

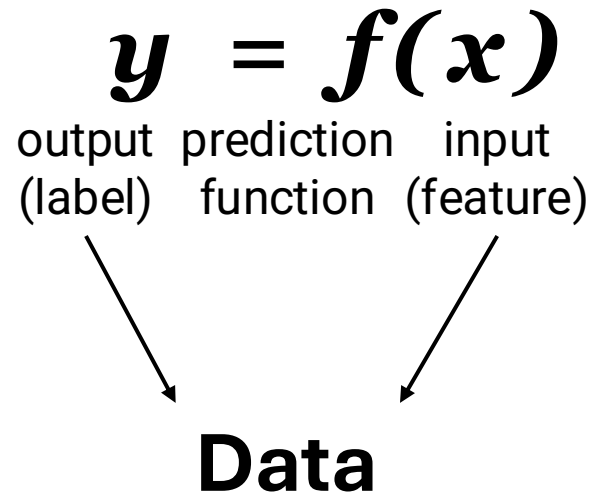
AI and ML for Chemists

Topic 3.1: Chemical Data

Pre-lecture Explore

Dr. Ganna (Anya) Gryn'ova

Data in ML



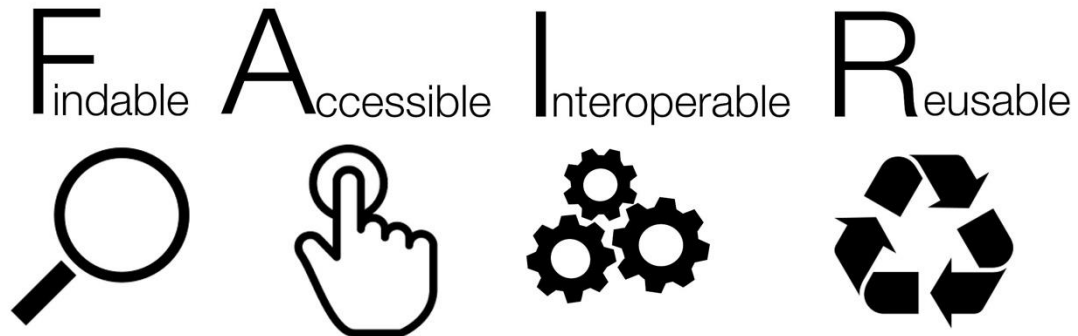
– information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer.

Cambridge Dictionary

Garbage In – Garbage Out

Good training data is:

- **Reliable**
- **Diverse**
- **Representative**
- **Large**
- **Correctly labelled**
- **Available**



Creating a *good* training dataset involves:

1. **Data collection**
2. **Data cleaning**
3. **Data labelling**
4. **Data preprocessing** (filtering, scaling, etc.)

How many molecules are there?

A popular estimate of 10^{60} refers to molecules composed of C, H, N, O, and H atoms, containing no more than 4 rings and weighting less than 500 Da.

Several other “theoretical” subspaces populated by small organic (often, drug-like) molecules have been enumerated, e.g., a more conservative estimate of 3.4×10^9 for molecules with ≤ 100 carbon atoms and the “Chemical Universe Database” GDB-17 with **166.4 billion** molecules with up to 17 C, N, O, S, and halogen atoms.

Across the entire chemical space, ca. **219 million** organic substances, alloys, coordination compounds, minerals, mixtures, polymers, and salts have been published and are recorded in the Chemical Abstracts Service (CAS) registry.

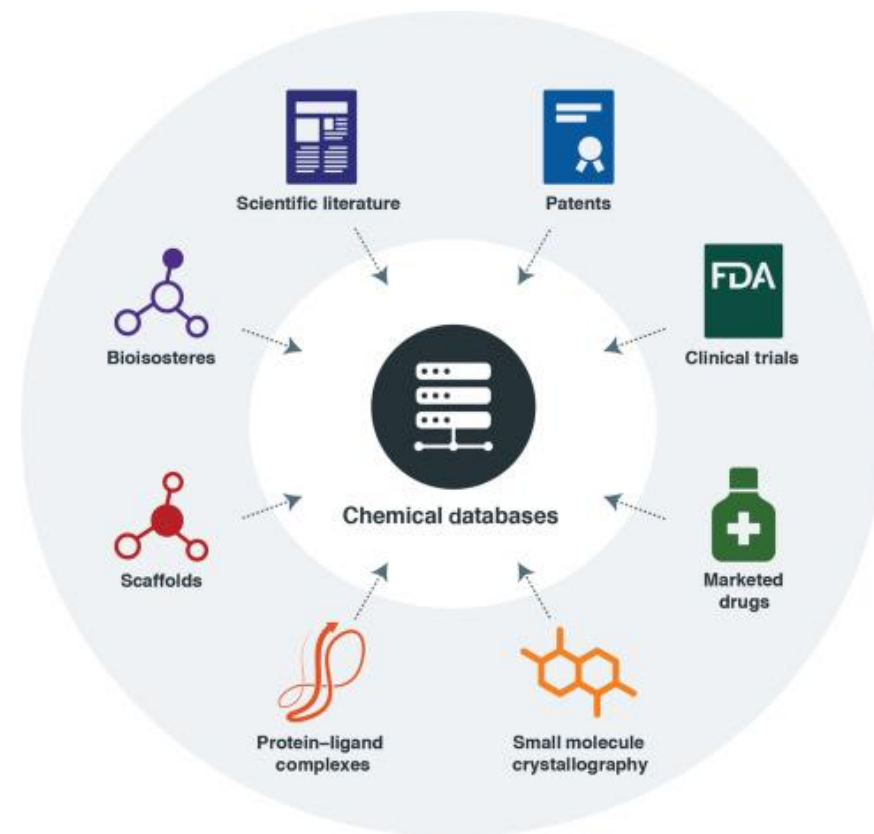
According to various estimates, the number of possible chemical compounds is in the range of 10^{18} – 10^{200} .

It is estimated that there may be 10^{78} to 10^{82} atoms in the known universe.

Experimental data

Numerous databases of chemical data (physical and chemical properties of molecules) exist – see https://en.wikipedia.org/wiki/Chemical_database and https://en.wikipedia.org/wiki/List_of_chemical_databases.

Is this data *good*? See “**Activity**”.



The situation is *even worse* for reaction data (e.g., yields, ee values, etc.):

- Good (large, reliable, and diverse) reaction datasets are **proprietary**.
- Publicly available datasets are either **small** or **biased** (or both).

Computed (synthetic) data

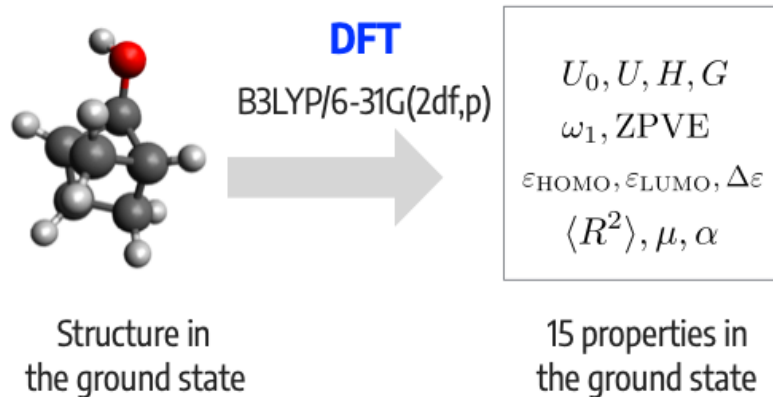
Curated datasets of computed molecular data have been created in the recent years.

For an up-to-date list, see the [awesome-chemistry-datasets](https://github.com/kjappelbaum/awesome-chemistry-datasets) page:

<https://github.com/kjappelbaum/awesome-chemistry-datasets>.

Example – QM7 dataset:

- 133,885 species with up to nine heavy atoms (C, O, N, and F);
- geometries, harmonic frequencies, dipole moments, polarizabilities, along with energies, enthalpies, and free energies of atomization;
- computed at the B3LYP/6-31G(2df,p) level of quantum chemistry.



1 HOMO: -0.1877 LUMO: 0.1171 Dipole_moment: 0.0	2 HOMO: -0.1157 LUMO: 0.8829 Dipole_moment: 1.6256	3 HOMO: -0.2028 LUMO: 0.6687 Dipole_moment: 1.8511	4 HOMO: -0.2845 LUMO: 0.9596 Dipole_moment: 0.0	5 HOMO: -0.3684 LUMO: 0.8191 Dipole_moment: 2.8937
6 HOMO: -0.267 LUMO: -0.8486 Dipole_moment: 2.1889	7 HOMO: -0.2385 LUMO: 0.1841 Dipole_moment: 0.0	8 HOMO: -0.2653 LUMO: 0.8784 Dipole_moment: 1.5258	9 HOMO: -0.2689 LUMO: 0.8513 Dipole_moment: 0.7356	10 HOMO: -0.3264 LUMO: 0.8376 Dipole_moment: 3.8266
11 HOMO: -0.254 LUMO: -0.8198 Dipole_moment: 2.5482	12 HOMO: -0.2543 LUMO: 0.8262 Dipole_moment: 1.7286	13 HOMO: -0.323 LUMO: 0.6949 Dipole_moment: 0.8597	14 HOMO: -0.2619 LUMO: 0.8798 Dipole_moment: 1.4131	15 HOMO: -0.2525 LUMO: 0.891 Dipole_moment: 1.1582
16 HOMO: -0.2888 LUMO: 0.1842 Dipole_moment: 5.96e-4	17 HOMO: -0.2682 LUMO: 0.1842 Dipole_moment: 1.7675	18 HOMO: -0.2431 LUMO: -0.8887 Dipole_moment: 2.7362	19 HOMO: -0.2436 LUMO: 0.8347 Dipole_moment: 3.3367	20 HOMO: -0.2495 LUMO: 0.8556 Dipole_moment: 3.4809
21 HOMO: -0.2267 LUMO: 0.8843 Dipole_moment: 0.8887	22 HOMO: -0.2612 LUMO: 0.874 Dipole_moment: 1.4259	23 HOMO: -0.2599 LUMO: -0.8214 Dipole_moment: 0.0	24 HOMO: -0.3182 LUMO: -0.6543 Dipole_moment: 3.792	25 HOMO: -0.3696 LUMO: -0.8926 Dipole_moment: 0.8823

Datasets of materials

