Analysing and Rationalising Molecular and Materials Databases Using Machine-Learning[1] SANDIP DE, MICHELE CERIOTTI, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland — Computational materials design promises to greatly accelerate the process of discovering new or more performant materials. Several collaborative efforts are contributing to this goal by building databases of structures, containing between thousands and millions of distinct hypothetical compounds, whose properties are computed by high-throughput electronic-structure calculations. The complexity and sheer amount of information has made manual exploration, interpretation and maintenance of these databases a formidable challenge, making it necessary to resort to automatic analysis tools. Here we will demonstrate how, starting from a measure of (dis)similarity between database items built from a combination of local environment descriptors, it is possible to apply hierarchical clustering algorithms, as well as dimensionality reduction methods such as sketchmap, to analyse, classify and interpret trends in molecular and materials databases, as well as to detect inconsistencies and errors. Thanks to the agnostic and flexible nature of the underlying metric, we will show how our framework can be applied transparently to different kinds of systems ranging from organic molecules and oligopeptides to inorganic crystal structures as well as molecular crystals.

Sandip De
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Date submitted: 11 Nov 2016   Electronic form version 1.4