Data

We need data to train models

- Experimental databases
  - Largest database for experimentally verified inorganic materials: Inorganic Crystal Structure Database (ICSD)
  - Crystallography Open database (COD)

- MGI spawned a number of DFT-generated databases
  - We list a few, work with 3
  - How to download database
  - How to extract data

Freely available (DFT) databases
Spawned by MGI
Awareness about license agreement

- **Materials genome initiative** ⬅
  - https://www.mgi.gov
- **Materials Project**
  - https://materialsproject.org ⬅
- NOMAD
  - https://nomad-coe.eu
- **Open quantum materials database**
  - http://oqmd.org ⬅
- **AFLOW**
  - http://www.aflow.org
- **Computational materials repository**
  - https://cmr.fysik.dtu.dk ⬅
- **novomag**
  - https://www.novomag.physics.iastate.edu/structure-database
- **Novamag**
  - https://zenodo.org/records/3241267
- **Alexandria**
  - https://figshare.com/articles/dataset/Alexandria_DB/27174897?file=49622718 ⬅
- JARVIS-DFT
  - https://jarvis.nist.gov/ ⬅

Machine intelligible representation of materials

Materials feature construction

- Human intelligible representation:
  - Lattice vectors and atomic basis
  - FCC crystal with one atom basis (Fe, …)
  - Zinc blend structure
  - Graphene structure

https://www.chemtube3d.com/

FCC

$$\vec{a}_1 = \frac{a}{2}(0,1,1); \quad \vec{a}_2 = \frac{a}{2}(1,0,1); \quad \vec{a}_2 = \frac{a}{2}(1,1,0);$$

1-atom basis: $\vec{t}_1 = (0,0,0)$

Zinc blend

$$\vec{a}_1 = \frac{a}{2}(0,1,1); \quad \vec{a}_2 = \frac{a}{2}(1,0,1); \quad \vec{a}_2 = \frac{a}{2}(1,1,0);$$

2-atom basis: S at $\vec{t}_1 = (0,0,0)$; Zn at $\vec{t}_2 = \frac{a}{4}(1,1,1)$

2-atom basis: S at $\vec{t}_1 = (0,0,0)$; Zn at $\vec{t}_2 = \frac{a}{4}(1,1,1)$

- This is not the best representation for ML if it has to predict material properties.
- Here is why …
  - What if the material is rotated?
  - What if positions of the Zn and S atoms are interchanged/permuted?
  - What if we displace all the atoms by the same vector (translation), or translate the origin of the lattice vectors?
    - Properties cannot change.
  - Material properties depend on the chemical properties of the constituents. Only Z values given here. IP, EA, electronegativity may be relevant.

- So what are the requirements (symmetry properties) that features should ideally posses?

Requirements of ideal feature vectors

1. **Completeness**

   A feature should contain all information that are required for determination of the property of interest. Graphite and diamond are both made of C. Yet, their properties (e.g., hardness) are very different. Structural information is important.

2. **Non-degeneracy** (for *first principles* features)

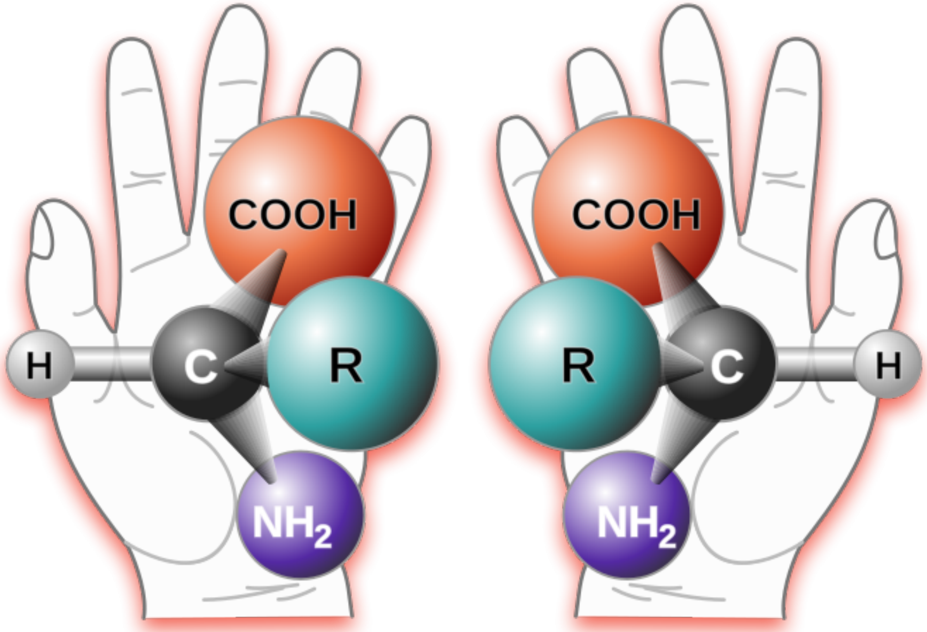
$$\hat{H}_{el} = -\frac{1}{2}\sum_{i}^{N_e} \nabla_i^2 - \sum_{i,I}^{N_e,N_a} \frac{Z_I}{|\vec{r}_i - \vec{R}_I|} + \sum_{i<j}^{N_e} \frac{1}{|\vec{r}_i - \vec{r}_j|}.$$

- Two materials M_1 and M_2 should produce different feature vectors $\vec{x}_1$ and $\vec{x}_2$. If M_1, M_2 both $\rightarrow \vec{x}$, they will be predicted to have the same property. Not generally correct.
- More formally, M_1, M_2 different materials, so have different Hamiltonians.
- Hence they have different ground state wave functions, $\Psi_1 \neq \Psi_2$. Observable O corresponding to property of interest: $\langle \Psi_1 | \hat{O} | \Psi_1 \rangle \neq \langle \Psi_2 | \hat{O} | \Psi_2 \rangle$.
- Therefore, the predictions contradict our quantum mechanical understanding.

3. **Compactness**

- Features should contain the minimal amount of information relevant to the target property. Having more information will make the training process complicated.

Chiral amino acid



Wikipedia

4. **Uniqueness**

- One-word description for rotational, translational and permutation invariance.

5. **Meaningful**

- For a given feature, **Feature → Property** relation should be simple enough for the model to learn it within its scope of complexity.

- Example: If a property depends on the square of a feature x, and we are using a linear model, it is not meaningful.

  Using x^2 is obviously more meaningful. If our model is non-linear, and can capture quadratic dependence, then using x as a feature is meaningful.

- Complexity of a model depends on a number of factors: model type, number of parameters, number of hyper-parameters etc.

6. **Continuous**

- Features should change continuously with change in composition.

- Based on our chemical intuition that small changes in composition should change property by a small amount.

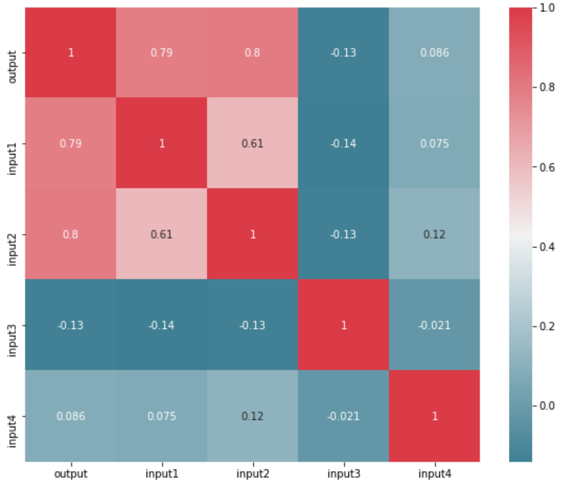
  - Change FeCo to $\mathrm{Fe_{0.499}Co_{0.501}}$, properties cannot change much.

7. Differentiable

- Small changes in positions of atoms expected change property by small amounts
- Hence, features should also change by small amounts in a continuous way.
- So they should be differentiable wrt the coordinates of the atoms.

8. Uncorrelated

- If two or more features are strongly correlated, keeping both of them is not going to add independent information about the material. It can, in fact, complicate model training.
- We can keep one among the correlated ones.
- There are systematic ways of dimensional reduction.



Z. Luna, Medium

9. Easy to compute

- Constructing the feature vectors should not take more time than training the ML model

Some commonly used feature

Features encoding only stoichiometry

1. **Stoichiometry norm** L_p, ($p = 1, 2, 3$).

- Stoichiometry norm for a particular p: $||x_p|| = \left(\sum_i |x_i|^p\right)^{1/p}$.
- x_i is the atomic fraction of the i-th elemental species in the material.
- Example: Fe_3O_4. $x_{\mathrm{Fe}} = 3/7$; $x_{\mathrm{O}} = 4/7$. Hence,
   - $||x_2|| = \left((\frac{3}{7})^2 + (\frac{4}{7})^2\right)^{1/2} = 0.071$; & $||x_3|| = \left((\frac{3}{7})^3 + (\frac{4}{7})^3\right)^{1/3} = 0.064$.

2. **Stoichiometry entropy**

- $S_e = = -\sum_i x_i \ln x_i$.
- With Fe_3O_4, $S_e = -\frac{3}{7}\ln(3/7) - \frac{4}{7}\ln(4/7) = 0.68$.

3. Atomic fraction vector

- $\mathbf{v}_{\mathrm{af}} = \{x_{\mathrm{H}}, x_{\mathrm{He}}, x_{\mathrm{Li}}, \cdots, x_{\mathrm{Z_{max}}}\}$
- For VO_2, $\mathbf{v}_{\mathrm{af}} = \{0,0,\cdots, x_{\mathrm{O}} = 2/3,\cdots, x_{\mathrm{V}} = 1/3,\cdots, x_{\mathrm{Z_{max}}}\}$.

Features encoding stoichiometry & chemical properties

1. If Q is an atomic property, then

- $\bar{Q} = \sum_i x_i Q_i$ configuration weighted average.
- $|\delta Q| = \sum_i x_i |Q_i - \bar{Q}|$ configuration weighted absolute deviation.
- Composition weighted mode $cwMO(Q) = Q_{\max}$, the Q value for the constituent i where $x_i = \max(\{x\})$, the constituent with the largest atomic fraction. If two (or more) elements have the same atomic fraction, then $cwMo(Q)$ is defined as their average value: $cwMo(Q) = \dfrac{Q_{\mathrm{A}} + Q_{\mathrm{B}}}{2}$, if $x_{\mathrm{A}} = x_{\mathrm{B}}$.
- Atomic properties (Q) used

  (i) atomic number Z,

  (ii) period number P,

  (iii) group number G,

  (iv) electronegativity ϵ,

  (v) number of valence electrons n_v.

Nelson & Sanvito, PRMat 3, 104405 (2019)