Enumeration and analysis of a small molecule universe

Stefan O. Gugler

June 19, 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).

1 Enumeration

??

1.1 Introduction and Specifications

Even though our chemical space is truncated to the first and second coordination shell, which allows us to use a maximum of two heavy atoms, $e \in \{C, N, O, P, S\}$ per ligand, it is still enormously big. In a first step we assemble the ligands and in a second step we attach them combinatorially to a metal center to produce octahedral complexes.

For the ligand design, a number of parameters are introduced to later impose constraints on them to further decrease the vastness of the space. These parameters include the number of H atoms bond to any atom, h^i_j, the charge c^i_j, the number of lone pairs l^i_j, the number of valence electrons v^i_j, where i and j are the ligand's denticity and the atom index inside the ligand, respectively. We only build mono-heavy-atomic ligands (MHALs), di-heavy-atomic ligands (DHALs), and bidentate tetra-heavy-atomic ligands (THALs), which are two identical DHALs bonded together, imposing a strong symmetry constraint.

The full combinatorial space is first reduced via heuristics from classical chemistry like charge, sterics, satisfaction of the octet rule, whether the ligand is open or closed shell and bond orders. The most unfeasible candidates are not even scored all others are scored according to the heuristics mentioned above. The next reduction step is a cutoff over a certain score to obtain a space of desired and sensible structures, yet still too large to even enumerate, let alone calculate. To obtain a feasible space to enumerate, we sample a certain percentage from the the desired space weighted over the distribution of isoelectronic stuctures. This space is small enough to be enumerated. Samples are drawn randomly (?) to obtain DFT results and to interpolate the rest of the enumerated space (see Figure 1).

All structures are stored as SMILES strings.

Box 1.1 Valency (IUPAC Definition)

The maximum number of univalent atoms (originally hydrogen or chlorine atoms) that may combine with an atom of the element under consideration, or with a fragment, or for which an atom of this element can be substituted.

1.2 Spanning the full space

For each element CNOPS we allow a charge $c \in \{-2, -1, 0, 1, 2\}$ (we suppress the denticity and atom indices i and j where it is clear from the context) and a number of attached hydrogen atoms $h \in \{0, 1, 2, 3, 4\}$.

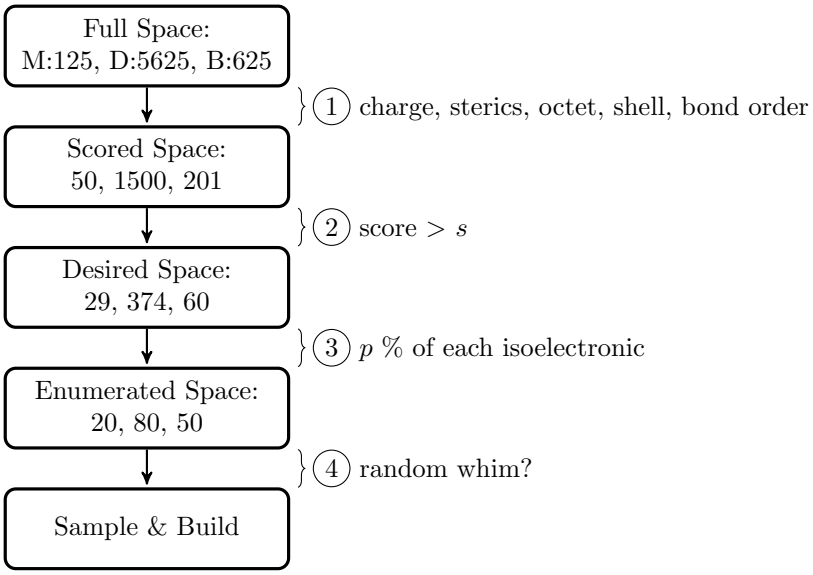

Figure 1: A diagram delineating the steps taken to truncate the full space into a space feasible for enumeration. The three numbers under the space label denote the number of mono-heavy-atomic, di-heavy-atomic, and tetra-heavy-atomic bidentate ligands obtained after the reduction step. On the right, the reduction criteria are listed.

For the **MHALs**, we combinatorially cycle through all combinations of elements, of charge, and of number of H atoms, resulting in $5^3 = 125$ candidates.

The **DHALs** were enumerated similarly, iterating through e_1, e_2, h_1, and h_2. Since, from a DFT perspective, charge is not a local property, we don't iterate over all combinations of c_1 and c_2, since it would produce many degeneracies (i.e. $-1+1=0$ is the same overall charge as $-2+2=0$) and it would not be clear, which charge distribution would be the right one, from a classical picture. For that reason, we iterate over a total charge c_{tot}. For every instance of c_{tot}, we call a subroutine that exhaustively checks all charge combinations that could result in c_{tot}. For each combination of charges c_1 and c_2, we calculate the valency ν (see Box 1.1) of the respective atom as follows, given the lone pairs, l_i and valence electrons v_i from previous knowledge:

$$\nu_i = v_i - c_i - 2 \cdot l_i - 2 \cdot h_i \ . \tag{1}$$

We set the bond order b to be the minimum of the valencies of the two atoms in the DHAL,

$$b = \min\{\nu_1, \nu_2\} \ , \tag{2}$$

if the absolute difference of ν_1 and ν_2 is smallest over all charge combinations and $0 \leq b \leq 4$. The optimal charge distribution over the two atoms is assumed to be the c_1 and c_2 fulfilling the requirements above. If none of these criteria is fulfilled, we set a zero bond order, denoted as #0 in the SMILES formalism. This resulted in 5625 DHALs.

Since the **THALs** are composed of two bonded DHALs (labeled 1 and 2 for the first DHAL and $1'$ and $2'$ for the second DHAL, 1 and $1'$ being the connecting atom), instead of building them from scratch, we decided to combine all the DHALs and exclude the ones that do not bond effectively. As before, we iterate through e_1, e_2, h_1, and h_2 of the building block DHAL. Similarly as with the regular DHAL enumeration described above, we assign a total charge c_{tot} to one of the building block DHALs. We cycle then through all combinations of c_1 and c_2 and calculate the valencies of each atom according to Eq. (1). If $\nu_1 > 0$, it means that it has at least one electron to bond with atom 2 and if $\nu_2 > 1$, it means that it has at least two electrons, one to bond with atom 1 and one to bond with atom $2'$ of the other DHAL. Since the problem is symmetrical, the valencies $\nu_{1'} > 0$ and $\nu_{2'} > 0$ are exactly the same and only labeled differently to speak about them more effectively. The bond order $b_{12} \in \{1, ..., \nu_2\}$ is chosen as

$$\min_b \{|\nu_2 - \nu_2 - b_{12}|\} \tag{3}$$

from which we can simply calculate the second bond, b_{23}, as

$$b_{23} = \min\{\nu_1, \nu_2 - b_{12}\} \ . \tag{4}$$

We have two possibilities to choose the best charge and bond distribution: Either we require to have minimal bond orders or minimal absolute charges. We

decided in favor of minimal charges, since this promises better stability for the resulting complexes. So, if $|c_1|$ and $|c_2|$ are smallest, we choose the corresponding bond orders and charges as the best ones. If none of these criteria is fulfilled, we set a zero bond order, denoted as #0 in the SMILES formalism. This resulted in a total of 5625 THALs.

1.3 The scored and the desired space

In all three sets, we discard open shell compounds, since it leads to difficulties with the DFT calculations.

In the set of **MHALs**, we discard all compounds with a charge outside the boundaries $1 \geq c \geq -3$. We will see this asymmetric restriction towards positive charge throughout this paper, since we know that positive charge usually aggregates on the metal center which leads to instability and negative charges are rather common in ligands. For the remaining compounds, we impose the following scores, s, for charges

$$s_{\text{charge}} = \begin{cases} 0 & \text{if } c = 1 \\ 3 & \text{if } 0 \geq c \geq -2 \\ 0 & \text{if } c = -3 \ , \end{cases} \tag{5}$$

for the number of H atoms on the connecting atom (CA), which measures sterics,

$$s_{\text{sterics}} = \begin{cases} 0 & \text{if } h = 4 \\ 3 & \text{if } h \leq 3 \ , \end{cases} \tag{6}$$

and for the octet rule,

$$s_{\text{octet}} = 4 - (|8 - v|) \ . \tag{7}$$

This leads to a total score,

$$s = s_{\text{charge}} + s_{\text{sterics}} + s_{\text{octet}} \ . \tag{8}$$

A total of 50 compounds were scored, spanning the scored space. From these, we took all with a score $s > 7$ to span the desired space (see Figure 2). A 2D histogram of score vs. valence electrons can be seen in Figure 3 In Figure 4 we can see a distribution of the charges over the desired space of MHALs. No positive charges were selected.

The **DHALs** are restricted in a similar way as the MHALs: We discard all compounds with charge $c > 1$ and a number of H atoms on the CA, $h > 3$, since it is not sterically accessible. Lastly, we remove the compounds with a bond order $b = 0$. For the remaining compounds, we impose the following scores, s, for polarization

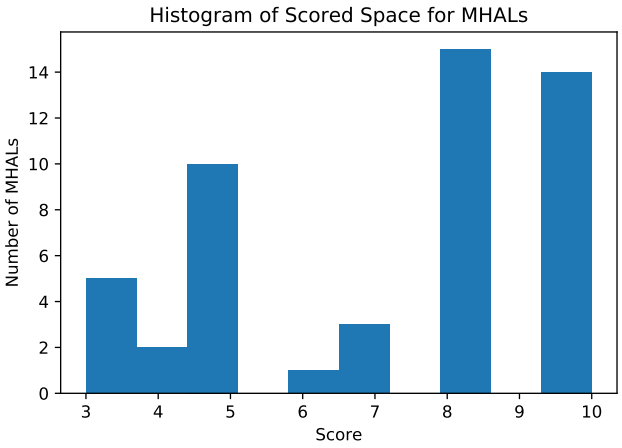

Figure 2: Histogram of all scored MHALs. The ones excluded in the beginning are not shown.

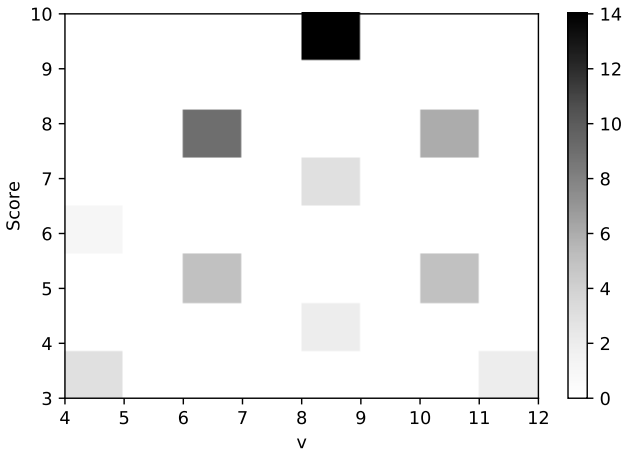

Figure 3: A 2D histogram to show the relationship between score and valence electrons.

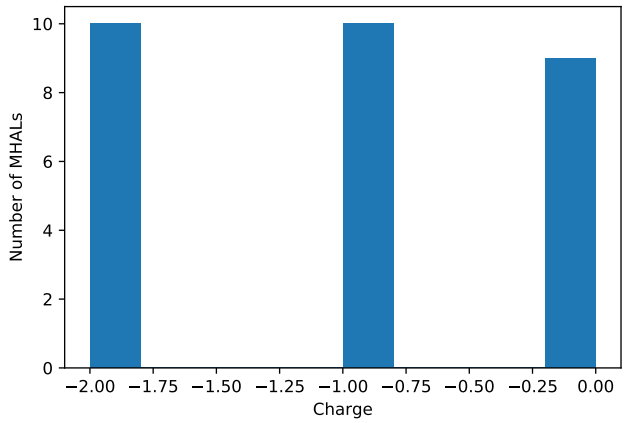

Figure 4: A histogram of the desired MHALs's charges. No positive charges were selected.

$$s_{\text{polarization}} = \begin{cases} 0 & \text{if } |c_1| + |c_2| = 4 \\ 1 & \text{if } |c_1| + |c_2| = 3 \\ 3 & \text{if otherwise ,} \end{cases} \tag{9}$$

to punish high polarization between atom 1 and 2. We slightly prefer single bonds over other types of bonds,

$$s_{\text{bond}} = \begin{cases} 3 & \text{if } b = 1 \\ 2 & \text{if otherwise .} \end{cases} \tag{10}$$

Charge is scored as follows, with $c_{tot} = c_1 + c_2$,

$$u_{\text{charge}} = \begin{cases} 0 & \text{if } c_{tot} = 1 \\ 3 & \text{if } c_{tot} = 0 \\ 2 & \text{if } -1 \geq c_{tot} \geq -2 \\ 1 & \text{if } c_{tot} = -3 \\ 0 & \text{if } c_{tot} = -4 \text{ ,} \end{cases} \tag{11}$$

valency as follows,

$$s_{\text{valency}} = 5 - |\nu_1 - \nu_2| \text{ ,} \tag{12}$$

and the sterics as

$$s_{\text{sterics}} = \begin{cases} 0 & \text{if } h = 3 \\ 3 & \text{if } h \leq 2 \text{ .} \end{cases} \tag{13}$$

The total score is then again the sum over all scores,

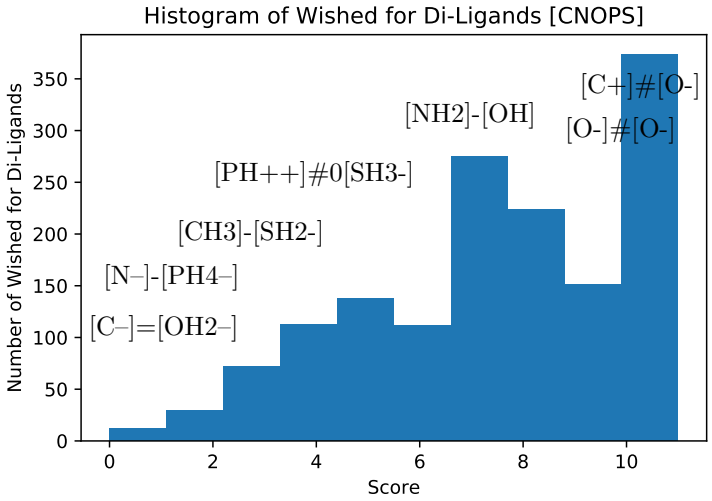

Figure 5: Histogram of all scored DHALs. The ones excluded in the beginning are not shown. Some examples are shown, representing bins at certain scores.

$$s = s_{\text{polarization}} + s_{\text{bond}} + s_{\text{charge}} + s_{\text{valency}} + s_{\text{sterics}} \ . \tag{14}$$

In Figure 5 a histogram of the scores is shown. In total, 1500 compounds were scored, spanning the scored space. If we truncate the space at a score $s > 9$ we can reduce the space to the desired space, which only contains 372 remaining compounds. In Figure 5 we see a 2D histogram. A trend is visible that a higher score favors more neutral charge. The bond orders decrease the larger they get (see Figure 7) and the charge is biased towards neutral (see Figure 8).

Lastly, we exclude from the scoring all compounds with triple bonds, allenic compounds, and compounds outside the charge range $2 \geq c_{tot} \geq -4$ from the set of **THALs**. The bonds are scored with

$$s_{\text{bond}} = \begin{cases} 3 & \text{if } b_{12} = b_{12} = 1 \\ 3 & \text{if } b_{23} = 2 \\ 0 & \text{if } b_{12} = 2 \ . \end{cases} \tag{15}$$

where the double bond in one place determines the other bond to be a single bond. Valency is scored as

$$s_{\text{valency}} = 5 - |\nu_1 - \nu_2| \ , \tag{16}$$

sterics on the CA as

$$s_{\text{sterics}} = \begin{cases} 0 & \text{if } h_1 = 3 \vee h_2 = 3 \\ 3 & \text{if otherwise} \ , \end{cases} \tag{17}$$

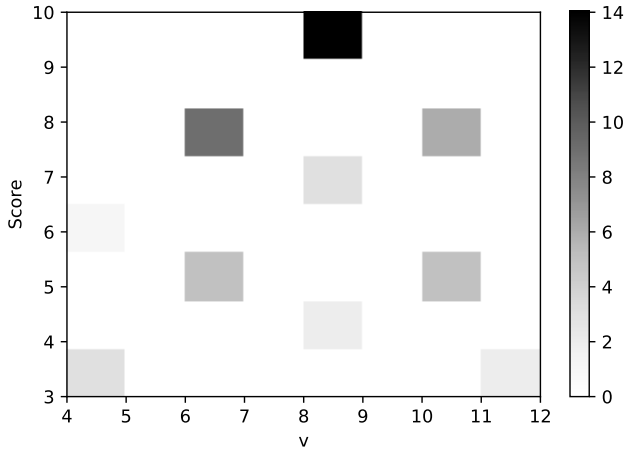

Figure 6: A 2D histogram to show the relationship between score and valence electrons of the DHALs.

where the $h_1 = 3 \wedge h_2 = 3$ case never happens. Finally, the charge is scored as

$$u_{\text{charge}} = \begin{cases} 5 & \text{if } c_1 = 1 \vee c_2 = 0 \\ 3 & \text{if } (c_1 = 1 \vee c_2 = -1) \wedge (c_1 = -1 \vee c_2 = 1) \\ 3 & \text{if } (c_1 = -1 \vee c_2 = -1) \\ 0 & \text{if } (c_1 = -2 \vee c_2 = -2) \\ 3 & \text{if } (c_1 = 0 \vee c_2 = -1) \wedge (c_1 = -1 \vee c_2 = 0) \\ 0 & \text{if } (c_1 = 2 \vee c_2 = -1) \wedge (c_1 = -1 \vee c_2 = 2) \\ 2 & \text{if } (c_1 = 0 \vee c_2 = -2) \wedge (c_1 = -2 \vee c_2 = 0) \\ 1 & \text{if } (c_1 = -2 \vee c_2 = -1) \wedge (c_1 = -1 \vee c_2 = -2) \\ 2 & \text{if } (c_1 = -2 \vee c_2 = 0) \wedge (c_1 = 0 \vee c_2 = -2) \ . \end{cases} \tag{18}$$

The total score for the THALs is then the sum over all subtotals:

$$s = s_{\text{bond}} + s_{\text{valency}} + s_{\text{sterics}} + s_{\text{charge}} \ . \tag{19}$$

In total, 1635 THALs were scored which we reduced to 148 THALs by selecting only the ones with a score higher than 13. In Figure 9 we see a 2D histogram of charge vs. isoelectronics. The bond orders decrease the larger they get (see Figures 11 and 12) and the charge is biased towards neutral (see Figures 13, 14, and 15).

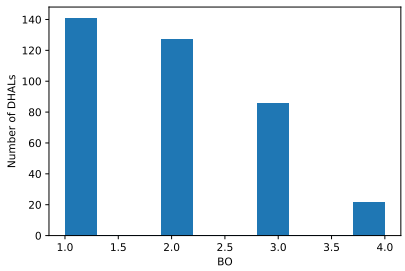

Figure 7: A histogram of the desired DHALs's bond orders. We see that we have much more low bond orders than high bond orders..

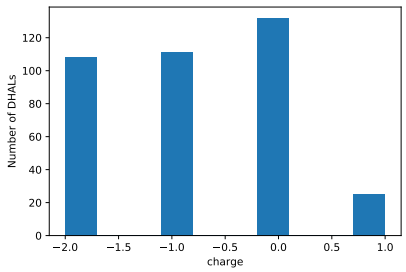

Figure 8: A histogram of the desired DHALs's charges. We can observe a bias towards neutral ligands as well as a disfavor of positive ligands.

1.4 The realm of the possible: The enumerated space

The enumerated space, $\mathcal{S} \subset \mathbb{R}^{155}$, will be the space in which our solutions live, accounting for the 155 RACs. Our constraints allow us to assemble around 1.5 M complexes. We choose 20 MHALs, 80 DHALs and 50 THALs. Constraining on weak symmetry (axial positions distinguishable, symmetrical equatorial positions), we obtain $(20 + 80)^3 + 50^2 + 50 \cdot 100^2 = 1,502,500$ compounds.

The 20 MHALs were chosen uniformly proportionally to their isoelectronic distribution, around 66 % of each bin of isoelectronics. The DHALs were chosen accordingly, around 20 % of each bin. The bidentates were chosen without regard on the isoelectronic distribution.

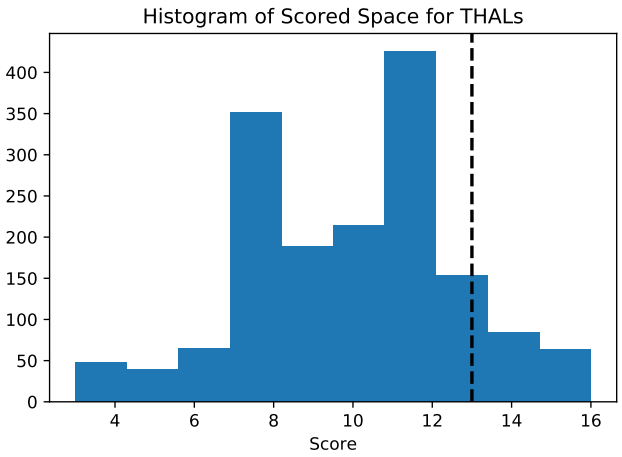

Figure 9: Histogram of all scored THALs. The ones excluded in the beginning are not shown.

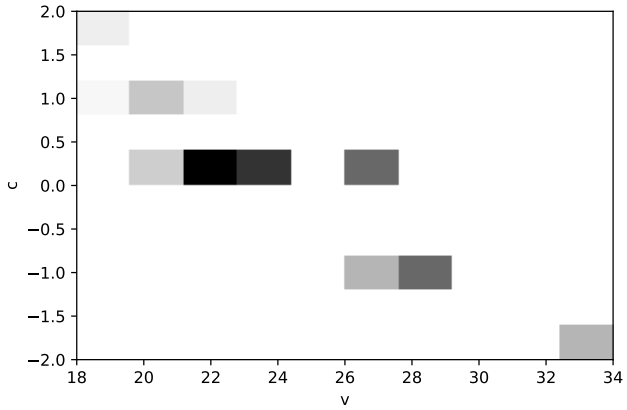

Figure 10: A 2D histogram to show the relationship between score and valence electrons of the THALs.

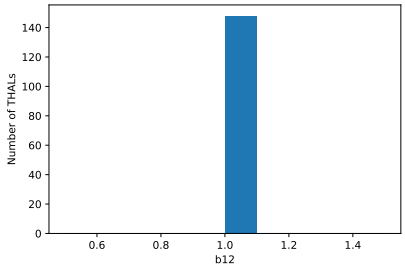

Figure 11: A histogram of the desired THALs's bond orders in bond b_{12}. We only see single bonds, as is expected.

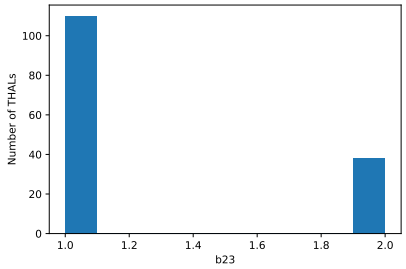

Figure 12: A histogram of the desired THALs's bond orders in bond b_{23}. Wesee a few double bonds, as might be sterically expected.

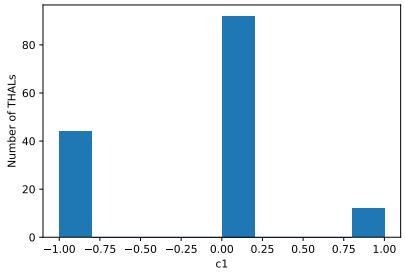

Figure 13: A histogram of the desired THALs's charges on c_1. We can observe a bias towards neutral ligands as well as a slight disfavor of positive ligands.

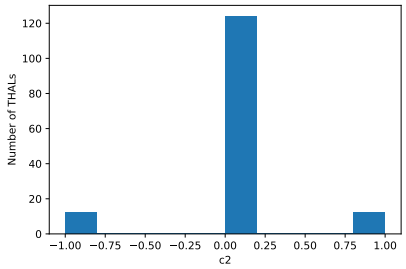

Figure 14: A histogram of the desired THALs's charges on c_2. We can observe a bias towards neutral ligands as well as a slight disfavor of positive ligands.

2 Ligand Field Assembly

2.1 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

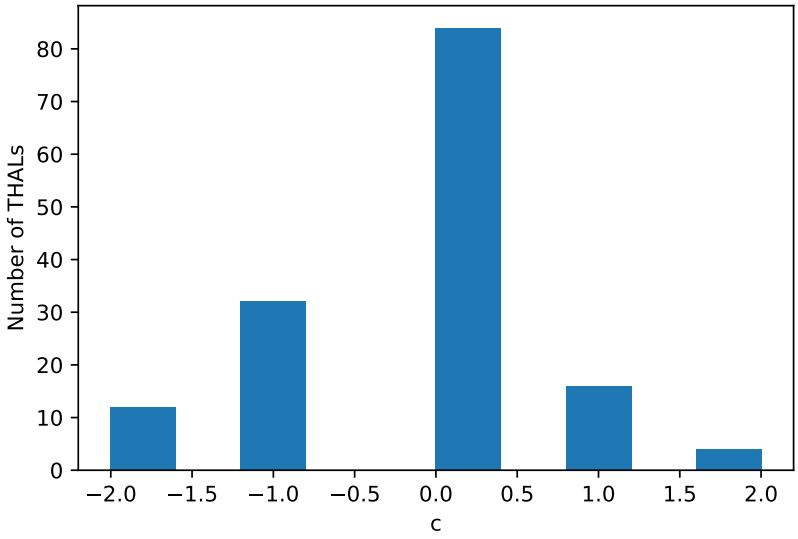

Figure 15: A histogram of the desired THALs's charges on c_{tot}. We can observe a bias towards neutral ligands as well as a slight disfavor of positive ligands.

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.

2.2 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 16). The connecting atom is colored according to the CPK coloring with a gradient transition from oxygen (red) to phosphorus(orange). The connecting 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

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	$> 1.8 \cdot 10^{14}$

find more fine grained complexes.

2.3 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 (KDE) distrbution and Scott's rule to estimate the bandwidth of the KDE:

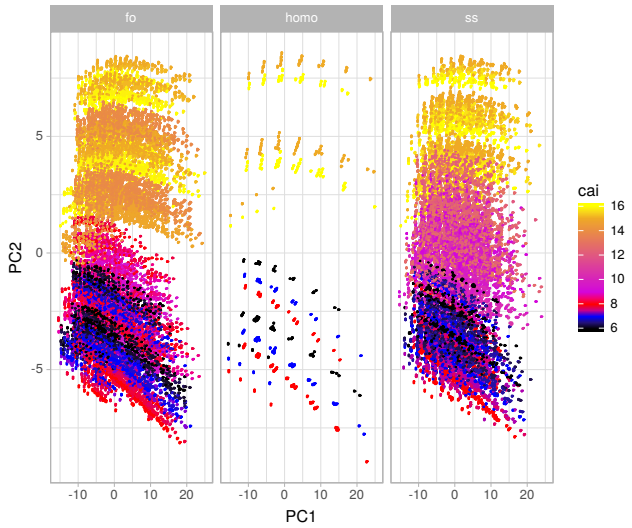

Figure 16: A PCA on the homoleptic subspace. The strongly symmetric (ss) and "5+1" symmetric (fo) subspace were projected onto it. The connecting atom identity (cai) is encoded in the color following CPK coloring with a gradient transition from oxygen (red) to phosphorus (orange).