Enumeration and analysis of a small molecule universe

Stefan O. Gugler

April 30, 2018

Abstract

ACS Boston Abstract: Enumerating the inorganic universe of small complexes for machine learning. Transition metal complexes form promising functional inorganic materials due to their wide range of tunable electronic properties. However, exhaustive enumeration and calculation of all possible ligand fields is clearly intractable due to the vast nature of chemical space. Virtual high-throughput screening with density functional theory (DFT) allows us to harvest leads with desired properties but is severely constrained by 1) long calculation times and 2) variable accuracy. More accurate correlated methods are available to address 2) but drastically worsen 1). Machine learning techniques potentially allow us to address both issues simultaneously. Our group has previously developed data-driven models based on DFT results which have highlighted the dominant role of metal-proximal atoms (i.e. from the first and second coordination shell) in predicting spin state ordering, bond lengths, and ionization potential of the metal center. This motivates a systematic exploration of the space of octahedral complexes made of organic ligands with up to two heavy atoms (CNOPS), representing the metal-proximal environment. Even in this limited space, the number of potential candidate complexes is infeasible to calculate and so we propose a family of scoring functions that are used to extract mono- and bidentate ligands that most likely form stable complexes based on valency, net charge, and steric effects. The resulting organic ligand universe is then compared to similar studies of small organic molecules. Exploiting isoelectronic structure and empirical stability learned from previous studies, we sample the most promising compounds from this space with high-throughput DFT. We assess DFT performance selectively with more accurate correlated wavefunction calculations using domain-based local pair-natural orbital coupled cluster (DLPNO-CCSD(T)) and apply machine learning to model the difference between correlated wavefunction and DFT results in a composition-dependent manner. By doing this, we hope to learn property estimates for the full space of possible metal-proximal environments along with estimates of DFT reliability relative to DLPNO-CCSD(T).

Table 1: The sizes of the selected subsets of octahedral space.

Set	description	size
Homoleptics	eq = ax	553
"5+1" symmetric	eq = ax1 $\neq$ ax2	163,620
"4+2" symmetric	eq1 $\neq$ eq2 = ax	185,376
Strongly symmetric	eq $\neq$ ax	245,316
Equatorially asymmetric	eq1 $\neq$ eq2 $\neq$ ax	15,924,796
Weakly symmetric	eq $\neq$ ax1 $\neq$ ax2	45,077,310
Complete Heteroleptics	$L_i \neq L_j$	$\approx 5.9 \cdot 10^{12}$
Octahedral Space	all	$> 1.8 \cdot 10^{14}$

1 Ligand Field Assembly

1.1 Subsets of octahedral space

The full combinatorial space of all ligands generated in Section **??** is vast with a lower estimated bound of $> 1.8 \cdot 10^{14}$, calculated from cube coloring theorem. The difficulty is to include bidentate symmetry into the calculations. In the following, we will motivate why only a fraction of the full space is of interest. This will give us the possibility of actually enumerating the ligand fields and calculate properties of the subsets. One goal of this work is to be able to fine tune molecular properties such as oxidation energy or spin splitting energy. It seems natural to assume that a full set of all possible ligand fields gives the most fine grained raster over all possible properties, a complex could have. This might be statistically true but also prevent rigorous analysis. The set is too dense to traverse on a computationally feasible time scale. We propose the homoleptic subset of the octahedral space as a backbone. Conceptually, homoleptic octahedral compounds must be on the edge of the actual octahedral space, since the properties of single complexes do not get more distinct than in homoleptic ones. From the homoleptic complex space it is possible to go to lower symmetry classes and get closer to a complex with desired properties without analyzing all complexes on the way. This reduces computational cost by several orders of magnitude. Our selection of complexes can be found in Table 1. They were chosen to be of highest possible symmetry to result in the smallest number of complexes and to be synthetically interesting.

1.2 Subset analysis

We did a principal component analysis on the homoleptic subspace and projected the strongly symmetric and "5+1" symmetric subspace onto it (see Figure 1). The connecting atom is colored according to the CPK coloring with a gradient transition from oxygen (red) to phosphorus(orange). The connect-

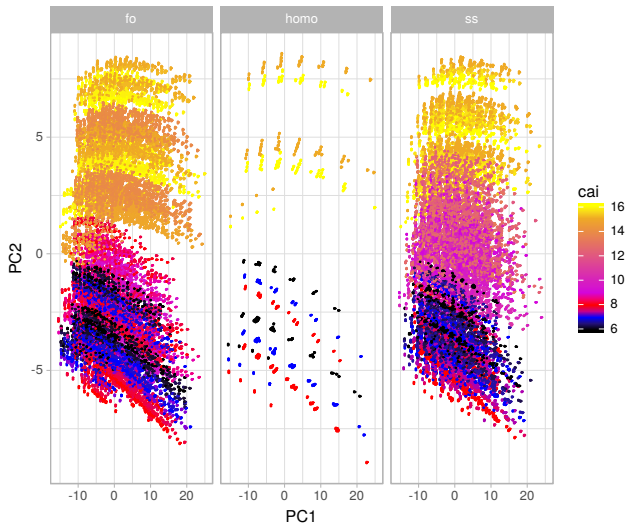

Figure 1: A PCA on the homoleptic subspace. The strongly symmetric (ss) and "5+1" symmetric (fo) subspace were projected onto it. The connecting atom identity (cai) is encoded in the color following CPK coloring with a gradient transition from oxygen (red) to phosphorus (orange).

ing atom is the average over all 6 connecting atoms. We can see that the gap on PC2 separates first period elements from second order elements in the homoleptic case. Going to lower symmetry, we see that this gap is filled up with fractional element types. This sonsolidates our hypothesis that the homoleptics build the backbone of the full space and that lowering symmetry allows us to find more fine grained complexes.

1.3 Entropy of the subsets

To compare the different subsets, we devised a non-unqiue footprint to characterize the ligand fields in 5 dimensions:

- total charge
- total valence electrons
- electronegativity of the connecting atom
- ${}^{\mathrm{lc}}_{\mathrm{ax,eq}}\chi_1 = \sum EN_{\mathrm{CA}} \cdot EN_i$
- ${}^{\mathrm{lc}}_{\mathrm{ax,eq}}\chi'_1 = \sum EN_{\mathrm{CA}} - EN_i$

We then calculate the entropy, H_{KDE}, of the Kernel Density Estimated distrbution with Scott's parameter.