GRAPH-BASED METHOD FOR STRUCTURE GENERATION

SCOTT HABERSHON

In this research note, we describe a simple method for *de novo* generation of molecular structures based on connectivity matrices.

Background

Many aspects of *molecular design* require one to be able to generate new molecular structures and then assess their properties.[1] As a result, a number of algorithms have been developed to generate new molecular structures.

[1] Such as absorption spectra, dipole moment, thermal stability, and so on.

However, an important problem with many of these structure-generation methods is that they often rely on a *fragment library*; in other words, one assumes that an existing set of molecular components[2] is available which can be "glued together" to make new structures.

[2] Such as phenyl rings, methyl groups, carboxylic acids, and so on...

So, there is a need for an unbiased *de novo* structure generation method which does not require input of molecular fragments or definition of a structural motif library; in this note, we show how ideas of connectivity graphs ca be used to meet this goal.

Connectivity matrices

A connectivity matrix (or *graph*) for an n-atom molecule is simply an $n \times n$ matrix with entries which are 0 if two atoms are not bonded, and 1 if two atoms are bonded.

So, the element G_{ij} of the connectivity matrix $\mathbf{G}$ are:

$$G_{ij} = \begin{cases} 1 & \text{if } r_{ij} < r_{ij}^{cut}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here, r_{ij}^{cut} is a distance cut-off value which indicates whether or not two atoms i and j are bonded. For example, this might be defined as:

$$r_{ij}^{cut} = \alpha(R_i + R_j), \tag{2}$$

where R_i and R_j are the covalent radius for atom types i and j, and $alpha$ is a parameter which would be expected to take a value of, say, 1.1. With this definition, two atoms would be identified as being bonded if they lie at a distance which is less than the sum of the two covalent radii, plus some small additional shift of about 10% of this sum of radii.

Graph-enforcing potentials

Once we have defined a graph **G**, it is straightforward to generate molecular structures which conform to the bonding pattern encoded in the graph. In particular, we can use the idea of a graph-enforcing potential, as introduced in reference [1], to impose a target connectivity graph on a molecular structure. In other words, given a target graph **G**, the graph-enforcing potential enables generation of the Cartesian coordinates (x, y, z) for each atom such that the bonding in generated structure agrees with the graph.

The graph-enforcing potential $W(\mathbf{r}, \mathbf{G})$ used in [1, 2] has the following functional form:

$$W(\mathbf{r}, \mathbf{G}) = \sum_{j>i} \Big[\delta(G_{ij} - 1)[H(r_{ij}^{min} - r_{ij})\sigma_1(r_{ij}^{min} - r_{ij})^2 + H(r_{ij} - r_{ij}^{max})\sigma_1(r_{ij}^{max} - r_{ij})^2] + \delta(G_{ij})\sigma_2 e^{-r_{ij}^2/(2\sigma_3^2)} \Big] + V_{mol}(\mathbf{r}, \mathbf{G}) \tag{3}$$

The sum at the start of Eq. 3 runs over all pairs of atom (*i.e.* it is a pairwise potential). The different contributions are as follows:

The term in red is a harmonic restraining force which acts on pairs of atoms which are *bonded* in order to keep their bond-length between the limits r_{ij}^{min} and r_{ij}^{max}. The first part is a "delta" function, defined such that

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

So, the term $\delta(G_{ij} - 1)$ implies that $\delta = 1$ when atoms i and j are bonded, and $\delta = 0$ otherwise. As a result, the term in red only operates on bonded pairs of atoms.

The first part of the red term reads as follows:

$$H(r_{ij}^{min} - r_{ij})\sigma_1(r_{ij}^{min} - r_{ij})^2.$$

Here, $H(x)$ is the Heaviside function, defined as:

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x > 0. \end{cases} \tag{5}$$

So, $H(r_{ij}^{min} - r_{ij})$ is zero as long as $r_{ij} > r_{ij}^{min}$; in this case, this first term does not contibute anything to the potential energy value. If, on the other hand, $r_{ij} < r_{ij}^{min}$, then $H(r_{ij} - r_{ij}^{min}) = 1$ and the first term does contribute to the potential energy. In particular, a harmonic term of the following form

$$\sigma_1(r_{ij}^{min} - r_{ij})^2$$

is applied to the system, where σ_1 is a user-defined constant. The effect of this potential energy term is to push the atoms i and j to bond lengths such that $r_{ij} > r_{ij}^{min}$. In other words, this term pushes the atoms apart until they are some minimum distance r_{ij}^{min} away from each other.

The second part of the red term has a similar effect, but instead of maintaining some *minimum* distance, it makes sure that a *bonded* pair of atoms always remain closer than some maximum allowed distance r_{ij}^{max}:

$$H(r_{ij} - r_{ij}^{max})\sigma_1(r_{ij}^{max} - r_{ij})^2.$$

Together, the influence of the pair-potential term in red is to ensure that a pair of *bonded* atoms[3] *always* have bond-lengths between the pre-defined limits r_{ij}^{min} and r_{ij}^{max}. The behaviour of this pair potential term is highlighted in Fig. 1.

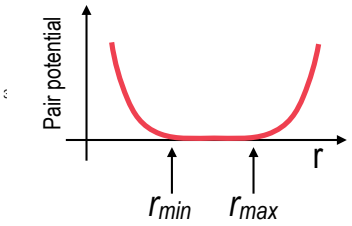

Figure 1: Functional form of the graph-enforcing potential energy term in red.

The term in blue acts as a repulsive potential between pairs of atoms which are *not* bonded. The $\delta(G_{ij})$ term makes sure that this term only applies to pairs of atoms for which $G_{ij} = 0$. The remainder of this term is a simple Gaussian repulsive potential with a strength parameter σ_2 and a range parameter σ_3.

The term in orange is a *molecular* term; it only operates between distinct molecular species.[4] The $V_{mol}(\mathbf{r}, \mathbf{G})$ term has the following form:

[4] However, note that a single atom, on its own and not bonded to anything, is also classed as a "molecule" in this term.

$$V_{mol}(\mathbf{r}, \mathbf{G}) = \sum_{J>I} [H(R^{min} - R_{IJ})\sigma_4(R^{min} - R_{IJ})^2, + H(R_{IJ} - R^{max})\sigma_4(R^{max} - R_{IJ})^2]. \tag{6}$$

Here, R_{IJ} is the distance between the centres-of-mass of two molecules I and J, and R^{min} and R^{max} are user-defined minimum and maximum distances between any pair of molecules.

By comparing to the bonding term in Eq. 3, we see that the molecular term V_{mol} is designed to make sure that distinct molecules are "kept apart" from each other; the functional form of V_{mol} is the same as that shown in Fig. 1, but it acts on the centre-of-mass of molecules, rather than single atoms. As a result, V_{mol} ensures that distinct molecules, as defined in the graph $\mathbf{G}$ for the system, are kept apart and are not allowed to approach each other.

Using $W(\mathbf{r}, \mathbf{G})$ to generate chemical structures

The function $W(\mathbf{r}, \mathbf{G})$ acts as a potential energy term which is zero if, and only if, the connectivity graph calculated for the configuration $\mathbf{r}$, labelled $\mathbf{G}(\mathbf{r})$, exactly matches the target connectivity graph $\mathbf{G}$. If there

are any differences between $\mathbf{G}(\mathbf{r})$ and $\mathbf{G}$, the total force on the atoms arising from $W(\mathbf{r}, \mathbf{G})$ will be non-zero.

As a result, optimizing the geometry under the action of the graph-enforcing potential $W(\mathbf{r}, \mathbf{G})$ will move the atoms to a configuration such that $W(\mathbf{r}, \mathbf{G}) = 0$. In other words, using geometry optimization under $W(\mathbf{r}, \mathbf{G})$, the atoms will move so that their connectivity graph $\mathbf{G}(\mathbf{r})$ exactly matches the target graph $\mathbf{G}$.

This simple idea is used in our graph-driven sampling (GDS) method to generate chemical reaction paths; however, as described below, the same approach can be used to automatically generate new molecular structures.

Graph-based structure generation algorithm

Based on optimization under $W(\mathbf{r}, \mathbf{G})$, one can construct an algorithm which can be used to create molecular structures.[**?**, **?**] The approach we take has two distinct parts:

1. Generate a chemically-sensible graph G^0 using an Monte Carlo optimization procedure;
2. Generate a set of Cartesian atomic coordinates which minimize $W(\mathbf{r}, \mathbf{G}^0)$ and hence match the target graph $\mathbf{G}^0$.

Here, we consider these two parts independently.

Generating chemically-sensible molecular graphs

Our first goal is to generate a connectivity graph for an n-atom molecule which is *chemically-sensible*; by this, we mean a chemical graph $\mathbf{G}^0$ which obeys common atomic valence rules, such as:

- carbon atoms are usually bonded to between 1 and 4 other atoms;
- oxygen atoms are usually bonded to between 1 and 2 other atoms;
- ...

We note that these constraints are completely general and generic; in our current molecular structure generation code, these constraints are defined by the user, so they can be changed as desired.

In addition, we might be interested in avoiding creating of some molecular structure motifs which, although "chemically-sensible", would not be expected to commonly feature in typical molecular structures which we might be interested in. For example,

- We might want to avoid creation of structures containing 3-membered or 4-membered rings;

- We might want to avoid creating structures with single oxygen-oxygen bonds;
- ....

Again, these constraints can be flexibly defined, or can even be ignored completely.

Once we have defined a set of valence constraints, and perhaps additional structural constraints, we can then define a penalty function which can be directly evaluated for a given trial graph $\mathbf{G}^0$. We can define this penalty function as

$$P(\mathbf{G}^0) = \alpha N_{val}(\mathbf{G}^0) + \gamma N_{str}(\mathbf{G}^0), \tag{7}$$

where N_{val} is the number of atoms which violate a defined valence constraint, N_{str} is the number of violations of any user-defined structural constraints, and α and γ are user-defined parameters. Both N_{val} and N_{str} can be evaluated directly given only knowledge of the graph $\mathbf{G}^0$. Note that the penalty function is zero only if the graph $\mathbf{G}^0$ does not violate any of the constraints.

As a concrete example, suppose we wanted to generate a simple hydrocarbon molecule[5] such that:

[5] ...containing only carbon and hydrogen...

- all carbon atoms have atomic valences of 1,2,3 or 4;
- all hydrogen atoms have atomic valence of 1;
- there are no 3-membered rings.

Our penalty function could be written as:

$$P(\mathbf{G}^0) = \alpha N_{val}(\mathbf{G}^0) + \gamma N_{3-rings}(\mathbf{G}^0), \tag{8}$$

where N_{val} is the number of atoms which violate one of the valence constraints and $N_{3-rings}$ is the number of 3-membered rings in the structure (as identified from the graph $\mathbf{G}^0$). The penalty function would only adopt a value of zero if there are no atoms which violate the defined valence ranges *and* there are no 3-membered rings in the structure.

Once we have defined our graph-penalty function $P(\mathbf{G}^0)$, generating chemically-sensible graphs is relatively straightforward; we simply need to search for a graph $\mathbf{G}^0$ for which $P(\mathbf{G}^0) = 0$.[6]

[6] In other words, we need to find a graph which has no errors, as defined by our input constraints.

Our current algorithm to automatically generate a suitable molecular graph is as follows:[7]

[7] This is the method currently implemented in CDE.

1. Select a number of of *heteroatoms* n;
2. Generate a random initial graph, $\mathbf{G}^0$ by randomly assigning 0 or 1 to off-diagonal elements.[8]

[8] However, note that the graph must be symmetric.

3. Assign atom labels to each of the n atoms using a simple probabilistic algorithm, as follows:

   (a) Each atom-type is given a user-defined probability $p(X)$, where X is an atom-type;

   (b) For each atom, select an atom-type at random;

   (c) Accept the atom-type assignment if $ran(0,1) < p(X)$.[9] Otherwise, reject the assignment and go back to step (b).

4. Using the graph $\mathbf{G}^0$ and the set of atom-label assignments, evaluate the initial penalty function $P(\mathbf{G}^0)$;

5. Now perform a Monte Carlo simulation, at low temperature, in order to minimize the penalty function $P(\mathbf{G}^0)$. Here, we use the following algorithm:

   (a) Store the current penalty function value as $P_{old}(\mathbf{G}^0)$;

   (b) Randomly change either a graph-element G^0_{ij} *or* the atomic label of one of the atoms. This change generates a new graph $\mathbf{G}'$;

   (c) Calculate the new penalty function $P_{new}(\mathbf{G}')$;

   (d) Accept the new move with a probability:

$$w_{accept} = \min\left[1, e^{-\frac{(P_{new}(\mathbf{G}') - P_{old}(\mathbf{G}^0))}{k_B T_{eff}}}\right]$$

   where T_{eff} is some effective temperature.

   (e) Go back to (b).

   The Monte Carlo simulation will naturally explore possible graphs in a biased search such that it is driven towards graphs which have low (ideally zero) penalty functions. After sufficient Monte Carlo annealing simulation, the penalty function should be zero.

[9] $ran(0,1)$ means draw a random number between 0 and 1.

Generating coordinates from graphs

Once a zero-penalty heteroatom graph $\mathbf{G}^0$ has been generated, we need to generate the atomic Cartesian coordinates of the heteroatoms. To achieve these, we use the following approach:

1. Given the target graph $\mathbf{G}^0$, we perform optimization of the Cartesian coordinates of the n heteroatoms under the graph-enforcing potential $W(\mathbf{r}, \mathbf{G}^0)$, as described above.[10]

2. After generation of suitable atomic coordinates for the heteroatoms, hydrogen atoms are added to satisfy local atomic valences; we use the standard hydrogen-addition algorithm in *OpenBabel* to achieve this.[11]

[10] In our current strategy, the initial heteroatom coordinates lie on a simple cubic lattice.

[11] An alternative would be to add hydrogens in a similar manner to the Monte Carlo scheme above.

3. Finally, starting from the atomic Cartesian coordinates generated by optimization under $W(\mathbf{r}, \mathbf{G})$ and addition of hydrogen atoms, we perform a geometry optimization calculation.

Summary

The above two steps, namely identification of a graph followed by determination of coordinates matching the graph, enables generation of

References

[1] S. Habershon. Sampling reactive pathways with random walks in chemical space: Applications to molecular dissociation and catalysis. *J. Chem. Phys.*, 143(9):094106, 2015.

[2] S. Habershon. Automated prediction of catalytic mechanism and rate law using graph-based reaction path sampling. *J. Chem. Theory Comput.*, 12(4):1786–1798, 2016.