

Given that analysis of reconstruction error (Fig. 1D) argues for a choice of 10 dimensions as a cutoff point, and cophenetic correlation analysis suggests there are many well-motivated choices of subspace choices $>=6$ (Fig. S2), we therefore settled on a subspace size of 10 for all further analyses. Visualizations of NMF reconstruction quality for different choices of subspace size are provided in figure S3, which shows that most of the global and local structure of the original data is explained with 10 NMF basis vectors. We wish to note, however, that in general there is no single exact criterion for NMF model selection. There are multiple justifiable choices of subspace size, each of which may lead to different insights about the data, or be useful for different goals.

Sparseness of basis vectors

An immediate consequence of the non-negativity constraint is sparseness of the basis vectors. As seen in Figure 2, the basis vectors consist of a handful of large values, with the remaining values near or equal to zero. Intuitively, a given basis vector

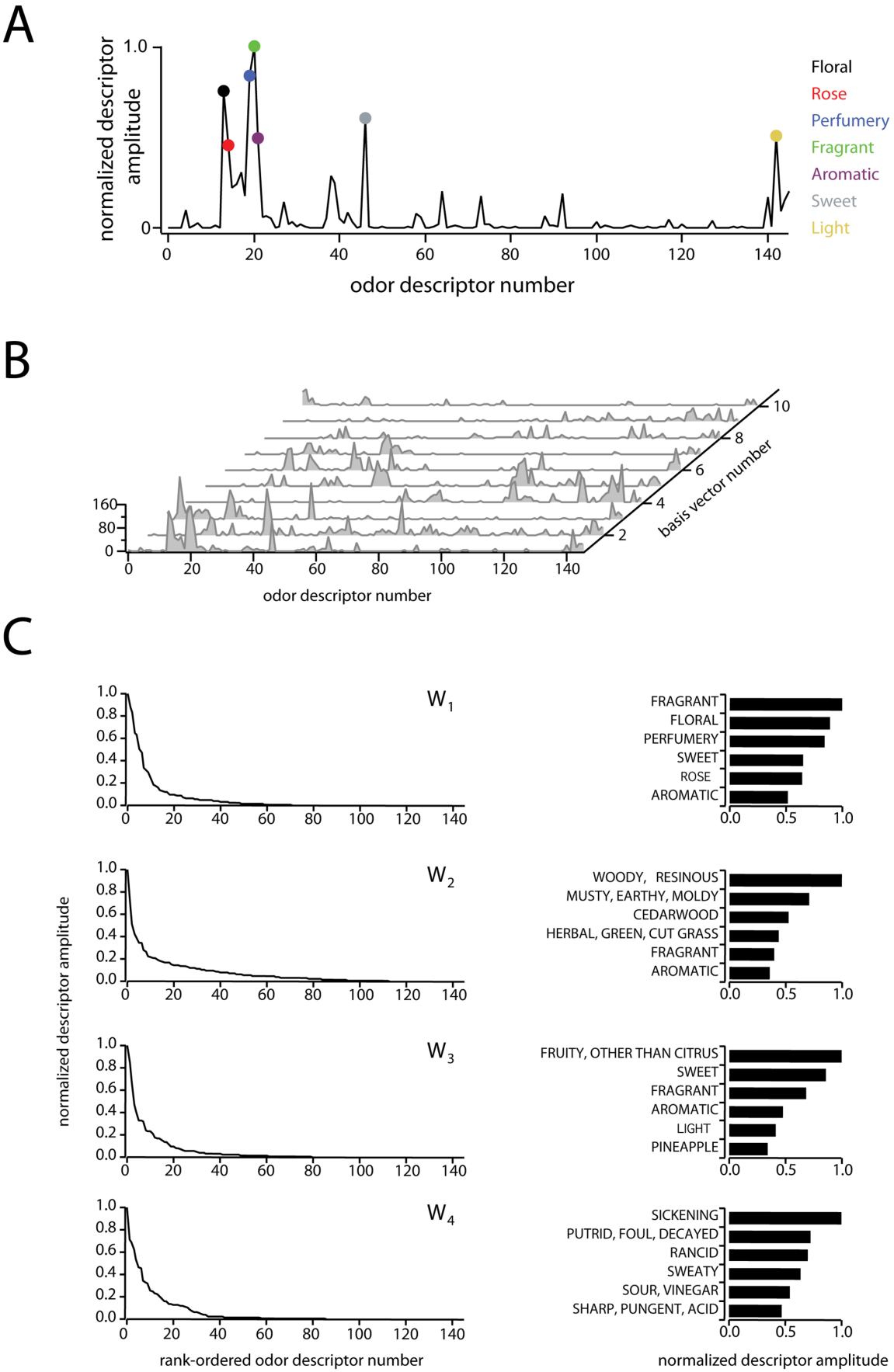

A
normalized descriptor amplitude
1.0
0
0
20
40
60
80
100
120
140
odor descriptor number
Floral
Rose
Perfumery
Fragrant
Aromatic
Sweet
Light
B
160
80
0
10
8
6
4
2
basis vector number
odor descriptor number
C
normalized descriptor amplitude
W1
W2
W3
W4
FRAGRANT
FLORAL
PERFUMERY
SWEET
ROSE
AROMATIC
WOODY, RESINOUS
MUSTY, EARTHY, MOLDY
CEDARWOOD
HERBAL, GREEN, CUT GRASS
FRAGRANT
AROMATIC
FRUITY, OTHER THAN CITRUS
SWEET
FRAGRANT
AROMATIC
LIGHT
PINEAPPLE
SICKENING
PUTRID, FOUL, DECAYED
RANCID
SWEATY
SOUR, VINEGAR
SHARP, PUNGENT, ACID
rank-ordered odor descriptor number
normalized descriptor amplitude

Figure 2. Properties of the perceptual basis set W. Plot of normalized odor descriptor amplitude vs. odor descriptor number for the basis vector $\mathbf{W}_1$. Each point along the x-axis corresponds to a single odor descriptor, and the amplitude of each descriptor indicates the descriptor's relevance to the shown perceptual basis vector. Colored circles show the 7 largest points in the basis vector, and descriptors corresponding to these points are listed to the right. (B) Waterfall plot of the 10 basis vectors constituting **W**, used in subsequent analyses. Note that each vector contains many values close to or equal to zero. (C) Detailed view of the first four basis vectors and their leading values. Left column: peak-normalized, rank ordered basis vectors, illustrating their sparseness and non-negativity. Right column: semantic descriptors characterizing the first four basis vectors. Bars show the first six rank-ordered, peak-normalized components of basis vectors 1 through 4 (subset of data from left column). The semantic label for each component is show to the left.
doi:10.1371/journal.pone.0073289.g002

indicates a subset of descriptors that are related and particularly informative (Fig. 2 A), while the set of all basis vectors (Fig. 2B) defines a library of such aggregate descriptors that span the space. Figure 2C shows the first four basis vectors, which have been normalized and ranked in decreasing order to highlight their sparseness. The six most heavily weighted descriptors for each basis vector are shown to the right. Together, these vectors define 4 descriptor axes that can be roughly labeled as 'fragrant', 'woody', 'fruity', and 'sickening.' We note that these labels are for purposes of concision only, as each axis is actually a meta-descriptor consisting of a linear combination of more elementary descriptors. A list of rank-ordered descriptors for all 10 dimensions is shown in Table 1.

To ensure that the sparse basis vectors we obtained were not an artifact of the NMF procedure, but rather depended on correlations in the data, we repeated the calculation of W for three shuffled versions of the profiling data (Fig. 3). In the 'full shuffle' condition, all elements of the data matrix A were randomly permuted, eliminating all correlations. In the 'descriptors-shuffled' conditions, the elements of each column of A were randomly permuted, while in the final 'odorants-shuffled' conditions, the elements of each row of A were randomly permuted. In agreement with the idea that the sparseness obtained by NMF is data dependent, sparseness was drastically reduced in the basis sets obtained from all sets of shuffled data (compare Fig. 3C with Fig. 2B).

In histograms of basis vectors obtained from the full-shuffled and descriptor-shuffled data (Fig. 3A), it was evident that both basis sets contained fewer zero-valued elements than the unshuffled basis set. Interestingly, the long-tail behavior of the histogram was preserved (even enhanced) in the odorants-shuffled condition (Fig. 3B). While this does indicate that a small number of basis vector elements did have very large values in the odorant shuffle cases, this was notably at the expense of peak behavior at zero (Fig. 3A, green). Moreover, basis vectors derived from a given odorant-shuffled matrix were highly inconsistent across repetitions of the factorization, which we assessed by computing consensus matrices (see Methods) documenting the stability of clusters across different iterations of NMF (Figure 4). In brief, we found that only the original data had clusters that were consistent across iterations.

While these first several NMF dimensions (Fig. 2, and Table 1) define a perceptual descriptor space reminiscent of that observed previously with PCA, we note that variance is distributed somewhat differently in the NMF vs PCA basis sets. In essence, we have traded degrees of freedom for increased interpretability of individual perceptual dimensions. Interestingly, despite the fact that NMF imposes no formal orthogonality constraint on basis vectors, the perceptual basis set discovered by NMF was still near-orthogonal (Fig. 5); that is, most pairwise comparisons among the basis vectors in **W** subtend an angle close to $\pi/2$ (median angle = 72.9 degrees).

Distribution of odors in the new perceptual descriptor space

We next asked how the 144 individual odor profiles (that is, columns of **H**) are distributed in the new 10 dimensional perceptual descriptor space spanned by **W**. One possibility, for example, is that many of the descriptor space dimensions are redundant, resulting in odors being confined to a thin, low-dimensional slice of the full space. At the other extreme, odors may densely occupy descriptor space, indicating that dimensions contain non-redundant features, with all dimensions necessary to fully characterize odors.

To investigate these and other possibilities, we first examined the structure of **H**, the matrix of odor weights obtained from NMF (recall that each column of **H** corresponds to an odor, and defines a point in 10-dimensional descriptor space spanned by **W**; Fig. 1A). We took the Euclidian norm of each column of **H**, and then sorted all columns into 10 groups defined by their largest coordinate in descriptor space. More explicitly, the 144 columns of **H** were scanned left to right until one was found with a largest coordinate in dimension 1. This was then assigned as the first column of the re-ordered matrix. The remaining set of columns was similarly scanned, until all columns with a largest first-coordinate had been found. This procedure was then iterated on the remaining dimensions 2–10. Note that this is just a cosmetic reordering of columns that preserves row orderings – no new structure has been added, and no existing structure been destroyed.

Intriguingly, this procedure revealed a prominent block diagonal structure to the full matrix **H** (Fig. 6A) indicating that: 1) a given odor tends to be characterized by a single prominent dimension, and 2) all 10 dimensions are occupied. Furthermore, this suggests that a given odor percept may be considered an instance of one of several fundamental qualities (see discussion).

These two properties can be alternatively visualized when odors (columns of **H**) are plotted as points in the 10 dimensional perceptual space spanned by basis set **W**. Because this perceptual space is high-dimensional and difficult to represent geometrically, we show a representative 3 dimensional subspace of **W**. We note that this is not a projection of the data, but rather a selective visualization of a subspace. Figure 6B shows all 144 odors in the space spanned by perceptual dimensions 1–3. Most odors are clustered diffusely near the origin (gray points in Fig. 6B), since their peak coordinates do not reside in this particular 3-D subspace. By contrast, when odors are separated into groups defined by peak coordinate (as in Fig. 6A), it is evident that a given odor tends to be best defined by a single perceptual dimension. The black, red, and blue points in figure 6B, for example, are those points with largest coordinates occurring in the first, second, and third dimensions respectively. While there was notable structural homology among the odors in a given diagonal block of **H** (Fig. 6C), we did not quantify this further in the present work. Figure S4 shows additional representations of odorants distributed in descriptor space, and further highlights the categorical nature of the perceptual space derived from NMF.

Table 1. 10 largest-valued descriptors for each of the 10 basis vectors obtained from non-negative matrix factorization.

W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
FRAGRANT	WOODY, RESINOUS	FRUITY, OTHER THAN CITRUS	SICKENING	CHEMICAL	MINTY, PEPPERMINT	SWEET	POPCORN	SICKENING	LEMON
FLORAL	MUSTY, EARTHY, MOLDY	SWEET	PUTRID, FOUL, DECAYED	ETHERISH, ANAESTHETIC	COOL, COOLING	VANILLA	BURNT, SMOKY	GARLIC, ONION	FRUITY, CITRUS
PERFUMERY	CEDARWOOD	FRAGRANT	RANCID	MEDICINAL	AROMATIC	FRAGRANT	PEANUT BUTTER	HEAVY	FRAGRANT
SWEET	HERBAL, GREEN, CUT GRASS	AROMATIC	SWEATY	DISINFECTANT, CARBOLIC	ANISE (LICORICE)	AROMATIC	NUTTY (WALNUT ETC)	BURNT, SMOKY	ORANGE
ROSE	FRAGRANT	LIGHT	SOUR, VINEGAR	SHARP, PUNGENT, ACID	FRAGRANT	CHOCOLATE	OILY, FATTY	SULFIDIC	LIGHT
AROMATIC	AROMATIC	PINEAPPLE	SHARP, PUNGENT, ACID	GASOLINE, SOLVENT	MEDICINAL	MALTY	ALMOND	SHARP, PUNGENT, ACID	SWEET
LIGHT	LIGHT	CHERRY (BERRY)	FECAL (LIKE MANURE)	PAINT	SPICY	ALMOND	HEAVY	HOUSEHOLD GAS	COOL, COOLING
COLOGNE	HEAVY	STRAWBERRY	SOUR MILK	CLEANING FLUID	SWEET	CARAMEL	WARM	PUTRID, FOUL, DECAYED	AROMATIC
HERBAL, GREEN, CUT GRASS	SPICY	PERFUMERY	MUSTY, EARTHY, MOLDY	ALCOHOLIC	EUCALIPTUS	LIGHT	MUSTY, EARTHY, MOLDY	SEWER	HERBAL, GREEN, CUT GRASS
VIOLETS	BURNT, SMOKY	BANANA	HEAVY	TURPENTINE (PINE OIL)	CAMPHOR	WARM	WOODY, RESINOUS	BURNT RUBBER	SHARP, PUNGENT, ACID

doi:10.1371/journal.pone.0073289.t001

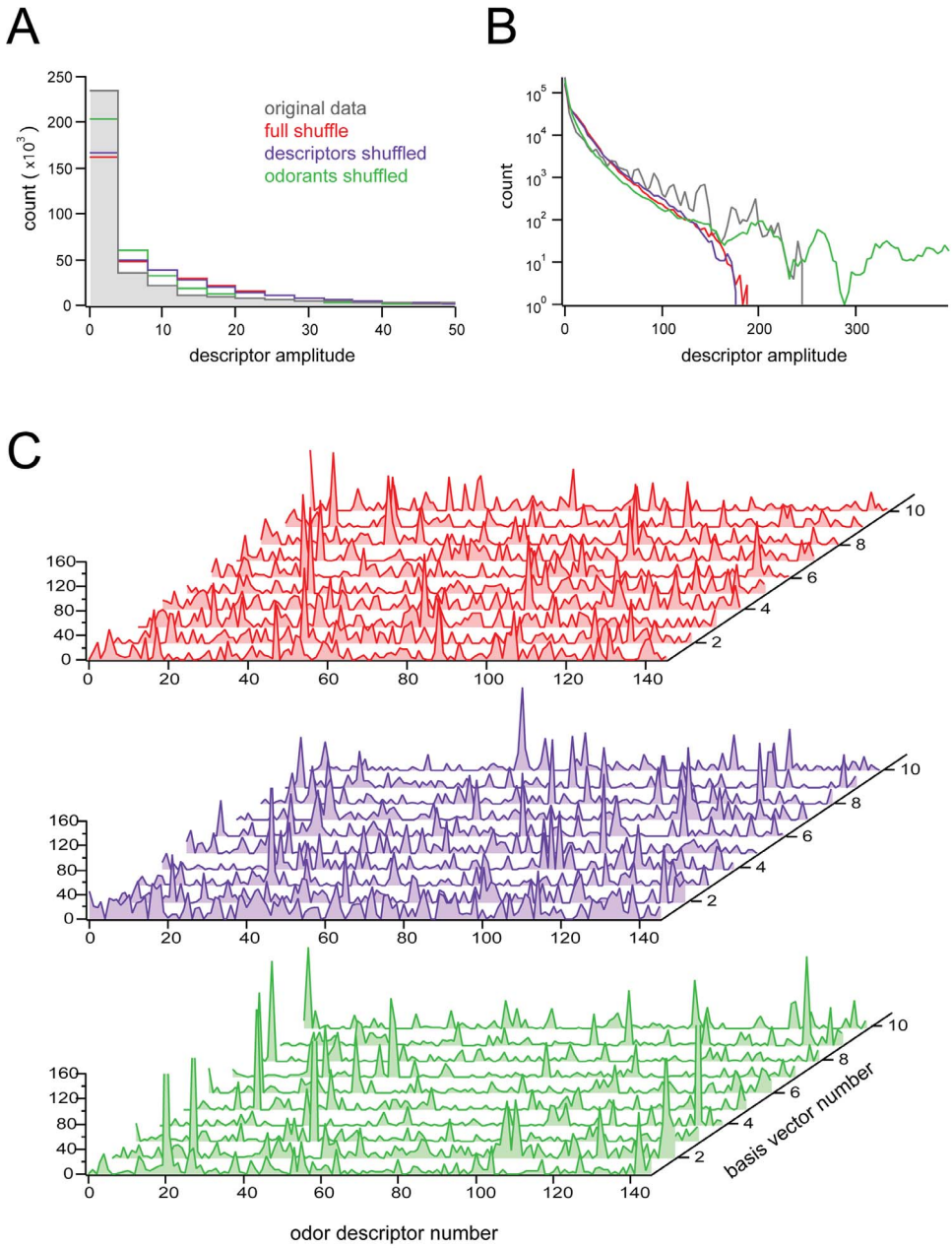


Figure 3. NMF on full, descriptor-only, and odor-only shuffled versions of the data. (A) Peak behavior of histograms obtained from NMF performed on shuffled data, for each of the various shuffling conditions (see text for descriptions). (B) Tail behavior of histograms, same procedure and conditions as in (A); note difference in scaling of axes between (A) and (B). (C) Waterfall plots of basis sets obtained when NMF was applied on shuffled data, for various shuffling conditions. Note the comparative lack of sparseness, relative to the basis set shown in Fig. 3A. Reproducibility of basis vectors across iterations of NMF for shuffled data sets was eliminated, or severely compromised, as shown in Fig. 4.
doi:10.1371/journal.pone.0073289.g003

As a final means for investigating whether odorants are smoothly vs. discretely arranged in descriptor space, we constructed two-dimensional embeddings for the matrices **W** and **H** using the stochastic neighbor embedding (SNE) algorithm. Briefly, this technique provides a planar representation of all pairwise distances between odors in the original high dimensional space, such that relative neighbor relations are preserved (e.g. odors that are close together in the original space are also close together in the embedding). Applying SNE to the descriptor space (**W**), we obtained 8 discrete and non-overlapping clusters of the 146 descriptors, which are shown in Figure 7. Similarly, applying SNE to the space of odorants (H), we obtained 10 discrete and non-overlapping clusters of the 144 odors (Figure 8). In sum, the perceptual descriptor space derived from NMF is not smoothly occupied.

Bi-clustering of descriptors and odors

The perceptual space, **W**, discovered by NMF can be considered a set of 10 meta-descriptors, each of which is a linear combination of more elementary descriptors. While these dimensions are compact and categorical in that a given odor tends to have a prominent single coordinate (Figs. 6 and S4), this may also

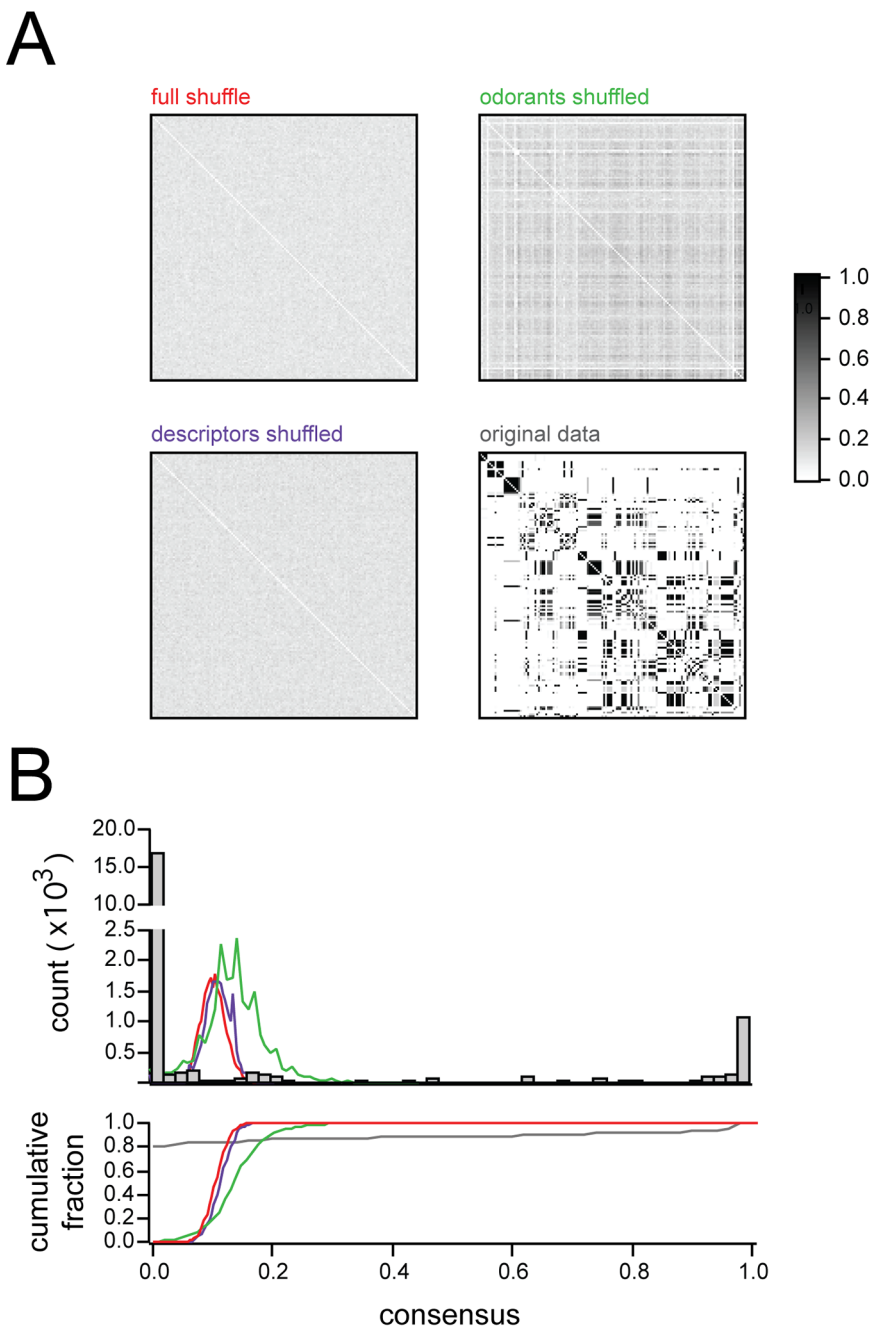


Figure 4. Consensus Matrices for odor-shuffles, descriptor-shuffles, and full-shuffles. (A) Consensus matrices (see text) showing reliability of basis sets when NMF is applied to various shuffled versions of the data. Only the original data shows the bimodal distribution of 1s and 0s characteristic of highly reliable clustering. Image ranges and colorscale same for all 4 matrices. (B) Top: Histograms of consensus matrix values for the three shuffling conditions, and the original data, confirming that only the original data shows a bimodal distribution of 1s and 0s (line colors correspond to labels in (A)). Bottom: Cumulative histograms, same data as above.
doi:10.1371/journal.pone.0073289.g004

obscure interesting details about the organization of the descriptor space. For example, within a dimension there may be correlations between specific descriptors and specific odors.

To explore this potential fine-scale structure wherein subsets of odorants show distinct correlations among subsets of descriptors, we sought submatrices of **WH** (the NMF approximation to the original data matrix **A**) with large values in both the descriptor and odorant dimensions (Fig. 9). Briefly, we did this by performing 10-reorderings (one for each perceptual dimension) of rows and columns of **WH** via the process illustrated in figure 9A. Rank-ordering the first column of **W**, for example, aggregates the peak valued descriptors for the first perceptual dimension, $\mathbf{W}_1$. Similarly, rank ordering the first row of H aggregates those odorants with largest weights in $\mathbf{W}_1$. Applying these row and column re-orderings simultaneously to the matrix **WH** gives a matrix whose largest values are in the upper-left corner.

The clear upper-left organization of these submatrices illustrates that there are sets of odors to which distinct odor descriptors apply.

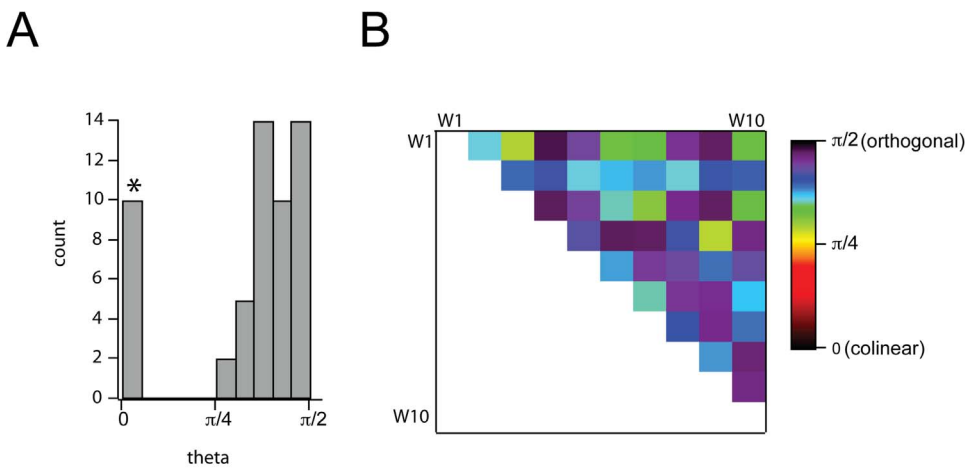
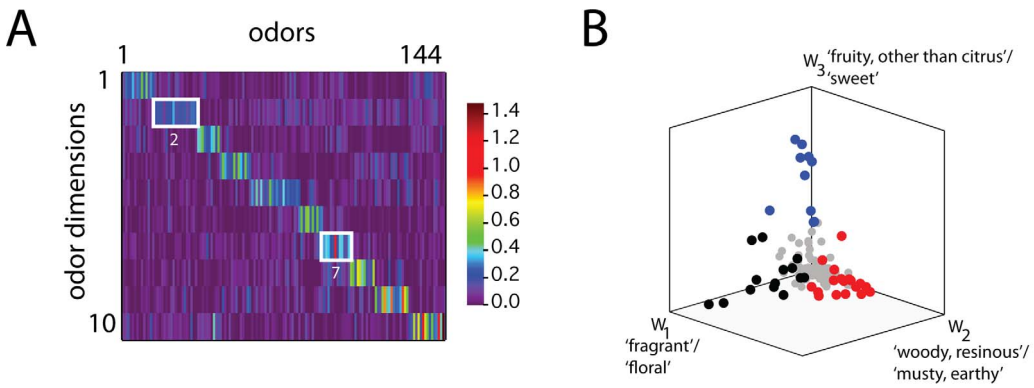


Figure 5. Approximate orthogonality of the NMF basis vectors. (A) Histogram of angles subtended by all pairs of basis vectors, **W**. Histogram was constructed for all pairwise comparisons between dimensions, excluding self-comparisons. Bar with (*) denotes self-comparisons. (B) Matrix of pairwise comparisons of angles between dimensions.
doi:10.1371/journal.pone.0073289.g005



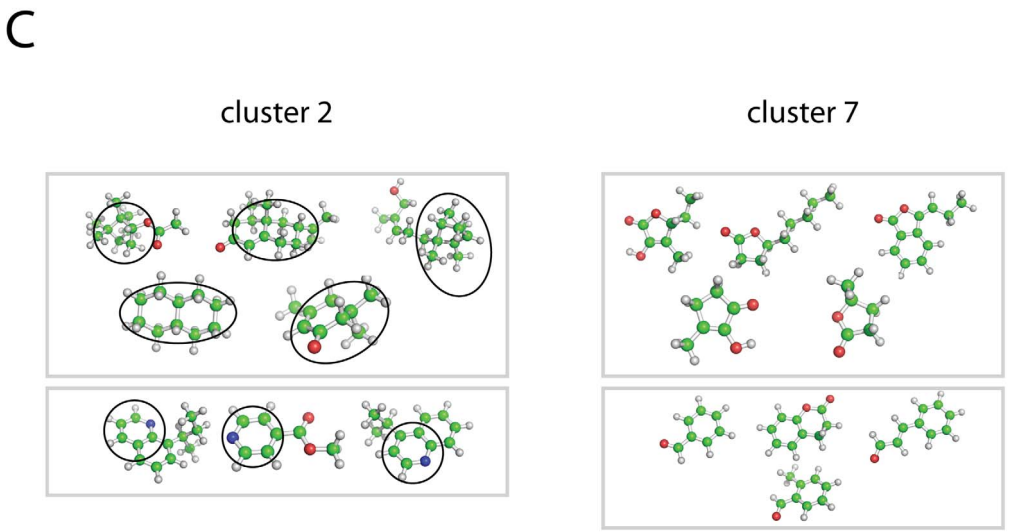
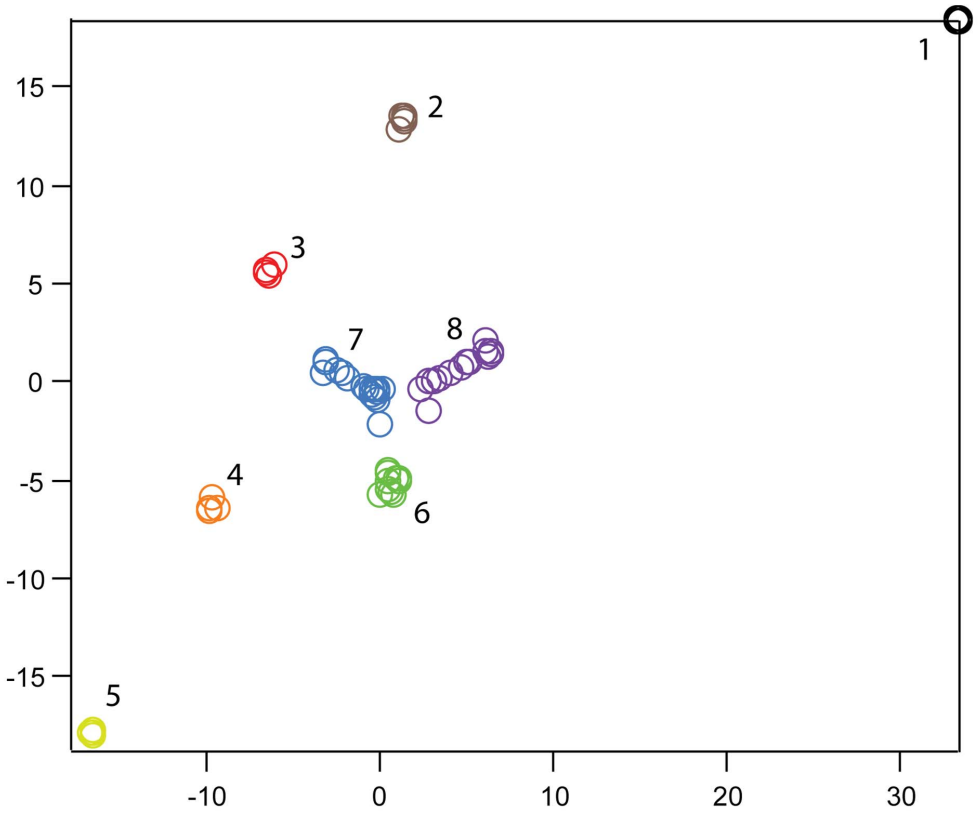


Figure 6. Visualization of odors expressed in coordinates of the new basis. (A) The weight matrix, **H**, discovered by NMF. Columns of **H** (each column corresponds to a different odor), are normalized and sorted into groups defined by peak coordinate (1–10). (B) Plot of all 144 odors (each point is a column of **H**) in the space spanned by the first 3 basis vectors, $\mathbf{W}_1$,$\mathbf{W}_2$, and $\mathbf{W}_3$. Black, red, and blue points are those with peak coordinates in dimensions 1, 2, and 3 respectively. Gray points are all remaining odors. (C) Chemical structures of representative odorants from the second and seventh diagonal blocks of the sorted matrix **H** (panel (A)).
doi:10.1371/journal.pone.0073289.g006



Cluster 1

FRUITY/CITRUS, LEMON, GRAPEFRUIT, ORANGE

Cluster 2

NAIL POLISH REMOVER, MOTHBALLS, ALCOHOLIC, ETHERISH, ANAESTHETIC, CLEANING FLUID, GASOLINE, SOLVENT, TURPENTINE (PINE OIL), LEATHER, TAR, CREOSOTE, DISINFECTANT, CARBOLIC, MEDICINAL, CHEMICAL, AMMONIA, NEW RUBBER, KEROSENE PAINT, VARNISH, METALLIC

Cluster 3

TEA LEAVES, CARAWAY, MINTY, PEPPERMINT, CAMPHOR, EUCALIPTUS, ANISE (LICORICE), CHEESY, COOL, COOLING

Cluster 4

FLORAL, ROSE, VIOLETS, LAVENDER, COLOGNE, MUSK, PERFUMERY, FRAGRANT, AROMATIC, SOAPY, INCENSE, LIGHT

Cluster 5

FRUITY/OTHER THAN CITRUS, PINEAPPLE, GRAPE JUICE, STRAWBERRY, PEAR, CANTALOUPE, HONEY, DEW MELON, PEACH (FRUIT), BANANA, CHERRY (BERRY), SWEET, RAISINS

Cluster 6

LAUREL LEAVES, BLACK PEPPER, GREEN PEPPER, DILL, OAK WOOD, COGNAC, WOODY, RESINOUS, CEDARWOOD, GERANIUM LEAVES, CELERY, FRESH GREEN VEGETABLES, CRUSHED WEEDS, CRUSHED GRASS, HERBAL, GREEN, CUT GRASS, RAW CUCUMBER, CARDBOARD, ROPE, WET PAPER, MUSTY, EARTHY, MOLDY, RAW POTATO, MUSHROOM, BEANY, BARK, BIRCH BARK, CORK, DRY, POWDERY, CHALKY

Cluster 7

HONEY, ALMOND, NUTTY (WALNUT ETC), SPICY, CLOVE, CINNAMON, CHOCOLATE, VANILLA, MAPLE SYRUP, CARAMEL, MALTY, MOLASSES, COCONUT, HAY, BAKERY (FRESH BREAD), PEANUT BUTTER, BURNT, SMOKY, FRESH TOBACCO SMOKE, COFFEE, STALE TOBACCO SMOKE, BURNT PAPER, BURNT MILK, BURNT CANDLE, OILY, FATTY, BUTTERY, FRESH BUTTER, POPCORN, FRIED CHICKEN, WARM

Cluster 8

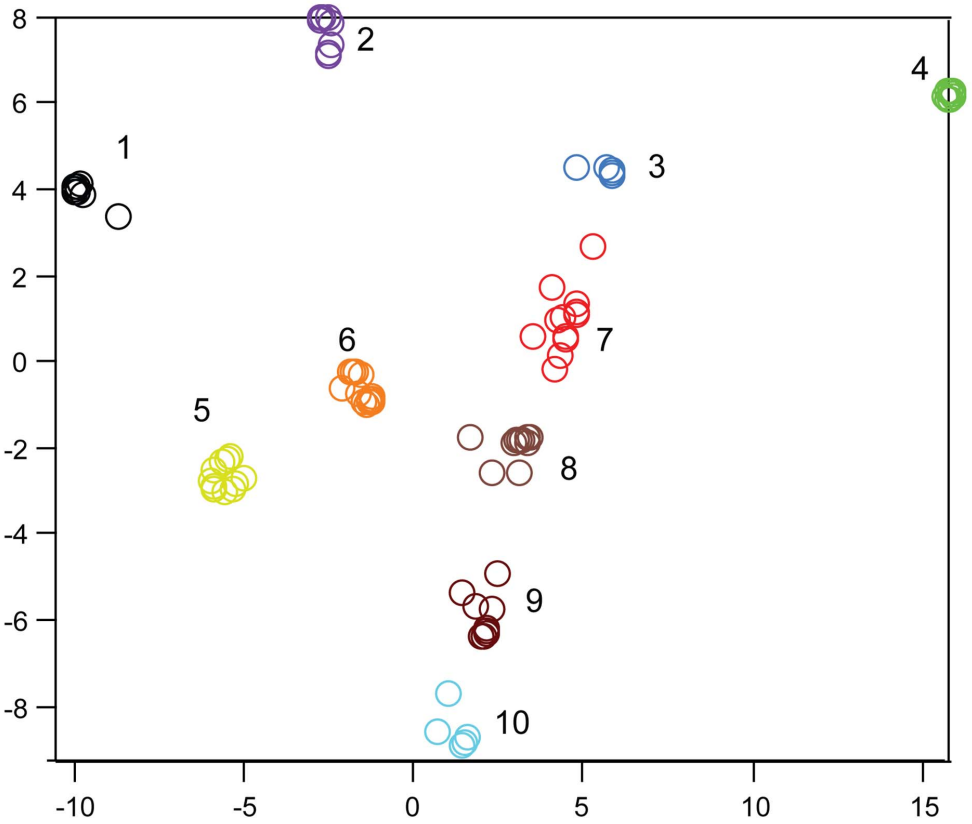
APPLE (FRUIT), SEASONING (FOR MEAT), GRAINY (AS GRAIN), YEASTY, SOUR MILK, FERMENTED (ROTTEN) , FRUIT, BEERY, WET WOOL, WET DOG, DIRTY LINEN, STALE, MOUSE, EGGY (FRESH EGGS), BURNT RUBBER, BITTER, SHARP, PUNGENT, ACID, SOUR, VINEGAR, SAUERKRAUT, URINE, CAT URINE, FISHY, KIPPERY (SMOKED FISH), SEMINAL, SPERM, SOOTY, MEATY (COOKED, GOOD), SOUPY, COOKED VEGETABLES, RANCID, SWEATY, HOUSEHOLD GAS, SULFIDIC, GARLIC, ONION, BLOOD, RAW MEAT, ANIMAL, SEWER, PUTRID, FOUL, DECAYED, FECAL (LIKE MANURE), CADAVEROUS (DEAD ANIMAL), SICKENING, HEAVY

Figure 7. Two-dimensional embedding of the descriptor-space, W. Results of stochastic neighbor embedding (see text) applied to the similarity matrix for **W**. Axis units are arbitrary, but preserve neighbor relations present in the higher dimensional space, **W**. Note that discrete clusters are clearly evident. Clusters were identified by eye, and descriptors composing each cluster are listed in the table below.
doi:10.1371/journal.pone.0073289.g007

Members of all clusters, as defined by their peak coordinate in the new 10 dimensional descriptor space, are given in Table 2.

Discussion

We have applied non-negative matrix factorization (NMF) to odor profiling data to derive a 10-dimensional descriptor space for human odor percepts. For the data set investigated, individual odor profiles are well-classified by their proximity to a single one of these dimensions, with all 10 dimensions being approximately equally expressed across the set of odors. This is consistent with the notion that olfactory space is high-dimensional [27], and not smoothly occupied [14,28]. More speculatively, the observation that odors tend to be confined to a single best dimension of the NMF basis (Figure 6, and Figure S4 in supporting information) suggests that a given olfactory percept can be described as an 'instance' of one of several fundamental qualities. Whether these proposed qualities are innate or the product of learning is, naturally, an important question, but one that is beyond the scope



Cluster 1

Aldehyde C-16, Allyl Caproate, iso-Amyl Acetate, Amyl Butyrate, Dimethyl Benzyl Carbinyl Butyrate , Ethyl Butyrate, Ethyl Propionate, Fructone, Methyl Anthranilate , Undecylenic Acid, gamma-Valerolactone

Cluster 2

Amyl Phenyl Acetate, Auralva, iso-Bornyl Acetate, Cashmeran, Dimethyl Phenyl Ethyl Carbinol, Hydroxy Citronellal, Indolene, beta-Ionone, alpha-Irone, Lyral, Methoxy-Naphthalene: 2-Methoxy, Naphthalene , Methyl Acetaldehyde Dimethyl Acetal, Musk Galaxolide, Musk T onalid, Phenyl Ethanol, Sandiff, Santalol

Cluster 3

dl-Camphor, l-Carvone, p-Cresyl Acetate, Eucalyptol, l-Menthol, Methyl Salicylate, Safrole

Cluster 4

Amyl Cinnamic Aldehyde Diethyl, Acetal, Citral, Citralva, Floralozone, Hexyl Cinnamic Aldehyde, Linalool, d-Limonene , Melonal, Myracaldehyde

Cluster 5

Anisole, 1-Butanol, m-Cresol, p-Cresol, p-Cresyl-iso-Butyrate, p-Cresyl Methyl Ether , Cyclohexanol , 2,5-Dimethyl, Pyrazine , Diola, Diphenyl Oxide, 1-Heptanol, 1-Hexanol, 3-Hexanol, Iodoform, Methyl Furoate, para-Methyl Quinoline , Nonyl Acetate, 1-Octanol, Phenyl Acetylene, Terpineol, Tetraquinone, Thymol, Toluene

Cluster 6

Adoxal, Andrane, iso-Butyl Quinoline Chlorothymol, Iso-Cyclocitral, Cyclotropal, Decahydro Naphthalene, Dibutyl Amine, Grisalva, Hexanal, Hydratropic Aldehyde Dimethyl Acetal, 2-Methyl-iso Borneol, Methyl iso-Nicotinate , Nootkatone, 1-Octen-3-OL, iso-Phorone, alpha-Pinene, iso-Propyl Quinoline, Propyl Sulfide, gamma-Undecalactone

Cluster 7

Abhexone, Acetophenone, Aldehyde C-18, Anethole, Benzaldehyde, Dihydro Pyrone, Caryophyllene (beta and gamma Isomers) , Celeriax, Cinnamic Aldehyde, Coumarin, Cuminic Aldehyde , Eugenol, Furfural, trans-1-Hexenal, ortho-Tolualdehyde, Vanillin

Cluster 8

ortho-Acetyl Pyridine, Cyclotene, 2,4-trans-trans-Decadienal , 2,3-Dimethyl Pyrazine , 2,5-Dimethyl Pyrrole , 2-Ethyl Pyrazine, Furfuryl Mercaptan, Guaiacol, Heptanal, Thienopyrimidine, Zingerone

Cluster 9

Butyl Sulfide, Cyclodithalfarol, 2-Cyclohexanedione , Diethyl Sulfide, Dimethyl Trisulfide, Hexyl Amine , Pyridine, Tetrahydro Thiophene, Thioglycolic Acid, Thiophene

Cluster 10

Amyl Valerate, Butanoic Acid, Hexanoic acid, Hexyl Amine, Indole, Maritima, Methyl Thiobutyrate, Pentanoic Acid, 4-Pentenoic Acid, Phenyl Acetic Acid , Propyl Butyrate, Skatole, Trimethyl Amine, iso-Valeraldehyde, iso-Valeric Acid

Figure 8. Two-dimensional embedding of the odorant-space, H. Results of stochastic neighbor embedding (see text) applied to the similarity matrix for **H**. As in figure 7, axis units are arbitrary, but preserve neighbor relationships observed in the full-dimensional space, **H**. Clusters were identified by eye, and odorants composing each cluster are listed in the table below.
doi:10.1371/journal.pone.0073289.g008

of this study. In addition, we note two important caveats of the present work. First, the fundamental odor qualities we propose are necessarily provisional, given the limitations of the Dravnieks data set in size and odorant diversity. Second, constraining perceptual judgments to a fixed and possibly limited lexicon (i.e. the 146 descriptors) may obscure the true complexity of odor space.

The perceptual dimensions obtained from NMF identify descriptors that are salient in several previous analyses of odor space [9–11,13,27], and commonly applied in ratings of odor quality. Moreover, these dimensions are consistent with a broad ecological perspective on olfactory function [29,30] which emphasizes the importance of chemosensation in coordinating

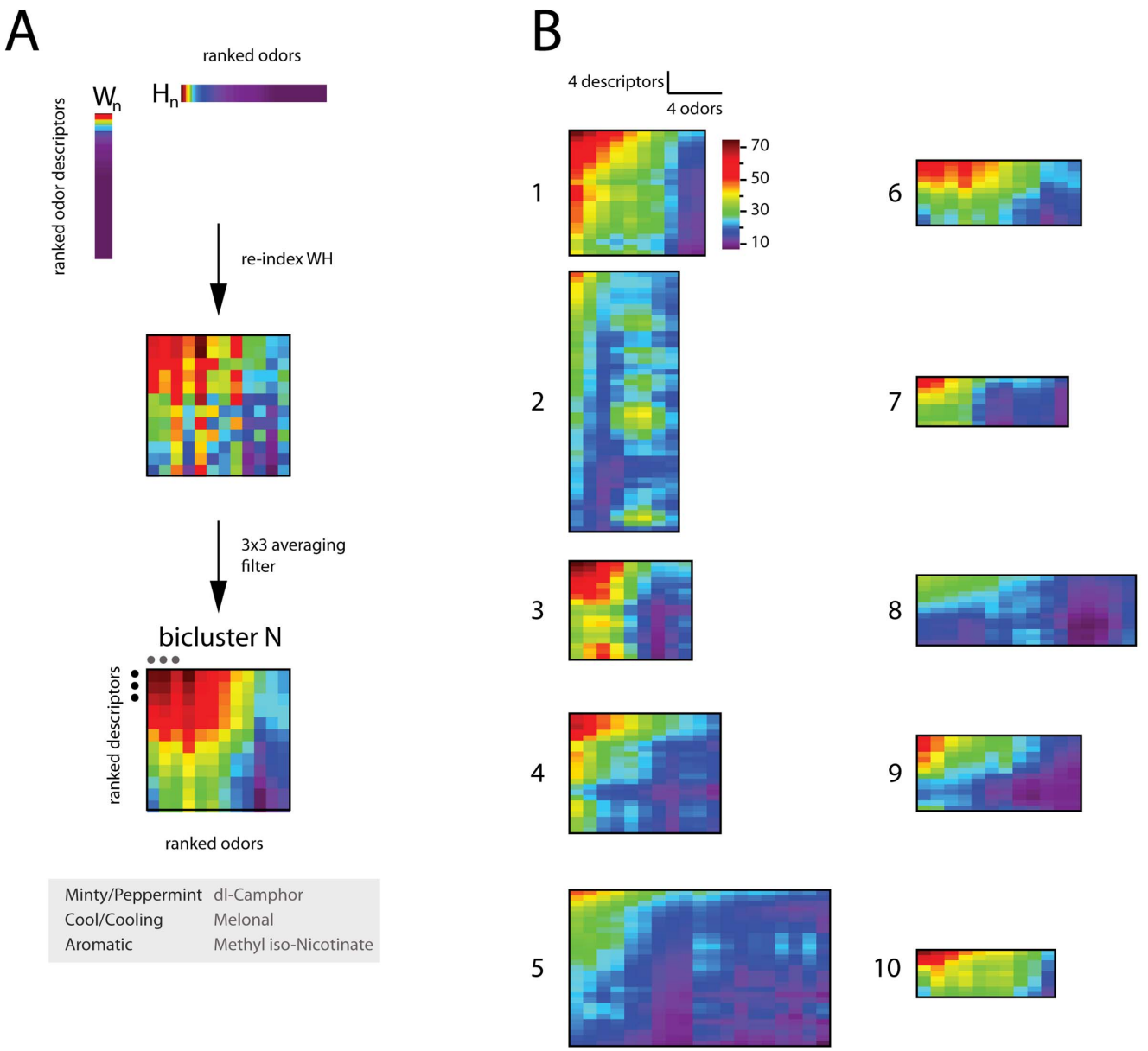


Figure 9. Co-clustering of descriptors and odors. (A) Overview of method used for defining a bicluster (see text for definition). A column k of **W** (descriptors), and the corresponding k^{th} row of **H** (odors) are rank ordered. The indices derived from the rank-ordering are used to re-order rows and columns of **WH** (accomplished by computing the outer product between the rank-ordered k^{th} column of **W** and rank-ordered k^{th} row of **H**), producing a submatrix with high correlation among both odors and descriptors. By the nature of the sorting procedure, these matrices – biclusters – will have their largest values in the upper-left corner. For purposes of visualization, biclusters were convolved with an averaging filter. (B) The 10 biclusters defined by NMF on odor perceptual data.
doi:10.1371/journal.pone.0073289.g009

approach, withdrawal, and the procurement of safe food. For example, we observe, as others have, dimensions corresponding to relative pleasantness ('fragrant' ($\mathbf{W}_1$), 'sickening' ($\mathbf{W}_4$)). In addition, most of the remaining dimensions identified appear to correspond to cues of potential palatability/nonpalatability: 'fruity, non-citrus' ($\mathbf{W}_2$), 'woody, resinous' ($\mathbf{W}_3$), 'chemical' ($\mathbf{W}_5$), 'sweet' ($\mathbf{W}_7$), and 'lemon' ($\mathbf{W}_{10}$). We hasten to note that the labels applied above are only an aid to intuition, as each perceptual basis is really a meta-descriptor consisting of linear combinations of more elementary descriptors. Moreover, it is possible that such linear combinations obscure interesting details about the exact positions of these more elementary descriptors. For a thorough treatment of this issue, one should consult Zarzo et al [9,31].

While several of these same principal qualities have been identified before, NMF describes a notably different representation of the space in which they reside. Specifically, NMF leads to a description of odor space defined by dimensions that apply categorically. By contrast, odors in PCA space are more diffusely distributed across dimensions. Moreover, odors in PCA space (as well as spaces derived from multidimensional scaling and factor analysis) tend to be smoothly distributed in subspaces that span multiple axes, though heirarchical applications of PCA have identified several quality-specific clusters [9]. Naturally, these

Table 2. List of compounds in every cluster identified from NMF.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1. Isoamylphenylacetate, 2. Aurantiol, 3. 6,7-dihydro-1,1,2,3,3-pentamethyl-4-(5H)indanone, 4. Indol-hydroxycitronellal, 5. beta-ionone (low concentration), 6. beta-ionone (high concentration), 7. N′-[(E)-3-(5-methoxy-2,3-dihydro-1,4-benzodioxin-7-yl) prop-2-enoyl]-2,3-dihydro-1,4-benzodioxine-3-carbohydrazide, 8. hydroxyisohexyl 3-cyclohexene carboxaldehyde, 9. 2-methoxynaphthalene, 10. Diethoxymethane, 11. Galaxolide, 12. ethylenebrassylate, 13. Phenylethyl Alcohol (low concentration) 14. Phenylethyl Alcohol (high concentration)	15. Cedrene epoxide, 16. bornyl acetate, 17. 8-sec-Butylquinoline, 18. 2,4,6-trimethylcyclohex-3-ene-1-carbaldehyde, 19. decalin, 20. dibutylamine, 21. Synthetic amber, 22. 1,1-Dimethoxy-2-phenylpropane, 23. Methyl isonicotinate, 24. Nootkatone, 25. 1-octen-3-ol, 26. isophorone (low concentration), 27. isophorone (high concentration), 28. Isopropyl quinolone, 29. Argeol, 30. Gamma-undecalactone, 31. 10-undecenoic acid	32. ethylmethylphenylglycidate (low concentration) 33. ethylmethylphenylglycidate (high concentration) 34. allylcaproate, 35. isoamyl acetate, 36. n-amyl butyrate, 37. Dmbc butyrate, 38. ethyl butyrate, 39. ethyl propionate, 40. Fructone, 41. methylanthranilate, 42. Pentylvalerate	43. Butyric Acid 44. hexanoic acid 45. indole 46. methylthiolbutyrate 47. n-pentanoic acid 48. 4-pentenoic acid 49. 50. . phenylacetic acid 51. Propyl butyrate 52. Skatole (3-Methyl-1H-indole) 53. Isovalerylaldehyde 54. isovaleric acid	55. Acetophenone 56. Anisole 57. 1-Butanol 58. 4-cresol 59. p-Tolylisobutyrate 60. 4-methyl anisole 61. cyclohexanol 62. 2,5-dimethylpyrazine 63. methyl hexyl ether 64. 1-hexanol 65. 3-hexanol 66. iodoform 67. methyl furan-3-carboxylate 68. 4-methylquinoline 69. phenylacetylene 70. alpha-terpineol 71. 6-methyl-1,2,3,4-tetrahydroquinoline 72. Thymol 73. Toluene 74. 3-Methyl-1H-indole

Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
1. Anethole 2. 8-sec-Butylquinoline 3. carvone 4. caryophyllene 5. 4-cresyl acetate 6. eucalyptol 7. Eugenol 8. Menthol 9. methyl salicylate 10. Safrole	11. Abhexon 12. Gamma-nonalactone 13. Benzaldehyde 14. 3,4-dihydrocoumarin 15. 3-Propylidene phthalide 16. cinnamic aldehyde 17. coumarin 18. cyclotene 19. Furaldehyde 20. 2-hexenal 21. 2-methylbenzaldehyde 22. gamma-valerolactone	23. Vanillin 24. 2-acetylpyridine 25. 2,4-decadienal 26. Pyrazine 27. methyl hexyl ether 28. 2,5-dimethylpyrrole 29. Ethylpyrazine 30. Ethylpyrazine 31. Heptanal 32. n-hexanal 33. 1-Octanol 34. 2-methyl-5,7-dihydrothieno[3,4-d]pyrimidine	35. Zingherone 36. dibutyl sulfide 37. Chlorothymol 38. 2-Mercaptopropanone 39. 1,2-cyclohexanedione 40. diethyl sulfide 41. dimethyltrisulfide 42. furfurylmercaptan 43. Guaiacol 44. Hexylamine 45. Hexylamine 46. AC1L18DS	47. polythiophene 48. Adoxal 49. Amyl cinnamic aldehyde diethyl acetal 50. Citral 51. Geranonitrile 52. Cuminaldehyde 53. 4-Methyl-2-(1-phenylethyl)-1,3-dioxolan 54. 2-Methyl-4-phenylbutan-2-ol 55. phenyl ether 56. Floralozone 57. Heptanol 58. hexylcinnamic aldehyde 59. hydroxycitronellal 60. linalool 61. limonene 62. Melonal 63. Myrac aldehyde 64. n-Nonyl acetate

doi:10.1371/journal.pone.0073289.t002

differences in the representation of odor space are a consequence of the different constraints applied when obtaining a basis from PCA vs NMF. Whereas PCA basis vectors are chosen to be orthogonal, and allow any linear combination of variables, NMF basis vectors are constrained to be non-negative, allowing only positive combinations of variables. It is worth noting, however, that the NMF basis set is still approximately orthogonal (mean pairwise angle between different basis vectors is 72.9 degrees (Fig. 5)). Moreover, NMF is capturing structure in the data beyond simple first-order statistics, as applying NMF to scrambled versions of **A** fails to produce sparse and perceptually meaningful basis vectors (Fig. 3).

Intuitively, the non-negativity constraint produces NMF basis vectors defined by subsets of descriptors that are weighted and co-applied in particularly informative combinations, defining dimensions that range from absence to presence of a positive quantity. This contrasts to basis vectors and dimensions derived from other techniques, which extend from one quality to that quality's presumed opposite. Such dimensions have intuitive interpretations in some cases, for example, the experimentally supported 'pleasantness' dimension corresponding to principal component 1 (PC1), which ranges from 'fragrant' to 'sickening'. Interestingly, constraining the NMF subspace to 2 shows that most odors fall homogeneously along a continuum reminiscent of the first principal component (Fig. S5 in supporting information). However, second and higher order PCs become progressively more difficult to interpret, spanning such qualities as 'woody, resinous' → 'minty, peppermint' (PC2), and 'floral' → 'spicy' (PC3). Whether odor percepts are more accurately represented as residing in dimensions that span oppositely valenced qualities, or dimensions that represent only a single quality will depend on whether there is systematic opponency in peripheral or central odor representations.

It may be possible to observe physiological properties of odor representations indicative of one kind of representation vs. another. If the underlying perceptual dimensions of odor space are categorical, one would expect relative similarity between odor representations for odors occupying the same putative perceptual dimension. Similarly, one would expect abrupt, state-like transitions in neural representations of slowly morphing binary mixture

stimuli whose component odors nominally 'belong' to different perceptual dimensions. Consistent with these criteria, a recent study has shown discrete transitions in the ensemble activity of the zebrafish olfactory bulb during such odor morphs ([14], but see [32]).

Our study has some limitations that should be noted. Chief among these is the small size of the odor profiling data set used relative to the much larger set of possible odors, which may limit the generality of our findings. In future studies, it will be necessary to extend the NMF framework to larger sets of odors than the 144 investigated presently, such that a more complete and representative sample from odor space is obtained. Another limitation pertains to the 'subjective' nature of odor profiling data. While profiles are quantitative in the sense that they are stable and reliable across raters [33], it is clearly important to corroborate profiling-derived estimates of the intrinsic dimensionality of odor space, as well as proposals for how this space is structured, with psychophysical tests of discriminability [34]. It would be interesting, for example, to test whether the approximately orthogonal axes we observe are recapitulated in data derived from tests of pairwise discriminability. Finally, our analysis cannot distinguish between perceptual vs. cognitive influences on the organization of human odor space. One possibility is that the coarse division of odor-space into quality-specific axes reflects the existence of fixed points or attractors [14,28] that guide odor processing dynamics; similarly, there may exist a set of especially stable, prototypical glomerular maps that serve a related functional role. Another possibility is that early olfactory processing only resolves odor quality to a degree sufficient to rank relative pleasantness, with further parsing of this percept into discrete categories occurring through mechanisms involving learning and context.

In summary, we have shown that olfactory perceptual space can be spanned by a set of near-orthogonal axes that each represent a single, positive-valued odor quality. Odors cluster predominantly along these axes, motivating the interpretation that odor space is organized by a relatively large number of independent qualities that apply categorically. Independently of whether our description of odor space identifies innate or 'natural' axes determined by receptor specificities, it provides a compact description of salient, near-orthogonal odor qualities, as well as a principled means for identifying and rating odor quality. Finally, our study has identified perceptual clusters that may help elucidate a structure-percept mapping.

Supporting Information

Figure S1 Comparison of PCA and NMF. Plot of cumulative fraction of variance explained for PCA and NMF, for various choices of subspace size.
(TIF)

Figure S2 Cophenetic correlation vs. choice of subspace size. Cophenetic correlation obtained for NMF representations of increasing subspace size. Procedure is defined in the text.
(TIF)

Figure S3 NMF-derived approximations of odor profiles (2*A*) Image of original data (left) and NMF-derived approximations **WH** for subspaces of 5 (center) and 10 (right). Same range and color scale for all images. Because the data matrix contains many small and zero-valued entries among sparse, large-valued entries, the colorscale has been gamma-transformed ($2\gamma = 1.8$) for better visualization and comparisons. Arrowheads indicate columns shown in more detail in panel below. (2*B*) Detailed representation of columns 70–74 of original data matrix **A** (black traces) and NMF approximations to those columns by **WH** for a 10 dimensional subspace (red traces).
(TIF)

Figure S4 Representations of odorants distributed in perceptual space. (2*A*) Star plots of odorants (columns of **H**). Odorant weight vectors are wrapped on $[0,22\pi]$ for visualization purposes. Left: three example odorants and their distributions in perceptual space, showing that a given odorant tends to occupy a single one of ten perceptual dimensions, to the exclusion of others. Right: star plot of all 144 odorants in the perceptual space. Colors indicate odors with a common peak coordinate in the 10-D descriptor space. (2*B*) Visualizations of various three-dimensional subspaces of the matrix **H**, as in Figure 6.
(TIF)

Figure S5 NMF reveals hedonic valence of odors. For a choice of subspace 2, NMF reveals the hedonic valence of odors. (2*A*) left column: basis vectors returned for NMF with subspace 2. right column: normalized amplitudes and descriptors for leading values of rank-ordered basis vectors. (2*B*) Plot of all 144 odors in the space spanned by $\mathbf{W}_1$, $\mathbf{W}_2$ (analogous to plots shown in Fig. 6 in the main manuscript). Colors indicate classification based on largest coordinate (black, $\mathbf{W}_1$, gray, $\mathbf{W}_2$), showing coarse categorization into good-vs-bad smelling odors.
(TIF)

Acknowledgments

We thank Dr. Alexei Koulakov for kindly providing an electronic copy of the Dravnieks odor database, and Dr. Nathan Urban for initial help on the project. We thank Drs. Rick Gerkin, and Krishnan Padmanabhan for helpful feedback on an earlier manuscript.

Author Contributions

Conceived and designed the experiments: JBC AR CSC. Performed the experiments: JBC AR CSC. Analyzed the data: JBC AR CSC. Wrote the paper: JBC AR CSC.

References

1. Arzi A, Sobel N (2011) Olfactory perception as a compass for olfactory neural maps. Trends Cogn Sci (Regul Ed) 15: 537–545.
2. Lotto RB, Purves D (2002) A rationale for the structure of color space. Trends Neurosci 25: 84–88.
3. Lennie P, D'Zmura M (1988) Mechanisms of color vision. Crit Rev Neurobiol 3: 333–400.
4. Henning H (1916) Der Geruch. Leipzig.
5. Amoore JE (1974) Evidence for the chemical olfactory code in man. Ann N Y Acad Sci 237: 137–143.
6. Amoore JE (1967) Specific anosmia: a clue to the olfactory code. Nature 214: 1095–1098.
7. Schiffman SS (1974) Physicochemical correlates of olfactory quality. Science 185: 112–117.
8. Schiffman SS (1974) Contributions to the physicochemical dimensions of odor: a psychophysical approach. Ann N Y Acad Sci 237: 164–183.
9. Zarzo M, Stanton DT (2006) Identification of latent variables in a semantic odor profile database using principal component analysis. Chem Senses 31: 713–724.
10. Khan RM, Luk CH, Flinker A, Aggarwal A, Lapid H, et al. (2007) Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. J Neurosci 27: 10015–10023.
11. Koulakov AA, Kolterman BE, Enikolopov AG, Rinberg D (2011) In search of the structure of human olfactory space. Front Syst Neurosci 5: 65.
12. Lapid H, Shushan S, Plotkin A, Voet H, Roth Y, et al. (2011) Neural activity at the human olfactory epithelium reflects olfactory perception. Nat Neurosci 14: 1455–1461.
13. Haddad R, Weiss T, Khan R, Nadler B, Mandairon N, et al. (2010) Global features of neural activity in the olfactory system form a parallel code that predicts olfactory behavior and perception. J Neurosci 30: 9017–9026.
14. Niessing J, Friedrich RW (2010) Olfactory pattern classification by discrete neuronal network states. Nature 465: 47–52.

15. Laing D (1991) The Human Sense of Smell. Philadelphia: Springer.
16. Paatero P, Tapper U (1994) Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5: 111–126.
17. Paatero P (1997) Least squares formulation of robust non-negative factor analysis. Chemometrics Int Lab Sys 37: 23–35.
18. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788–791.
19. Berry M, Browne M, Langville A, Pauca V, Plemmons R (2007) Algorithms and Applications for Approximate Nonnegative Matrix Factorization. Computational Statistics & Data Analysis 52: 155–173.
20. Dravnieks A (1985) Atlas of Odor Character Profiles. Philadelphia: ASTM Data Series ed. ASTM Committee E-18 on Sensory Evaluation of Materials and Products. Section E-18.04.12 on Odor Profiling.
21. Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. SIGIR '03, pp. 267–273.
22. Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering – a resampling-based method for class discovery and visualization of gene expression microarray data. In: Machine Learning, Functional Genomics Special Issue. 91–118.
23. Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA 101: 4164–4169.
24. van der Maaten L, Hinton G (2008) Visualizing high-dimensional data using t-SNE. J Mach Learn Res 9: 2579–2605.
25. Hinton G, Roweis S (2002) Stochastic neighbor embedding. Advances in Neural Informational Processing Systems 15: 833–840.
26. Kim PM, Tidor B (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 13: 1706–1718.
27. Mamlouk A, Chee-Ruiter C, Hofmann U, Bower JM (2003) Quantifying olfactory perception: mapping olfactory perception space by using multidimensional scaling and self-organizing maps. Neurocomputing 52: 591–597.
28. Laurent G (2002) Olfactory network dynamics and the coding of multidimensional signals. Nat Rev Neurosci 3: 884–895.
29. Bargmann CI (2006) Comparative chemosensation from receptors to ecology. Nature 444: 295–301.
30. Gottfried JA (2009) Function follows form: ecological constraints on odor codes and olfactory percepts. Curr Opin Neurobiol 19: 422–429.
31. Zarzo M, Stanton DT (2009) Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps. Atten Percept Psychophys 71: 225–247.
32. Khan AG, Thattai M, Bhalla US (2008) Odor representations in the rat olfactory bulb change smoothly with morphing stimuli. Neuron 57: 571–585.
33. Dravnieks A (1982) Odor quality: semantically generated multidimensional profiles are stable. Science 218: 799–801.
34. Wise PM, Olsson MJ, Cain WS (2000) Quantification of odor quality. Chem Senses 25: 429–443.