Last Updated: January 10, 2017

Module 2, Part 2

Chemical Representations on Computer: Connection Tables

Learning Objectives

- Understand the principles behind connection table representation of chemical structures
- Translate structural formulas into simplified connection tables and vice-versa
- Recognize the parts of a MOL file, a common connection table file format
- Map the correspondence between features of a structural formula and entries in a MOL file
- Adjust connection tables to make simple modifications to chemical structures
- Track how changes in a chemical sketch program and the underlying connection table data relate to each other.

Contents

Connection tables are a cluster of related file formats specifically designed to render chemical structure in machine-readable form. No matter what form of input and output a computer program uses, anytime a computer does any analysis on chemical structure, it most likely makes use of connection tables. Connection tables are typically employed behind the scenes of chemical computer programs, out of the user's view. Whether you are involved in sophisticated cheminformatics analysis of chemical structure data or you just want to be able to search chemical databases confidently, it's important to have a general idea of how connection tables work.

Like structural formulas, connection tables represent the basic building blocks of structural organic chemistry: atoms and bonds. They encode information and rules about chemical structures in the form of a table that a computer algorithm can parse. A program can then use this information in many ways, to present a visual image to a user, to compare structural features among many compounds and associated data, to provide a link between chemical records in different databases, and much more. Connection tables enable chemical structure information to be programmatically accessible and are the key to most cheminformatics applications.

Introduction to Connection Tables

Connection Tables: Background

Before we dig into how connection tables work, it's useful to know a little bit about where they came from. An early version of the connection table was developed by the scientific information specialist Calvin Mooers as a purposefully *ineffective* form of chemical representation. Mooers argued that explicit tabular representations of every atom and bond in a structural formula were not very good for either machine handling (too large and difficult to parse) or communication (too difficult to read). Ironically, the development of hardware changed this calculus and the table form is quite amenable to parsing. However, it's useful to keep in mind that from the beginning, the virtue of connection tables was not their amenability to computation or to human interpretation, but the precision and explicitness with which they expressed chemical information.

Lesson 1: Connection tables were first and foremost about expressing chemically-meaningful information as extensively and explicitly as possible, which makes them ideal for informatics and other computing purposes.

Connection tables were first put to use in information systems at DuPont in the early 1960s. DuPont had **a lot** of chemicals and chemical information to keep track of across many divisions of the company. The connection table could be encoded by clerical workers rather than trained chemists, thereby saving company resources. One worker could number atoms (arbitrarily) and write down lists of atoms and bonds, and another worker could keypunch that list onto a punched card. Unlike systematic nomenclature or line notation, connection tables could be handled by workers who had little or no chemistry training.

Lesson 2: Connection tables express a lot of chemical information, but people without a lot of chemical training (but with the help of software) can still create and work with them.

Chemical Abstracts Service partnered with DuPont to develop the connection table in the mid-1960s. CAS was in the midst of developing the Registry: a computer-based index of all compounds referred to in the literature. One part of the Registry system was the Registry Number, a record number for each compound or substance reported, in any context; the other half was the

connection table, a general approach to chemical structure and relationships tied to each registry number, that related all of the compounds in the Registry structurally to one another.

Lesson 3: Connection tables were originally developed for the primary purpose of recordkeeping within a very large, but contained chemical data bank.

These three lessons are useful to keep in mind as you delve into cheminformatics:

- Don't expect connection tables to be especially elegant or compact;
- Be alert for chemically naïve uses and results;
- Watch out for divergent conventions to start popping up when you deal with connection tables outside of well-defined and curated settings.

Perhaps most importantly, connection tables represent the machine part of human-machine interactions around chemical structures. Connection tables are applied in conjunction with database ID numbers, systematic names, and line notation that are designed for managing communication with humans. In order for connection tables and these various identifiers to serve their complementary purposes as well as possible (tracking structural relationships and locating specific compounds, respectively), we need reliable interfaces to connect these two forms of chemical representation.

A Simplified Connection Table

As a starting point, this section will introduce a simplified form of connection table, which we'll call an "SCT". **This SCT does not correspond directly to any existing file format** (at least as far as we know!). Rather, it is a convenient model that we will use just for the purpose of this demonstration.

Like most connection table formats, our SCT is made up of two tables: an **atom table** and a **bond table**.

> Our ***atom*** table will consist of two fields: one an index number identifying the atom we're talking about, one indicating atom type (i.e. C, H, O, N, etc.).
>
> Our ***bond*** table will consist of three fields: two indicating the two atoms that the bond connects, and one indicating the bond order (1=single, 2=double, 3=triple).

As an example, take isopropyl alcohol. *SCT I* is a connection table representing this compound – or, more specifically, representing this structural formula.

Connection tables are not necessarily unique

We could draw up other tables of atoms and bonds that represent this compound as well: for example, *SCT II* and *SCT III*. This is an important point: **connection tables are not necessarily unique**. Different tables can represent the same chemical structure. (Of course, there are many situations in which it is useful to have a unique connection table for each chemical structure. There are algorithms for selecting such preferred or "canonical" connection tables.)

Connection tables may treat H implicitly or explicitly

Note that, in *SCT I*, only non-hydrogen atoms are specified. This follows the common practice of simply assuming that an organic compound contains as much hydrogen as the rules of valence suggest that it ought to. Sometimes, however, hydrogen atoms are explicitly included in connection tables, as in *SCT IV*. Many databases do not show hydrogens in the visual structures that you see, but some include explicit hydrogens in the underlying connection tables while others do not. Unfortunately, it can be difficult to determine what type of connection table is in use, and this confusion can lead to all sorts of trouble.

Imagine searching for compounds by the number of atoms that they contain. Should you look for isopropyl alcohol among those compounds containing 4 atoms (implicit H) or 12 atoms (explicit H)? You might even have to look out for compounds containing 5 atoms (explicit hydroxyl H and implicit alkyl H). Depending on how confident you are about how your data is structured, you might need to design your search to handle all of these cases. When you are setting up a connection table format, if you have the choice, it's probably smartest either to make all H explicit or to make all H implicit.

Connection tables might deal ambiguously with stereochemistry

For our next set of examples, we'll take a look at 2-butanol. We can form its connection table in just the same way as before (*SCT V*). However, the substituted carbon in 2-butanol is a stereocenter. We wouldn't know that from the plain SCT, though one could infer this using a clever algorithm based on the rules for naming R/S isomers. Even if we as human viewers recognize that the compound is chiral and can draw in a dashed or wedged bond, there is no way to represent this within the bare-bones SCT atom and bond tables. We would need to add an additional field to the atom and/or bond table to handle chirality (*SCT VI, VII*). We could do so either in a chemically sophisticated way, annotating the atom property, in a chemically-naive translation of a diagram feature, annotating the bond configuration, or both.

The same goes for the E/Z configuration of a carbon-carbon double-bond. (*SCT VIII, IX, X*)

Delocalization and other phenomena

Connection tables are built to represent compounds atom by atom and bond by bond. Higher-order structural features that affect chemical behavior and identity – especially phenomena like electron delocalization and tautomerism – can be difficult to capture. For instance, *SCT XI* represents the benzene ring as three single bonds and three double bonds. Of course, so does the Kekulé structure in the structural formula, but it is a challenge to find such a pattern within a table, especially since the bonds in question might not show up in consecutive order, depending on how the table is put together. Even identifying functional groups within a connection table can be a tricky proposition. We will discuss this further below and in later units of this course.

Atom coordinates

Note that the SCT atom table does not tell you anything about the relative position of atoms. (As we have seen, you often have to go to the bond table just to figure out which atom is which.)

Many connection table formats contain two- or three-dimensional spatial coordinates for each atom entry. These coordinates may simply record the relative position of atoms in a structural formula sketched in a chemical drawing program (*SCT XII*). They may also represent the calculated or measured three-dimensional positions of atoms. We will address atom coordinates in connection tables in more detail below.

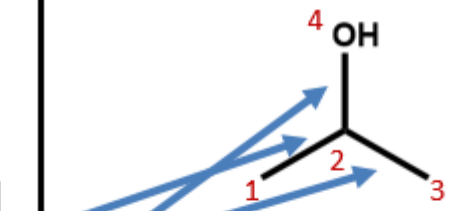
SCT I

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	O

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	1
2	4	1



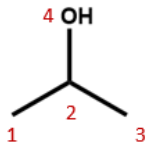
SCT I

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	O

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	1
2	4	1



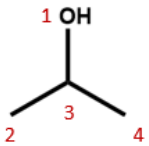
SCT II

Atom table

Atom #	Atom type
1	O
2	C
3	C
4	C

Bond table

Atom 1	Atom 2	Bond order
1	3	1
2	3	1
3	4	1



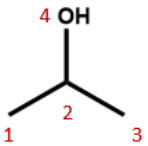
SCT III

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	O

Bond table

Atom 1	Atom 2	Bond order
2	3	1
1	2	1
4	2	1



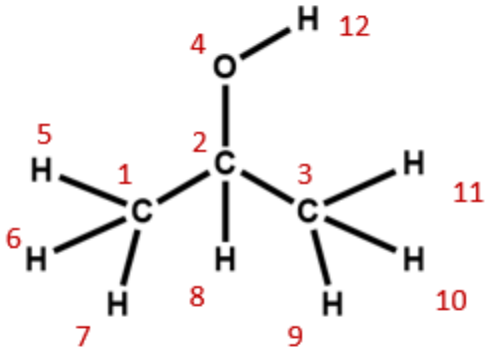
SCT IV

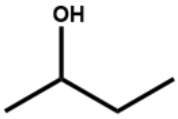
Atom table

Atom #	Atom type
1	C
2	C
3	C
4	O
5	H
6	H
7	H
8	H
9	H
10	H
11	H
12	H

Bond table

Atom 1	Atom 2	Bond order
1	3	1
2	3	1
2	4	1
1	5	1
1	6	1
1	7	1
2	8	1
3	9	1
3	10	1
3	11	1
4	12	1





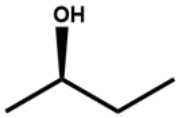
SCT V

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	C
5	O

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	1
3	4	1
2	5	1



SCT VI

Atom table

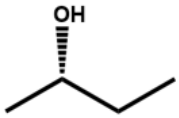
Atom #	Atom type
1	C
2	C
3	C
4	C
5	O

R

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	1
3	4	1
2	5	1

wedge



SCT VII

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	C
5	O

S

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	1
3	4	1
2	5	1

dash

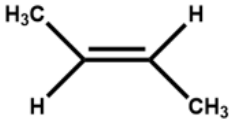
SCT VIII

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	C

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	2
3	4	1



SCT IX

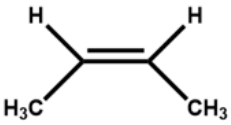
Atom table

Atom #	Atom type
1	C
2	C
3	C
4	C

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	2
3	4	1

E



SCT X

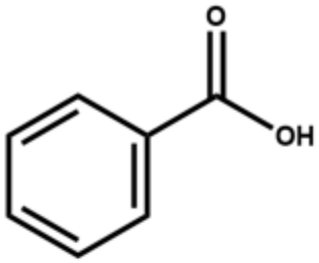
Atom table

Atom #	Atom type
1	C
2	C
3	C
4	C

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	2
3	4	1

Z



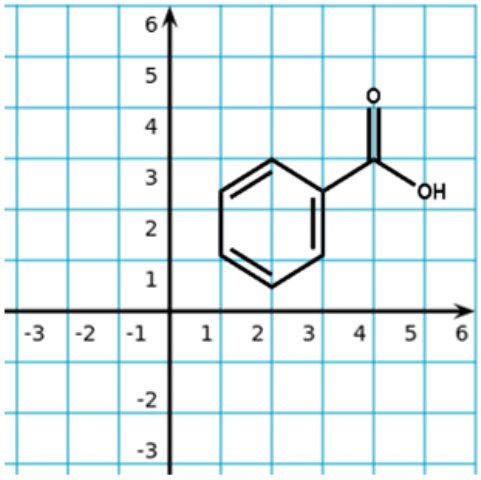
SCT XI

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	C
5	C
6	C
7	C
8	O
9	O

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	2
3	4	1
4	5	2
5	6	1
6	1	2
1	7	1
7	8	1
7	9	2



SCT XII

Atom table

Atom #	Atom type	X	Y
1	C	3	2.5
2	C	2	3
3	C	1	2.5
4	C	1	1
5	C	2	0.5
6	C	3	1
7	C	4	3
8	O	5	2.5
9	O	4	4

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	2
3	4	1
4	5	2
5	6	1
6	1	2
1	7	1
7	8	1
7	9	2

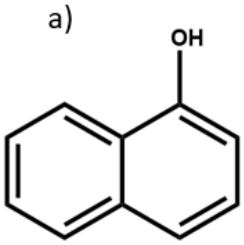
Exercises

1. Number each of the atoms in the structural formula for benzoic acid in SCT XI.

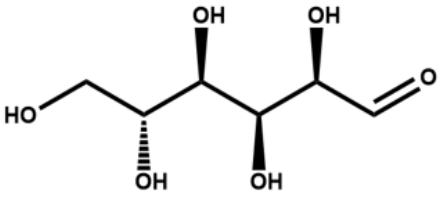
2. Write two different valid SCTs for each of structures a) - c).

2.

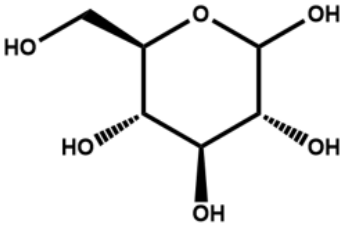
a)



b)



c)



3. Draw structural formulas for the compounds represented by SCTs a) – c).

3.

a)

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	C
5	C
6	C
7	C

Bond table

Atom 1	Atom 2	Bond order
1	5	2
2	7	1
3	5	1
4	5	1
6	2	2
3	7	2
6	1	1

b)

Atom table

Atom #	Atom type
1	O
2	C
3	C
4	C
5	C

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	5	2
3	4	2
4	1	1

c)

Atom table

Atom #	Atom type
1	C
2	C
3	O
4	O
5	C
6	N

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	1
2	4	2
1	5	1
1	6	1

4. Write a chemically-equivalent structural formula for 2a) that results in a non-equivalent SCT. Then, write that SCT.

Anatomy of a MOL file

Most connection table formats contain one or more of the following:

- A list of atoms, specifying the elemental identity of each atom
- A list of bonds, specifying the atoms that it connects and the bond multiplicity (single, double, triple)

- 2D or 3D spatial coordinates for each atom (sometimes measured, sometimes calculated; often, it's not clear which)
- Counts of the number of atoms and bonds in the molecule
- Attributes associated with atoms or bonds (e.g. R/S configuration of a stereocenter; dashed/wedged bond
- Attributes associated with an entire structure (e.g. net charge)

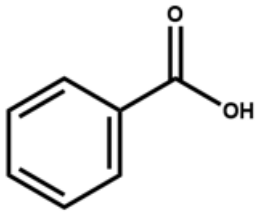
The MOL file, a widely-used chemical structure file format, contains all of these.

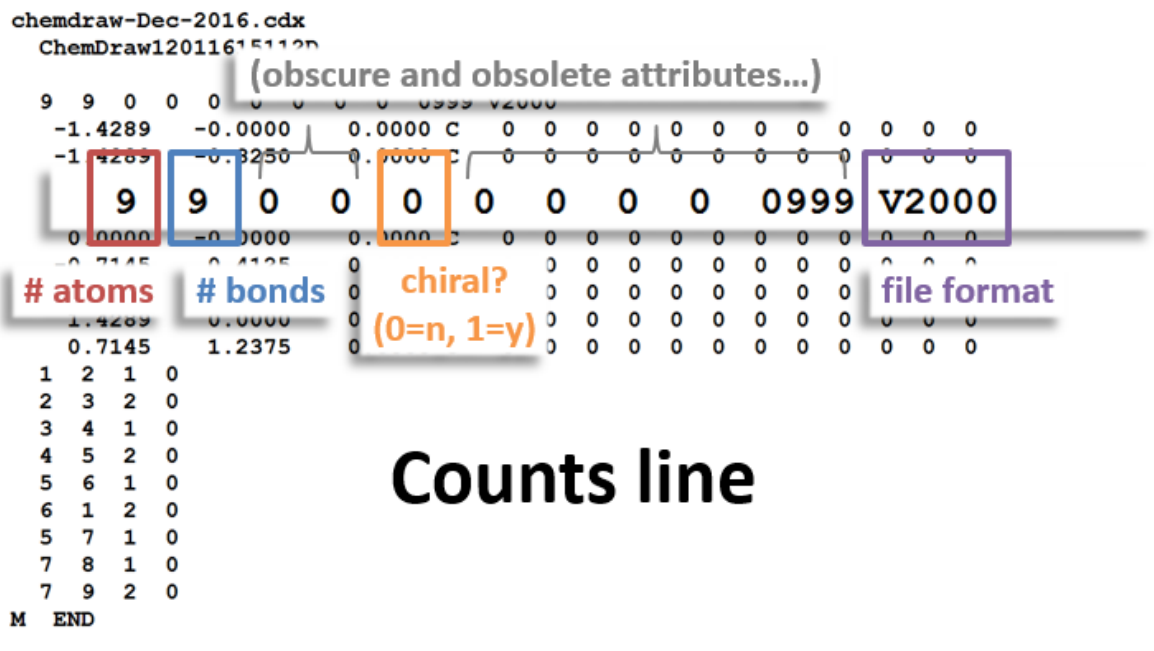
Here is a MOL file for benzoic acid, generated by ChemDraw, which provides options to save or to copy sketches in this file format. The following figures illustrate the anatomy of a MOL file (MOL v2000, to be specific): the counts line, the atoms block, the bonds block, and the properties block.

Note that benzoic acid has nine atoms and nine bonds, not counting hydrogen. If all explicit hydrogen was included in this connection table, there would be six more entries in the atoms and bonds blocks, and the counts line would show fifteen atoms and fifteen bonds.

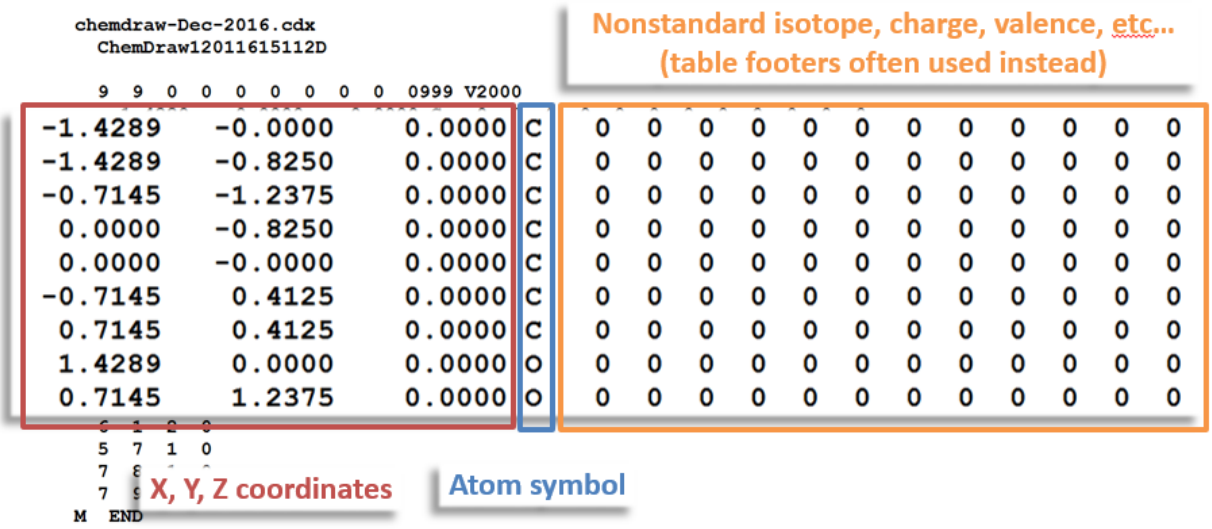
```
chemdraw-Dec-2016.cdx
  ChemDraw12011615112D

  9  9  0  0  0  0  0  0  0  0999 V2000
   -1.4289   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4289   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -1.2375    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289    0.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    1.2375    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0
  2  3  2  0
  3  4  1  0
  4  5  2  0
  5  6  1  0
  6  1  2  0
  5  7  1  0
  7  8  1  0
  7  9  2  0
M  END
```

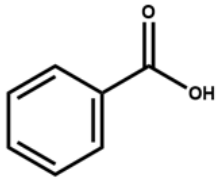


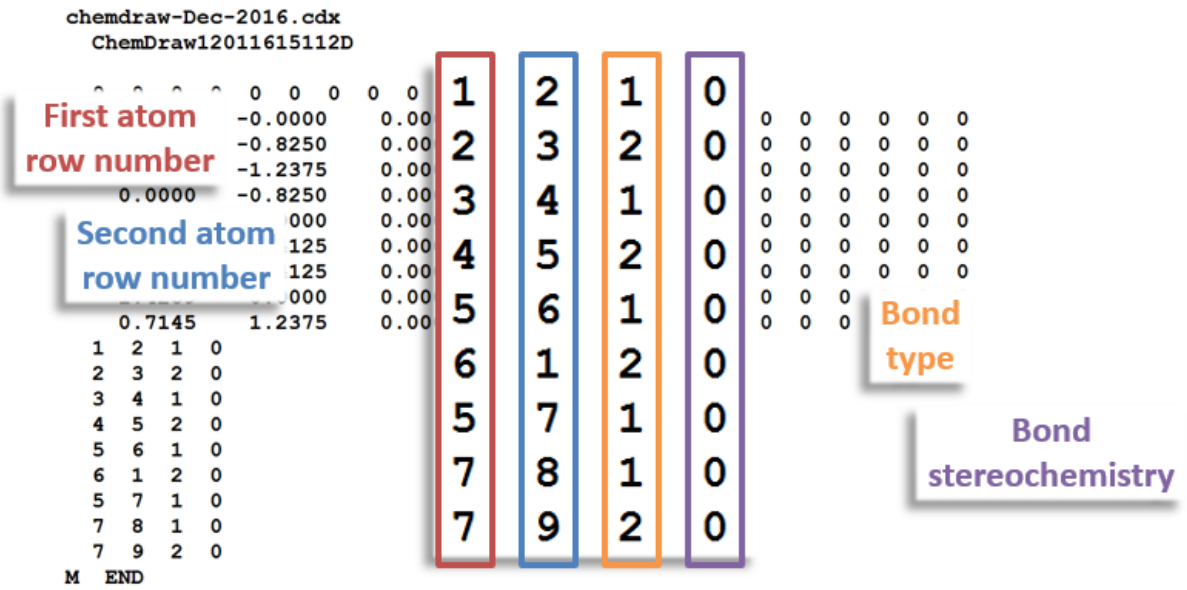


Counts line

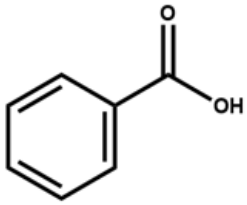


Atoms block




chemdraw-Dec-2016.cdx
ChemDraw12011615112D
0 0 0 0 0
-0.0000 0.00
-0.8250 0.00
-1.2375 0.00
0.0000 -0.8250 0.00
000 0.00
125 0.00
125 0.00
000 0.00
0.7145 1.2375 0.00
1 2 1 0
2 3 2 0
3 4 1 0
4 5 2 0
5 6 1 0
6 1 2 0
5 7 1 0
7 8 1 0
7 9 2 0
M END
First atom row number
Second atom row number
Bond type
Bond stereochemistry
1 2 1 0
2 3 2 0
3 4 1 0
4 5 2 0
5 6 1 0
6 1 2 0
5 7 1 0
7 8 1 0
7 9 2 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0
0 0 0

Bonds block


O
OH

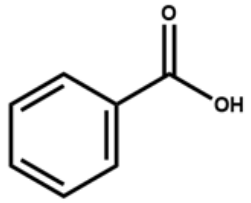
```
chemdraw-Dec-2016.cdx
  ChemDraw12011615112D

  9  9  0  0  0  0  0  0  0  0999 V2000
   -1.4289   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4289   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -1.2375    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    0.4125    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    0.4125    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289    0.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    1.2375    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0
  2  3  2  0
  3  4  1  0
  4  5  2  0
  5  6  1  0
  6  1  2  0
  5  7  1  0
  7  8  1  0
  7  9  2  0
M  END
```

M END

(Other properties specified in Properties Block footers following bond block, w/ prefix M)

END = end



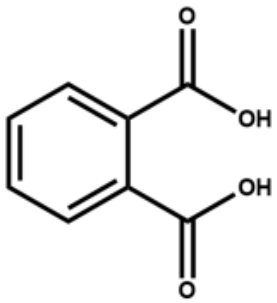
Multiple molecules

A connection table can represent multiple distinct compounds. Take a look at **MOL II**, phthalic acid. We can represent the stoichiometrically equivalent phthalic anhydride plus water by keeping the same atom block and changing a couple of entries in the bonds block. Now, we have one connection table (**MOL III**) representing two molecules. (Connection tables can also be used in representing reactions. For more on this, see online documentation on MOL and related file formats.)

MOL II

```
chemdraw-Dec-2016.cdx
  ChemDraw12231609352D

 12 12  0  0  0  0  0  0  0  0999 V2000
   -1.4289    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4289   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289    0.4125    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    1.6500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289   -0.4125    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -1.6500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0
  2  3  1  0
  3  4  2  0
  4  5  1  0
  5  6  2  0
  6  1  1  0
  5  7  1  0
  7  8  1  0
  7  9  2  0
  4 10  1  0
 10 11  1  0
 10 12  2  0
M  END
```



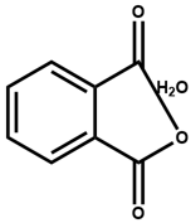
phthalic acid

MOL III

```
chemdraw-Dec-2016.cdx
  ChemDraw12231609372D

 12 12  0  0  0  0  0  0  0  0999 V2000
   -1.4289    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4289   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289    0.4125    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    1.6500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289   -0.4125    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -1.6500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0
  2  3  1  0
  3  4  2  0
  4  5  1  0
  5  6  2  0
  6  1  1  0
  5  7  1  0
  7  9  2  0
  4 10  1  0
 10 11  1  0
 10 12  2  0
  7 11  1  0
M  END
```

Same atom table!



phthalic anhydride + water

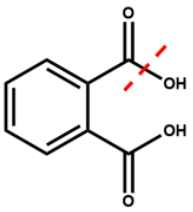
MOL II bond table

1	2	2	0
2	3	1	0
3	4	2	0
4	5	1	0
5	6	2	0
6	1	1	0
5	7	1	0
~~7~~	~~8~~	~~1~~	~~0~~
7	9	2	0
4	10	1	0
10	11	1	0
10	12	2	0

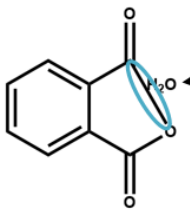
MOL III bond table

1	2	2	0
2	3	1	0
3	4	2	0
4	5	1	0
5	6	2	0
6	1	1	0
5	7	1	0
7	9	2	0
4	10	1	0
10	11	1	0
10	12	2	0
7	11	1	0

No bonds (except for implicit H) to/from atom 8, the water oxygen



Break a bond



Form a bond

Tricky features

Working with connection tables can become tricky when it comes to features of chemical identity that are not directly represented as a static collection of atoms and covalent bonds, such as:

- Aromaticity and delocalization
- Tautomerism
- Coordination

Sometimes these phenomena are not (or even cannot be) represented in the connection table at all. Other times, different file formats (or different users of the same file format) will adopt different conventions for indicating them. This can make things tricky for those who want to manipulate chemical structure data across the tens of millions of known chemical compounds (and the limitless space of possible compounds). However, it also means that there are ample opportunities for developing clever cheminformatic solutions to the limitations of connection tables.

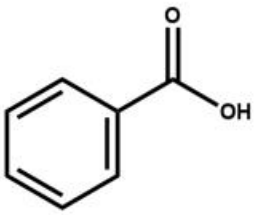
Few of these issues are likely to be solved completely. Think of the following examples, and the exercises that follow, as training in the sort of questions that you would be prudent to ask when it comes to working with digital data about chemical structures.

Aromaticity

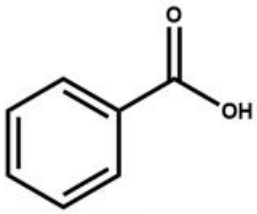
Structural formulas **I**, **IV**, and **V** all representing the same molecule: benzoic acid. However, remember: connection tables are typically correspond to structural formulas on an atom-by-atom, bond-by-bond level, not on a holistic level. Since these are three different patterns of atoms and bond, they correspond to three different MOL files. Each of the two Kekulé structures for the benzene ring shows up as a different set of single and double bonds (**MOL I**, **MOL IV**). The MOL file format uses the number 4 to indicate bonds that are explicitly labeled as aromatic (**MOL V**). This has the advantage of differentiating aromatic bonds from single and double bonds without requiring the chemist to write a script to identify and label the alternating single and double bonds of a Kekulé structure. However, some software may not be built to handle this convention. (You might even run into cases in which it's interpreted as a quadruple bond!)

Atom table for each of MOL I, MOL IV, and MOL V (same atom set!)

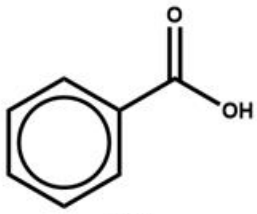
```
  9  9  0  0  0  0  0  0  0  0999 V2000
   -1.4289   -0.0000    0.0000 C   0  0  0  0
   -1.4289   -0.8250    0.0000 C   0  0  0  0
   -0.7145   -1.2375    0.0000 C   0  0  0  0   ...
    0.0000   -0.8250    0.0000 C   0  0  0  0
    0.0000   -0.0000    0.0000 C   0  0  0  0   ...
   -0.7145    0.4125    0.0000 C   0  0  0  0
    0.7145    0.4125    0.0000 C   0  0  0  0   ...
    1.4289    0.0000    0.0000 O   0  0  0  0
    0.7145    1.2375    0.0000 O   0  0  0  0
```



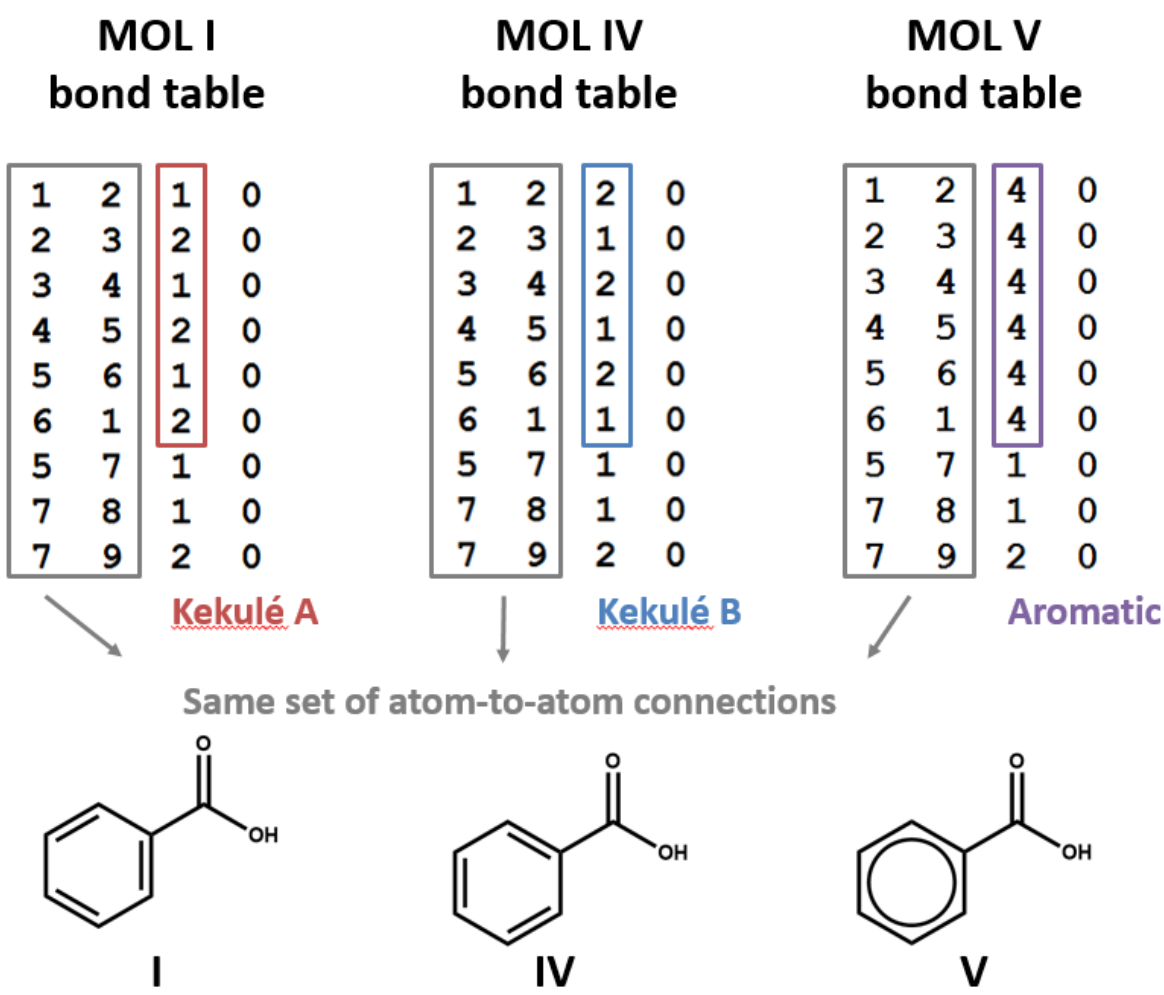
I



IV



V


MOL I
bond table
MOL IV
bond table
MOL V
bond table
1 2 1 0
2 3 2 0
3 4 1 0
4 5 2 0
5 6 1 0
6 1 2 0
5 7 1 0
7 8 1 0
7 9 2 0
1 2 2 0
2 3 1 0
3 4 2 0
4 5 1 0
5 6 2 0
6 1 1 0
5 7 1 0
7 8 1 0
7 9 2 0
1 2 4 0
2 3 4 0
3 4 4 0
4 5 4 0
5 6 4 0
6 1 4 0
5 7 1 0
7 8 1 0
7 9 2 0
Kekulé A
Kekulé B
Aromatic
Same set of atom-to-atom connections
O
OH
I
O
OH
IV
O
OH
V

Conjugate acids and bases

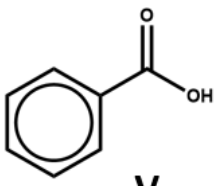
Two structural formulas may represent the same compound in different conditions. (E.g. conjugate acids/bases.) Again – keep in mind that, even though these structural formulas may refer to the same compound, they will be represented by different connection tables. You may need to choose one or the other of these connection tables / structural representations – or both – depending on your aims and the conventions of the database that you're using. (**V**, **VI**)

MOL V

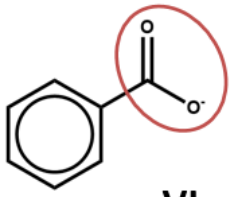
```
  9  9  0  0  0  0  0  0  0  0999 V2000
   -1.4289    0.0000    0.0000 C   0  0
   -1.4289   -0.8250    0.0000 C   0  0
   -0.7145   -1.2375    0.0000 C   0  0
    0.0000   -0.8250    0.0000 C   0  0
    0.0000    0.0000    0.0000 C   0  0
   -0.7145    0.4125    0.0000 C   0  0
    0.7145    0.4125    0.0000 C   0  0
    1.4289    0.0000    0.0000 O   0  0
    0.7145    1.2375    0.0000 O   0  0
  1  2  4  0
  2  3  4  0
  3  4  4  0
  4  5  4  0
  5  6  4  0
  6  1  4  0
  5  7  1  0
  7  8  1  0
  7  9  2  0
```

MOL VI

```
  9  9  0  0  0  0  0  0  0  0999 V2000
   -1.4289    0.0000    0.0000 C   0  0
   -1.4289   -0.8250    0.0000 C   0  0
   -0.7145   -1.2375    0.0000 C   0  0
    0.0000   -0.8250    0.0000 C   0  0
    0.0000    0.0000    0.0000 C   0  0
   -0.7145    0.4125    0.0000 C   0  0
    0.7145    0.4125    0.0000 C   0  0
    1.4289    0.0000    0.0000 O   0  5
    0.7145    1.2375    0.0000 O   0  0
  1  2  4  0
  2  3  4  0
  3  4  4  0
  4  5  4  0
  5  6  4  0
  6  1  4  0
  5  7  1  0
  7  8  1  0
  7  9  2  0
M  CHG  1   8  -1
```



V



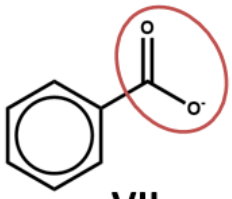
VI

Resonance

Run-of-the mill delocalization presents some of the same problems as aromaticity, but there is no conventional label for (non-aromatic) delocalized electrons, such as the delocalized negative charge and pi system in benzoate (**VII** and **VIII**). The connection tables will simply represent one resonance structure or another.

MOL VII

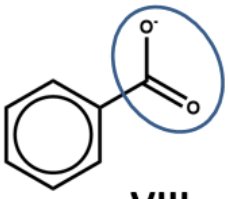
```
  9  9  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0
    0.7135    0.4119    0.0000 C   0  0
    1.4270    0.0000    0.0000 O   0  5
    0.7135    1.2358    0.0000 O   0  0
   -0.7135   -1.2358    0.0000 C   0  0
   -1.4270   -0.8239    0.0000 C   0  0
   -1.4270    0.0000    0.0000 C   0  0
   -0.7135    0.4119    0.0000 C   0  0
    0.0000   -0.8239    0.0000 C   0  0
  1  2  1  0
  2  3  1  0
  2  4  2  0
  5  6  4  0
  6  7  4  0
  7  8  4  0
  9  5  4  0
  1  8  4  0
  1  9  4  0
M  CHG  1   3  -1
```



VII

MOL VIII

```
  9  9  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0
    0.7135    0.4119    0.0000 C   0  0
    1.4270    0.0000    0.0000 O   0  0
    0.7135    1.2358    0.0000 O   0  5
   -0.7135   -1.2358    0.0000 C   0  0
   -1.4270   -0.8239    0.0000 C   0  0
   -1.4270    0.0000    0.0000 C   0  0
   -0.7135    0.4119    0.0000 C   0  0
    0.0000   -0.8239    0.0000 C   0  0
  1  2  1  0
  2  3  2  0
  2  4  1  0
  5  6  4  0
  6  7  4  0
  7  8  4  0
  9  5  4  0
  1  8  4  0
  1  9  4  0
M  CHG  1   4  -1
```



VIII

Tautomerism

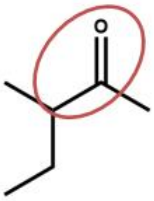
Connection tables don't link together tautomers in a straightforward way. You may need to work with multiple connection tables to account for different tautomers or to make sure that you have the most appropriate one for your purposes (**IX**, **X**).

MOL IX

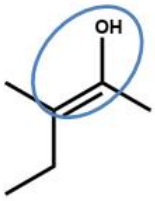
```
  7  6  0  0  0  0  0  0  0  0999 V2000
   -0.3572   -0.0000    0.0000 C   0  0
    0.3572    0.4125    0.0000 C   0  0
    0.3572    1.2375    0.0000 O   0  0
    1.0717   -0.0000    0.0000 C   0  0
   -1.0717    0.4125    0.0000 C   0  0
   -1.0717   -1.2375    0.0000 C   0  0
   -0.3572   -0.8250    0.0000 C   0  0
  1  2  1  0
  2  3  2  0
  2  4  1  0
  1  5  1  0
  6  7  1  0
  7  1  1  0
```

MOL X

```
  7  6  0  0  0  0  0  0  0  0999 V2000
   -0.3572   -0.0000    0.0000 C   0  0
    0.3572    0.4125    0.0000 C   0  0
    0.3572    1.2375    0.0000 O   0  0
    1.0717   -0.0000    0.0000 C   0  0
   -1.0717    0.4125    0.0000 C   0  0
   -1.0717   -1.2375    0.0000 C   0  0
   -0.3572   -0.8250    0.0000 C   0  0
  1  2  2  0
  2  3  1  0
  2  4  1  0
  1  5  1  0
  6  7  1  0
  7  1  1  0
```



IX



X

Chirality

MOL files do indicate chirality. However, they can do so in two ways. A "1" or "6" in the fourth field of the bonds table indicates wedged and dashed bonds, respectively. A "1" or "2" in the stereochemistry field of the atom table represents the chirality of a stereocenter. (To make things even more complicated, software may account for the chirality of a stereocenter atom when generating a MOL file but ignore it when rendering a MOL file!) (**XI**, **XII**)

MOL XI

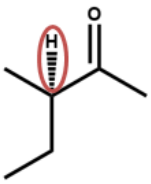
1 = "is chiral"

```
  8  7  0  0  1  0  0  0  0  0999 V2000
   -0.3572   -0.0000    0.0000 C   0  0
    0.3572    0.4125    0.0000 C   0  0
    0.3572    1.2375    0.0000 O   0  0
    1.0717   -0.0000    0.0000 C   0  0
   -1.0717    0.4125    0.0000 C   0  0
   -1.0717   -1.2375    0.0000 C   0  0
   -0.3572   -0.8250    0.0000 C   0  0
   -0.3572    0.8250    0.0000 H   0  0
  1  2  1  0
  2  3  2  0
  2  4  1  0
  1  5  1  0
  6  7  1  0
  7  1  1  0
  1  8  1  6
```

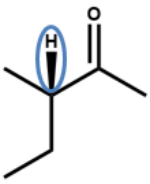
Note that counts line, atoms table, bond table reflect the explicit hydrogen drawn to indicate chirality

MOL XII

```
  8  7  0  0  1  0  0  0  0  0999 V2000
   -0.3572   -0.0000    0.0000 C   0  0
    0.3572    0.4125    0.0000 C   0  0
    0.3572    1.2375    0.0000 O   0  0
    1.0717   -0.0000    0.0000 C   0  0
   -1.0717    0.4125    0.0000 C   0  0
   -1.0717   -1.2375    0.0000 C   0  0
   -0.3572   -0.8250    0.0000 C   0  0
   -0.3572    0.8250    0.0000 H   0  0
  1  2  1  0
  2  3  2  0
  2  4  1  0
  1  5  1  0
  6  7  1  0
  7  1  1  0
  1  8  1  1
```



XI



XII

Hack-a-Mol

Here's a website where you can play with the relationship between 2D structures, 3D renderings, identifiers, and connection tables, courtesy of the cheminformatician Bob Hanson.

There's a link on the page to a document explaining "How it Works" (also linked here).

Let's take another look at benzoic acid. Clear the 2D sketch window using the white box button at the top, second from the left, and then draw benzoic acid. Click the right arrow button. That should render a 3D structure in the window to the right and generate a MOL file in the text window below. (For details on how where this data comes from, see "2D to 3D" and "3D to structure data" sections in "How it Works.")

Now, take a look at the MOL file in the text window. You will note that, as a default, Hack-a-Mol includes explicit H in the MOL files it generates. (See discussion of explicit and implicit H earlier in this module for more information.)

Identify the atoms and bonds that make up the ring. (These will vary depending on the way that you drew the molecule – the 2D sketch application numbers atoms and bonds in the order that they are drawn.) Remember, the first two columns in each bond table entry refer to rows in the atom table, and the third column gives the bond type (1=single, 2=double, etc.) connecting these two atoms. (You can check yourself by hovering over atoms in the 3D window or clicking the "labels" link above this window.)

Once you have identified the six ring bonds in the MOL file, manually adjust them to generate the other Kekulé structure of the ring. (That is, switch the 1's for 2's and the 2's for 1's in the bond type fields (third column) of the bond table entries for the six ring bonds.) With the cursor still in the text window, press enter. This should generate the other Kekulé structure for benzoic acid in both the 3D and 2D windows.

Just for kicks, let's generate a nonsense structure. Change all of the ring bonds to double bonds, and press enter. You should now have a chemically-offensive structure involving a cyclohexahexene ring with six positively charged carbon atoms violating valence rules. There's a lesson here – software won't tell you that your structure data is chemically nonsensical unless it is programmed to do so.

Revert to benzoic acid, either by changing the bonds back manually or just by clearing the 2D sketch window, re-drawing, and clicking the right arrow button again.

Now, let's stick a chlorine atom onto the benzene ring. Using the atom and bond tables, locate the atom table entry for a ring hydrogen ortho, meta, or para to the carboxyl group (your pick!). Change the atom symbol in this atom table entry from H to Cl, and press enter. You should now have the chlorobenzoic acid isomer of your choice in both 3D and 2D windows.

One more exercise: let's make our benzoic acid into pyridine-3-carboxylic acid – that is, benzoic acid with N in place of one of the ring carbons meta to the carboxylic group. This is the compound better known as niacin (vitamin B3).

(Tangential fun fact: niacin, discovered as an acidic reaction product of nicotine, was originally named nicotinic acid. In the 1930s, it was found to be the essential nutrient that prevented pellagra, a devastating disorder widely prevalent in the American South in the early twentieth century.

Public health officials promoted enriching flour with nicotinic acid, and the epidemic of pellagra began to disappear. However, physicians and scientists worried that the name "nicotinic acid" gave the impression that they were curing mass disease by putting tobacco into bread. A National Research Council committee decided to change the name of the substance to niacin, short for nicotinic acid vitamin.)

Anyway: locate the entry for a ring carbon meta to the carboxyl group. (Hint: 1) use the atom and bond tables to identify the carbon atom bonded to the two oxygen atoms; 2) find the ring carbon bonded to that carboxyl carbon; 3) find a ring carbon two bonds away from that carboxyl-substituted ring carbon.) Change that carbon to N, and press enter.

Now we have the N atom in our ring, but you will notice that it's positively charged. We didn't change any of the explicit hydrogens, so the N atom remains protonated, like the C atom that it replaced. Let's get rid of that hydrogen atom. Locate the entry for the N-H bond in the bond table and the entry for the corresponding H atom in the atom table, and delete both of them. Press enter.

Unless you were very lucky, you should now have a monstrous mess in the 3D window and nothing at all in the 2D window. Uh-oh. Go back to the MOL file window, press ctrl-Z twice to undo the deletion of those rows, and press enter. That will take you back to N-protonated niacin.

By deleting a row of the atom table, we renumbered all of the subsequent atom table entries. Since we didn't change the atom references in the bond table, this broke all of the bonds to these renumbered atoms.

Once again, delete that N-H bond from the bond table and the entry for that H atom in the atom table. However, now fix the bond table references by **decreasing the atom number by 1** for all atoms below the row that you deleted. (That is, if the hydrogen that you deleted was the 13th atom table entry, change each 14 in the first two columns of the bond table to a 13, and change each 15 in the first two columns of the bond table to a 14.)

Hit enter. Ugh – your structure is probably screwed up **again**, even if you did all of this renumbering correctly. You may even have lost your ring, for some reason.

Take a look at the counts line of the MOL file – the row above the atom table, just below the file headers. The first two numbers in this line refer to the number of atoms and bonds in the molecule. Since we deleted an atom and a bond, we need to decrease each of these from 15 to 14. Do so, and then press enter again. You should now have niacin.

Whew. **Thank goodness that connection table handling is so amenable to automation!**

Play around some more with Hack-a-Mol. Take a look at the "How it Works" page – a lot of the notations, apps, and processes referred to on this page will be covered in subsequent weeks. You may find it useful to continue to come back to this page and play around with it as you move on in this course.

Further Reading

1. https://en.wikipedia.org/wiki/Chemical_table_file
2. CTFile Formats, June 2005, Elsevier/MDL, https://web.archive.org/web/20070630061308/http://www.mdl.com/downloads/public/ctfile/ctfile.pdf (Documentation for v2000 MOL file and related chemical table file formats.)
3. Hack-a-Mol: https://chemapps.stolaf.edu/jmol/jsmol/hackamol.htm (Documentation: https://chemapps.stolaf.edu/jmol/docs/misc/hackamolworkings.pdf)

Exercises

1. Does Hack-A-Mol handle the number 4 for an aromatic bond? How can you tell? Can you create a chemically sound but non-aromatic structure using 4s in the bond field?

2. Perfluorinated octanoic acid (PFOA) is a surfactant that played a key role for a long time in the manufacture of fluorinated polymers including Teflon. Over the past decade, it has been the subject of significant public health concern and a whole bunch of litigation.

Pull PFOA into Hack-a-Mol by typing it into the text search box below the 3D window and clicking "search."

2a. Edit the mole file to defluorinate PFOA, converting it into octanoic acid.

2b. Now make it into acetic acid. (It is possible to do this in a way that yields correct-looking 2D and 3D renderings without changing any XYZ coordinately, but you have to be ***very*** careful about how you delete and relabel atoms and bonds.)