Search for H-Bonded Motifs in Liquid Ethylene Glycol Using a Machine Learning Strategy

Supporting Information

*Aman Jindal, Vaishali Arunachalam and Sukumaran Vasudevan**

Department of Inorganic and Physical Chemistry

Indian Institute of Science, Bangalore 560012,

INDIA

*Author to whom correspondence may be addressed. E-mail: svipc@iisc.ac.in. Tel: +91-80-2293-2661. Fax: +91-80-2360-1552/0683;

Contents

S1. Choice of map size: Trimer fragments

The SOM analysis for trimer fragments was carried out for three map sizes 5 × 5, 7 × 7 and 8 × 8. For each of the map sizes the U-matrix representation and the quantization error for selected nodes are shown in Figures S1-S3.

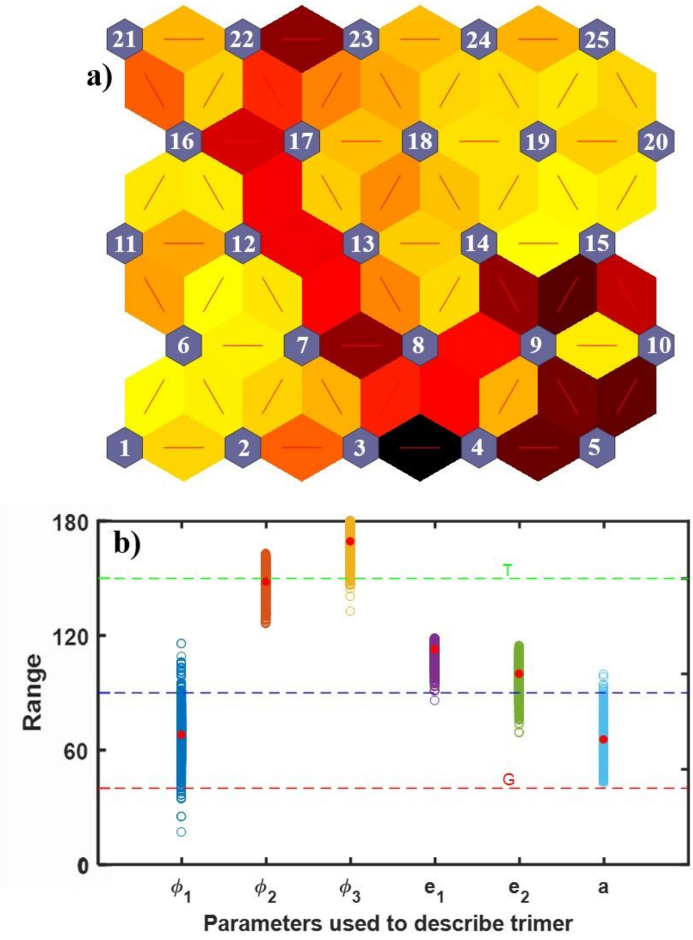


Figure S1: a) The 5 × 5 U-matrix for trimer EG fragments. The number on each node refers to the node ID. b) The quantization error, spread in the values of the input vectors of trimer fragment structures classified by node 10. The values of model vector corresponding to node 10 are shown as red filled circles.

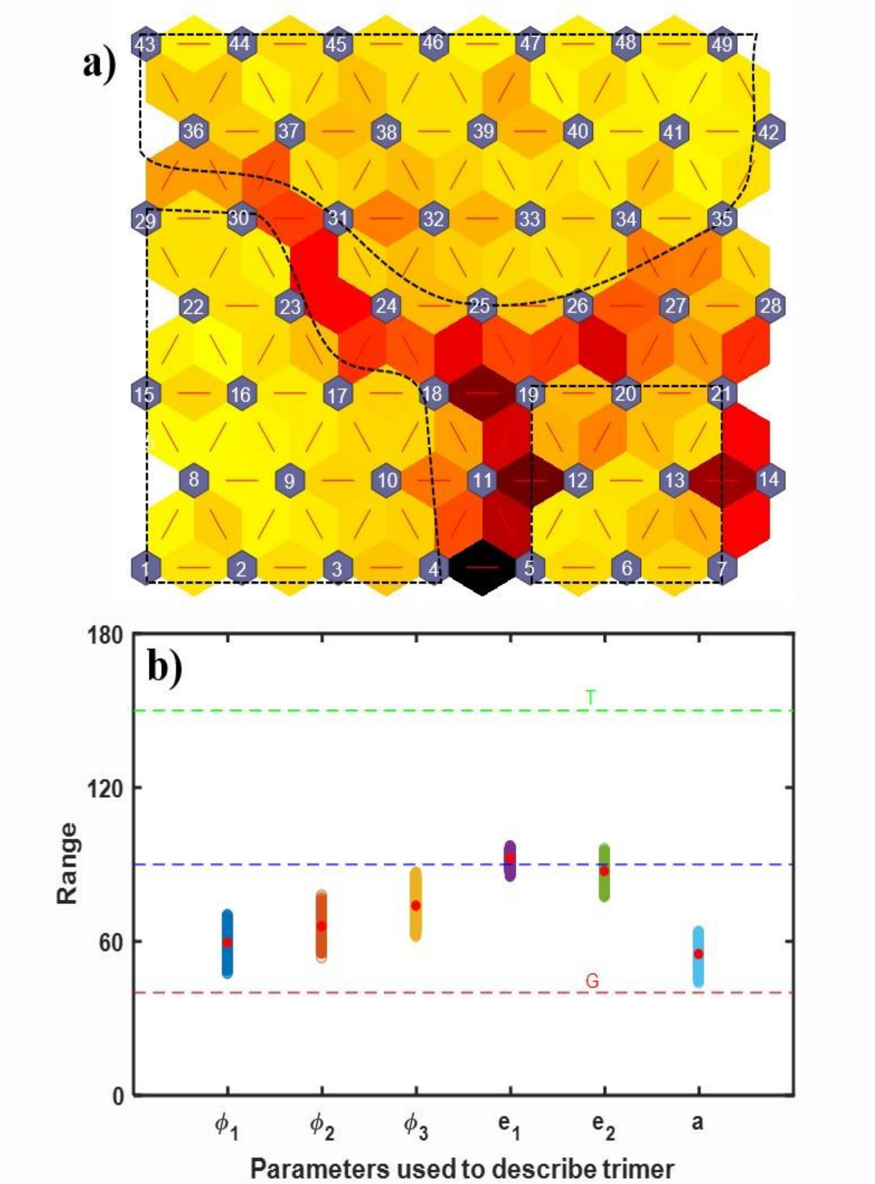


Figure S2: a) The 7 × 7 U-matrix for trimer EG fragments. The number on each node refers to the node ID. b) The quantization error, spread in the values of the input vectors of trimer fragment structures classified by node 10. The values of model vector corresponding to node 10 are shown as red filled circles.

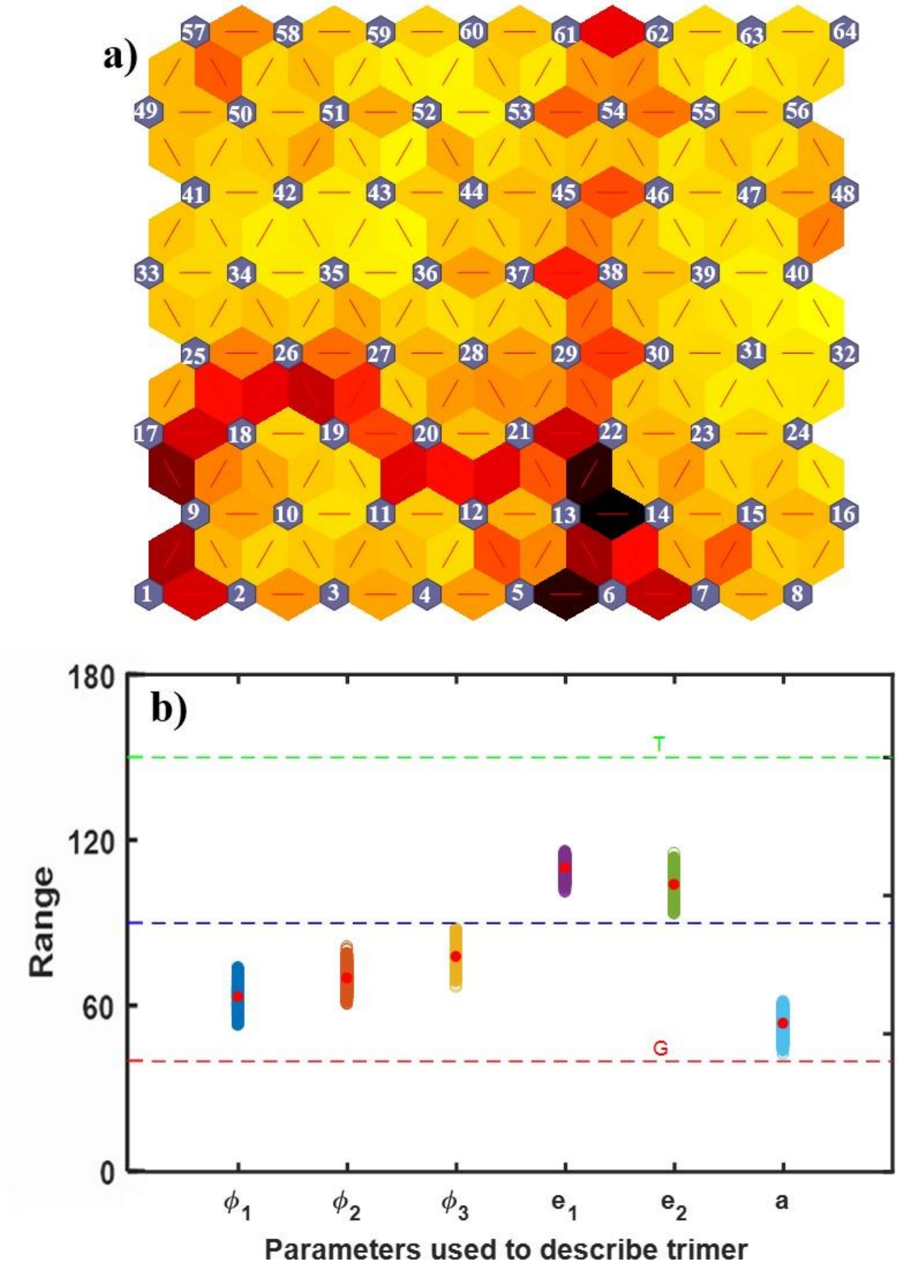


Figure S3: a) The 8 × 8 U-matrix for trimer EG fragments. The number on each node refers to the node ID. b) The quantization error, spread in the values of the input vectors of the trimer fragment structures classified by node 16. The values of model vector corresponding to node 16 are shown as red filled circles.

The reliability of the SOM was judged by the quantization error – the spread in the values of the input vectors with respect to the values of the model vectors. A large spread indicates that the clustering is inefficient and the map unreliable. It may be seen that for the 5 × 5 map the quantization error is large and hence the map unreliable. The quantization errors for the 7 × 7 and 8 × 8 maps are comparable and hence a 7 × 7 map size was considered adequate. A more detailed discussion is presented in the SOM methodology section, S2.

S2. SOM Methodology

In this study, a 7 × 7 hexagonal grid of neurons for the SOMs using a hexagonal layer topology function was used for fragments of size > 2 while for dimer fragments a 5 × 5 SOM was used. Two criteria are adopted to decide the map size, the distribution of clusters in the U-matrix and the quantization error - the measure of how similar the nodal center is to the input vectors in that node - at the end of the training. These two parameters give an idea about the reliability of the map. This is illustrated for the trimer fragment where the input vectors are 6-dimensional (ϕ_1, ϕ_2, ϕ_3, a, e_1 and e_2; ϕ's are the OCCO dihedral angle). For the trimer fragment a 7×7 map was found to be optimal. This size of the map was arrived at by a trial-and-error method by considering maps of smaller (5 × 5) and larger (8 × 8) sizes (Figures S1 and S3). Figures S1 and S3 gives the U-matrix and the quantization error for (5×5) and (8×8) map for the trimer fragments. For the 5×5 map (Figure S1) it may be seen that the clusters (sharing light color between the nodes in Figure S1a) are not well separated. Figure S1b shows the spread in the components of the input vectors, the 'quantization error' for node 10 of the map. The values of the representative vectors for the node are shown as red dots. The large spread in the input vectors clearly indicates that a (5 × 5) map is too small is size, resulting in poor resolution and fails in clustering the data into well-delineated clusters. The results of the (5 × 5) map may be contrasted with the results for the same data using a 8 × 8 map (Figure S2). The U-matrix representation of the map shows that three clusters of nodes are clearly delineated. As will be discussed later the nodes within a cluster are topologically related as may be seen from the light color between the nodes. Figure 3b shows that the quantization error, the spread in the values of the input vectors, is significantly reduced as compared to the 5 × 5 map (Figure S1b). The 8×8 map is unnecessarily complex and it was found that the similar results, clustering and quantization error, could be achieved using a 7×7 map (Figure S2). Consequently for the SOM analysis of trimer, tetramer and pentamer fragments a 7×7 map has been used. For the dimer, however, where the input vector is 5 dimensional (φ_1, φ_2, a, e_1 and e_2) a 5×5 map was found to suffice.

In the present study the layer-by-layer network initialization function, *initlay* available as part of the SOM toolbox in MATLAB was used to initialize the model vectors for each node. The *initlay* function does not require any initialization parameters. The Batch training algorithm, *trainbu* was used to train the model vectors associated with each node of the SOM. Each step of the training involves two processes. The competition in which the network computes the Best Matching Unit/Node (BMU) based on the Euclidean distance criteria and the cooperation where the BMU defines the neighborhood nodes whose weight are updated as per the Kohonen rule. In each epoch (one pass through the entire data set) not only the winning neuron but all neurons within its neighborhood are updated using the following Kohonen rule.

$$w_i(t + 1) = w_i(t) + \alpha(t)h_{ci}(t)[x(t) - w_i(t)] \qquad \text{(S1)}$$

where, $w_i(t)$ is the weight of i^{th} node at time t, $x(t)$ is the input vector randomly drawn from the input data set, α(t) the learning rate and, $h_{\mathrm{ci}}(t)$ the neighborhood function around the winning node c. The most widely used neighboring function is the

Gaussian neighborhood function, $h_{\mathrm{ci}}(t) = exp(-\delta_{ci}^2 /2r(t)^2)$, where δ_{ci} is the distance between nodes c and i in the SOM and $r(t)$ is the neighborhood radius at time t. In the batch unsupervised training, there is no time-variable learning rate parameter and $\alpha(t)$ has a fixed value of unity. It should be noted, however, that as the Self-Organizing Maps are trained with input vectors in a random order, starting with the same initial vectors does not guarantee identical training results.

In this study, the typical length of the training was 250 epochs. The length of the training was decided by the fact that the map showed no significant change. The weight learning function available in *trainbu* is *learnsomb*, where the initial neighborhood size is three and it is gradually reduced to one.

S3. Movie file (S1) for the cyclic dimer structure (D_1), corresponding to the node 25 of dimer fragment SOM (Figure 3) along the MD simulation trajectory is available at https://github.com/ipcaman/EG_dimer_node25/commit/67c749dff61d7aa36a5e0d6a-8e00b5f1e0d98b2e#commitcomment-46754591.

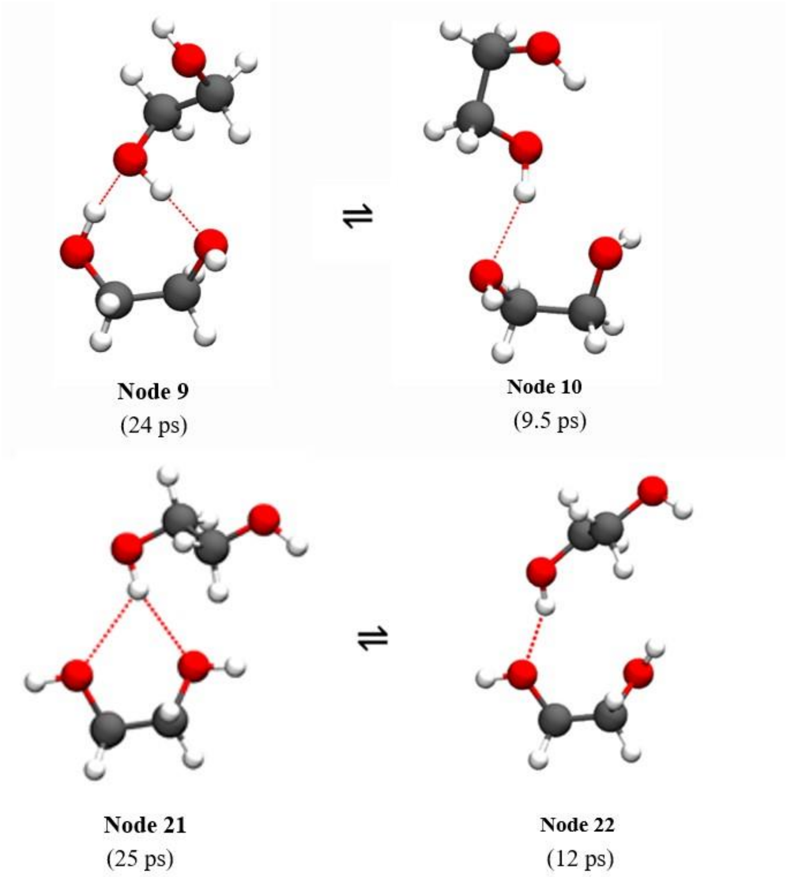
S4.

Table S1: Model vectors of the nodes of the trained dimer fragment SOM

Node	e_1	e_2	a	φ_1	φ_2	Conformation of EG_1-EG_2	Number of Hits
1	115.2	60.7	73.9	162.4	173.1	**TT**	431
2	113.9	88.1	64.8	160.8	172.9	**TT**	859
3	111.2	40.5	65.3	68.8	83.9	**GG**	417
4	109.7	40.3	61.6	57.1	68.2	**GG**	513
5	110.8	60.6	60.6	54.0	65.8	**GG**	903
6	113.6	90.5	62.9	82.6	170.1	**GT**	1198
7	113.7	67.9	68.0	90.9	168.1	**GT**	646
8	114.8	60.3	72.4	62.7	76.1	**GG**	878
9	108.3	64.5	56.1	66.6	78.8	**GG**	995
10	111.4	77.1	59.2	46.4	62.3	**GG**	730
11	112.8	44.6	67.7	67.0	170.3	**GT**	794
12	113.6	74.5	66.4	70.6	172.1	**GT**	1590
13	112.3	63.8	65.5	71.0	96.6	**GG**	449
14	115.0	79.1	70.7	70.7	81.6	**GG**	950
15	114.9	78.3	70.1	59.2	68.8	**GG**	1103
16	113.2	72.2	65.4	53.1	171.6	**GT**	1067
17	113.0	64.8	65.8	64.6	152.3	**GT**	610
18	112.6	84.5	60.9	52.9	87.6	**GG**	722
19	113.3	93.5	59.5	66.5	77.1	**GG**	1189
20	113.1	91.4	59.5	55.4	66.1	**GG**	1014
21	113.0	92.0	60.4	62.2	172.6	**GT**	1605
22	113.2	89.7	61.6	65.6	154.3	**GT**	961
23	113.0	82.8	63.1	70.5	126.7	**G***	339
24	111.7	88.4	57.8	75.7	91.1	**GG**	781
25	109.7	79.9	54.9	63.2	72.9	**GG**	1374

* values are outside the range defined for *gauche* (**G**) and *trans* (**T**) conformer

S5. Scheme for cyclic dimers D_2 and D_3



Scheme S1: Representative structure for nodes 9 and 21 of the dimer fragment SOM labeled as D_2 and D_3. The structures with slight variation are classified by neighboring nodes 10 and 22, respectively.

S6. Quantization error for trimer, tetramer and pentamer fragments.

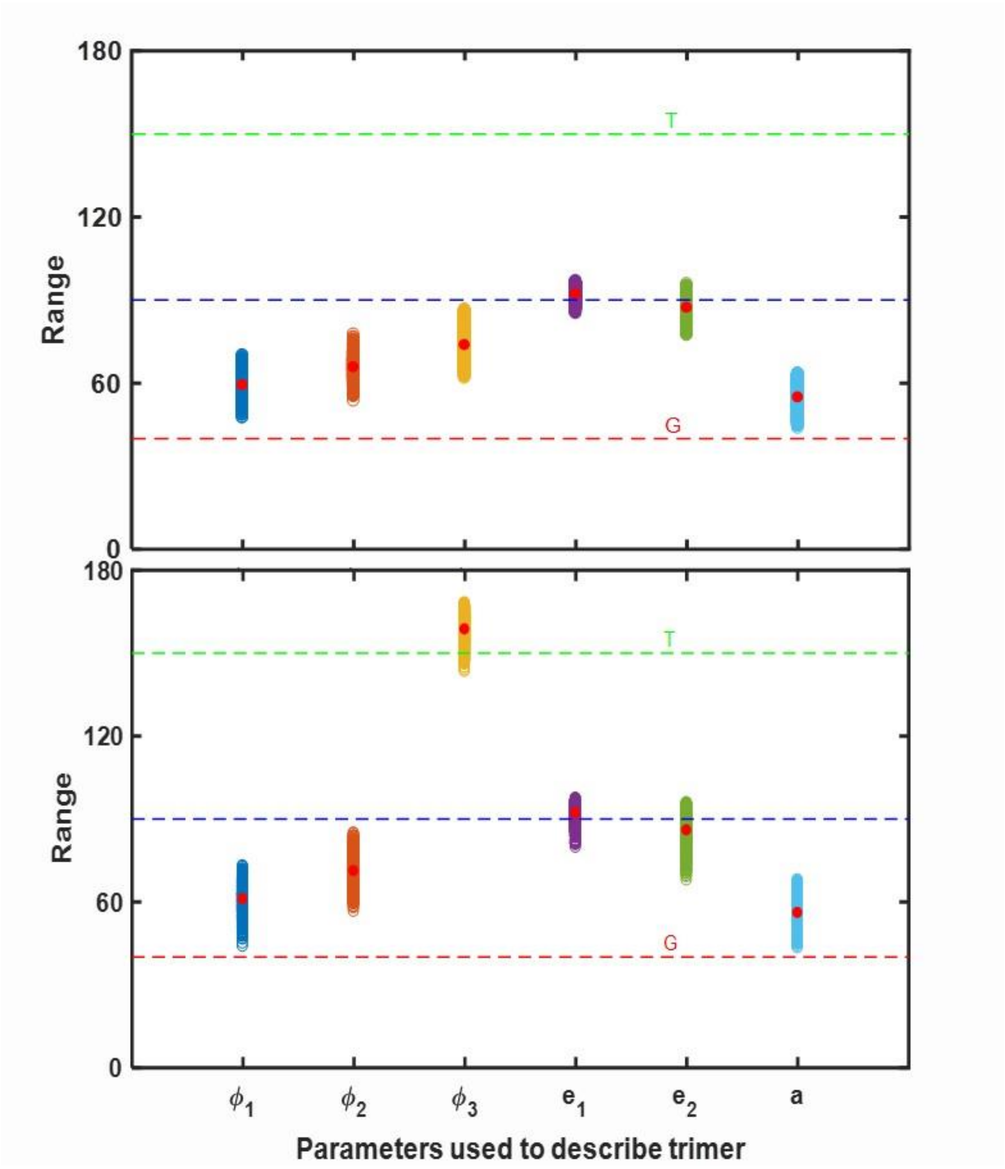


Figure S4: Spread in the values of the input vectors of fragment structures belonging to nodes 8 and 41 of the trimer fragments SOM. The values of model vector corresponding to these nodes are shown as red filled circles.

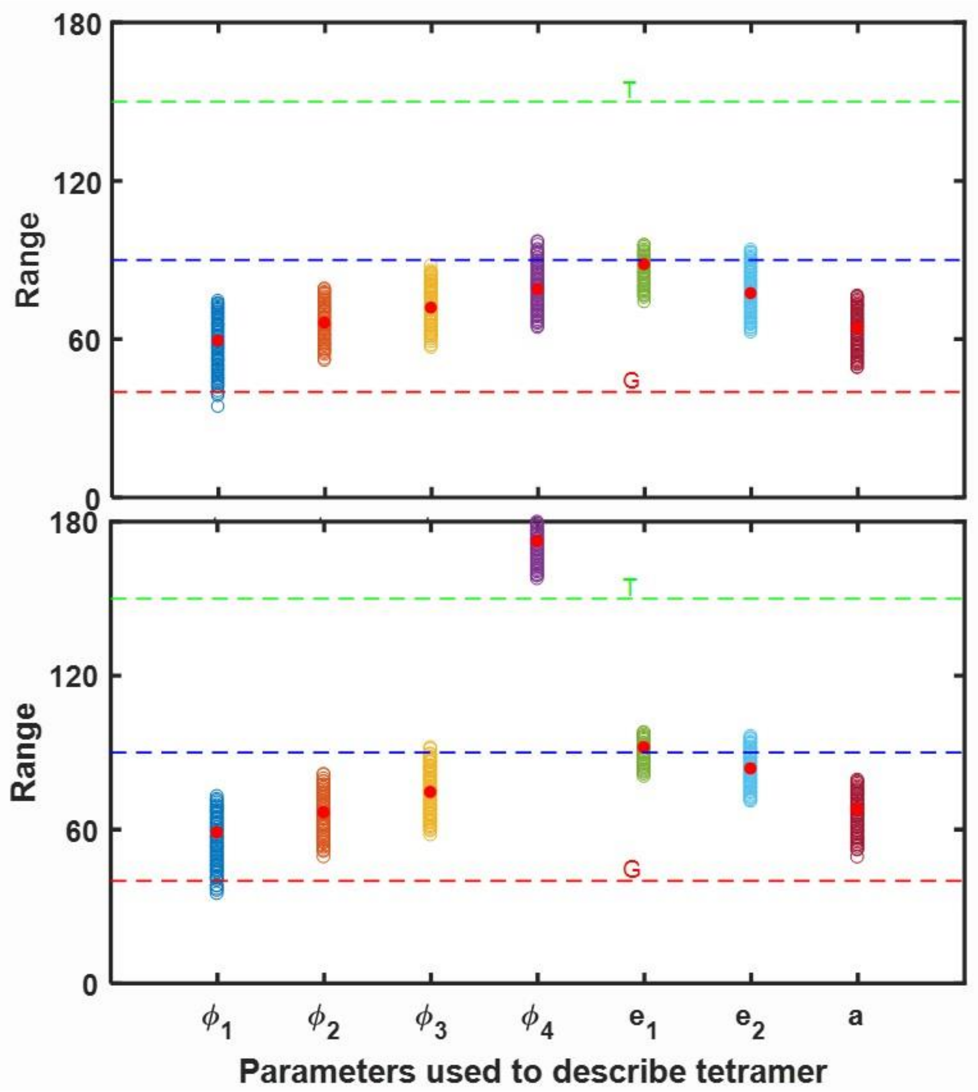


Figure S5: Spread in the values of the input vectors of fragment structures belonging to nodes 10 (**GGGG**) and 38 (**GGGT**) of the tetramer fragments SOM. The values of model vector corresponding to these nodes are shown as red filled circles.

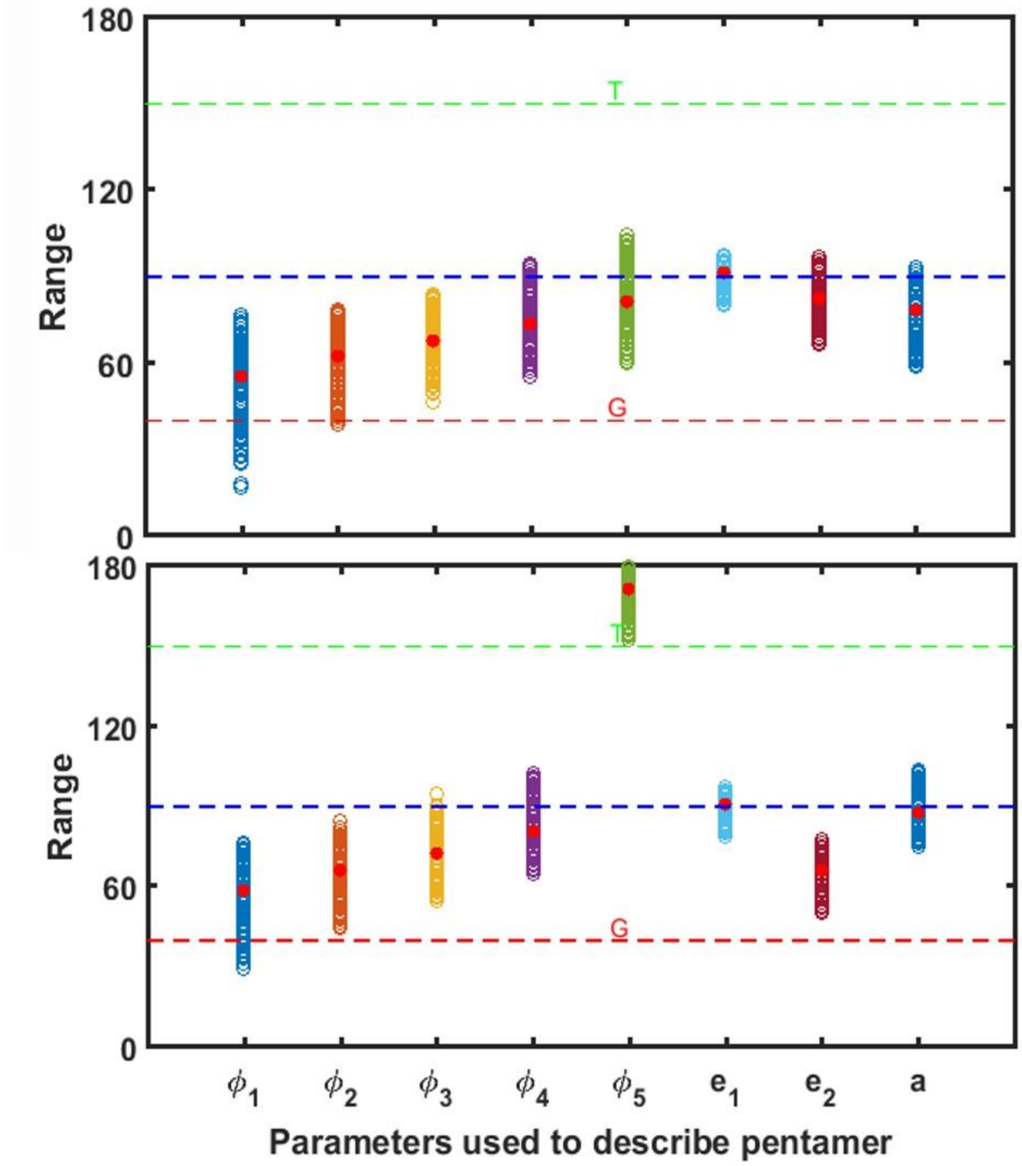


Figure S6: Spread in the values of the input vectors of fragment structures belonging to nodes 10 (**GGGGG**) and 38 (**GGGGT**) of the pentamer fragments SOM. The values of model vector corresponding to these nodes are shown as red filled circles.