**How to use the MATLAB code for the prediction of total electron ionization for the compounds.**

There are a few files that are needed to obtain the total electron ionization cross sections of new compounds. The descriptions of these files as follows:

**Descriptor\_input\_format.xlsx** -- This Excel file can be used to compose the input descriptors of a compound or of multiple compounds. The first column of the file is the name of the compound. This column is useful to track down a descriptor row and to understand the descriptors. After that there are 100 columns for the descriptors: Group, atom, and other lumped descriptors. The name of each descriptor and an explanation are written on the top of the column. One can obtain these descriptors after looking into the molecular structure of a compound.

For the ring strain, the values from Boyd et al.1 are used for five-membered rings in PAH: 6.5, 2, and 6.5 kcal/mol for acenaphthene-, acenaphthylene-, and fluoranthene-type rings, respectively. Otherwise, the ring-correction tables from Benson, *Thermochemical Kinetics, 2nd Ed.* 2 are consulted. A value is assigned after observing the closest ring structure from the latter tables.

Rows 4 to 18 are example compounds that can be used to test the code. Rows 19 to 414 are the 396 compounds in the database.

**Make\_a\_txt\_file\_after\_copying\_the\_descriptors.txt** -- Once the information is prepared with the previous Excel file, the descriptor columns and rows are copied. For example, in the current case, we have cut and pasted the C4 to CS18 cells from the Excel file into a text file with this name.

**Use\_this\_code\_to\_get\_TEICS\_at\_70\_75eV.m** -- This file is the MATLAB code, shown below, which uses the previous txt file as the input. The other input file is BoseWestmorelandNet.mat, which contains the ANN model’s parameters obtained after training the neural net model for 500 times. When the MATLAB code gets the inputs of the descriptors, it uses BoseWestmorelandNet.mat to obtain 500 different predicted values of the cross section. The average and 95% confidence gap from the mean from these 500 values are then calculated using t-statistics. Finally, these average values and the 95% confidence gap from the mean are recorded in Output.xlsx file.

**Output.xlsx** -- This results file has two columns without any headers. The first column contains the mean value and the second column contains the 95% confidence gap from the mean. The number of the rows are same as the number of compounds in the **Make\_a\_txt\_file\_after\_copying\_the\_descriptors.txt** file.

Reference:

(1) Boyd, R. H.; Christensen, R. L.; Pua, R. The Heats of Combustion of Acenaphthene, Acenaphthylene, and Fluoranthene. Strain and Delocalization in Bridged Naphthalenes. J. Am. Chem. Soc. 1965, 87 (16), 3554–3559. https://doi.org/10.1021/ja01094a004.

(2) Benson, S. W. Thermochemical Kinetics, 2nd Ed.; John Wiley & Sons, Inc.: New York, 1976.