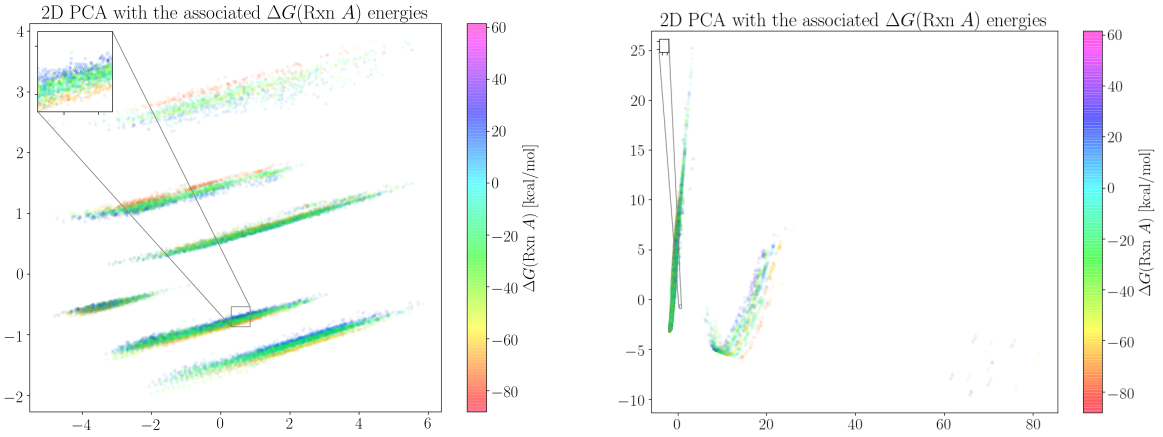
ML Project 2: Quantum machine learning - Plots appendix

Joachim Koerfer, Jannick Reichert, Luca Viano

I. 2D PCA PROJECTION

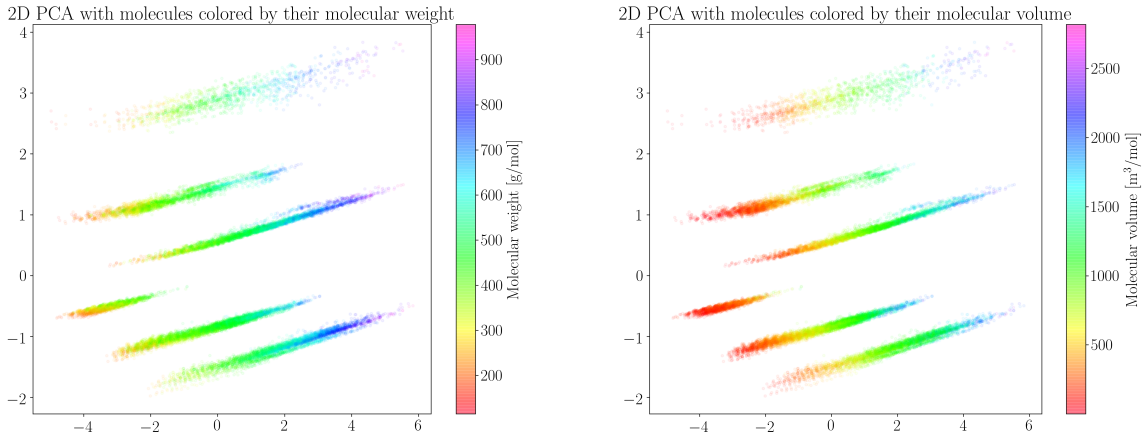


(a) Non expanded version

(b) Expanded version

Fig. 1: All the molecules with their properties (features) have been projected on the 2D plane with PCA. The energies ΔG(Rxn A) associated to the molecules are not present in the feature set, but have been added after the projection. In the expanded version the strongly elongated cluster along the y-axis hosts main of the points. As a consequence the colour trend, visible in each cluster of the non expanded version, can be appreciated only in the less populated cluster.

II. MOLECULES COLOURED BY THEIR MOLECULAR WEIGHT OR VOLUME

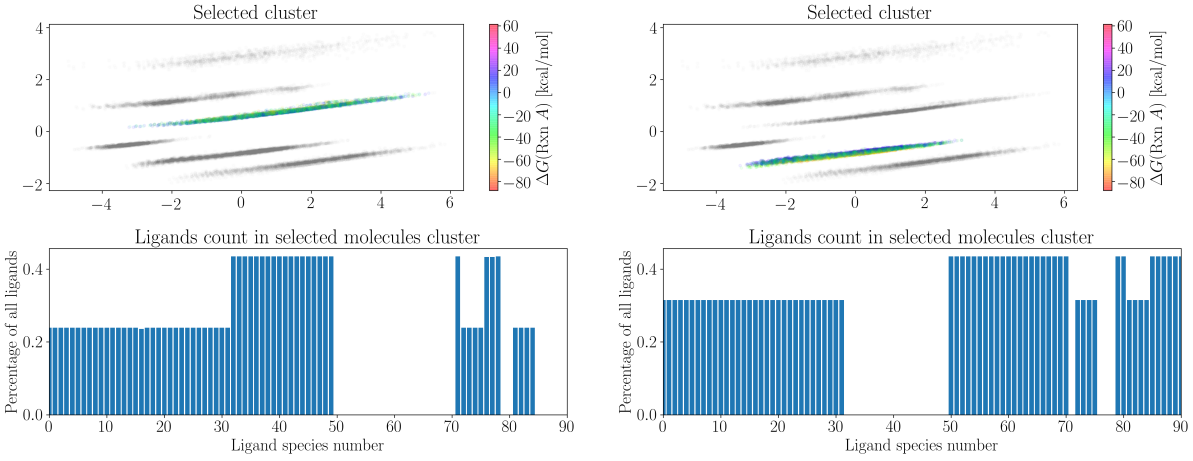


(a) Molecular weight

(b) Molecular volume

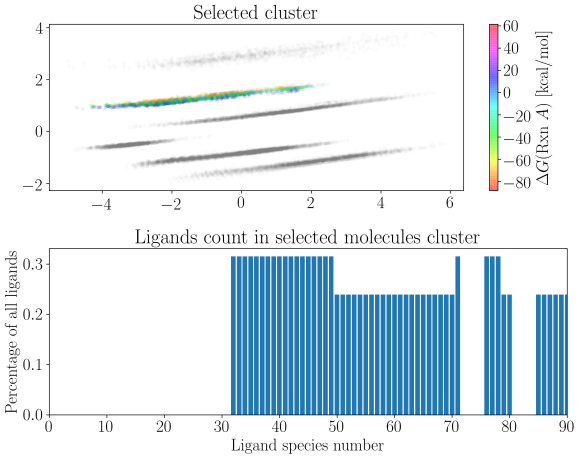
Fig. 2: This figure is similar to figure 1 except that the points have been coloured by their molecular weight or molecular volume (which are present in the data set). The same color trend along the x-axis can be observed since the molecular weight and volume are collinear quantities. This explains also why a dimensionality reduction can be taken in account.

III. Ligand count in cluster molecules

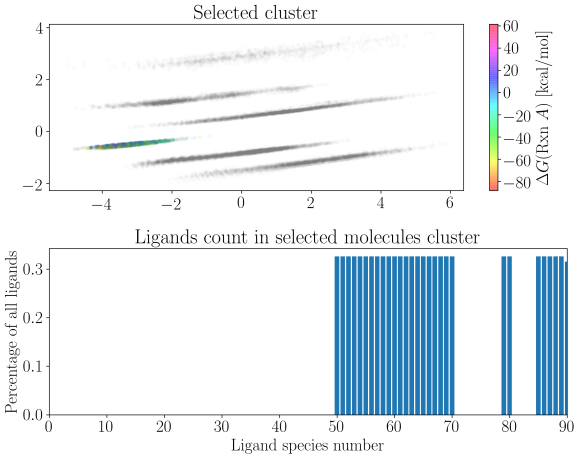


(a) First cluster

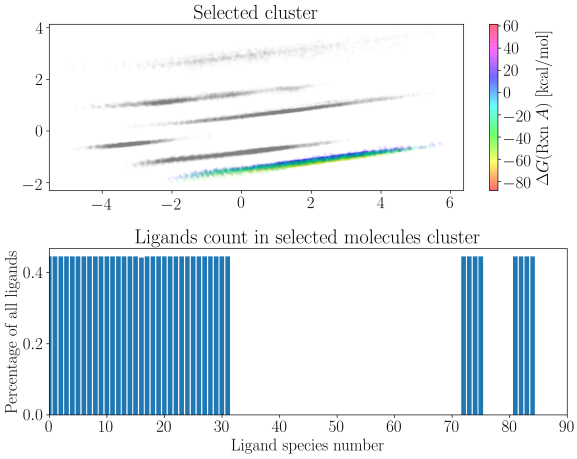
(b) Second cluster



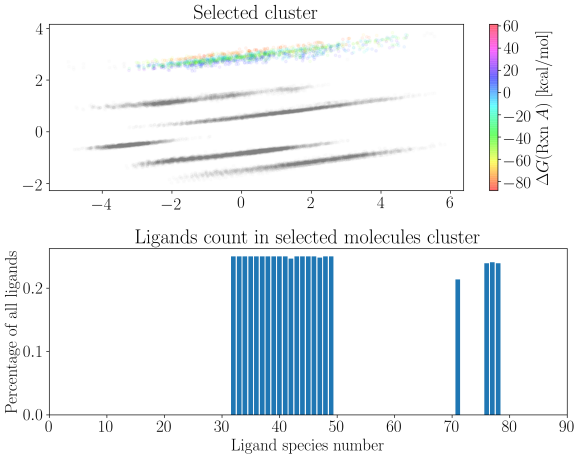
(c) Third cluster



(d) Fourth cluster



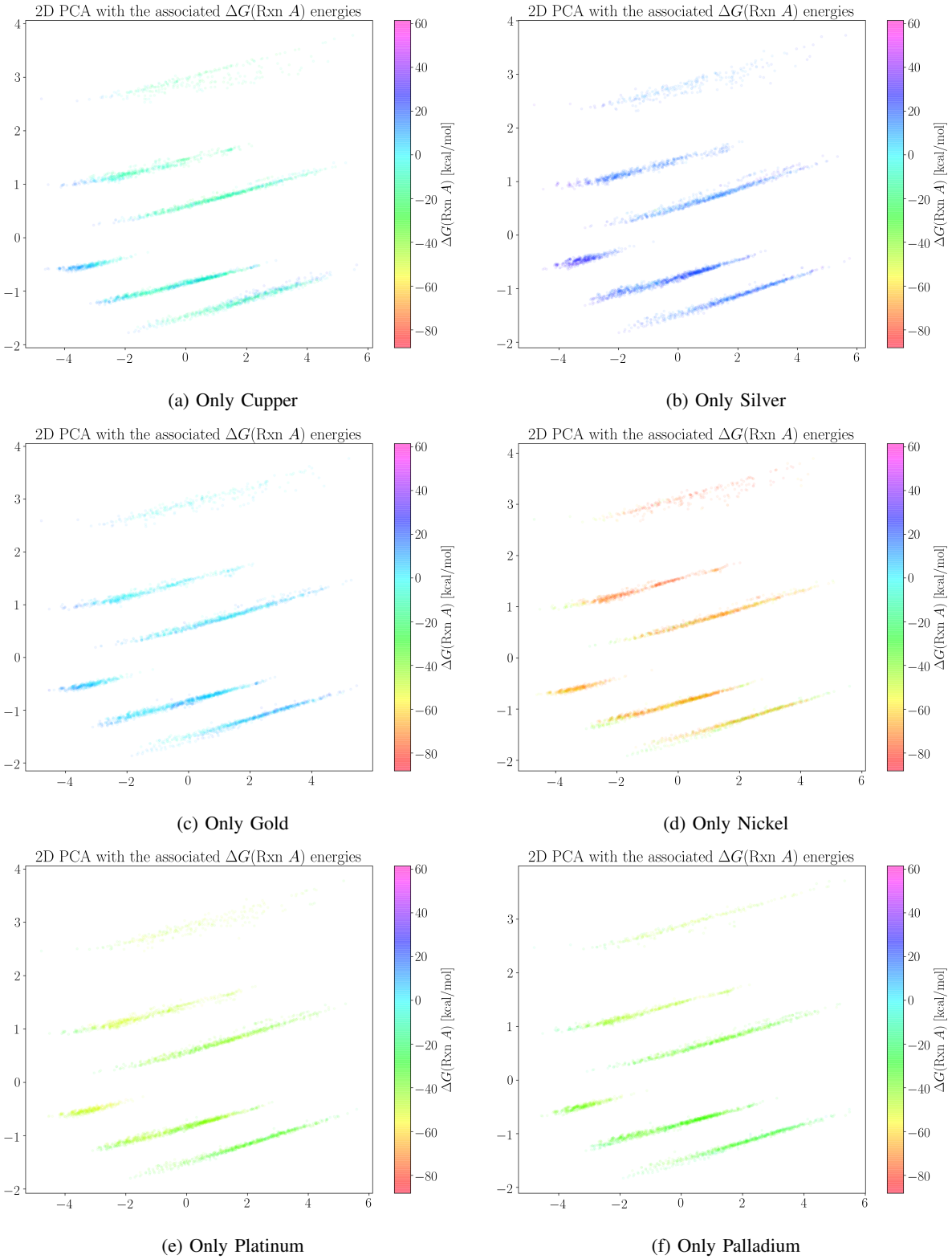
(e) Fifth cluster



(f) Sixth cluster

Fig. 3: This figure reports the ligand histogram for each of the 6 DBSCAN detected clusters. The fact that the ligands family is the factor causing the split is supported by the complementary shapes of the histograms, e.g., both the first and second cluster contain molecules that have a ligand with label between 0 and 31 but their separation is caused by second ligand. Indeed, it has label between 32 and 50 in the first case and between 50 and 70 in the second one.

IV. 2D PCA SELECTING ONLY ONE METAL IN THE MOLECULES



(a) Only Cupper

(b) Only Silver

(c) Only Gold

(d) Only Nickel

(e) Only Platinum

(f) Only Palladium

Fig. 4: The same figure as 1 but removing all metals except the one mentioned in the caption. The different colours of the clusters in the 6 different cases confirm that the catalyst energy is mainly determined by the metal.

V. Molecules coloured by a combination of their last 4 features in the data set

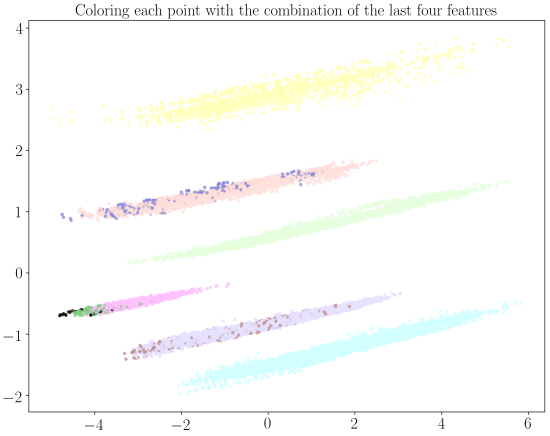


Fig. 5: The last four features of each molecule have been assigned to the CMYK colour space and then transformed to RGB for the final colour of each point. This means that the colours are the results of the combination of the last four features in the data set. This plot shows that each cluster has a distinct uniform colour and some outliers present in three of the six clusters.

VI. Gabriel prediction-error vs dimension dependence

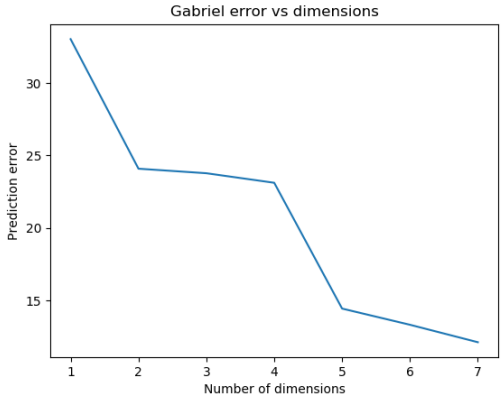


Fig. 6: It can be observed that the error is monotonously decreasing with the number of dimensions. Nevertheless the plateau between 2 and 5 dimension suggest that a model with two dimensions can explain the data as well as a five-dimensional model. This fact is confirmed by the number of p-values peak at $d = 3$ in figure 1 of the report.