SUPPLEMENTARY INFORMATION
(Lu, et al., "Chamber Evaluation…")

The use of an upstream separation column limits the number of vapors presented to the sensor array simultaneously. This, in turn, reduces the problem of vapor recognition to a series of simpler analyses applied to each retention time window individually. For this study, all of the vapors could be chromatographically resolved in ~ 6 minutes by adjusting the pressure tuning and temperature programming parameters of the separation module. Therefore the confirmation of vapor identity reduces to one of assessing the fidelity of the measured response pattern to that in the calibration library. That is, the goodness of fit of an unknown sample to its calibration set must be tested in order to conclude with confidence that the resolved peaks observed with the array are indeed attributable to the vapor expected to elute at the given retention time.

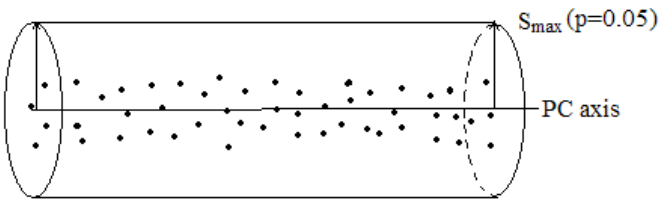

Figure S1. Graphical representation of the threshold distance defining vapor identity. A single principal component (PC) model is assumed for illustration and the threshold is established at a predefined significance level (e.g., 0.05) on the basis of response patterns determined during calibration (i.e., training). The central solid line is the PC axis, which is surrounded by calibration samples denoted by points. The contour of the constant distance from the PC axis defined by S_{max} (see text) is in the shape of a cylinder.

The approach can be illustrated by a one-principal-component classification model as shown in Fig. S1, constructed from a calibration data set using principal components

analysis.[1] The threshold of maximum distance from a subsequent sample to the centroid of the calibration set in multi-dimensional space can be computed at a certain significance level from the Mahalanobis distance (see below) after projection of the sample vector onto the principal component (PC) axes. If a new sample falls within the boundary established by this threshold, the sample is assigned with a known confidence level the identity of the vapor corresponding to that model. Otherwise it is rejected. We start with a calibration data set for a target vapor that can be expressed in matrix form as follows:

$$\underset{n\times m}{X} = \begin{bmatrix} x_{11} & x_{12} & . & x_{1m} \\ x_{21} & x_{22} & . & x_{2m} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{nm} \end{bmatrix} \tag{S1}$$

where x_{nm} is the response of the m^{th} sensor to the n^{th} calibration sample of the target vapor. PC modeling of this data set separates the data into a model matrix and a residual matrix. The latter is used to calculate a residual standard deviation (RSD, geometrically the Mahalanobis distance when dealing with vector quantities), which is used for determining the range of allowable variability in assessing the response pattern fidelity of subsequent test samples.

Cross-validation was used to obtain models with maximum predictive ability. A common cross-validation technique, which is well suited for small data sets, is the leave-one-out procedure in which one row at-a-time is deleted from the data matrix *X* and the PC(s) that best account for the variance in the remaining data are calculated. The resulting *n* PC models (the number of models is the same as the number of calibration samples) are used to predict the response patterns for the left-out samples (models may contain one or more PCs). A metric called the predicted residual error sum of squares (PRESS) measures the prediction

error of each PC model. The deviation between the actual and predicted values is used to estimate an overall prediction error. The model providing the minimum prediction error is finally calculated with all samples included. By this procedure, all samples are utilized both for calculating and for validating the model.

To assess the fidelity of a subsequent (unknown) sample to the established model, the following equation is used:

$$e_i(v) = X_i' - \bar{X}'(v) - \sum_{a=1}^{A} t_{ia} p_a' \qquad \text{(S2)}$$

where $e_i(v)$ is the residual (vector) of the unknown sample after being fitted to the model for vapor v; X_i' is the transpose of the unknown sample response vector of m elements (m is the number of sensors in the array); $\bar{X}'(v)$ is the transpose of the mean vector of the vapor v; A is the dimension of the model (i.e., the optimal number of principal components) determined by cross-validation; t_{ia} is the "score" of sample i on the principal component a, and p_a is the corresponding loading vector (weighting of all variables on principal component a). The "score" is the weight of each sample on the each PC, where a sample is the collection of responses on all sensors to a given concentration of a given vapor. The loading vector determines which sensors (referred to generically as variables) provide the greatest influence, or weight, on the magnitude and direction (in multi-dimensional space) of the PC.

The Mahalanobis distance, S_i, of an unknown sample i to the model under consideration, can be calculated via:

$$S_i = \sqrt{e_i'(v)\, e_i(v)/(m - A)} \qquad \text{(S3)}$$

where m is the number of sensors, $e_i(v)$ is the residual vector of the unknown sample after fitting to the model for vapor v and $e_i'(v)$ is the transposed vector of $e_i(v)$. The division by $(m-A)$ provides a distance measure that is independent of the number of variables and corrected for the loss of freedom due to the fitting of A principal components.

The S_i values for all n samples in the calibration set can also be calculated using Eq. S3 and collected in a distance matrix S of dimension $n \times 1$. The mean *RSD,* S_v, of the model for vapor v is defined by the following equation:

$$S_v = \sqrt{S'S/(n-A-1)} \tag{S4}$$

where S' is the transpose of distance matrix S. The division by $(n-A-1)$ gives a scale that is also independent of the number of samples and corrected for the loss in degrees of freedom due to the mean-centering and fitting of A principal components in Eq. S2.

Comparison of the RSD for an unknown sample (calculated by Eq. S3) to the mean RSD for the vapor v (calculated by Eq. S4) gives a direct measure of its similarity to the model. An F statistic was used for the comparison of S_i^2 and S_v^2. The degrees of freedom used to obtain the critical F-value are $(m-A)$ and $(m-A)(n-A-1)$, respectively, for S_i^2 and S_v^2. The upper limit of the RSD for any sample, S_{max}, can thus be calculated:

$$S_{\max}^2 = S_v^2 F_{crit} \tag{S5}$$

F_{crit} is usually determined at a significance level of 1% or 5% (i.e., p = 0.01 or 0.05). The latter level is a more stringent boundary, allowing 1 out of 20 samples that fit the model, on average, to be mis-identified as an outlier. In the former case, only 1 out of 100 samples that fit is rejected. If S_i of an unknown response pattern is less than or equal to S_{max}, then this

sample is assigned the identity of the vapor described by that particular model. If $S_i > S_{max}$, then the sample is rejected indicating that the response pattern has become distorted enough to suggest that the target vapor is contaminated with another vapor or that the response pattern is not due to the target vapor.

References

1. B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part B*, Elsevier, Amsterdam, 1998, pp. 228–232.