

GDB-9 & Small Molecule Universes

Stefan O. Gugler

Massachusetts Institute of Technology
Department of Chemical Engineering

May 31, 2018

① GDB-9 dataset

② Small Ligand Universe

General

Reduction Rules

SLU Analysis

③ Assembly of complexes

General

GDB-9

- short explanation and outline of some of the rules of GDB-17
- introduce GDB-9 as subset of GDB-17
- look at their two-heavy-atoms molecules and compare to our project

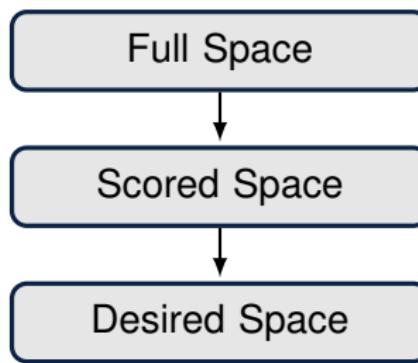
Motivation

- GDB-9 analogon for inorganic chemistry
- First and second shell complexes are representative
- Establish more informatics for inorganic chemistry

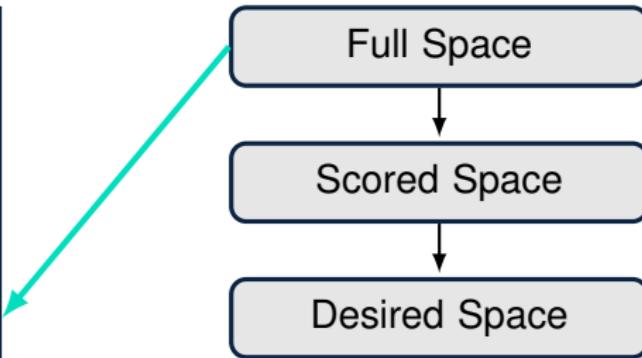
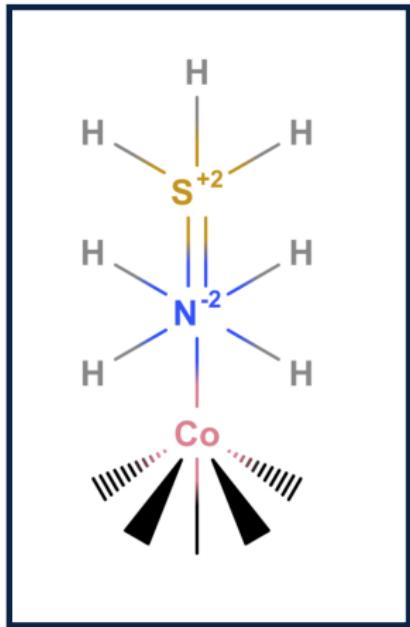
Procedure

- ① Generate small ligand universe (SLU)
- ② Assemble complexes with different symmetries from truncated SLU
- ③ Analysis of symmetry classes

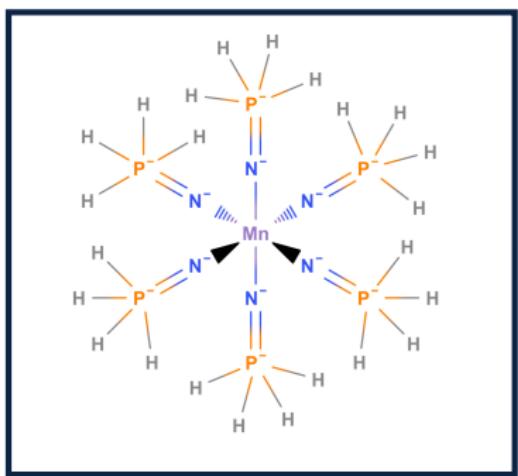
Reduction of the full space



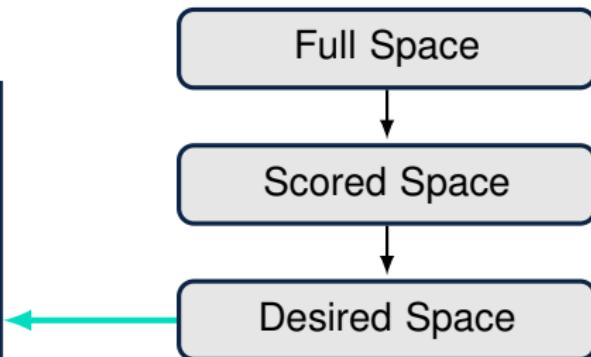
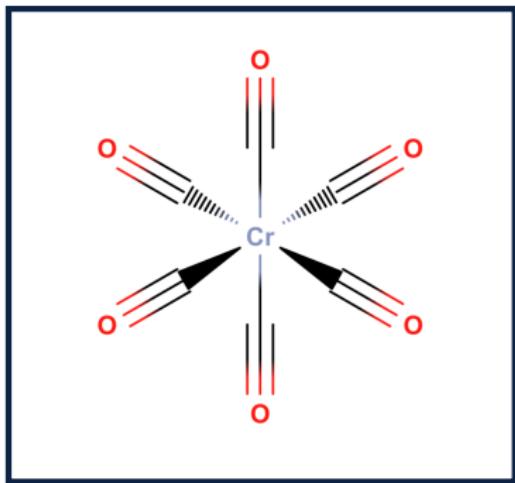
Reduction of the full space



Reduction of the full space



Reduction of the full space



Full Space → Scored Space

- charge $\in [-2, +2]$
- H atoms $\in [0, 4]$
- element $\in \{\text{C, N, O, P, S}\}$

Full Space:
M:125, D:5625,
B:5625



} charge, sterics, octet, shell, bond order

Scored Space:
50, 1171, 1635

These rules are demonstrated in the following for the di-heavy-atoms.

Rules to reduce Full Space → Scored Space

Constraints:

- Charge $c = c_1 + c_2 \leq 1$
- Sterics: H atoms < 4 on connecting atoms
- Closed shell (even number of electrons)

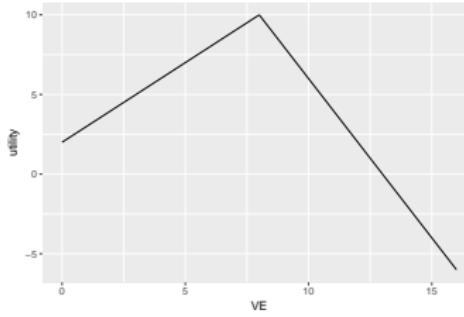
→ Reduces 5625 ligands to 1171.

Octet Rule

Molecules that fulfill the octet rule are more stable.

$$u_{\text{octet},i} = \begin{cases} 10 + 2 \cdot (8 - VE_i) & \text{if } 8 - VE_i < 0 \\ 10 - 1 \cdot (8 - VE_i) & \text{if } 8 - VE_i \geq 0 \end{cases} \quad (1)$$

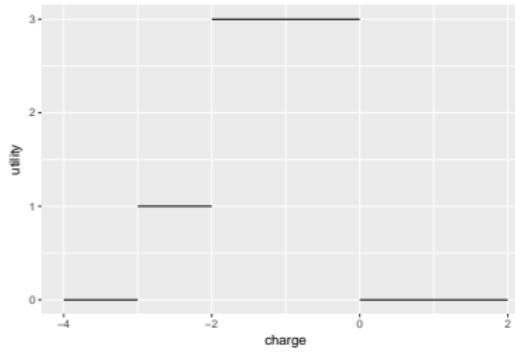
$$u_{\text{octet}} = \frac{1}{2} \sum_i^2 u_{\text{octet},1} + u_{\text{octet},2} \quad (2)$$



Charge Constraints

We reward only mildly negatively charged compounds.

$$u_{\text{charge}} = \begin{cases} 0 & \text{if } c_1 + c_2 > 0 \\ 3 & \text{if } 0 \geq c_1 + c_2 \geq -2 \\ 1 & \text{if } c_1 + c_2 = -3 \\ 0 & \text{if } c_1 + c_2 = -4 \end{cases} \quad (3)$$



VSEPR Constraints

If two atoms have the same amount of bond ready electrons, they are more stable.

$$U_{VSEPR} = 5 - \operatorname{Diff}_i (VE_i - 2 \cdot LP_i + c_i - 2 \cdot h_i) \quad (4)$$



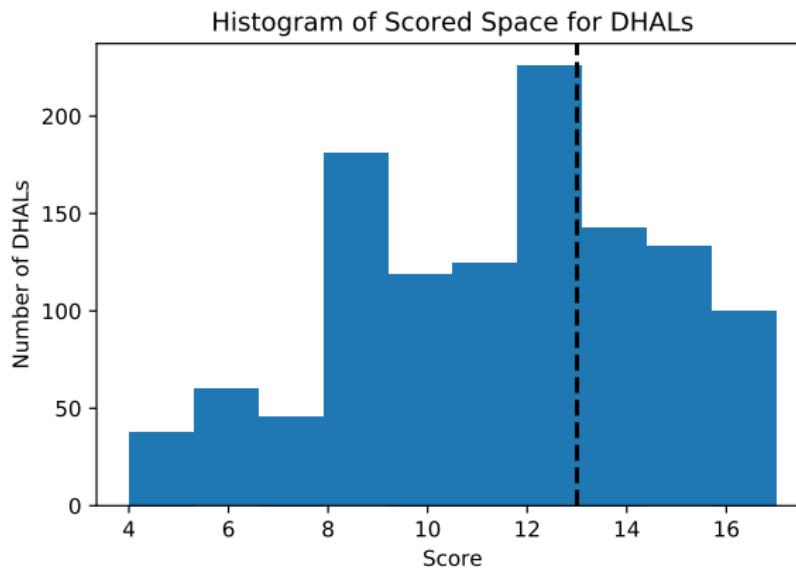
Sterics

All compounds with more than three H atoms on the connecting atom will be sterically hindered.

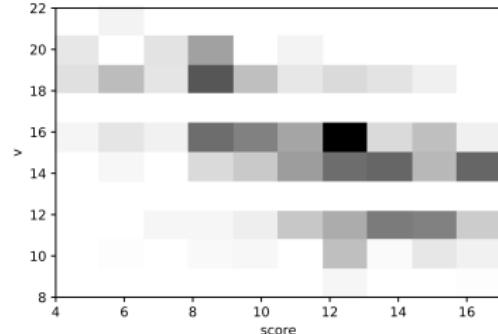
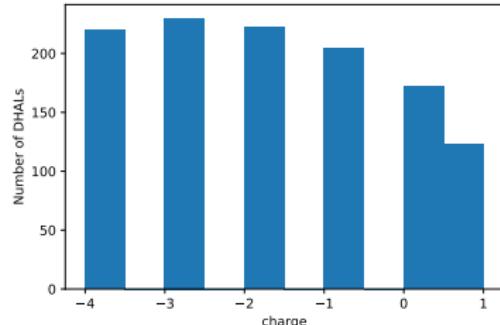
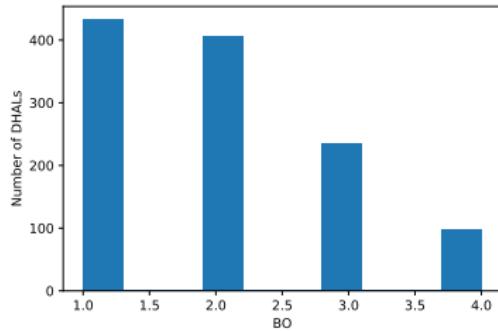
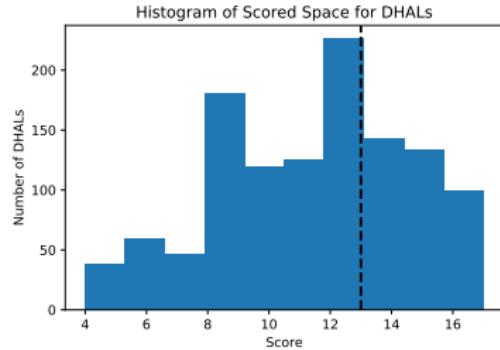
$$u_{CA} = \begin{cases} 2 & \text{if } h_1 = 3 \\ 3 & \text{if } h_1 < 3 \end{cases} \quad (5)$$

Global utility function

$$u_{\text{total}} = u_{\text{octet}} + u_{\text{charge}} + u_{\text{VSEPR}} + u_{\text{CA}} \quad (6)$$



Analysis: Score, bond order, charge, valence electrons



Recover spectrochemical series

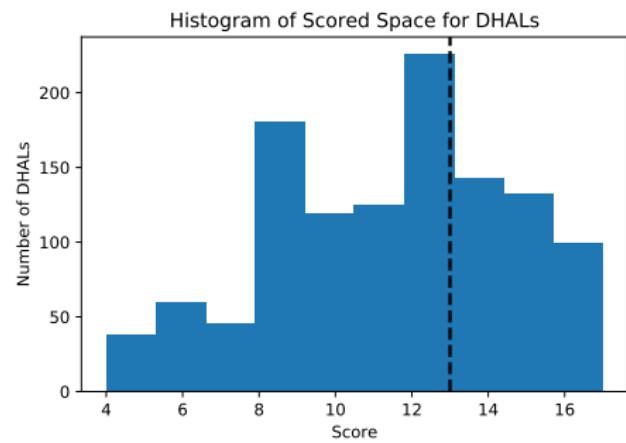
Spectrochemical series is recovered in the top scores.

Ligand	Score
[C]#[N-]	14
[C+]#[O-]	15
[N-]#[C]	15
[N+]#[O]	13
[O-]#[O-]	14
[S-]#[S-]	14

$s = 15$ also contains species like $[\text{CH}_2-]\#[\text{CH}]$, $[\text{OH}^-]=[\text{NH}]$, $[\text{NH}_2--]\#[\text{P}]$.

Desired space = top scoring ligands

We include all ligands with a score at least as high as the lowest scoring spectrochemical series ligand ($s = 13$)



This gives us 553 ligands in total.

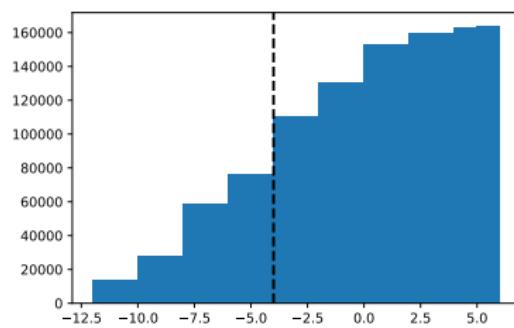
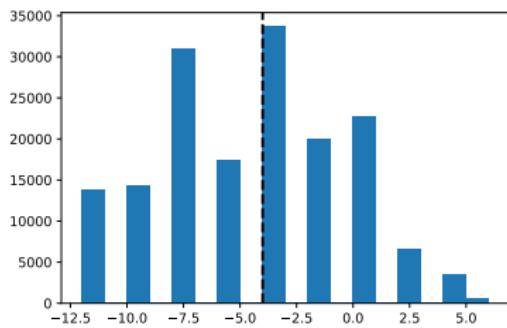
MOLECULE VIZ OF EACH CLASS

Subsets of octahedral space

Set	description	size
Homoleptics	$\text{eq} = \text{ax}$	553
"5+1" symmetric	$\text{eq} = \text{ax}_1 \neq \text{ax}_2$	163,620
"4+2" symmetric	$\text{eq}_1 \neq \text{eq}_2 = \text{ax}$	185,376
Strongly symmetric	$\text{eq} \neq \text{ax}$	245,316
Equatorially asymmetric	$\text{eq}_1 \neq \text{eq}_2 \neq \text{ax}$	15,924,796
Weakly symmetric	$\text{eq} \neq \text{ax}_1 \neq \text{ax}_2$	45,077,310
Complete Heteroleptics	$L_i \neq L_j$	$\approx 5.9 \cdot 10^{12}$
Octahedral Space	all	$> 1.8 \cdot 10^{14}$

Properties of the sets

- Reduce space to facilitate sampling from non-homoleptics
- Example: strongly symmetric, monodentate ligand fields (163,620)
- Exclude all with charge smaller than -4, which results in 87,150 ligand fields (53 %).



Principal Component Analysis

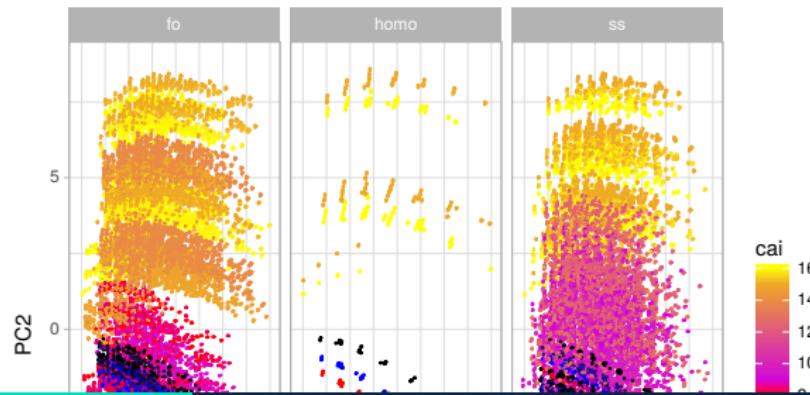
The homoleptics (ho) span the strong symmetry (ss) and "5+1" (fo) set.
what are the PCs, which explains which?

two slides: first of ho with rac155. new dataset into space, it spans the same idea.

just show only homo first colored by different props then show into SU.

rac39

show decay



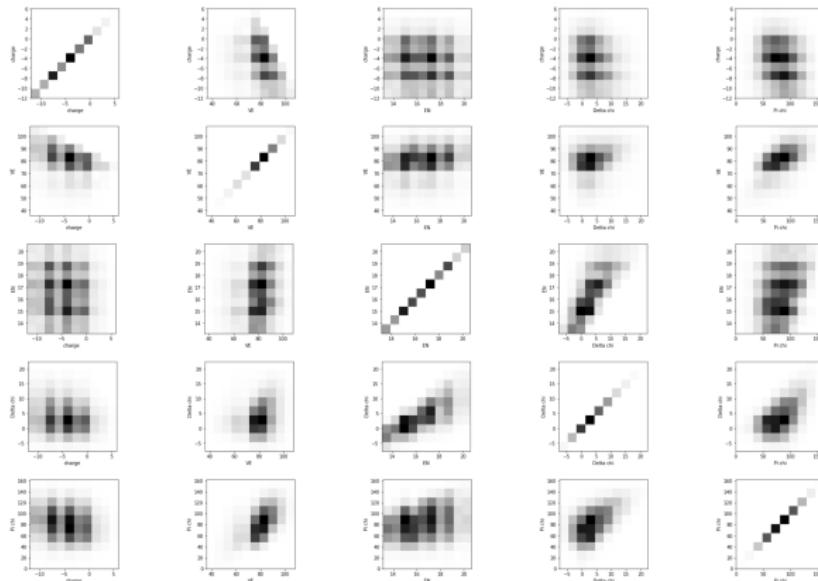
Footprint

We use five properties inspired by MCDL25 to characterize the ligand field and generate a five dimensional distribution:

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

Correlations for strongly symmetric monodentates

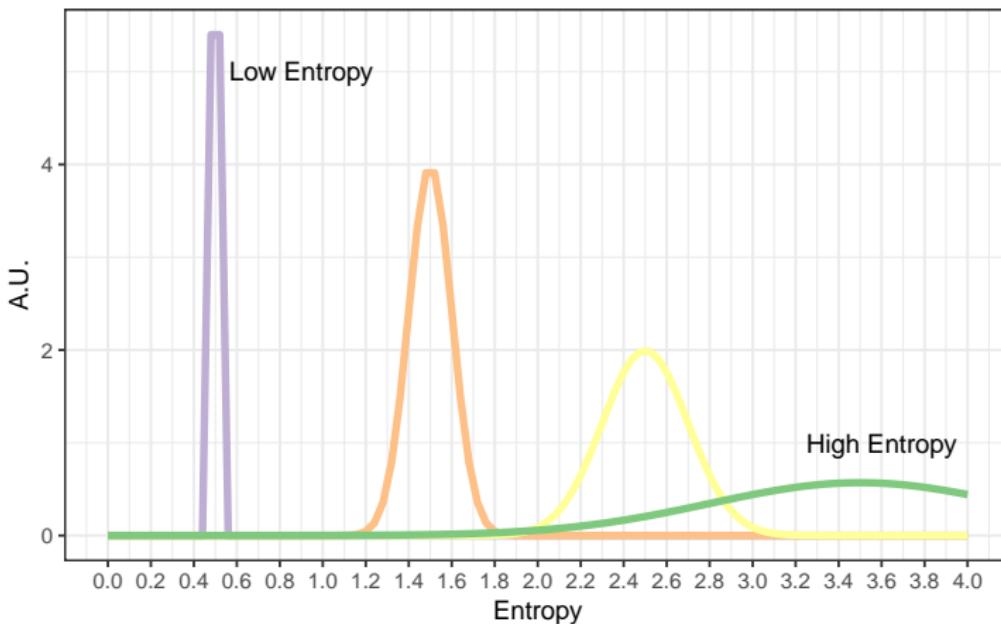
The less peaky the better.



Entropy and KDE

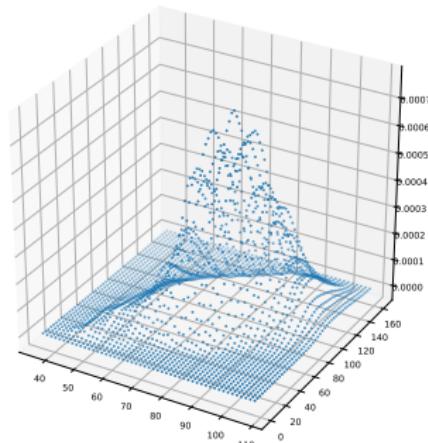
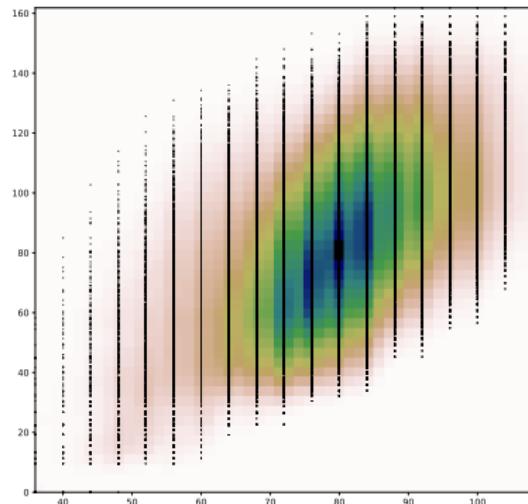
We then calculate the entropy, H_{KDE} , of the Kernel Density Estimated distribution. We want it to be uniform (high entropy) not to oversample.

Entropy Comparison



Example of KDE slice

Dimensions $_{\text{ax,eq}}^{\text{lc}} \chi_1$ vs. charge in H_{KDE} for strongly symmetric monodentates.

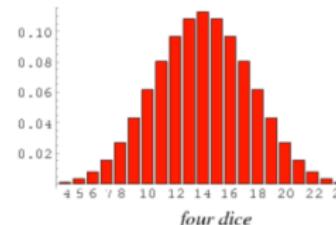
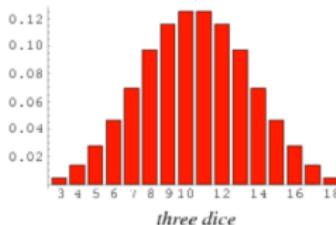
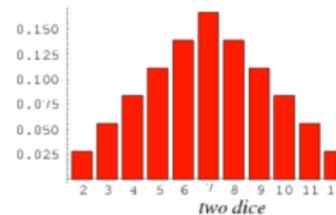
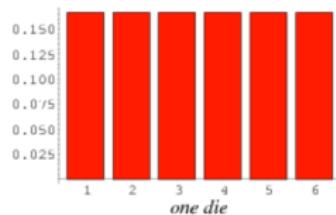


Monodentate Footprints

Table: Entropic footprint

Set	$H_{\text{KDE}}^{\text{monodent}}$	$H_{\text{KDE}}^{\text{bident}}$
Homoleptics	19.7	15.63
"5+1" symmetric	13.7	-
Strongly symmetric AC	-	9.47
Strongly symmetric ADC	12.70	5.53
"4+2" symmetric	12.70	9.47
Weakly symmetric	8.1	7.7
Equatorially asymmetric AC	-	10.04
Equatorially asymmetric ADC		

Multinomial peaking



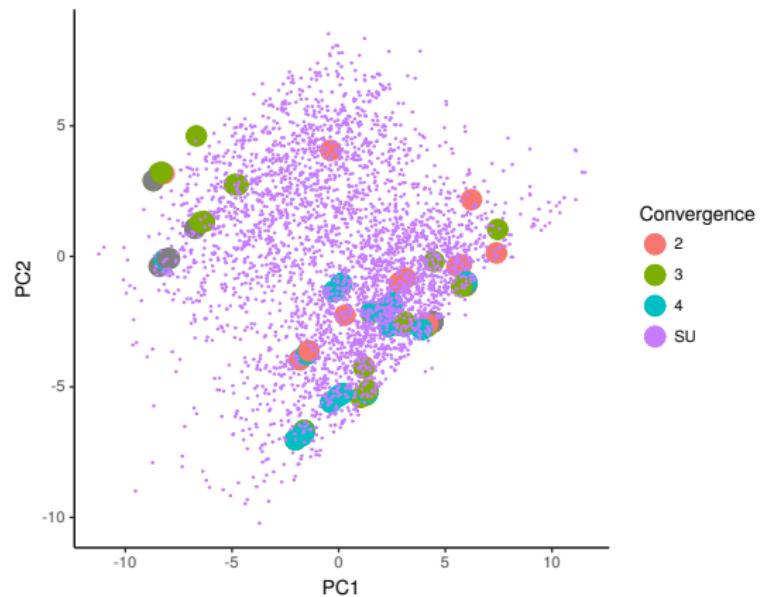
- The more dice, the more the distribution peaks which results in low entropy.
- Sample from low entropy uniformly and get similar molecules.

DFT calculations

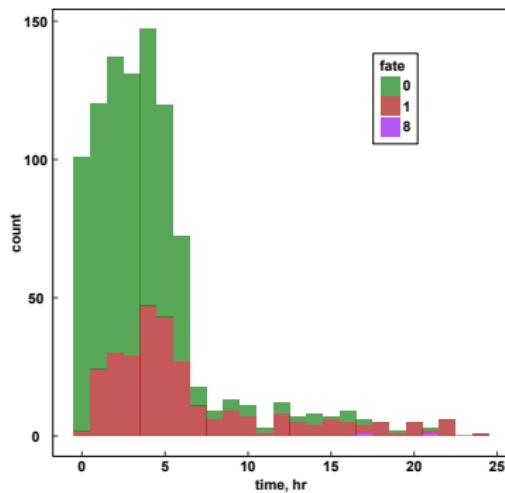
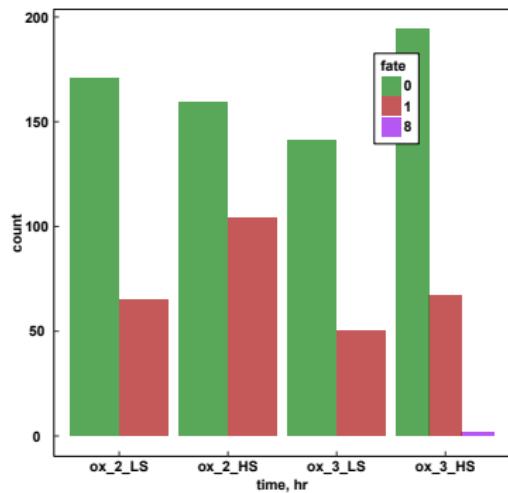
- Basis set: LACVPS (6-31G*)
- Effective core potential: LANL2DZ
- Functional: B3LYP

DFT calculations analysis

PCA of all sets with homoleptic results superimposed



DFT calculations analysis



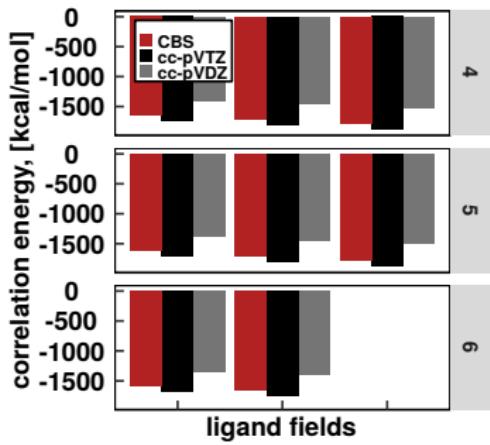
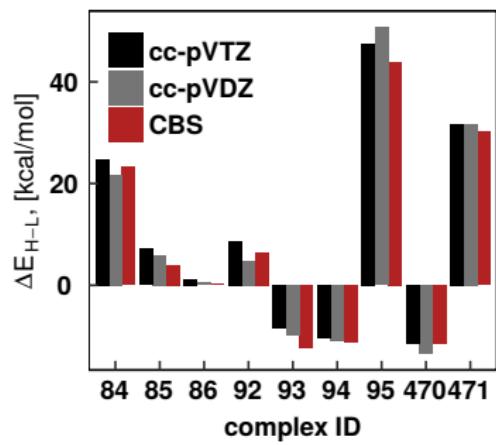
DFT calculations analysis

krr as small points and large points results from dft all colored by splitting energy

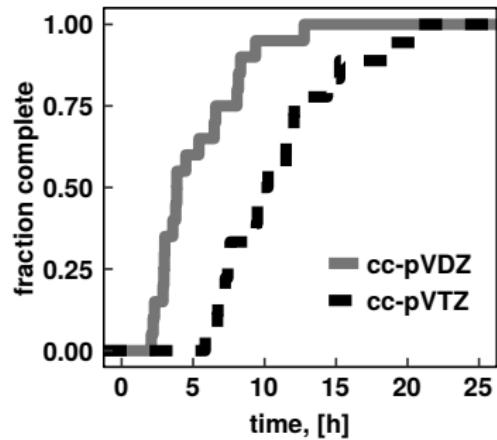
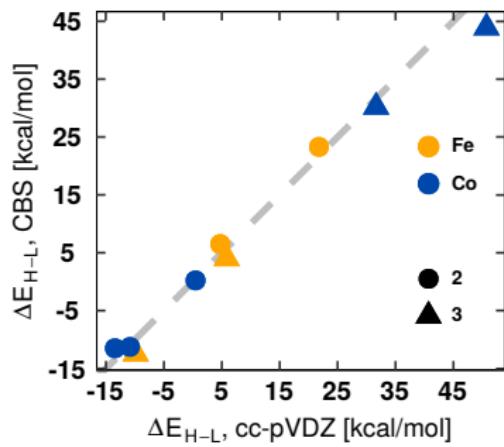
DLPNO calculations

- Basis set:
- Functional: B3LYP

Basis set comparison



CBS comparison and time analysis



Thanks for the opportunity to work in this lab!

