

Module 4a: Communicating chemical structure with formulas and names

Evan Hepler-Smith, Princeton University
Leah R. McEwen, Cornell University

Learning Objectives:

- To *recognize* various different kinds of chemical names, formulas, and other identifiers.
- What you *do and do not know* about a chemical compound based on one of these names, formulas, or identifiers.
- How one kind of chemical name, formula, or other identifier can be *translated* into another, and what sorts of information can be inadvertently lost or added in translation.
- How chemists *interpret* various kinds of chemical names, formulas, and other identifiers in chemically meaningful ways.

Table of Contents

[1.0 Communicating chemical structure with formulas and names](#)

Overview

Definitions

[1.1. Formulas](#)

[1.2. Names](#)

[1.3. Further Reading](#)

1.4. Chemical structure drawing programs

1.5. Exercises

Overview

Chemistry involves a lot of communication. In the classroom, in the laboratory, or at the computer screen, as a chemist, you are constantly referring to all sorts of different chemical substances and molecular entities. You do so using chemical *names*, *formulas*, and *notation*. You're probably already so accustomed to chemical names, formulas, and notation that you barely need to think about them when you use them, and can instead focus on the molecules that you're drawing, writing, or talking about. In this module, we're going to turn things around and think about chemical names, formulas, and notation themselves.

Why would we want to do that?

Where there's communication, there's always a danger of misunderstanding. Experienced human chemists are generally able to figure out when they've misunderstood each other over the identity of a particular compound. However, work in cheminformatics almost always involves communicating not just with other chemists, but with computer systems. Often, it also involves different computer systems communicating with each other. In these cases, it's often easier for miscommunication to go undetected. When it *is* detected, it's often difficult to figure out what went wrong.

You can minimize the impact of this kind of miscommunication by keeping in mind what various sorts of chemical names and formulas DO and DO NOT tell you about a particular compound, and by documenting the sources of the names and formulas that you use.

In **Part 1** of this module, we will dig into the most common kinds of chemical names, formulas, and notation to figure out a) how they work, b) why they work like they do, c) where they are most often used, and d) what they do and do not tell you about a chemical structure.

In **Part 2**, we'll introduce several chemical identifiers and representations developed specifically for use on computers.

Later modules of this course will focus on how these various sorts of identifiers are used in cheminformatics applications. In this module, we'll focus on the communications tasks that almost all chemists engage in. A convenient mnemonic for these tasks is "RSVP": Register, Search, View, Publish. Most forms of chemical representation were developed with these uses in mind.

(A quick note to reassure you before we dive in: **we're not going to be memorizing any nomenclature rules**. Systematic chemical nomenclature has become so complicated that even experts in the field use computer systems to review their work and catch their mistakes. In Part 2, we'll talk a little bit about how this has happened, since it will help you understand how to deal with some of the challenges that might come up when you have to deal with systematic chemical names in your own work.)

The ability to communicate effectively using chemical names, formulas, and notation is a kind of literacy. As with regular literacy, this chemical literacy is something that you will get better at with practice. The better you understand what's going on "under the hood" of various forms of chemical representation and the computer systems that make use of them, the better a chemical communicator you will become.

1.0.1. Definitions

Chemical identifiers and representations

There are lots of different kinds of chemical names and formulas. Confusingly, many of the terms that refer to them can be used in different ways.

Instead of trying to specify a single, unambiguous meaning for each term, we're going to lay out the various different things that people might mean when they're talking about, for example, an "empirical formula."

Formulas

A **structural formula** is any formula that indicates the connectivity of a compound – that is, which of its atoms are linked to each other by covalent bonds. There are various different kinds of structural formulas:

A **line formula** depicts connectivity but no three-dimensional structural information.

A **condensed formula** expresses the same information as a line formula using atomic symbols only.

A **Lewis formula** explicitly shows valence lone pairs in addition to bonds.

A **skeletal formula** is a simplified line formula in which carbon atoms are depicted as unlabeled vertices and hydrogen atoms bonded to carbon are suppressed. Skeletal formulas are the most common structural formulas.

Dash-wedge formulas use dashes and wedges to represent stereochemistry at sp^3 stereocenters.

Projection formulas indicate conformation.

These different ways of drawing structural formulas are often combined or used alongside one another, sometimes in different parts of the same formula. For this reason, it's not especially important or useful to memorize these terms and their definitions. Rather, you need to be able to interpret the kind of information that each of these formulas expresses. We'll discuss this in more detail below.

Empirical and molecular formulas indicate the composition of a compound only:

An **empirical formula** expresses the ratio of the elements (or sometimes polyatomic ions) that make up a compound, in lowest integer terms.

A ***molecular formula*** indicates the total number of atoms of each element in one molecule of a compound.

Names

A ***systematic name*** is a chemical name based on the structural formula of a compound. If you know the rules and vocabulary of the system in question, you should be able to write a name based on a structural formula and vice-versa. Chemists have developed various ways of translating formulas into names, so it is nearly always possible to write more than one systematic name for a given compound.

Locants and ***stereochemical descriptors*** are numbers, letters (such as R, S, E, and Z), and prefixes (cis, trans) that indicate how the molecular fragments indicated by different parts of a systematic name fit together in the named compound.

A ***trivial name*** is a relatively short, memorable name that identifies a chemical entity without describing its structure.

IUPAC nomenclature is a well-known international system of chemical names. In general, IUPAC nomenclature is systematic but flexible, offering several ways of writing a systematic name for any given compound. IUPAC nomenclature rules also allow the use of certain well-established trivial names as IUPAC names.

A ***preferred IUPAC name (PIN)*** is one of the possible IUPAC names for a compound, singled out as the name to be used in official contexts such as regulation.

Notation

Line notation expresses the structure of a compound using a string of characters. Line notation is designed to be easy for computers to process rapidly and reliably (and is usually not particularly legible to people). Currently, the most commonly used forms of line notation are ***SMILES/SMARTS*** and ***InChI***.

Registry numbers are unique identifiers for chemical substances. They are designed not to give you any information whatsoever about a compound's structure or its relationships to other compounds.

CAS Registry Numbers (CAS RNs) are the registry numbers used in the Chemical Abstracts Service Chemical Substance Registry, a major chemical database that can be searched with CAS applications including SciFinder and STN. They have often been used as official identifiers for chemical substances, especially in the US.

A **connection table** is a table listing all of the atoms and bonds in a molecule. It is the most common format used by computer programs to store, search, compare, and sort chemical structures. Connection tables are even harder for humans to read than line notation.

The **MDL Molfile (.mol file)** is a widely-used file format for connection tables.

1.1. Formulas

1.1.1. Structural formulas

“The purpose of a chemical structure diagram,” begins an article on how to draw these diagrams, “is to convey information—typically the identity of a molecule—to another human reader or as input to a computer program. Any form of communication, however, requires that all participants understand each other.”[\[1\]](#)

Below, we’ll go over the various ways in which structural formulas are most often drawn. Once again, our goal is to get you thinking about the kinds of structural formulas that you’ve gotten used to using without having to think too much about them. What could you possibly be misunderstanding in someone else’s structural formula? How could somebody misunderstand your structural formula? Is there a chemical feature in your head that didn’t make it into the formula that you drew? Is there *more* in the formula that you drew than you meant to express?

(We’ll be going over the ways in which formulas can be drawn. If you are interested in learning more about how formulas *should* be drawn, and in sharpening your own formula-drawing, we highly recommend checking out [this detailed guide](#). Here’s what you’ll find:

Production of good chemical structure depictions will likely always remain something of an art form. There are few cases where it can be said that a specific representation is “right” and that all others are “wrong”. These guidelines do not try to do that. Rather, they try to codify the sorts of general rules that most chemists understand intuitively but that have never been collected in a single printed document. Adherence to these guidelines should help produce drawings that are likely to be interpreted the same way by most chemists and, as importantly, that most chemists feel are “good-looking” diagrams.[\[2\]](#)

When chemists talk about “structure,” what do they mean? Chemical structure can mean several different things:

- **Connectivity** (also known as *constitution*): which atoms are linked to which by covalent bonds?
- **Stereochemistry**: what is the relative arrangement of these atoms and bonds in three-dimensional space? Are two groups across a double bond or ring *cis* or *trans* to each other? Is a stereocenter **R** or **S**?

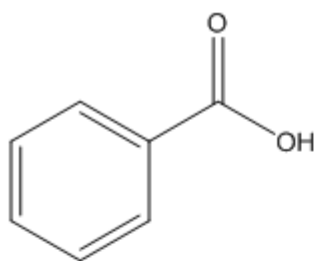
- *Conformation*: in which of the many configurations permitted by rotation around single bonds are all of the atoms of a compound arranged in space?
- *Crystal structure*: what is the precise position of each atom in the compound, in three-dimensional coordinates?

Structural formulas always express connectivity and often express stereochemistry. Both of these aspects of structure can usually be translated in a relatively straightforward way between different chemical formulas and names.

While structural formulas may also contain information about conformation, it is often more difficult to translate conformation from one formula to another or to a name. And while structural formulas may be drawn to suggest the shape of a molecule, they almost never contain reliable information about crystal structure.

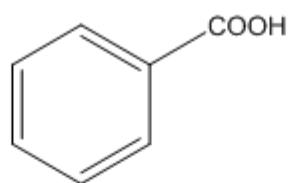
1.1.1.1. How do they work?

To get us started, here are some structural formulas:



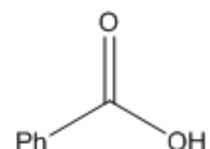
benzoic acid

LA



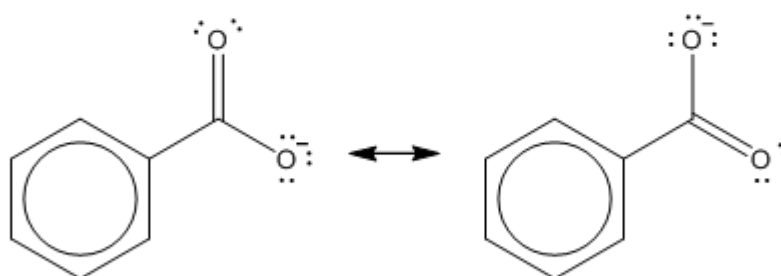
benzoic acid

LB



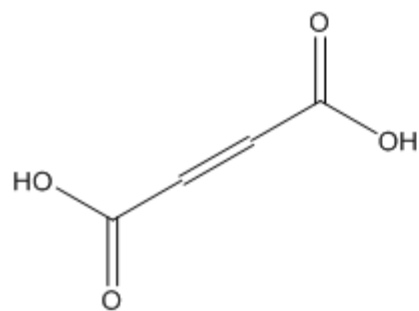
benzoic acid

LC



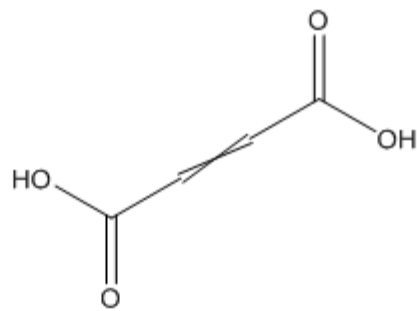
benzoate
(deprotonated benzoic acid)

LD



but-2-enedioic acid

II.A



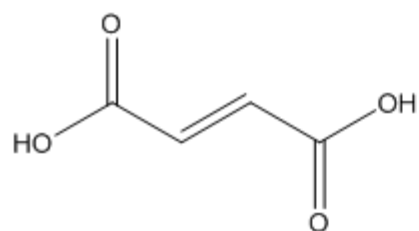
but-2-enedioic acid

II.B



but-2-enedioic acid

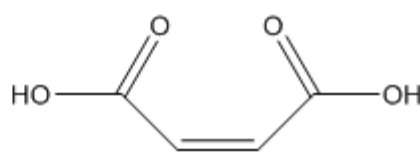
II.C



fumaric acid

(trans-but-2-enedioic acid)

II.D

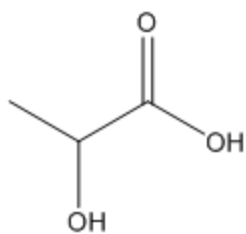


maleic acid

(cis-but-2-enedioic acid)

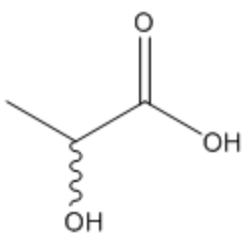
II.E

lactic acid



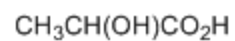
2-hydroxypropanoic acid

III.A



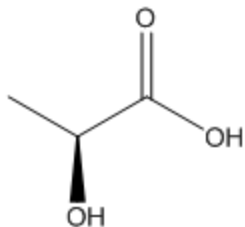
2-hydroxypropanoic acid

III.B



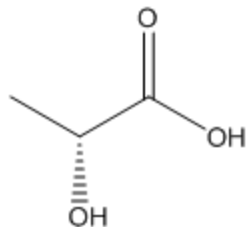
2-hydroxypropanoic acid

III.C



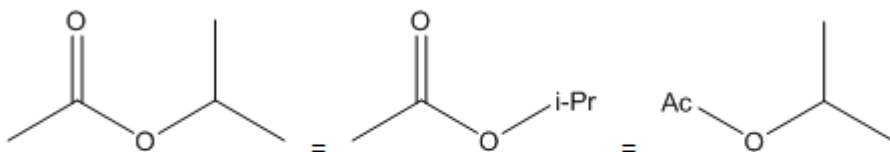
(*S*)-2-hydroxypropanoic acid

III.D



(*R*)-2-hydroxypropanoic acid

III.E

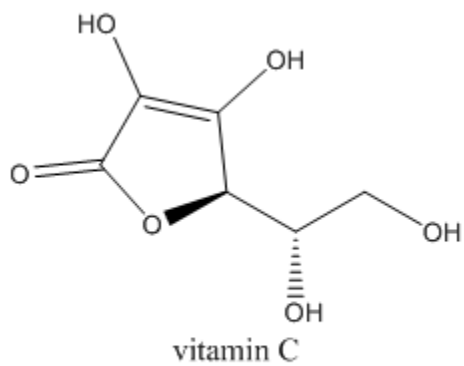


IV.A

IV.B

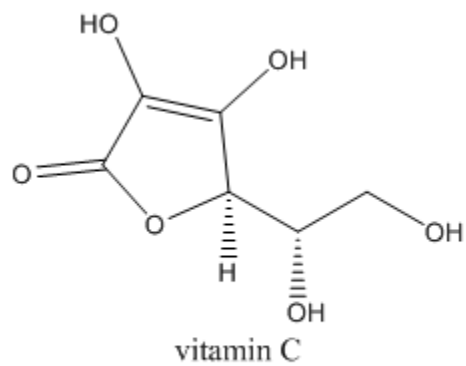
IV.C

isopropyl acetate

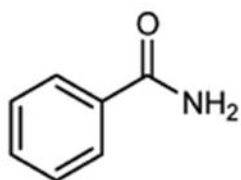


V.A

=



V.B

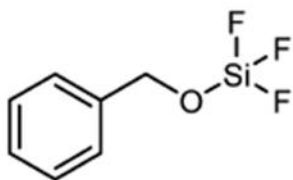


VI.A

=

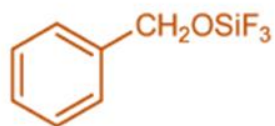


VI.B

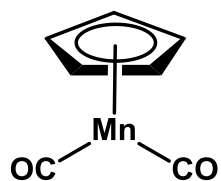


VI.C

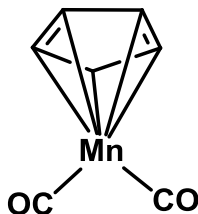
=



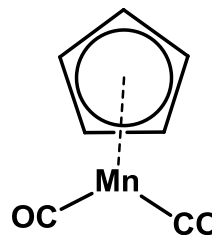
VI.D



VII.A



VII.B



VII.C

Compressing structural formulas: skeletal formulas, condensed formulas, and abbreviations.

In order to draw structural formulas more quickly and clearly, chemists typically draw carbon atoms as unlabeled vertices. We also typically leave out lone pairs and hydrogen atoms bonded to carbon. Formulas drawn in this way are sometimes called *skeletal formulas*.

Even skeletal formulas can take up a lot of space, and sometimes, you're only really interested in the structure of one part of a molecule.

Structural formulas often make use of abbreviations for common molecular subunits: *i*-Pr (isopropyl), Ph (phenyl), Me (methyl), Et (ethyl), Bu (butyl), *t*-Bu (*tert*-butyl), Ac (acetyl), among others. ([Here's a list.](#)) (I.C, IV.B, IV.C)

In order to abbreviate structural formulas even more, *condensed formulas* express structure without using any lines. A condensed formula can be written in place of an entire structural formula (II.C, III.C) or in place of a portion of a structural formula (I.B, VI. A-D).

These condensed formulas-within-a-structural formula are sometimes called "contracted atom labels." IUPAC guidelines for graphical representation provide the following specifications for how to write and interpret contracted labels:

Contracted atom labels attached to only one bond should be read outwards from that bond, usually from left to right if the bond is on the left of the label. If the bond is instead attached to the right of the label, the label will normally be read from right to left (313-14).

Parentheses are used when more than two non-hydrogen atoms are bonded to the same atom (e.g., branching; III.C).

The advantage of condensed formulas is that they can be written in normal type. However, it is often more difficult to perceive structural features in a condensed formula than in one of the graphical alternatives.

Stereochemistry

Structural formulas typically indicate cis-trans isomerism across double-bonds (**II.D, II.E**).

Structural formulas are sometimes drawn in a way that keeps this ambiguous. A crossed double bond and/or a double bond aligned linearly with its neighboring single bonds indicates unknown cis-trans configuration (**II.A, II.B**).

The configuration of chiral centers is shown using dashes and wedges. (**III.D, III.E**)

A wavy line explicitly indicates unspecified stereochemistry or a mixture of stereoisomers (**III.B**). A chiral carbon with only regular bond lines, on the other hand, could indicate that the chemist who drew the formula just didn't notice the stereocenter (**III.A**).

Condensed formulas usually do not show stereochemistry (**II.C, III.C**).

[Here's a little more on stereochemistry in structural formulas.](#)

Delocalization

Delocalization may be drawn via resonance structures (**I.D**), circles within aromatic rings (**I.D**, also). Dashed or dotted double bonds are also sometimes used to show delocalization. In some contexts, these can be confusing, since dotted and dashed bonds are also used to depict transition states, coordination relationships, hydrogen bonds, and other bonds that behave differently than covalent sigma and pi bonds.

Many chemistry databases index by structural formulas based on explicit connectivity for organic small molecules. However molecules such as coordination compounds and other delocalized systems do not fit easily into these conventions for representing bonds.

VII.A is the standard IUPAC graphical representation for publication, however, it is complex for software to interpret all the specialized notation, such as the wedged bonds and the bond into the middle of the ring. A human chemist would understand that this indicates a general relationship between the metal and the delocalized system. A computer program, however, might interpret this as a bond to a methyl group, unaffiliated with the ring.

VII.B is a more common representation for searching coordination compounds in chemistry indexes, with explicit bonding between the metal and all ring atoms. However, this is incorrect notation for publication as the association between the metal and the ring atoms are not covalent bonds.

VII.C is considered an acceptable alternative in the IUPAC standards. Not all databases will have provisions to interpret the circular bond notation or the dotted bond to confer a non-standard covalent system.

Delocalization is difficult to program, and almost all software applications do it differently. It is important to keep in mind the purpose for the formula, for human or computer readership.

1.1.1.2. Why do they work that way?

Structural formulas tell you a lot more about the atoms that make up a compound than its valence electrons. Bonds represent electrons, of course, and you can draw in lone pairs. You can add curved arrows to show electron movement, draw resonance structures or dotted bonds to show delocalization, and sketch in orbitals, of course. But in structural formulas themselves, as usually drawn, the electrons are mostly implicit – you know they're there, but they aren't actually what the drawing depicts.

This is peculiar, since the majority of chemical phenomena depend on interactions involving valence electrons.

There's a reason for this. Chemists began using structural formulas a hundred and fifty years ago, after they'd figured out the basic features of organic chemical structure (carbon atoms form chains, carbon forms four bonds, etc.) but *before* things like cis-trans isomerism, tetrahedral carbon, and even the electron itself had even been hypothesized.

By the time electrons and stereochemistry came along, structural formulas had come into general use, and chemists were quite familiar with them and fond of them. So they kept on using the same formulas, even though they'd been developed without electrons or stereochemistry in mind.

Eventually, chemists developed additional bits of notation – electron dots, dashes and wedges, and the like – to incorporate the electronic theory of bonding and stereochemistry into these familiar formulas. But even though electrons and stereochemical relationships became absolutely central features of how chemists think, it has always been a little bit difficult to represent them in structural formulas. It's just not what structural formulas were built for.

Of course, structural formulas continued and continue to be enormously productive ways of representing compounds. Chemists have learned to think of these formulas as expressions of contemporary chemical ideas. However, in some situations – cheminformatics among them – we sometimes run into an awkward disconnection between, on the one hand, the historical origin of structural formulas as maps of connections between atoms, and on the other hand, our present-day scientific understanding of the nature of chemical substances.

One more point: structural formulas were originally developed within the context of organic chemistry, and then applied in other fields such as coordination chemistry. Be aware that both people and computer programs will tend to assume, as a default, that structural formulas represent covalently bonded organic compounds. If you are working with structural formulas for complexes involving coordination or hydrogen bonding, make sure that these bonds aren't accidentally mistaken for the covalent bonds of organic compounds. (Some suggestions on how to avoid this pitfall are available on pages 292–295 [here](#).)

1.1.1.3. Where they are most often used?

Everywhere in which you're able to draw diagrams. Unfortunately, this excludes a lot of places, such as word processing programs, free-text search boxes, databases, and anytime you find yourself talking chemistry without a notepad in your pocket.

There's an easy solution for the last of these cases (keep a notepad in your pocket!); for the others, the solution is systematic nomenclature and notation, which we will discuss in the next two units.

One particularly useful feature of structural formulas is that you can easily draw a structural formula for a section of a molecule or identify one molecule as a section of another. Many applications for searching chemical databases (such as SciFinder and [PubChem](#)) allow you to perform substructure searches (for all molecules containing a certain structural formula subunit) and superstructure searches (for all molecules whose structural formulas can be found within a certain structural formula).

1.1.1.4. What questions should I ask?

Are we showing any implicit H's or lone pairs? Are we worried about the ones we aren't showing?

One or more H's can be drawn in when there's chemistry happening at an H, or if you want to indicate the configuration of a stereocenter (**V.B**). The same goes for lone pairs, when you have reason to call attention to them (**I.D**).

When you look at a skeletal formula, you know that all of the hydrogen atoms and valence lone pairs that you would expect to be there are in fact present, even though they aren't drawn in. Keep this in mind if you find yourself communicating with a human or a computer that you can't count on to fill in those missing H's and electrons.

How are we dealing with stereochemistry?

Structural formulas can specify stereochemistry (**II.D-E**, **III.D-E**, **V.A-B**) or leave it unspecified (**II.A-C**, **III.A-C**). In the latter case, it is typically impossible to tell from the structural formula alone whether you're dealing with a mixture of stereoisomers or unknown stereochemistry.

If you're concerned about stereochemistry – and in most cases, you probably are – be alert for stereocenters (including rings with multiple substituents) with unspecified stereochemistry.

Watch out for double-bonds just drawn on top of single bonds without considering cis-trans isomerism (and don't make this mistake yourself!).

Note that there is a chemical difference between substances of unspecified stereochemistry and mixtures of stereoisomers. (“What’s the stereochemistry? I don’t know!” vs. “What’s the stereochemistry? We’ve got both isomers!,” respectively.) However, when you’re dealing with a structural formula drawn without stereochemistry specified, it can be difficult to know which of these cases you’re dealing with.

How are we dealing with delocalization?

If there’s a delocalized π system in your molecule, think about whether you’ve chosen the appropriate resonance form, or whether it’s worth drawing multiple forms or indicating delocalization with a dotted bond.

How are we dealing with tautomers?

If your compound can tautomerize, think about whether you’ve chosen the appropriate tautomer for your purposes, whether it’s worth drawing both.

Keep in mind that both tautomerism and delocalization are much easier to recognize when you’re working with structural formulas than when you’re working with systematic names or other sorts of notation. (Delocalization is very difficult even to represent using any other form of name or notation.) When you translate structural formulas into another form, make sure delocalization and tautomerism don’t get lost in the shuffle.

1.1.2. Empirical and molecular formulas

You don’t always know, or need to express, or want to express the structure of a compound that you’re working with. In the case of inorganic salts, there’s little or no molecular structure (connectivity, that is) to represent.

In these cases, empirical and molecular formulas give you a way to identify the compound by its composition alone. And if you’re interested in a compound’s composition for its own sake, better to write down a molecular formula than keep counting each atom in a structural formula.

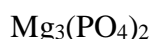
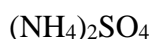
1.1.2.1. How do they work?

Empirical and molecular formulas are pretty straightforward: you just count the atoms or the ions.

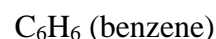
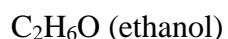
Empirical formulas are most often used to identify [salts](#). Empirical formulas typically express the relative [amount](#) of each element that the compound contains, in lowest integer terms.



Salts containing polyatomic ions are frequently represented with a formula expressing the relative amount of each *ion* that the compound contains, in lowest integer terms. Such formulas sometimes just referred to as “chemical formulas” and sometimes as empirical formulas.



To write a molecule formula, just count the atoms in one molecule of the compound.



Salts containing polyatomic ions are sometimes represented by a “molecular formula” expressing the total number of atoms of each element that are present when the ions combine in lowest integer terms.



1.1.2.2. Why do they work that way?

Empirical and molecular formulas predate structural formulas, but they actually became more important, not less, once structural formulas appeared on the scene. This was because:

Looking up chemical compounds was hard.

If you knew the structural formula for a compound and wanted to look it up in a big chemical dictionary, it was usually pretty easy to find it if the dictionary was organized by molecular formula. That way, you only had to look through the names for the couple dozen isomers that shared a molecular formula.

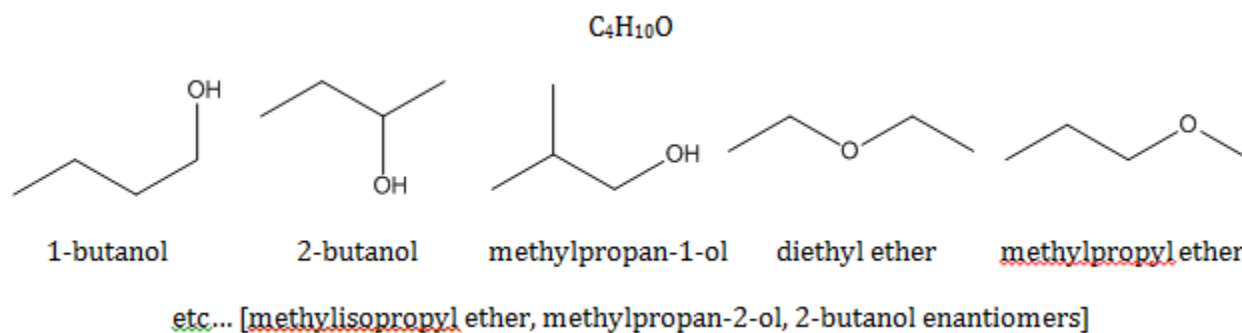
Sometimes chemists were wrong about their structure determinations.

It was helpful to have a formula that you could go on using unchanged if it turned out that the double bond wasn't where you thought it was, for multiple tautomeric forms of a compound

It's useful to think of a molecular formula as a general label for a compound rather a specific one, especially when you're dealing with an organic small molecule that almost certainly has a bunch of isomers.

That is: don't think to yourself "diethyl ether is $C_4H_{10}O$ " but rather "diethyl ether is one of the things in the $C_4H_{10}O$ box, along with 1-butanol, 2-butanol, etc."

$C_4H_{10}O$



1.1.2.3. Where they are most often used?

Empirical formulas are often used to represent the empirically-determined composition of an unknown sample.

Molecular formulas are most often used to identify molecular entities (organic compounds, covalently bonded inorganic compounds, coordination complexes) and their salts.

They usually show up in database entries for compounds, so you can use them to search for compounds (particularly useful if you suspect that tautomers might be throwing off your search).

You can also use them as a check to reveal if you

1.1.2.4. What questions should I ask?

Are we grouping atoms into ions or just listing them element by element?

How are we ordering the atoms?

The two most common ways in which empirical and molecular formulas are ordered are:

- From electropositive to electronegative / cation to anion

NaCl CaCO₃ AlCl₃ Fe₂O₃

SO₂ H₂O

NH₄C₂H₃O₂ NH₄NO₂

Exceptions:

NH₃

- In the order C, then H, then everything else, in alphabetical order. (This is sometimes called [Hill system order](#).)

ClNa CCaO₃ AlCl₃ Fe₂O₃

O₂S H₂O

C₂H₇NO₂ H₄N₂O₂

H₃N

Is this an empirical formula (a ratio of lowest terms) or a molecular formula (the total count of atoms in a particular chemical structure)?

C₂H₂ (ethylene) C₆H₆ (benzene)

...or

CH (ethylene and benzene)

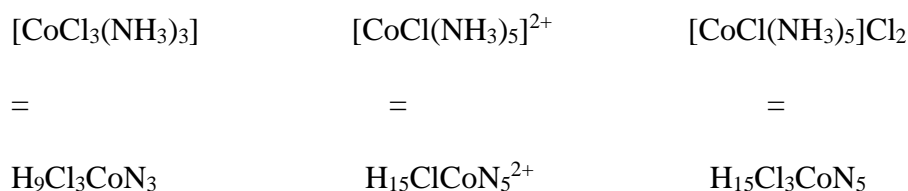
Are we referring to a specific isomer, and how do we know which one?

This almost goes without saying: for organic compounds and many inorganic ones, there are almost always a bunch of isomers that share the same molecular formula.

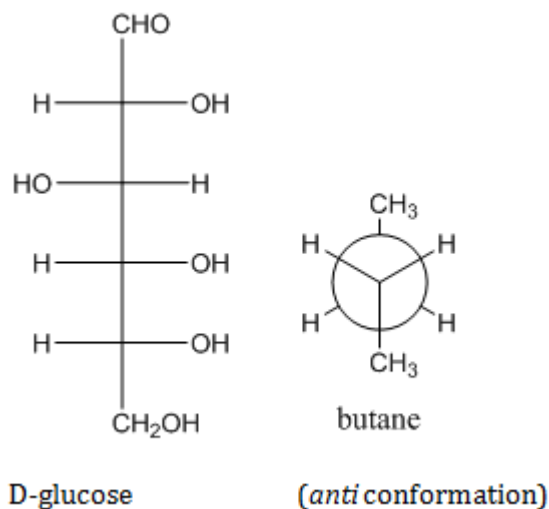
1.1.3. Other kinds of compounds and formulas

As we mentioned above, most of the general principles behind chemical names and formulas were originally worked out for organic compounds and were later adapted to other sorts of chemical entities.

Molecular formulas for coordination complexes are often written in brackets, in the order [central atom (usually a metal), then negative ligands, then neutral ligands]. They may also be written in Hill system order.



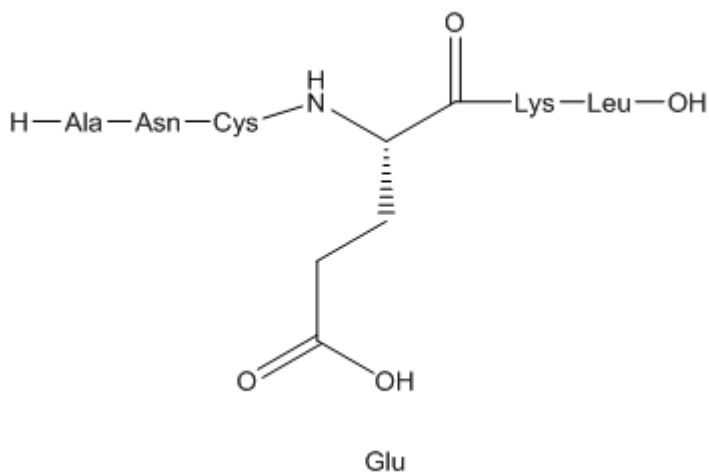
Projection formulas indicate stereochemistry or relative conformation.



The sequence of a fragment of biological polymer (a polypeptide or nucleic acid) is similar to a condensed formula, since it represents a linear chain of chemical units.



You may come across formulas in which one of these units is expanded.



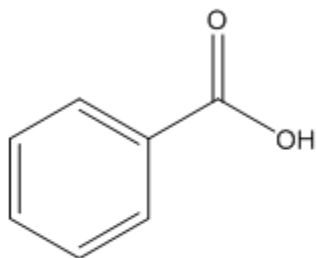
1.2. Unit 2: Names

1.2.1. How do they work?

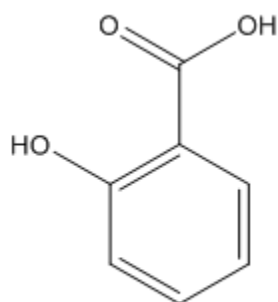
There are two kinds of chemical name: trivial names and systematic names. *Trivial names* identify a compound (or sometimes a few closely related compounds), but provide little or no information about its structure and its relationships to other compounds. A trivial name may be a technical chemical term, or it may be a common name taken from regular, nonscientific language. You can think of acronyms for systematic names (THF, DMSO, and so forth) as a kind of trivial name.

Systematic names indicate the complete constitution of the compound. Systematic names are based on structural formulas. Writing a systematic name involves taking apart a structural formula into subunits, finding the appropriate term for each subunit, and putting those terms together to form the name. You should therefore be able to draw a structural formula for a compound based on its systematic name, by taking apart the name into its subunits and writing down the structural formulas for each of these subunits, connecting them as specified in the name.

Trivial:



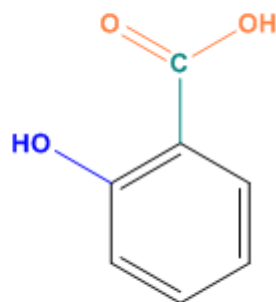
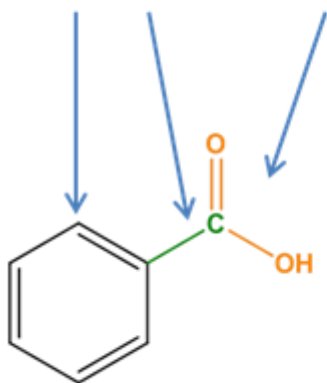
benzoic acid



salicylic acid

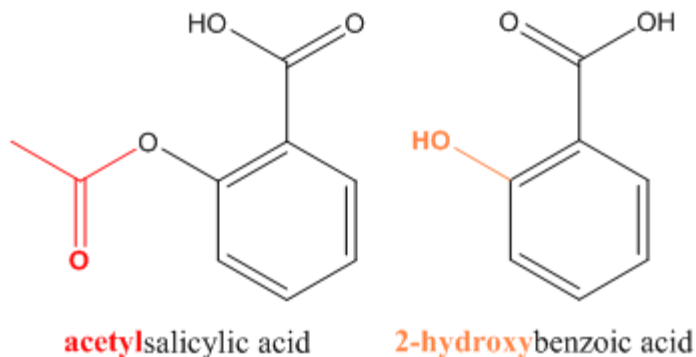
Systematic:

benzenemethanoic acid



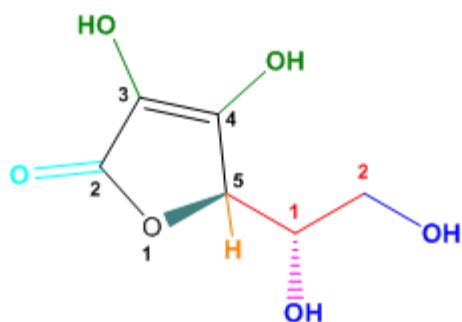
2-hydroxybenzenemethanoic acid

Semi-systematic names take a trivial name of a related compound as a root and name the compound systematically as a derivative of that compound.



The root of a systematic name indicates the compound's primary chain or parent compound, and prefixes and suffixes indicate the atoms or groups that are attached to that parent compound.

Locant numbers (and occasionally letters) indicate where these substituent groups are attached to the parent compound (or "substituted" for hydrogen atoms of this parent, hence the term "substituent.") Stereochemical prefixes – cis and trans, E and Z, R and S – are used to indicate stereochemistry. (If you need a refresher on assigning E/Z and R/S, [here's a primer.](#))



(R)-5-((S)-1,2-dihydroxyethyl)-3,4-dihydroxyfuran-2(5H)-one
(vitamin C)

Several different forms of systematic nomenclature have been used both in the past and in the present. Furthermore, the best-known nomenclature system, that of the International Union of Pure and Applied Chemistry (IUPAC), provides various options for how to name a compound. Therefore, most compounds have more than one systematic name. Fortunately, most systems of nomenclature in wide use are based on more or less the same principles and the same vocabulary.

L-*threo*-Hex-2-enonic acid, γ -lactone

L-3-Keto-*threo*-hexuronic acid lactone

2-oxo-L-*threo*-hexono-1,4-lactone-2,3-enediol

(*R*)-3,4-dihydroxy-5-((*S*)-1,2-dihydroxyethyl)furan-2(*5H*)-one

(*R*)-5-((*S*)-1,2-dihydroxyethyl)-3,4-dihydroxyfuran-2(*5H*)-one

Five systematic names for vitamin C (L-ascorbic acid)

IUPAC recently published rules for determining one Preferred IUPAC Name (PIN) for each compound. The rules for determining these names are rather complicated; however, as we will see shortly, other forms of notation are often used when you need a unique identifier for a compound.

1.2.2. Why do they work that way?

Systematic names were originally designed primarily for use in alphabetical indexes of chemical substances. However, the effort to make these names both unambiguous and canonical (see Unit 3.b in this module) for this purpose made many of these names extraordinarily difficult to read, let alone say out loud. Chemists came up with different approaches to systematic nomenclature tailored for different sorts of compounds and different ways of organizing a chemical index; that's how we ended up with so many different systematic names for the same compound.

Though some chemists initially predicted that systematic nomenclature would completely replace trivial names, this never happened. Trivial names convey little or no chemical information, but they have the advantage over systematic names in many of the qualities that we usually associate with good names: they are short, memorable, pronounceable, and easy to distinguish from other names.

1.2.3. Where they are most often used?

Trivial names are used constantly in informal chemical communication. Chemists working together on specific complex compounds will typically develop their own trivial "nicknames" for their compounds of interest.

Systematic names are often required if you want to register a new compound and for compounds discussed in publications. They are typically listed in database records accessible through search applications like PubChem, SciFinder, Reaxys, and ChemSpider, as well as on the Wikipedia pages for chemical substances. However, because of the various different systems of nomenclature in use, because IUPAC names are not unique, because names

formed according to now-defunct rules often stick around, and because of human error (a particularly issue in a crowd-curated site like Wikipedia), the systematic names that you find in these locations can sometimes vary.

Sections of systematic and semi-systematic names corresponding to a substructure of interest can be useful in searching for compounds containing that substructure, particularly in non-chemical settings like Google. However, this approach is generally less reliable than substructure searches that accept structural formulas as input.

1.2.4. What questions should I ask?

Are there structural ambiguities that the structural formula would clearly indicate but that the systematic name obscures?

When you're dealing with systematic names rather than structural formulas, it's much harder to recognize when you need to pay attention to delocalization, stereochemistry, and tautomerism. You may wish to sketch a structural formula based on the name (or make use of a computer program that does so) to determine whether any of these factors – particularly stereochemistry – apply.

What system of nomenclature does the name fit within?

Are you dealing with an IUPAC name? A Preferred IUPAC Name (PIN)? A CAS index name? A name that describes a structural formula without quite following any specific set of nomenclature rules?

Why am I using a systematic name, anyway?

Systematic names are difficult to read and to write. Before you decide to use them, make sure that there isn't a different chemical identifier that serves your purposes better. (See Unit 3.c below.)

1.3. Further reading & references

Formulas

Jonathan Brecher, *Pure and Applied Chemistry* 80, no. 2 (January 1, 2008), 227–410. URL: <http://pac.iupac.org/publications/pac/pdf/2008/pdf/8002x0277.pdf> (accessed Sept. 15).

Antony Williams, "[Chemical Structures,](#)" in *The ACS Style Guide* (American Chemical Society, 2006), 375–83. URL: <http://dx.doi.org/10.1021/bk-2006-STYG.ch017> (accessed Sept. 2015).

Neil G. Connelly and Ture Damhus, eds., *IUPAC Nomenclature of Inorganic Chemistry* (Cambridge: Royal Society of Chemistry, 2005), 53–67. (The “Red Book”). URL: http://old.iupac.org/publications/books/rbook/Red_Book_2005.pdf (accessed Sept. 2015).

Wikipedia entry on the Red Book. URL: https://en.wikipedia.org/wiki/IUPAC_nomenclature_of_inorganic_chemistry_2005 (accessed Sept. 2015).

Compound Interest, <http://www.compoundchem.com/> (accessed Sept. 2015).
(good examples of effective communication using formulas)

Names

ACS/CAS

“Names and Numbers for Chemical Compounds,” in *The ACS Style Guide* (American Chemical Society, 2006), 233–54. URL: <http://dx.doi.org/10.1021/bk-2006-STYG.ch012> (accessed Sept. 2015).

American Chemical Society, *Naming and Indexing of Chemical Substances for Chemical Abstracts, 2007 Edition* (Columbus, OH: American Chemical Society, 2008). URL: <http://www.cas.org/File%20Library/Training/STN/User%20Docs/indexguideapp.pdf> (accessed Sept 2015).

IUPAC

Henri A. Favre and Warren H. Powell, eds., *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013* (Cambridge: Royal Society of Chemistry, 2014). (The “Blue Book”). URL: <http://pubs.rsc.org/en/content/ebook/9780854041824> (accessed Sept. 2015).

Wikipedia entry on the Blue Book. URL: https://en.wikipedia.org/wiki/IUPAC_nomenclature_of_organic_chemistry (accessed Sept. 2015).

Neil G. Connelly and Ture Damhus, eds., *IUPAC Nomenclature of Inorganic Chemistry* (Cambridge: Royal Society of Chemistry, 2005), 53–67. (The “Red Book”). URL: http://old.iupac.org/publications/books/rbook/Red_Book_2005.pdf (accessed Sept. 2015).

[Wikipedia entry on the Red Book](https://en.wikipedia.org/wiki/IUPAC_nomenclature_of_inorganic_chemistry_2005). URL: https://en.wikipedia.org/wiki/IUPAC_nomenclature_of_inorganic_chemistry_2005 (accessed Sept. 2015).

[1] Jonathan Brecher, “[Graphical Representation Standards for Chemical Structure Diagrams \(IUPAC Recommendations 2008\)](#),” *Pure and Applied Chemistry* 80, no. 2 (January 1, 2008), 278. URL: <http://pac.iupac.org/publications/pac/pdf/2008/pdf/8002x0277.pdf> (accessed Sept. 2015).

[2] *Ibid.*, 280.

1.4. Chemical structure drawing programs

- 1- ChemDraw – requires subscription and download
 - a. Basic drawing package: http://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemDraw/ChemDrawPrime/Default.aspx
 - b. Free Trial Windows: <http://scistore.cambridgesoft.com/ScistoreProductPage.aspx?ItemID=8186>
 - c. Free Trial Mac: <http://scistore.cambridgesoft.com/ScistoreProductPage.aspx?ItemID=8185>
- 2- ChemDoodle – requires purchase and download
 - a. <https://www.chemdoodle.com/>
 - b. Free Trial: <https://www.chemdoodle.com/free-trial/>
- 3- ChemSketch – requires purchase and download
 - a. http://www.acdlabs.com/products/draw_nom/draw/chemsketch/
 - b. Free Trial at above link
- 4- BIOVIA Draw – free for academic use
 - a. <http://accelrys.com/products/collaborative-science/biovia-draw/draw-no-fee.php>
- 5- ChemSketch Freeware - free for academic and personal use, requires registration and download:
 - a. <http://www.acdlabs.com/resources/freeware/chemsketch/>

Other significant options (although they may not be as fully-fledged as the previous ones):

- 6- KnowItAll Academic edition - free for academics, requires registration and download:
 - a. <http://www.knowitall.com/academic/>

- 7- MarvinBeans - some part sort of the product including MarvinSketch are free to use:
 - a. <http://www.chemaxon.com/download/marvin-suite/#mbeans>
- 8- JSME - open source (BSD license):
 - a. <http://peter-ertl.com/jsme/>
- 9- JChemPaint - open source (LGPL license):
 - a. <https://jchempaint.github.io/>
- 10- Bioclipse - open source (EPL license):
 - a. <http://www.bioclipse.net>
 - b. Note: it is more than a drawing program but it does support molecule edition

For online use, one may want to check:

- 11- Xemistry Web Sketcher:
 - a. <http://www.xemistry.com/edit/frame.html>
- 12- JSME demo page:
 - a. http://peter-ertl.com/jsme/JSME_2015-06-14/JSME.html

An always nice and updated comparison is the one from Dr. Tamas Gunda at:

- <http://www.gunda.hu/dprogs/index.html>