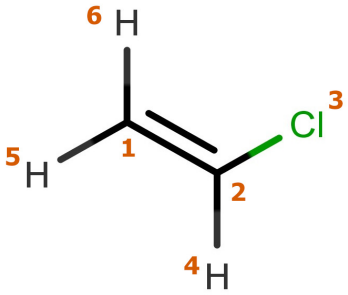
PREDICTION OF TOXICITY OF BENZENE DERIVATIVES USING LAPLACIAN MATRIX-BASED GRAPH THEORETICAL MOLECULAR DESCRIPTORS DERIVED BY GRAPH CONVOLUTION

Supporting Information 1

This file describes the mathematical procedure for development of Laplacian matrix based convolutional descriptors. The procedure will be illustrated using vinyl chloride molecule labeled as shown in the following figure.



The adjacency matrix (**A**) for this molecule is:

$$A := \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad i := 1 .. \mathrm{rows}(A)$$

The property matrix (**P**) for the vinyl chloride is:

$$P := \begin{pmatrix} 6 & 12.011 & 2.55 & 1.8 & 1.70 & 1.26 & 11.260 \\ 6 & 12.011 & 2.55 & 1.8 & 1.70 & 1.26 & 11.260 \\ 19 & 35.45 & 3.16 & 2.2 & 1.75 & 3.61 & 12.968 \\ 1 & 1.008 & 2.2 & 0.7 & 1.20 & 0.75 & 13.598 \\ 1 & 1.008 & 2.2 & 0.7 & 1.20 & 0.75 & 13.598 \\ 1 & 1.008 & 2.2 & 0.7 & 1.20 & 0.75 & 13.598 \end{pmatrix}$$

The number of rows is the same as the number of atoms in vinyl chloride while the number of columns is the same as the number of properties. The properties in this matrix are:

1. Atomic number (Z)
2. Relative atomic mass
3. Electronegativity (Pauling)
4. Polarizability ($Å^3$)
5. Van der Waals radius (Å)
6. Electron Affinity (eV)
7. First ionization potential (eV)

Now we have to define the degree matrix (**D**) using adjacency matrix (**A**). **D** is a diagonal matrix with the degree of the vertices in the main diagonal.

The degree matrix (**D**) for this molecule is:

$$D_{i,i} := \left(\sum_{j=1}^{\text{cols}(A)} A_{i,j} \right) \qquad D = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

With **A** and **D** calculated we can calculate signless Laplacian matrix (**L**):

$$L := D + A = \begin{pmatrix} 3 & 1 & 0 & 0 & 1 & 1 \\ 1 & 3 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

It is better to work with normalized Laplacian matrix. The normalization helps in reducing the influence of the nodes with large vertex degrees to dominate the properties of the Laplacian

matrix.

In order to perform the normalization we need to calculate $\mathbf{D}^{-1/2}$. Here, this matrix will be labeled as $\mathbf{D}_{half}$: The elements of this matrix are defined as:

$$\mathrm{D_{half}}_{i,i} := \frac{1}{\sqrt{\mathrm{D}_{i,i}}}$$

while for the matrix we have:

$$\mathrm{D_{half}} \rightarrow \begin{pmatrix} \frac{\sqrt{3}}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\sqrt{3}}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Finally, for the *normalized signless Laplacian matrix* ($\mathbf{L}_{norm}$) we have:

$$\mathrm{L_{norm}} := \mathrm{D_{half}} \cdot \mathrm{L} \cdot \mathrm{D_{half}}$$

The values of $\mathbf{L}_{norm}$ for the vinyl chloride are:

$$\mathrm{L_{norm}} \rightarrow \begin{pmatrix} 1 & \frac{1}{3} & 0 & 0 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ \frac{1}{3} & 1 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & 0 & 0 \\ 0 & \frac{\sqrt{3}}{3} & 1 & 0 & 0 & 0 \\ 0 & \frac{\sqrt{3}}{3} & 0 & 1 & 0 & 0 \\ \frac{\sqrt{3}}{3} & 0 & 0 & 0 & 1 & 0 \\ \frac{\sqrt{3}}{3} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Now we can pre multiply the property matrix (**P**) with the normalized Laplacian matrix ($\mathbf{L}_{norm}$). Using this we will obtain *Laplacian-transformed property matrix* ($\mathbf{M}_1$):

$$M_1 := L_{norm} \cdot P = \begin{pmatrix} 9.155 & 17.179 & 5.94 & 3.208 & 3.652 & 2.546 & 30.715 \\ 19.547 & 37.064 & 6.495 & 4.074 & 3.97 & 4.197 & 30.351 \\ 22.464 & 42.385 & 4.632 & 3.239 & 2.731 & 4.337 & 19.469 \\ 4.464 & 7.943 & 3.672 & 1.739 & 2.181 & 1.477 & 20.099 \\ 4.464 & 7.943 & 3.672 & 1.739 & 2.181 & 1.477 & 20.099 \\ 4.464 & 7.943 & 3.672 & 1.739 & 2.181 & 1.477 & 20.099 \end{pmatrix}$$

Here the index "1" in $\mathbf{M}_1$ represents how many times **P** was pre multiplied.

Also, if one analyses different column of $\mathbf{M}_1$ it could be noticed that the values, for example in the first column,are not the same as atomic numbers (*Z*) of the elements in **P**. They are changed as a result of the “interaction” of the neighboring atoms (vertices) using the Laplacian matrix. That is why $\mathbf{M}_1$ is called Laplacian-transformed property matrix.

In the matrix $\mathbf{M}_1$ the index “1” also means that only the “interaction” between selected atom and its first neighbors are considered. Often for similar molecules, the descriptors extracted from $\mathbf{M}_1$ could not distinguish between different structures. For example, when using substituted benzene derivatives with two identical substituents the descriptors obtained for meta and para substituted structures will be the same. Having this in mind, it is better to pre multiply **P** two or three times with $\mathbf{L}_{norm}$.

Here we will calculate the final atom-based descriptors using three pre multiplication of P with the normalized Laplacian matrix:

$$M_3 := L_{norm} \cdot L_{norm} \cdot L_{norm} \cdot P = \begin{pmatrix} 44.798 & 83.277 & 24.969 & 13.395 & 15.123 & 11.856 & 128.87 \\ 73.666 & 138.513 & 26.509 & 15.8 & 16.005 & 16.442 & 127.86 \\ 55.773 & 105.264 & 16.043 & 10.221 & 9.656 & 11.612 & 73.616 \\ 37.773 & 70.822 & 15.083 & 8.721 & 9.106 & 8.752 & 74.246 \\ 21.773 & 40.207 & 14.23 & 7.387 & 8.617 & 6.21 & 74.806 \\ 21.773 & 40.207 & 14.23 & 7.387 & 8.617 & 6.21 & 74.806 \end{pmatrix}$$

Finally, the descriptors developed using this procedure could be developed using (1) sum of the columns or (2) arithmetic mean of the columns that correspond to different atomic properties:

$$\text{sum\_Atomic\_No} := \sum_{j=1}^{\text{rows}(A)} M3_{j,1} = 255.555 \qquad \text{sum\_Rel\_atom\_mass} := \sum_{j=1}^{\text{rows}(A)} M3_{j,2} = 478.289$$

$$\text{sum\_Electroneg} := \sum_{j=1}^{\text{rows}(A)} M3_{j,3} = 111.063 \qquad \text{sum\_Polariz} := \sum_{j=1}^{\text{rows}(A)} M3_{j,4} = 62.911$$

$$\text{sum\_vdW\_radius} := \sum_{j=1}^{\text{rows}(A)} M3_{j,5} = 67.124 \qquad \text{sum\_El\_affinity} := \sum_{j=1}^{\text{rows}(A)} M3_{j,6} = 61.083$$

$$\text{sum\_1st\_ion\_pot} := \sum_{j=1}^{\text{rows}(A)} M3_{j,7} = 554.203$$

The meaning of the abbreviation for these descriptors is the following. First three characters ("sum") means that the descriptors are obtained using sums over the corresponding columns. The meaning of the remaining part of the abbreviation is described here:

1. **sum_Atomic_No** – the descriptor is based on the *atomic numbe*r (*Z*);
2. **sum_Rel_atom_mass** – the descriptor is based on the *relative atomic mass*;
3. **sum_Electroneg** – the descriptor is based on the *electronegativity* (Pauling);
4. **sum_Polariz** – the descriptor is based on the *polarizability* (Å^3);
5. **sum_vdW_radius** – the descriptor is based on the *van der Waals radius* (Å);
6. **sum_El_affinity** – the descriptor is based on the *electron affinity* (eV);
7. **sum_1st_ion_pot** – the descriptor is based on the *first ionization potential* (eV).