Continuous maps of molecules and atomic clouds in large databases

D. Widdowson, Y. Elkin, V. Kurlin

Computer Science department, Materials Innovation Factory, University of Liverpool, United Kingdom,
vitaliy.kurlin@gmail.com

Keywords: molecular configurations, atomic clouds, rigid motion, isometry invariants, distance metric, continuous maps

Mendelev arranged all known chemical elements into a spatial map - the periodic table parametrized by discrete coordinates - the period and group number. The periodic table was initially half-empty but effectively guided the search for new chemical elements.

The modern databases such as QM9 [1] and CSD contain millions of experimental and simulated molecules and solid crystalline materials that are not yet organized into a spatial map with unique and complete coordinates. The first question we should ask about real data is same or different [2]? Any answer to this question should provide an equivalence relation satisfying the three axioms: (1) reflexivity: any object is equivalent to itself A~A, (2) symmetry: if A~B then B~A, (3) transitivity: if A~B and B~C, then A~C. The transitivity axiom guarantees a well-defined classification when all data objects (molecules or crystals) are split into disjoint classes of equivalence. For instance, any object A generates its class [A]={B : B~A} of all objects B equivalent to A.

Though one can define many equivalences on molecules by symmetry group or chemical composition, the resulting classifications are weak in the sense that many molecules and materials (graphite and diamond) within the same class have different properties.

The strongest equivalence on molecules and crystals is rigid motion, which is a composition of translations and rotations, because any physical object preserves its properties under rigid motion in the same ambient conditions such as temperature and pressure.

Rigid shapes of molecules can be distinguished only by invariants that are descriptors preserved under rigid motion. Noisy data motivates as to require continuity under atomic displacements. Considering only atomic centers without chemical elements allows us to continuously quantify differences between close molecules with distinct compositions such as benzene and chlorobenzene.

The recent invariants invariants of periodic crystals [3,4] were adapted to fully complete and continuous invariants of clouds of m unlabeled atoms, which can be computed in a polynomial time in m [5]. These invariants distinguished all molecules in QM9 [1]. Fig. 1 compares a discrete map by chemical composition with a continuous map in the average lengths of bonds and no-bonds.

Figure 1. Visualization of 134K molecules in QM9 [1]. Left: x is an integer index counting elements C,N,O,F; y is the number of hydrogens; the color reflects the number of molecules with fixed (x,y). Right: continuous map of QM9 by (no-)bond lengths.