Electronic Supplementary Information for

Quantum of selectivity testing: detection of isomers and close homologs by AZO based E-nose without *a prior* training

Boris V. Goikhman,[a] Fedor S. Fedorov,*[a] Nikolay P. Simonenko,[b] Tatiana L. Simonenko,[b] Nikita A. Fisenko,[b,c] Tatiana S. Dubinina,[a] George Ovchinnikov,[a] Anna V. Lantsberg,[d] Alexey Lipatov,[e] Elizaveta P. Simonenko,[b] Albert G. Nasibulin*[a,f]

[a] Skolkovo Institute of Science and Technology, 3 Nobel Str., Moscow, 121205, Russian Federation
[b] Kurnakov Institute of General and Inorganic Chemistry of the Russian Academy of Sciences, 31 Leninsky pr., Moscow 119991, Russian Federation
[c] D. Mendeleev University of Chemical Technology of Russia, 9 Miusskaya sq., Moscow 125047, Russian Federation
[d] Bauman Moscow State Technical University, 5/1 Baumanskaya 2-ya Str., Moscow, 105005, Russian Federation
[e] South Dakota School of Mines and Technology, 501 E. Saint Joseph St., Rapid City, SD 57701, USA
[f] Aalto University, Kemistintie 1, P.O. Box 16100, 00076 Aalto, Finland

Corresponding authors: f.fedorov@skoltech.ru, a.nasibulin@skoltech.ru

Note 1. Chip preparation.

The preparation of employed chips included several steps (**Figure S1**). Briefly, the AZO material was synthesized by programmed co-precipitation; the sediment was separated by centrifugation and further dried. The AZO powder was further annealed at 350 °C for 1 h. A droplet, 1 μm, of dispersion made with AZO powder, ethanol, and water was placed at the multielectroded chip. The prepared chip was heated to 350 °C (1 h) followed by stabilization at 300 °C for 24 h in air. The image of the prepared chip is presented in **Figure S1b**.

Note 2. Setup for testing sensor performance.

Analyte concentration in the mixture with air was adjusted using a home-made gas mixing setup with diffusion vials DYNACAL® (VICI Metronics Inc., USA). The gas mixing setup is presented in **Figure S2**. It includes a source of pure dry air (i), mass flow controllers (ii), a container with a diffusion vial (iii) in a constant air flow, switching valve (iv) to direct the constant flows from the line with pure dry air or line with air containing analyte vapors either to a chamber (v) with a chip or to exhaust. The switching valve was controlled using home-made electronics based on relays enabling automatic switching of the flows using @LabView software. The pump (vi) connected via the mass flow meter was used to enable pressure difference to drive the mass flow meter to adjust the concentration of an analyte. The electric board controlled the chip temperature and was used for data acquisition powered by a power source (vii). All setup, including valve, board, and mass flow meters was controlled by PC (viii) employing @LabView software.

Figure S3 shows the concentration range assessed for each analyte by gravimetric analysis.

Note 3. DecisionTreeRegressor parameters.

We've used DecisionTreeRegressor from the Scikit-learn package with the following parameters:[s1]

(1) criterion is set to 'squared_error' — the function to measure the quality of a split is set to the mean squared error;

(2) splitter is set to 'best'. It is the strategy used to choose the split at each node, and we use best split as an alternative to random one;

(3) max_depth is set to None, which is the maximum depth of the tree. The nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples which are set to 2;

(4) min_samples_leaf is set to 1. It is the minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches;

(5) min_weight_fraction_leaf is set to 0. It is the minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Samples have equal weight when sample_weight is not provided;

(6) max_features is set to None, which is the number of features to consider when looking for the best split. None means max_features is set to n_features;

(7) random_state is set to 0. It means that we've fixed random seed to 0;

(8) max_leaf_nodes is set to None. The algorithm grows a tree with max_leaf_nodes in the best-first fashion. Best nodes are defined as relative reduction in impurity (which is computed

as described in ref. s1). Selected parameter value None means an unlimited number of leaf nodes;

(9) min_impurity_decrease is set to 0. The node will be split if this split induces a decrease of the impurity greater than or equal to this value;

(10) ccp_alpha is set to 0. It is a complexity parameter used for Minimal Cost-Complexity Pruning. We selected a value of zero meaning that in our case no pruning is performed.

Note 4. Characterization of synthesized material by SEM and TEM.

Figure S4 shows the morphology and structure of aluminum-doped zinc oxide (AZO) assessed by SEM and TEM. **Figure S4a** shows AZO distributed on the surface in a form of agglomerated flakes. TEM BF image confirms the porous structure of AZO; the corresponding DF image suggests the crystal size to be about 20 nm (**Figure S4b,c**).

Note 5. ToF-SIMS spectrum.

Representative ToF-SIMS spectrum is given in **Figure S5**. The observed signal of ions, i.e. m/z (Th), suggest the following species: Ga^+ = 69; $^{71}Ga^+$ = 71; $^{64}Zn^+$ = 64; $^{66}Zn^+$ = 66; $^{68}Zn^+$ = 68, Al^+= 27, $^{27}Si^+$=27, $^{28}Si^+$=28, K^+=39, Na^+=23, where Ga are ions used in FIB gun, and Si signal originates from the surface. K and Na ions are impurities. There are non-marked signals to be related to organic compounds present on the surface.

Note 6. Machine learning with Mol2vec "fingerprints".

We prepared other chip in a similar way to test alcohols and other volatile compounds. Employing this chip, we measured responses towards methanol, ethanol, 1-propanol, 2-propanol, 1-butanol, isobutanol, tert-butanol, isoamyl alcohol, methyl t-butyl ether, 2-methoxyethanol, acetone (ca. 0.9 ppm in the mixture with air), amyl alcohol (ca. 0.65 ppm, in the mixture with air) at 300 °C. The measurement protocols were similar. The signals from 11 sensors at the multisensor chip were recorded.

The sensor responses of the tested multisensor chip are presented in **Figure S8**. The sensor responses magnitude correlates well with the trends discussed in the manuscript. We also notice a distinct pattern of responses depending on the analyte type. We employed a Mol2vec database of chemical "fingerprints" that are embedding of the molecular structure obtained with deep learning methods[s2,s3] on pair with support vector regression (SVR)[s4] machine learning algorithm to evaluate the possibility of selective determination of analyte without prior training of e-nose. The sensor responses are projected in 2D PCA (**Figure S8m**), while Mol2vec molecular "fingerprints" reduced to 2D PCA are given in **Figure S8n**. From a data-science perspective, we are using small datasets, hence we expect some variation of quality with the addition of new data. The utilized Mol2vec "fingerprints" are usually considered superior to older PubChem "fingerprints", but on the smaller dataset, there is no better gain in metrics. SVR acknowledges the presence of non-linearity in the data by solving optimization problems not precisely, but up to a certain threshold. We use Scikit-learns[s1] realization of SVR with the following parameters:

kernel is set to 'rbf' — meaning using radial basis function as kernel functions similar to the ones used for SVM, gamma is set to 'scale' meaning it uses 1 / (number of features * variance(X)) as the value of gamma for radial basis functions, tol is set to 0.001 — stopping

tolerance for optimization procedure, C is set to 1.0 — the penalty is a squared l2 penalty, epsilon is set to 0.1 — which is the epsilon in the epsilon-SVR model: it specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value

Applying the SVR under a leave-one-out cross-validation strategy, we obtained an R^2 value of 0.55. The results are projected to a 1D PCA plot (**Figure S8o**), characterized by explained variance = 0.577. We found a good prediction for amyl alcohol, 1-propanol, ethanol, tert-butanol, isobutanol.

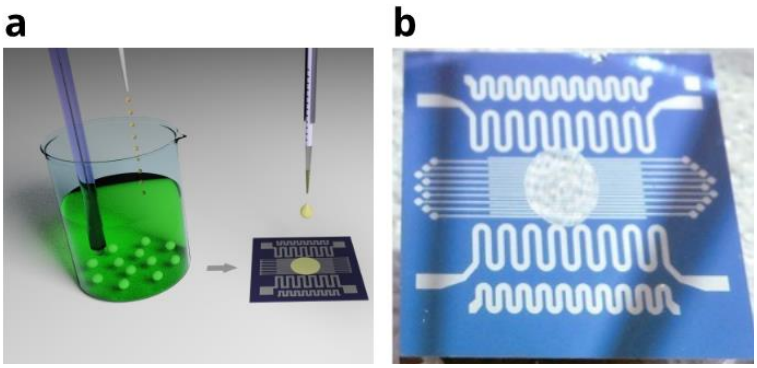

Figure S1. Schematic representation of the chip preparation. (a) Programmed co-precipitation and drop-casting, (b) image of the chip with AZO layer over coplanar Pt/Ti strip electrodes.

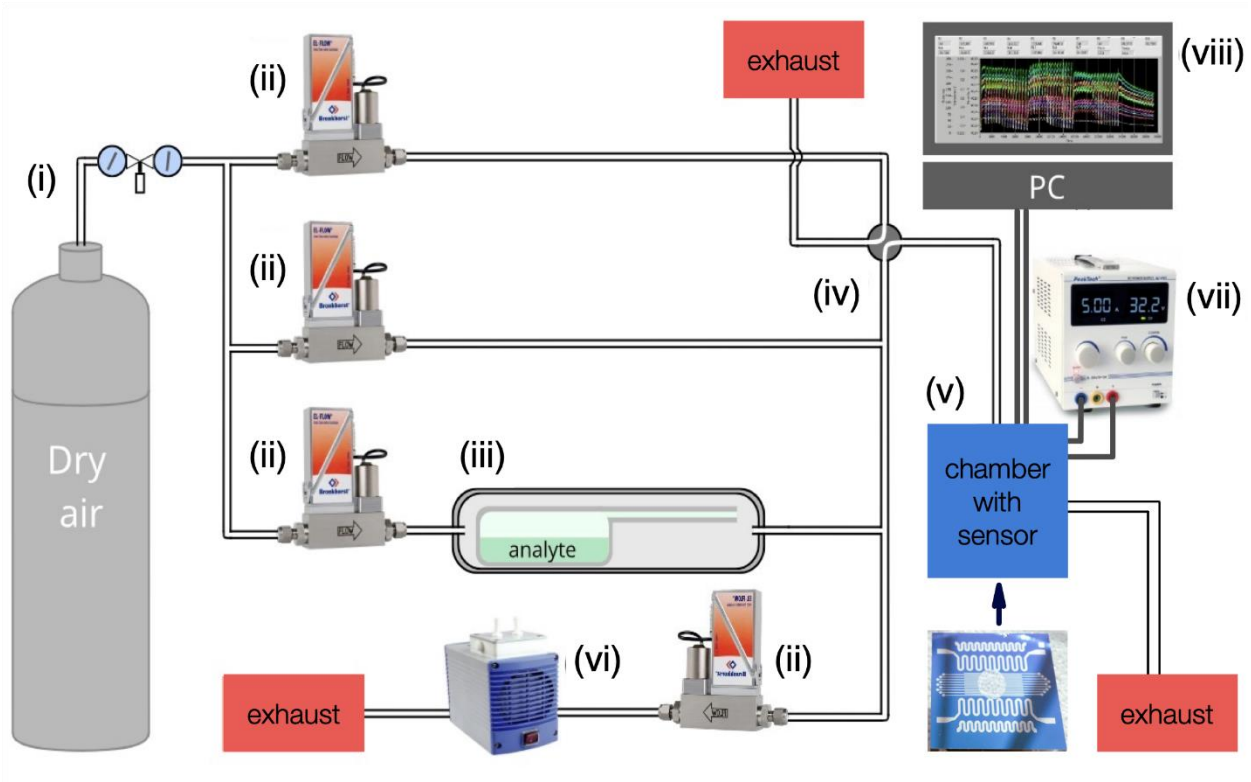

Figure S2. Scheme of experimental setup: (i) pure dry air source; (ii) MFCs; (iii) container with a diffusion vial; (iv) switching valve; (v) chamber with e-nose; (vi) air pump; (vii) power source, (viii) PC.

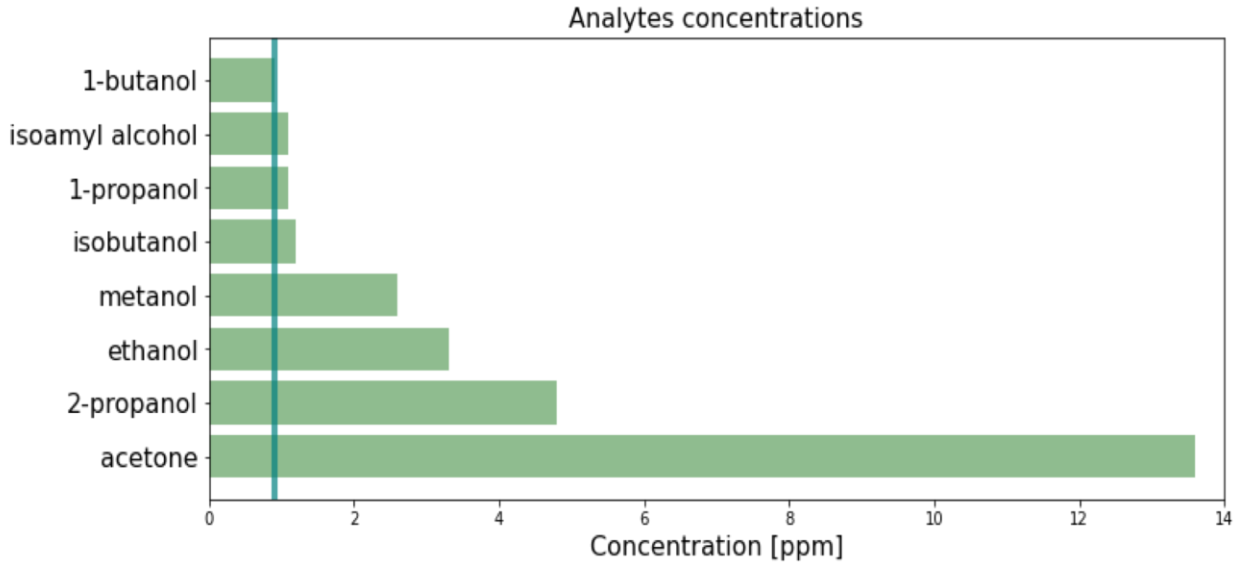

Figure S3. Concentrations of analytes provided by the diffusion vial B(A).

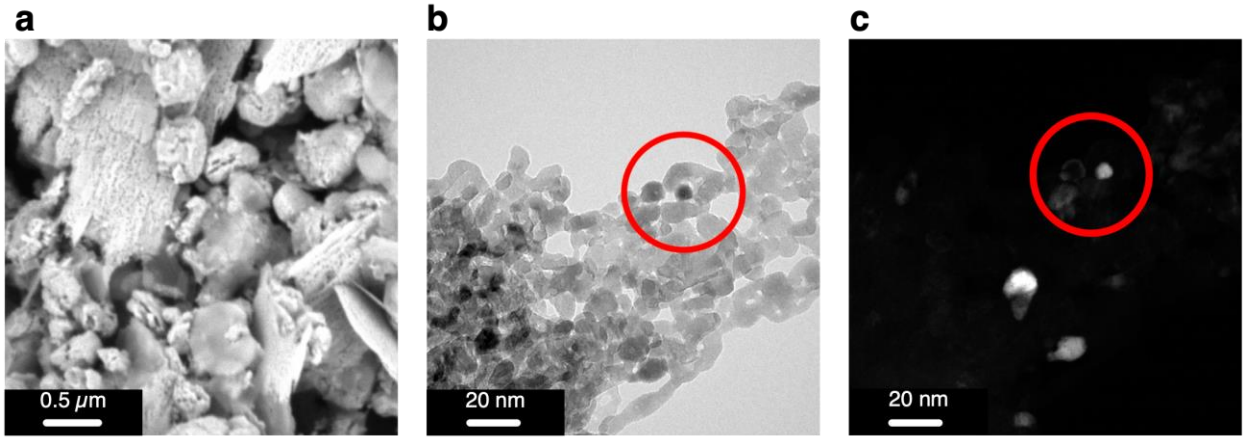

Figure S4. (a) SEM high-resolution image of AZO layer; (b) BF TEM image and (c) DF TEM image of the selected area of AZO petal.

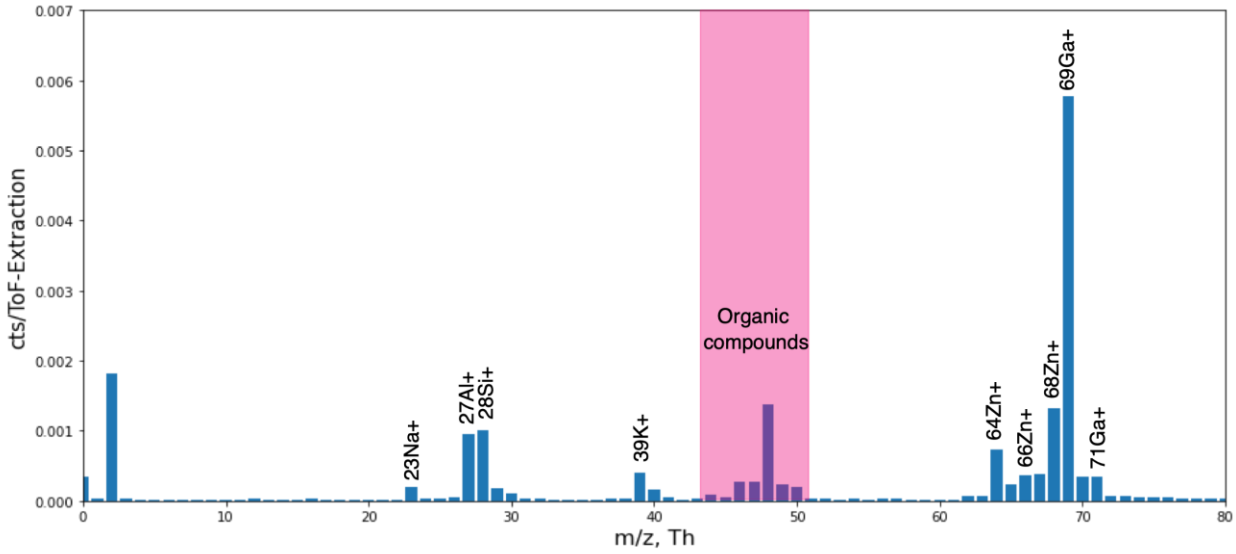

Figure S5. Representative positive mass spectrum of the AZO layer.

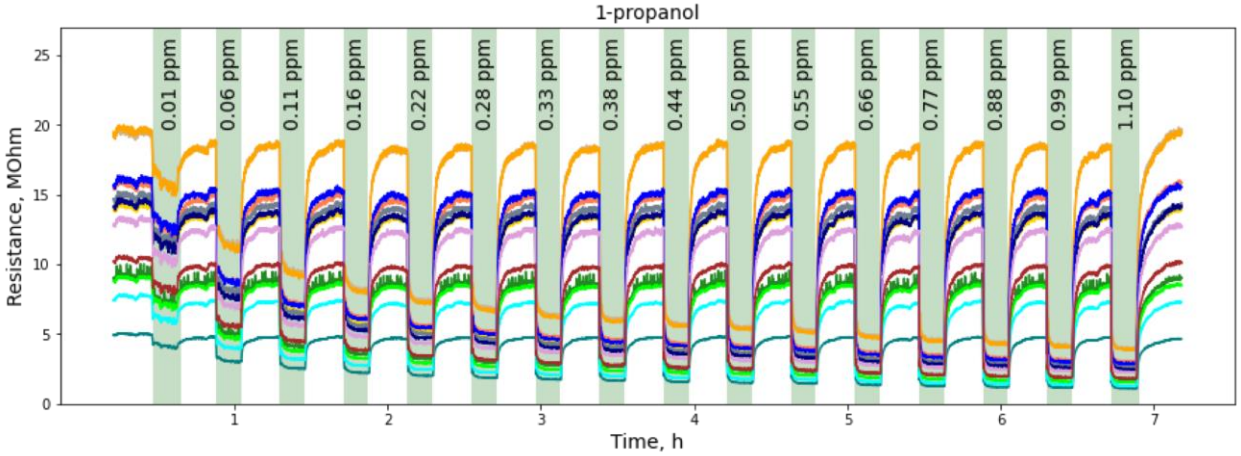

Figure S6. Resistance transients of all sensors at different concentrations of 1-propanol [0.01; 1.10] ppm mixed with air at 300 °C.

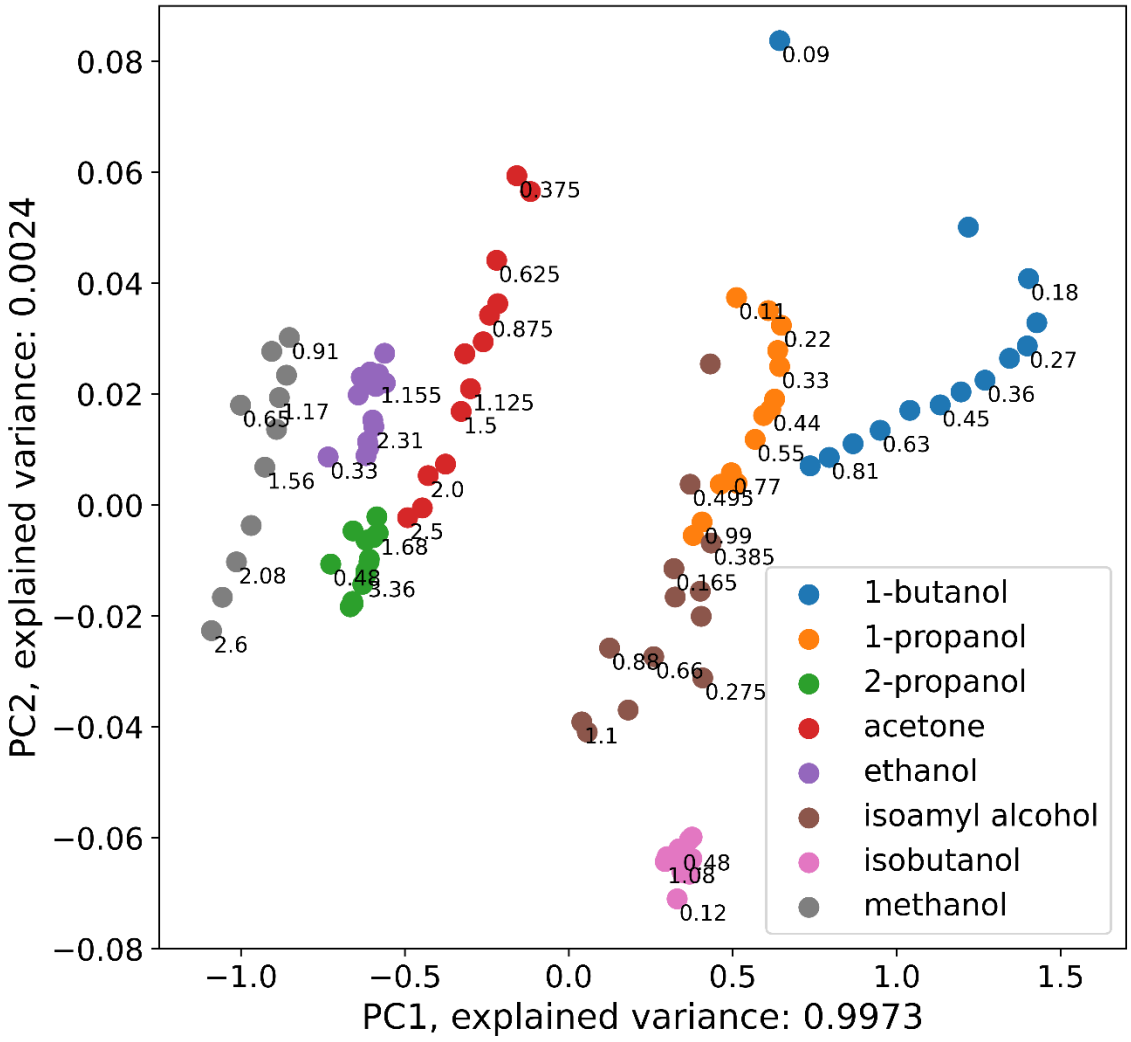

Figure S7. PCA of the sensor responses towards methanol, ethanol, 1-propanol, 2-propanol, 1-butanol, isobutanol, and isoamyl alcohol vapors in the mixture with air at 300 °C normalized on concentration dependence (C^n). Clustering is well-presented for some analytes, i.e. isobutanol, ethanol, 2-propanol. Concentrations of analyte in the mixture with air are given in *ppm*.

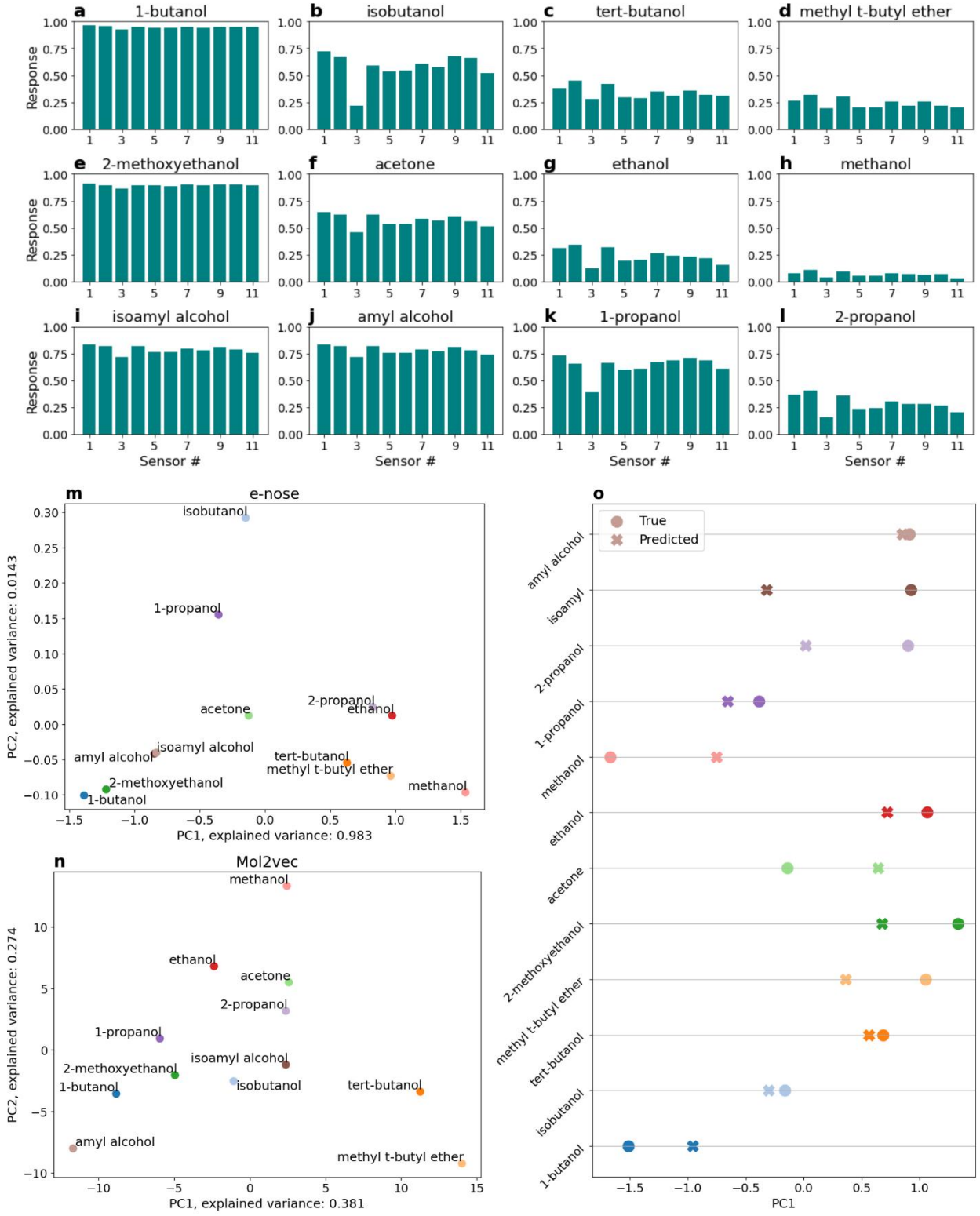

Figure S8. Selective recognition and analysis of “fingerprints” of the VOCs using Mol2vec molecular “fingerprints”. The presented data belong to other chip prepared using the same method. (a - l) Chemoresistive responses of sensors of the multisensor chip towards methanol, ethanol, 1-propanol, 2-propanol, 1-butanol, isobutanol, tert-butanol, isoamyl alcohol, methyl t-butyl ether, 2-methoxyethanol, acetone (all ca. 0.9 ppm in the mixture with air), amyl alcohol (ca. 0.65 ppm, mixed with air), at 300 °C, (m-o) PCA representation of data acquired from the

multisensor system, and from Mol2vec “fingerprints” of chosen analytes with the corresponding prediction of “unknown” analyte in the mixture with air: (m) PCA of vector signal recorded by multisensor system for tested analytes in the mixture at concentration ca. 0.9 ppm in the mixture with air (ca. 0.65 ppm for amyl alcohol); (n) PCA projection of Mol2vec “fingerprints”; (o) SVR prediction of “reduced sensor data" from Mol2vec “fingerprints” by training using PCA data of multisensor vector signal shown in (m) to 1D PCA accordingly, circles denote true values, crosses — predicted ones.

Notes and references:

s1 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

s2 B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing and Z. Wu, *Deep Learning for the Life Sciences*, O’Reilly Media, 2019.

s3 S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model.*, 2018, **58**, 27–35.

s4 J. C. Platt, in *Advances in Large Margin Classifiers*, MIT Press, 1999, pp. 61–74.