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

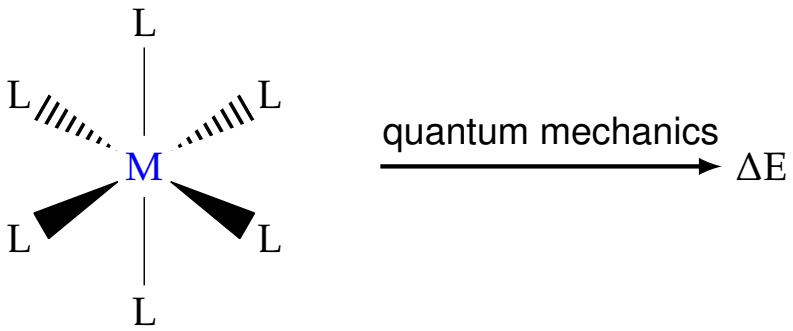
- short explanation and outline of some of the rules of GDB-17
- introduce GDB-9 as subset of GDB-17
- look at their two-heavy-atoms molecules and compare to our project

Motivation

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

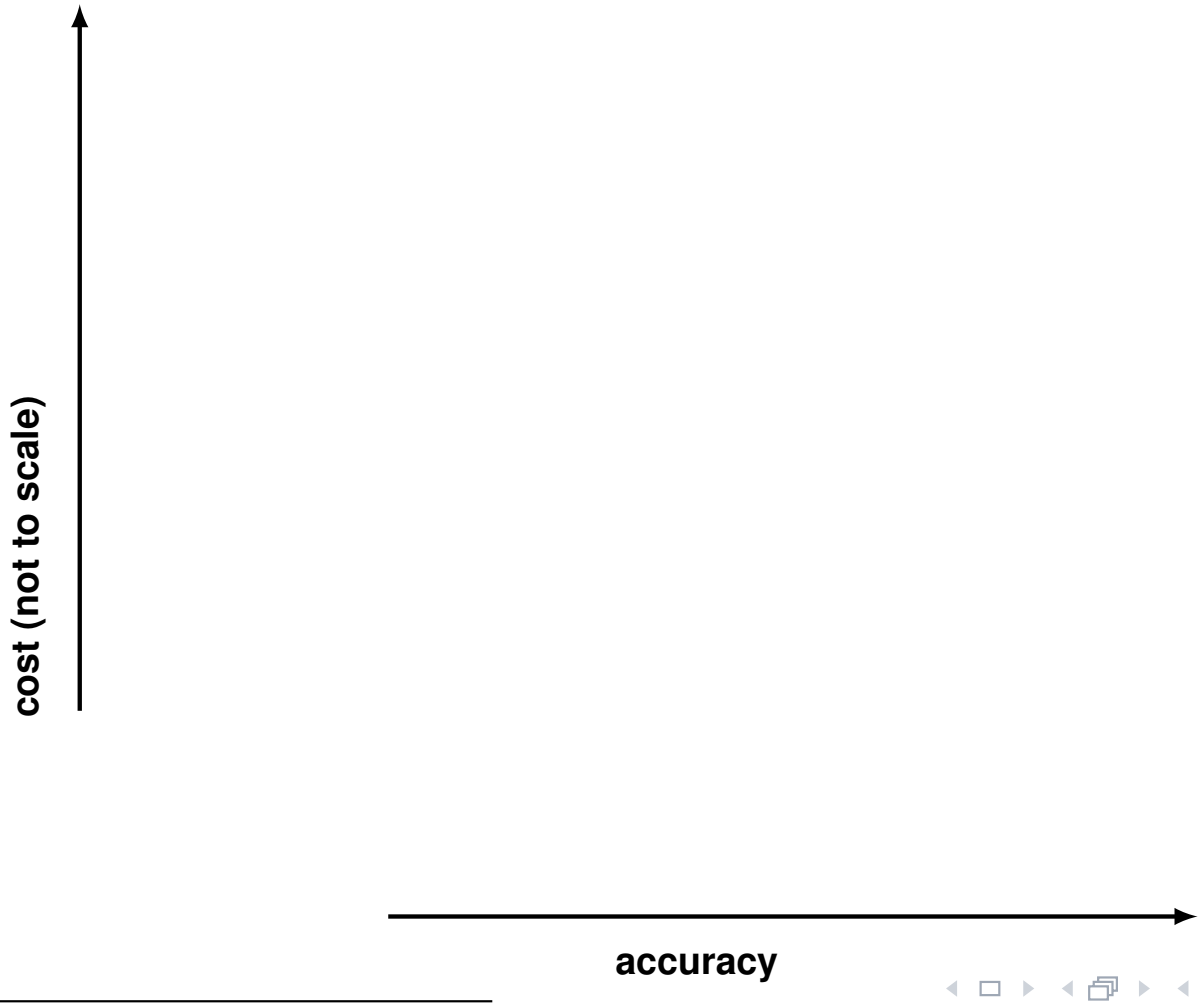
Connecting Structure and Activity

How to calculate properties?[1]



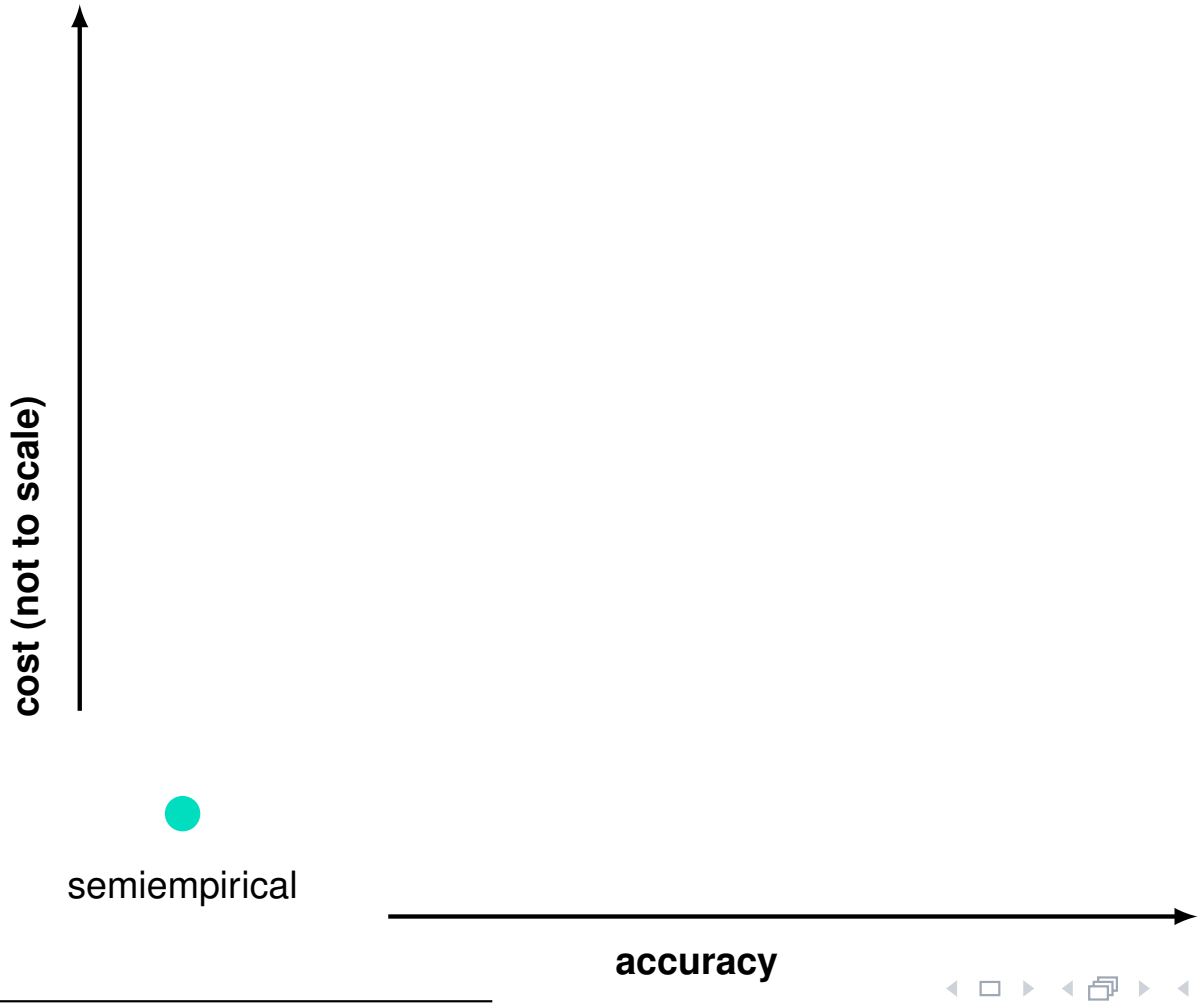
Connecting Structure and Activity

How to calculate properties?[1]



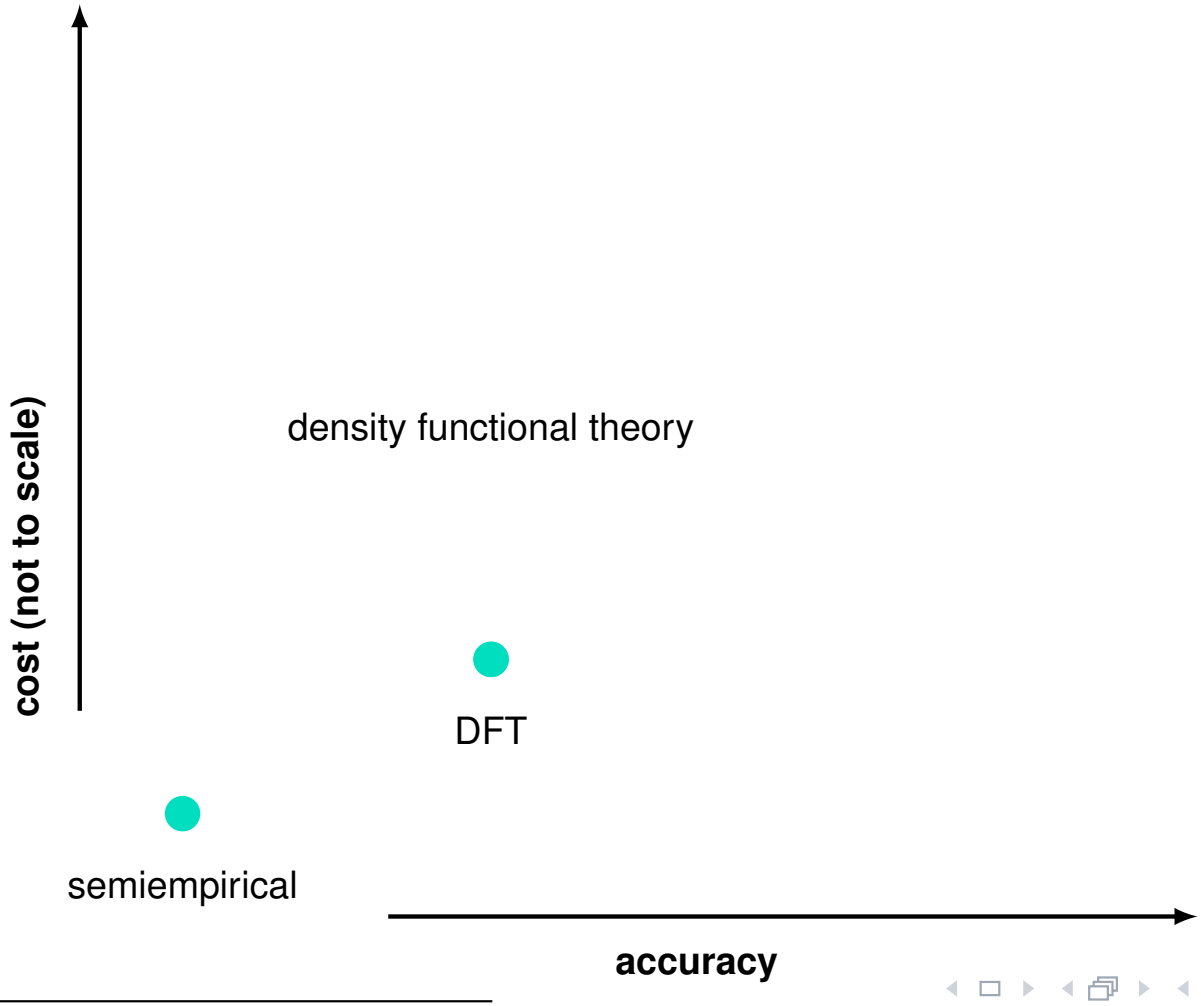
Connecting Structure and Activity

How to calculate properties?[1]



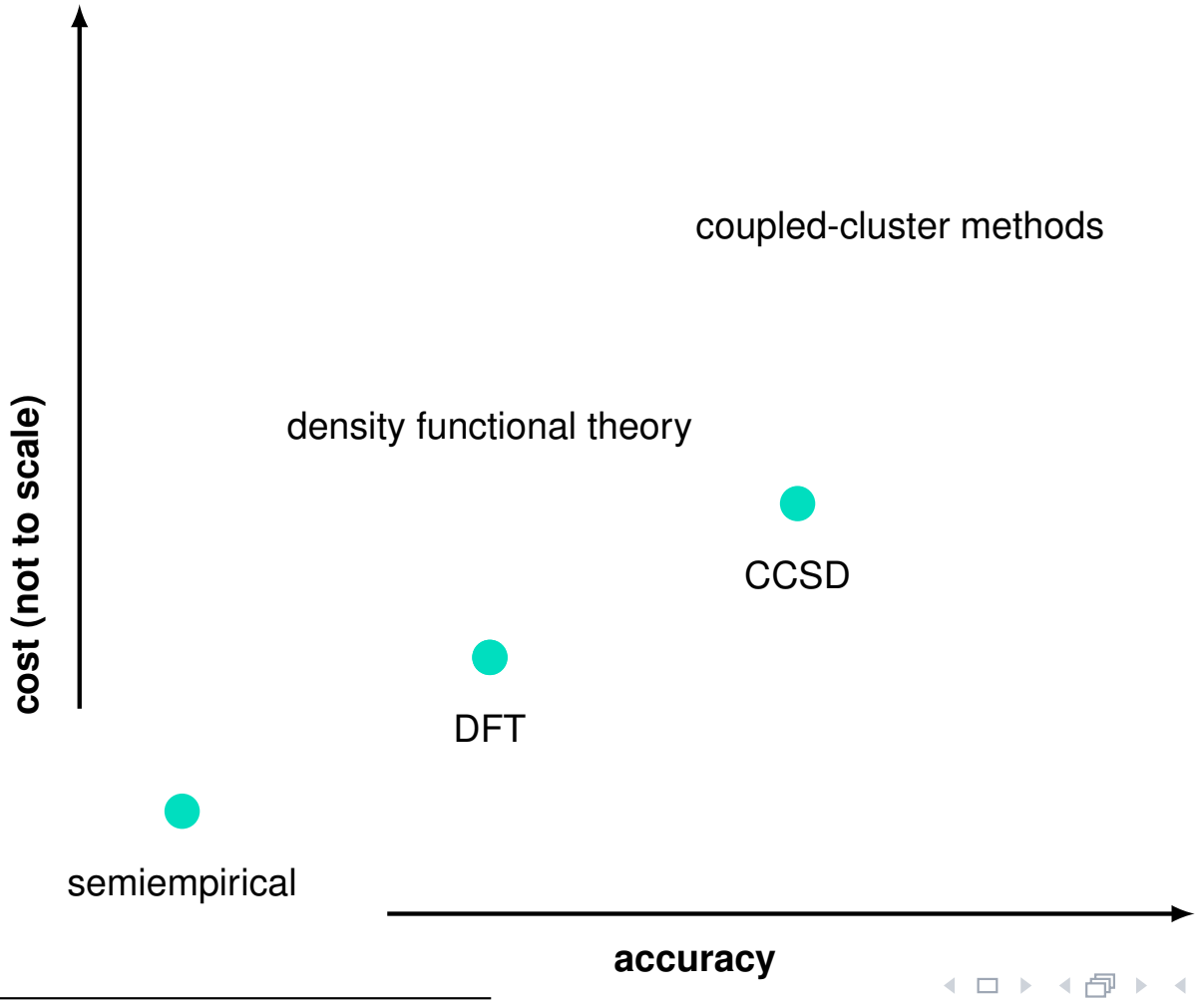
Connecting Structure and Activity

How to calculate properties?[1]



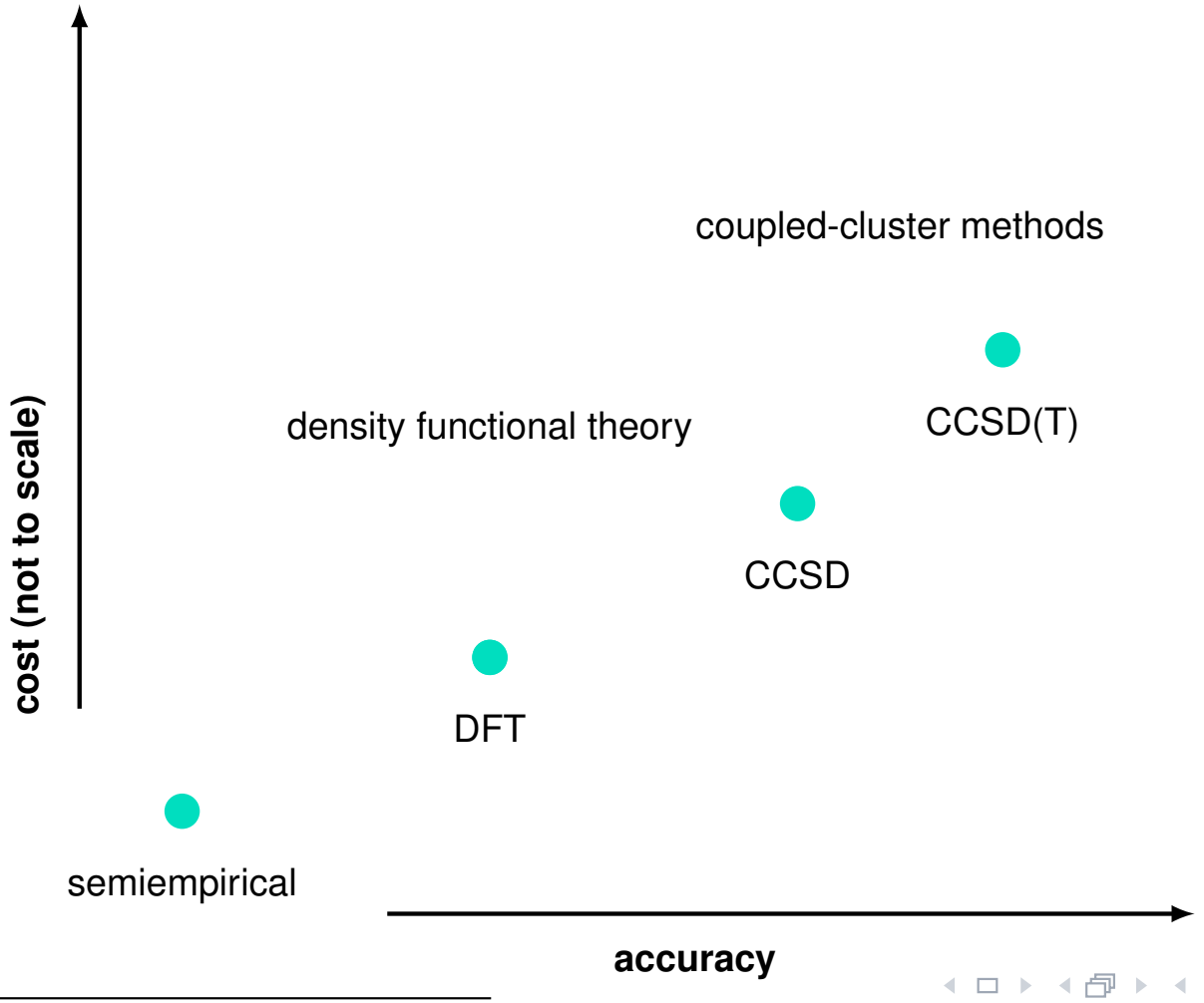
Connecting Structure and Activity

How to calculate properties?[1]



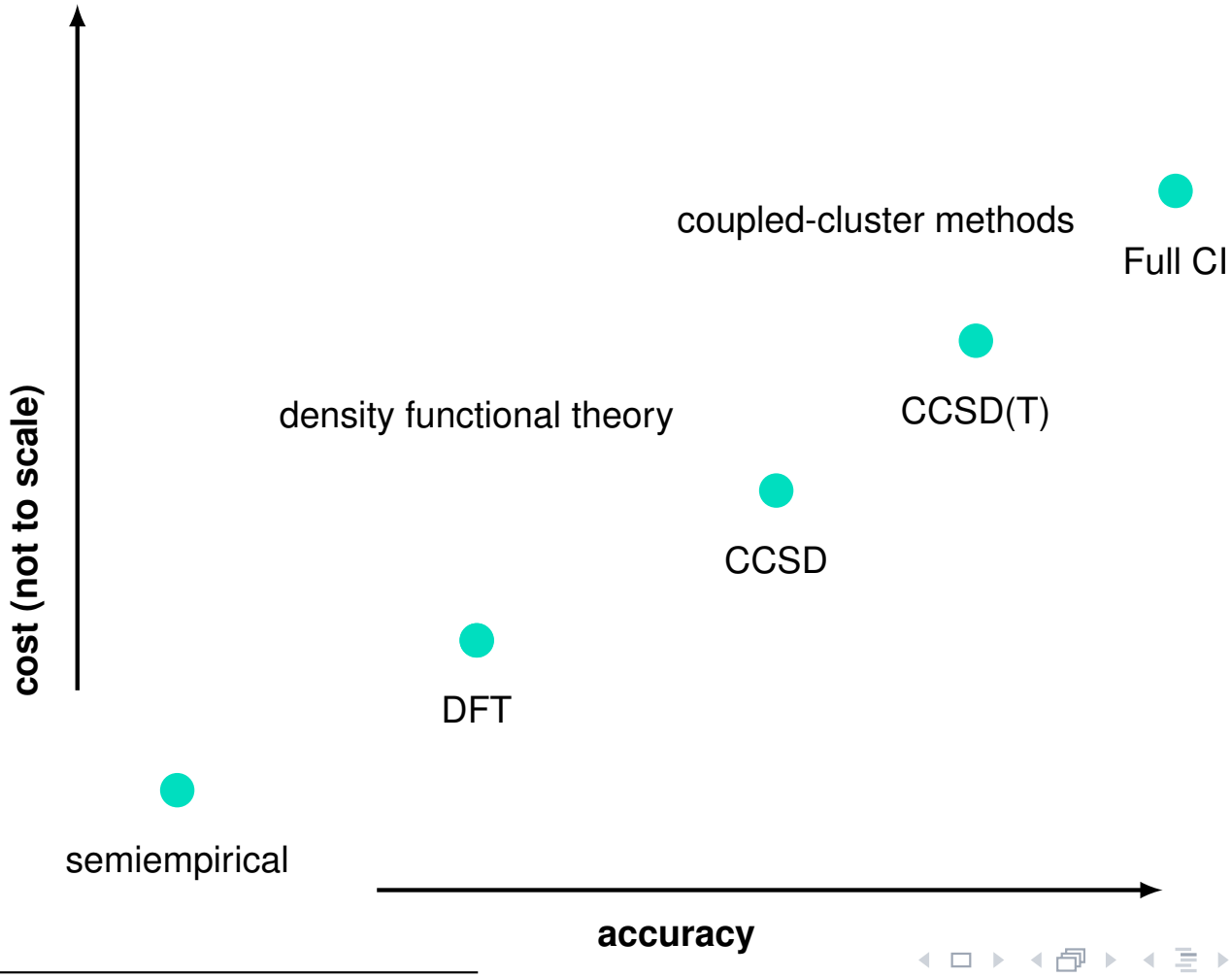
Connecting Structure and Activity

How to calculate properties?[1]



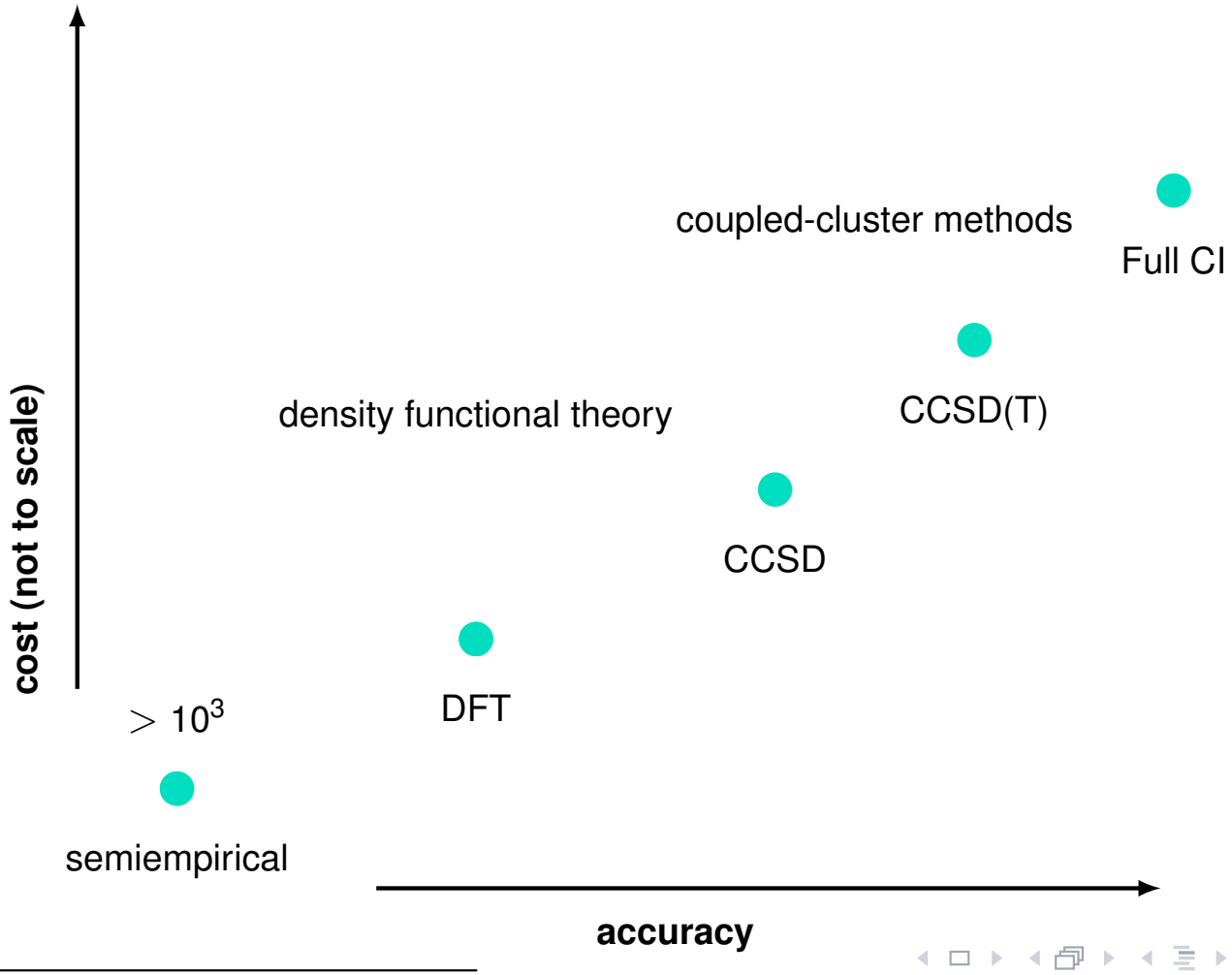
Connecting Structure and Activity

How to calculate properties?[1]



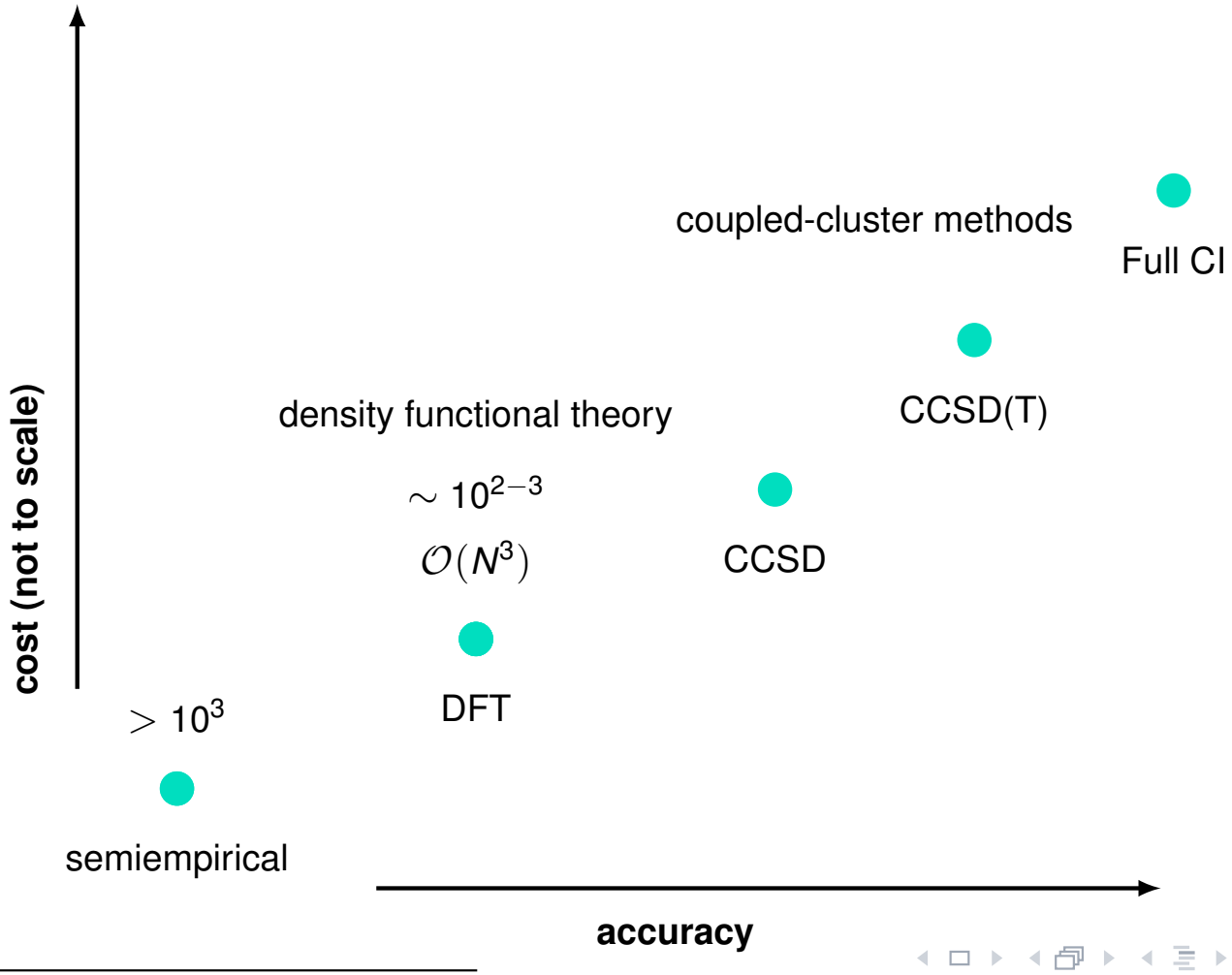
Connecting Structure and Activity

How to calculate properties?[1]



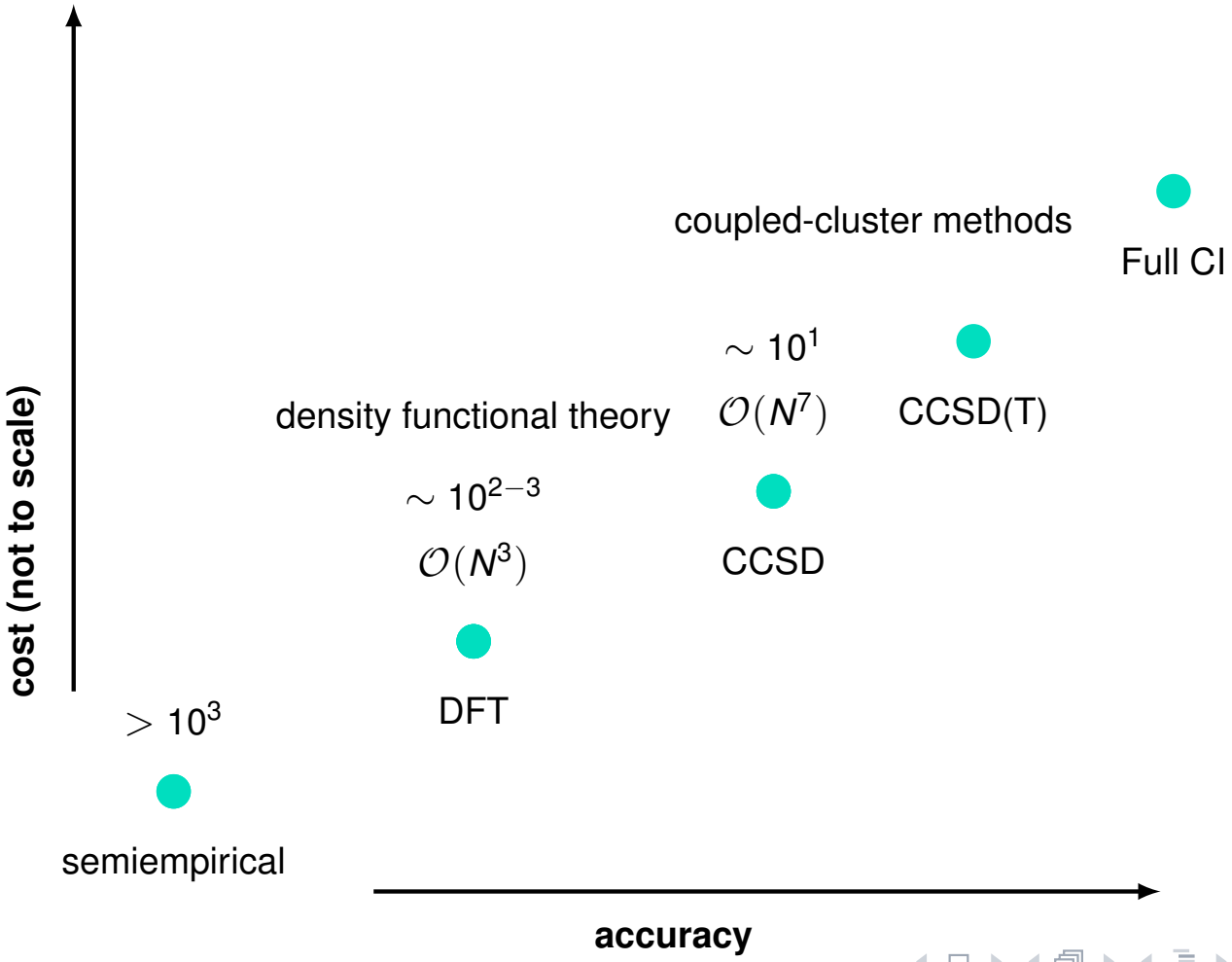
Connecting Structure and Activity

How to calculate properties?[1]



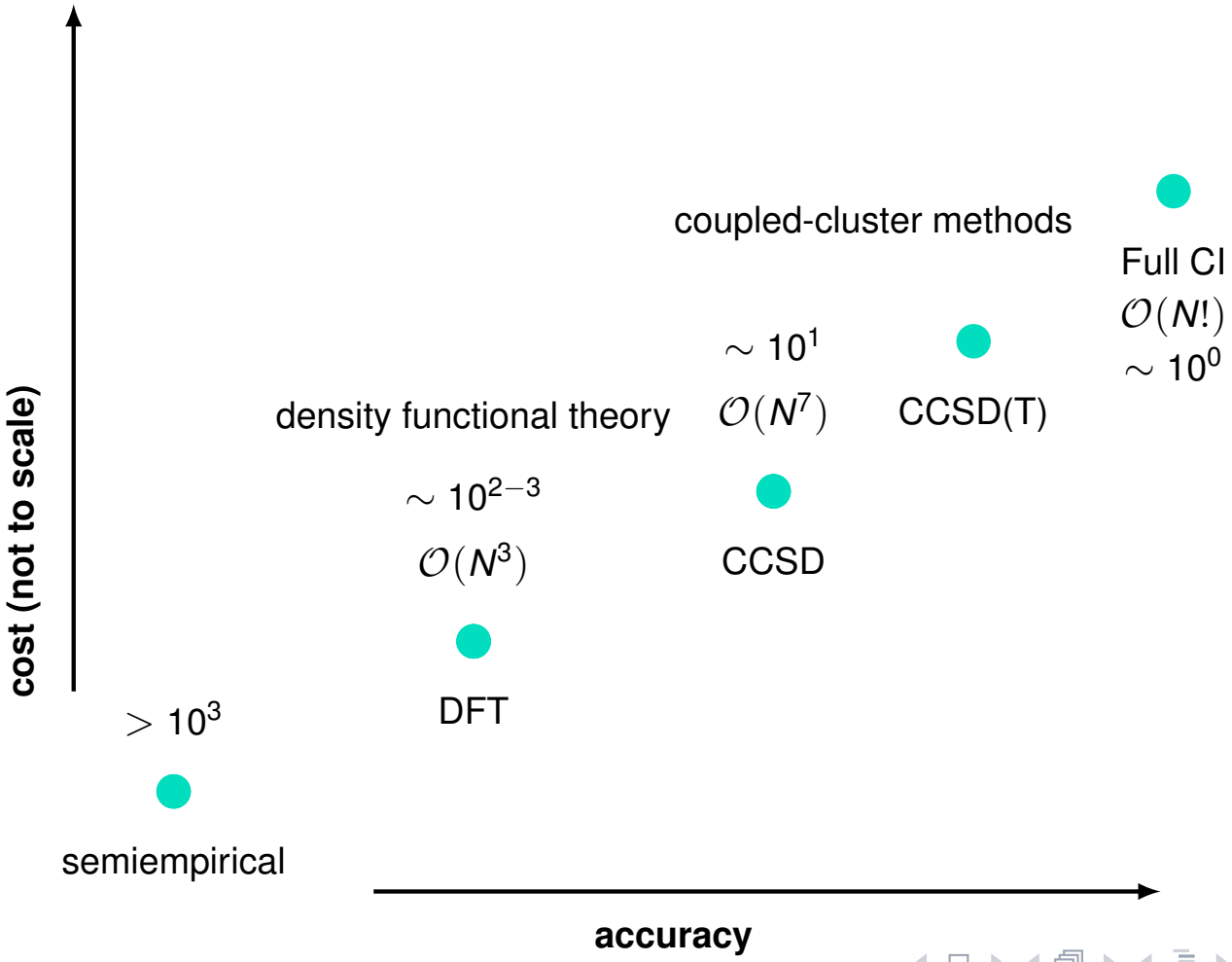
Connecting Structure and Activity

How to calculate properties?[1]



Connecting Structure and Activity

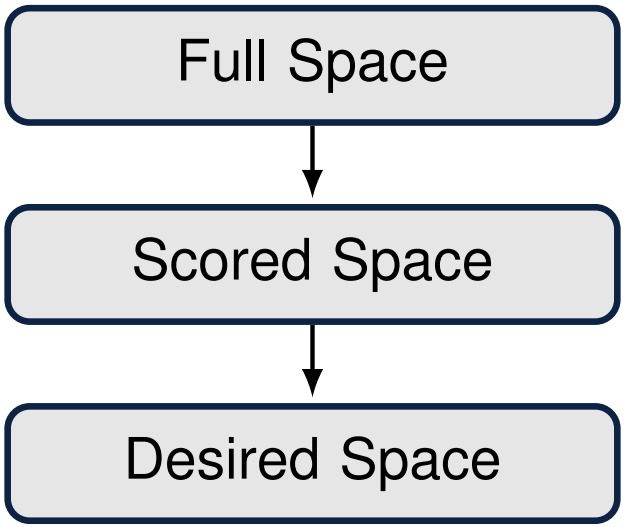
How to calculate properties?[1]



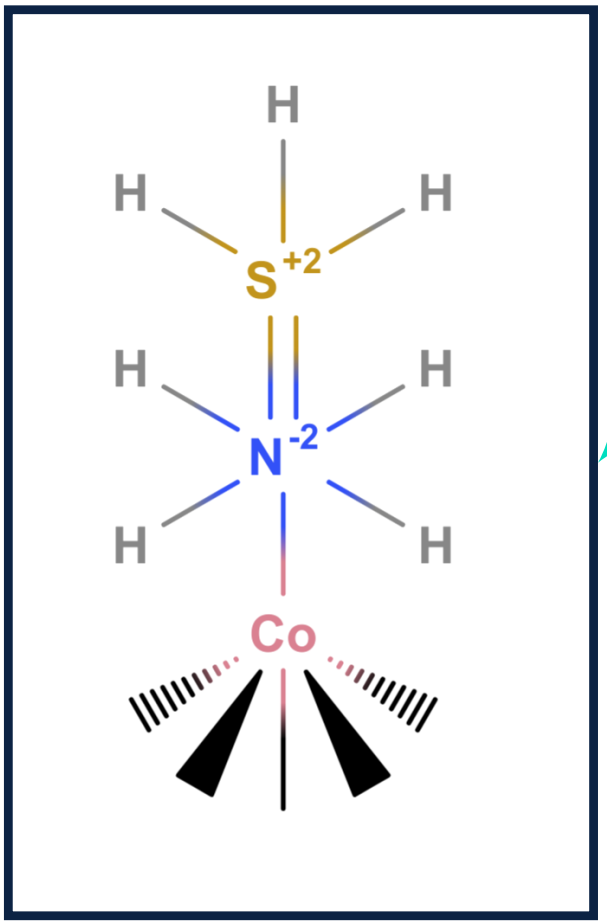
Procedure

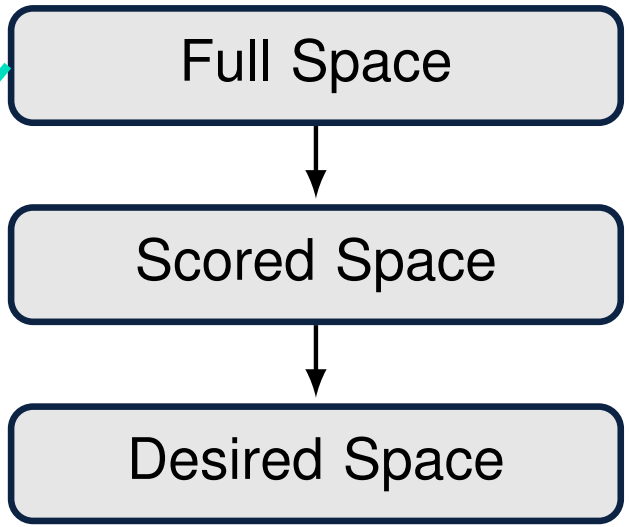
1. Generate small ligand universe (SLU)
2. Assemble complexes with different symmetries from truncated SLU
3. Analysis of symmetry classes

Reduction of the full space

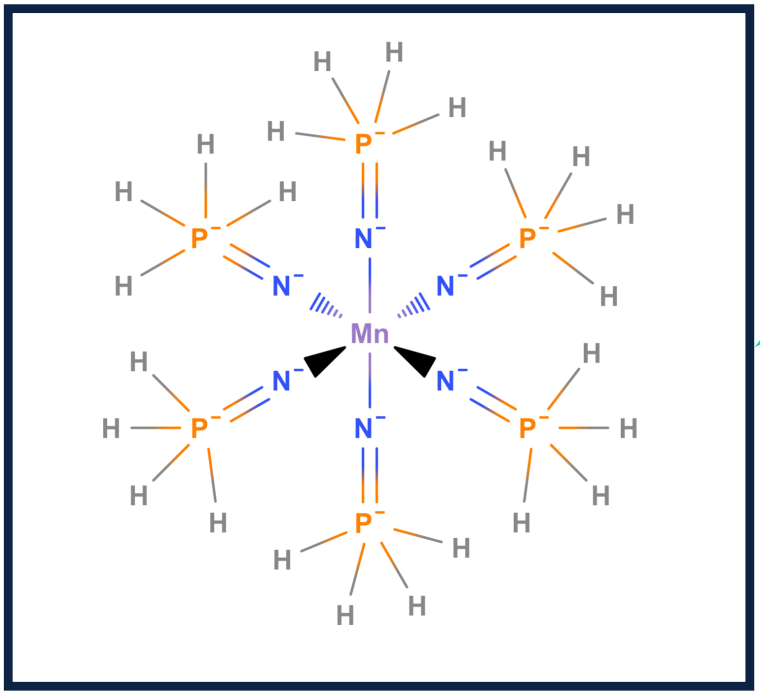


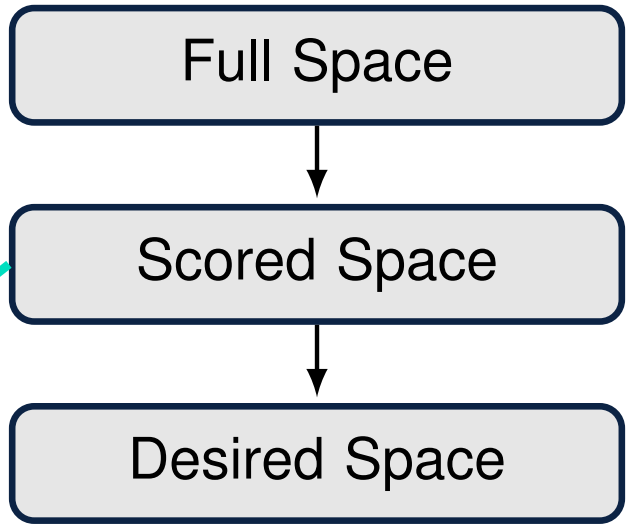
Reduction of the full space



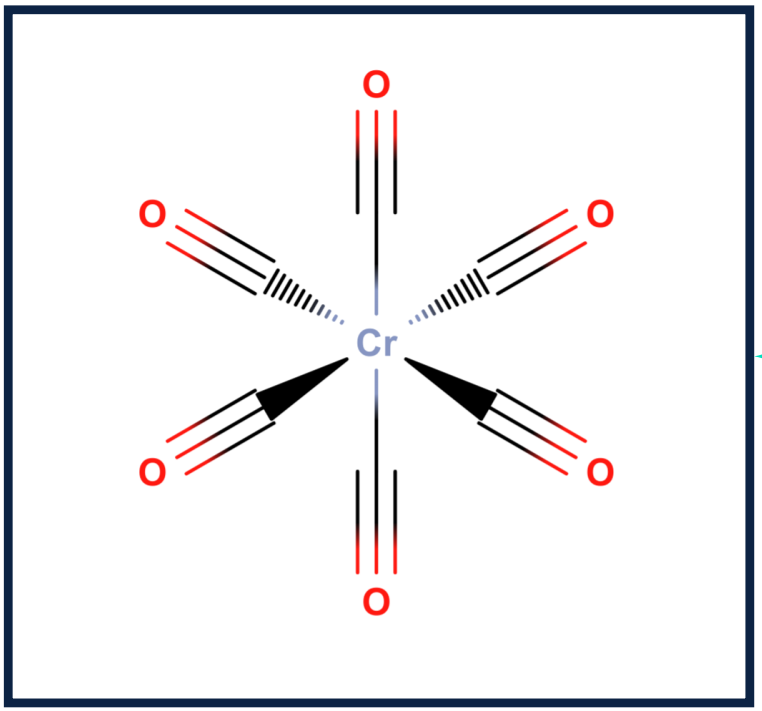


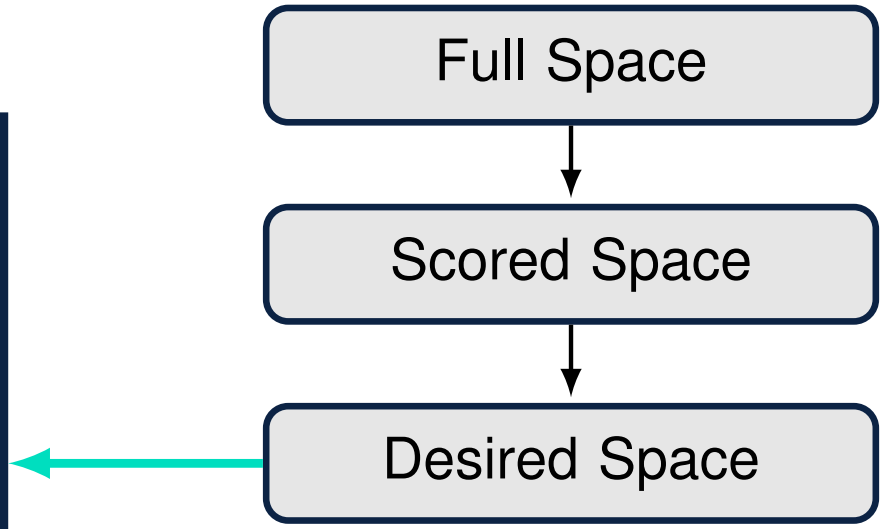
Reduction of the full space





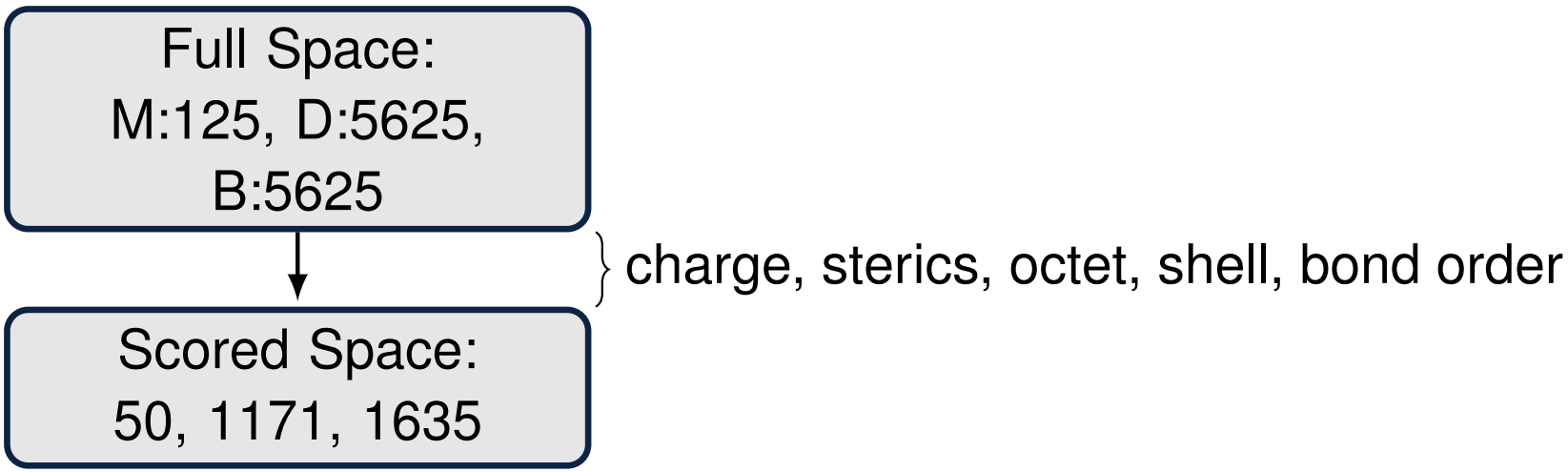
Reduction of the full space





Full Space $\rightarrow$ Scored Space

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



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

Rules to reduce Full Space $\rightarrow$ Scored Space

Constraints:

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

$\rightarrow$ Reduces 5625 ligands to 1171.

Octet Rule

visualization

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

$$u_{\text{octet}} = \frac{1}{2} \sum_{i}^{2} u_{\text{octet,1}} + u_{\text{octet,2}} \tag{2}$$

Charge Constraints

visualization

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

VSEPR Constraints

visualization

$$u_{\mathrm{VSEPR}} = 5 - \underset{i}{\mathrm{Diff}} \left(VE_i - 2 \cdot LP_i + c_i - 2 \cdot h_i \right) \tag{4}$$

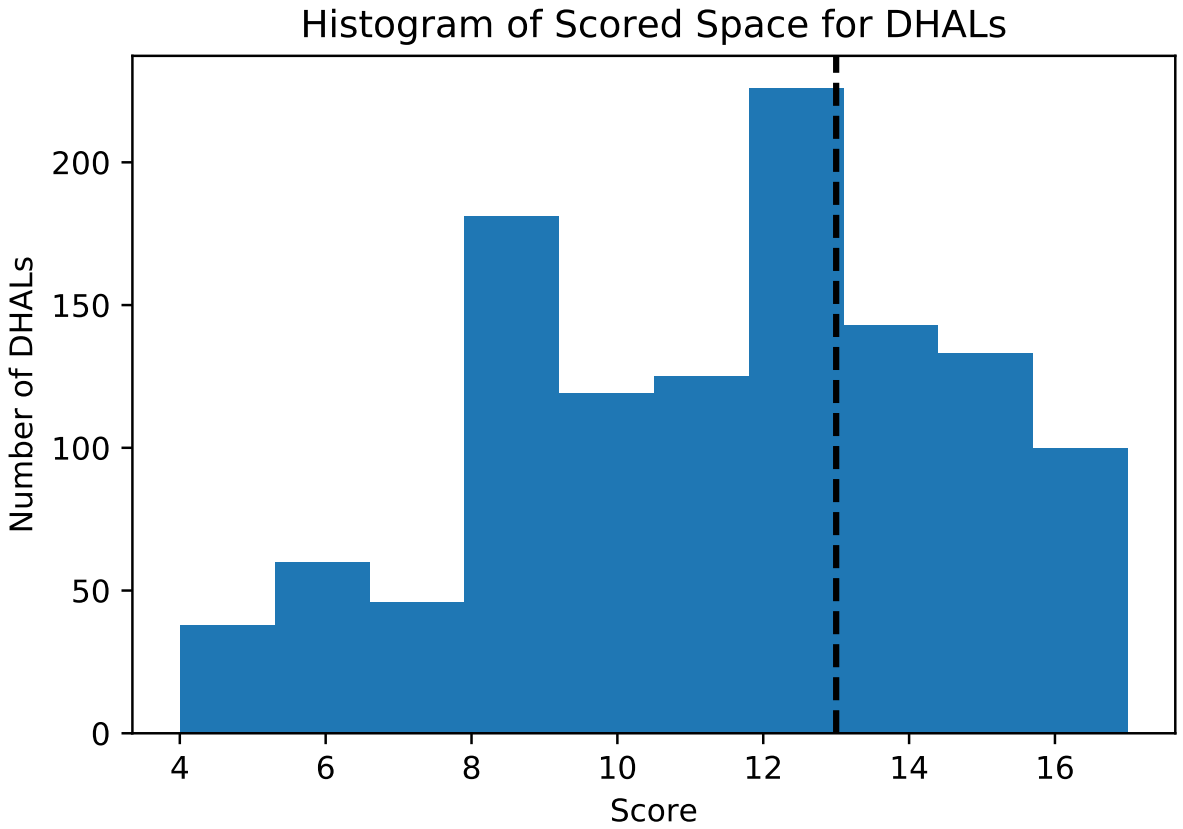
Sterics

visualization

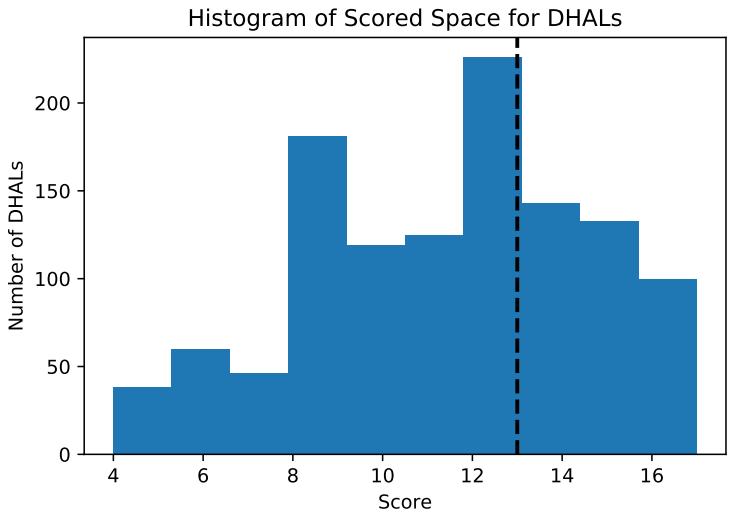
$$u_{\mathrm{CA}} = \begin{cases} 1 & \text{if } h_1 = 4 \\ 2 & \text{if } h_1 = 3 \\ 3 & \text{else} \end{cases} \tag{5}$$

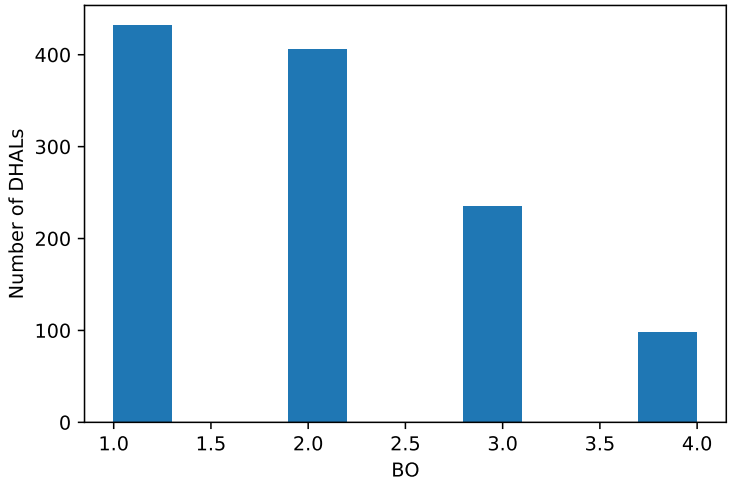
Global utility function

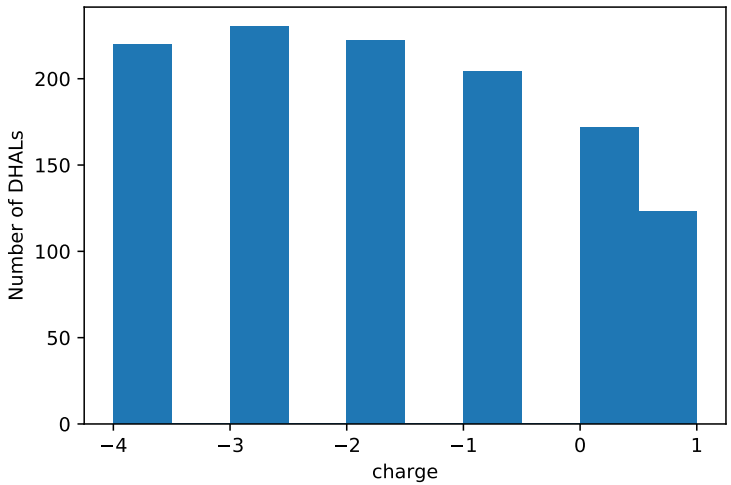
$$u_{\text{total}} = u_{\text{octet}} + u_{\text{charge}} + u_{\text{VSEPR}} + u_{\text{CA}} \tag{6}$$

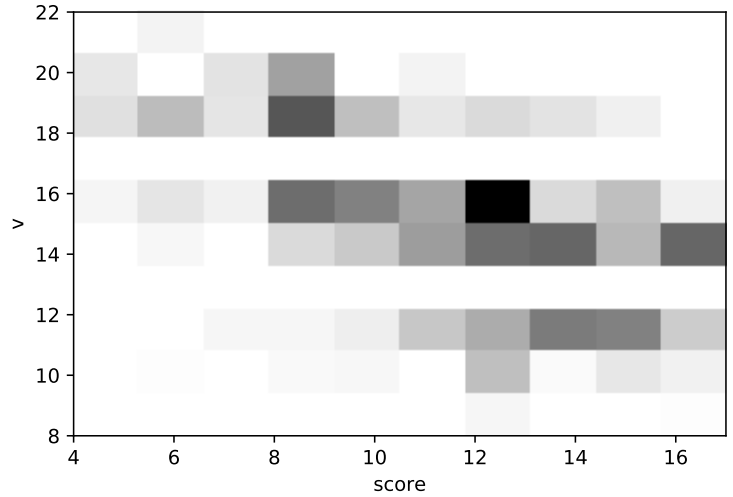


Analysis: Score, bond order, charge, valence electrons









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 [CH2-]#[CH], [OH-]=[NH], [NH2--]#[P].

Desired space = top scoring ligands

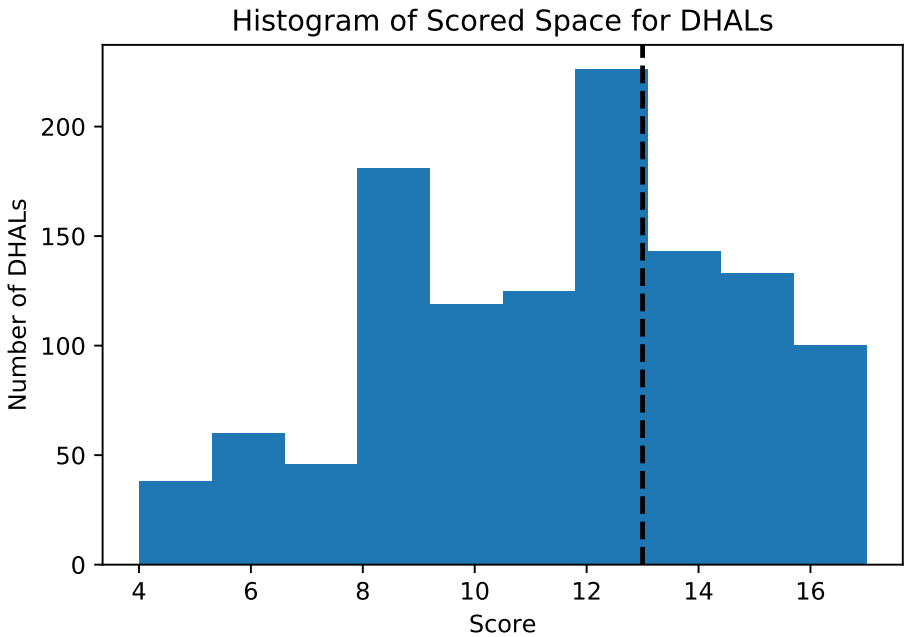
We include all ligands with a score at least as high as the lowest scoring spectrochemical series ligand ($s = 13$)

Scored Space:
50, 1171, 1635

↓

Desired Space:
29, 374, 148

} score $\geq s$



This gives us 553 ligands in total.

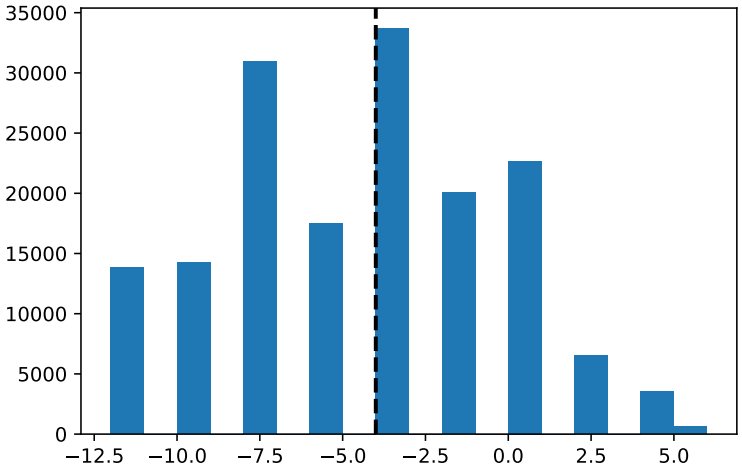
example molecule for each symmetry class for visualization that the next slide is more graspable.

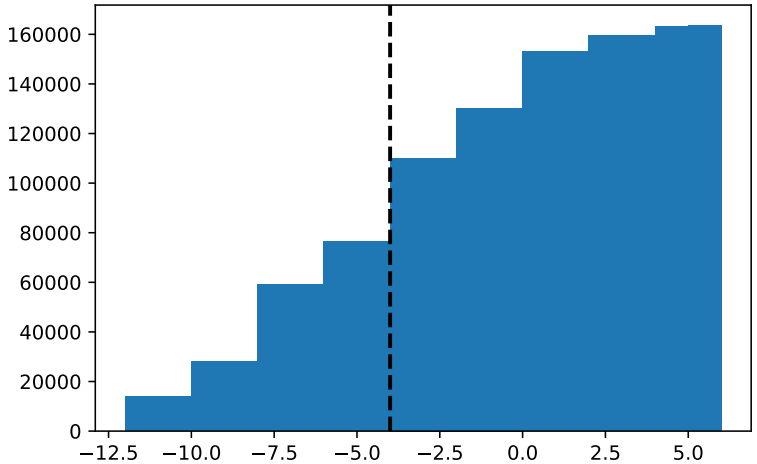
Subsets of octahedral space

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	$> 1.8 \cdot 10^{14}$

Properties of the sets

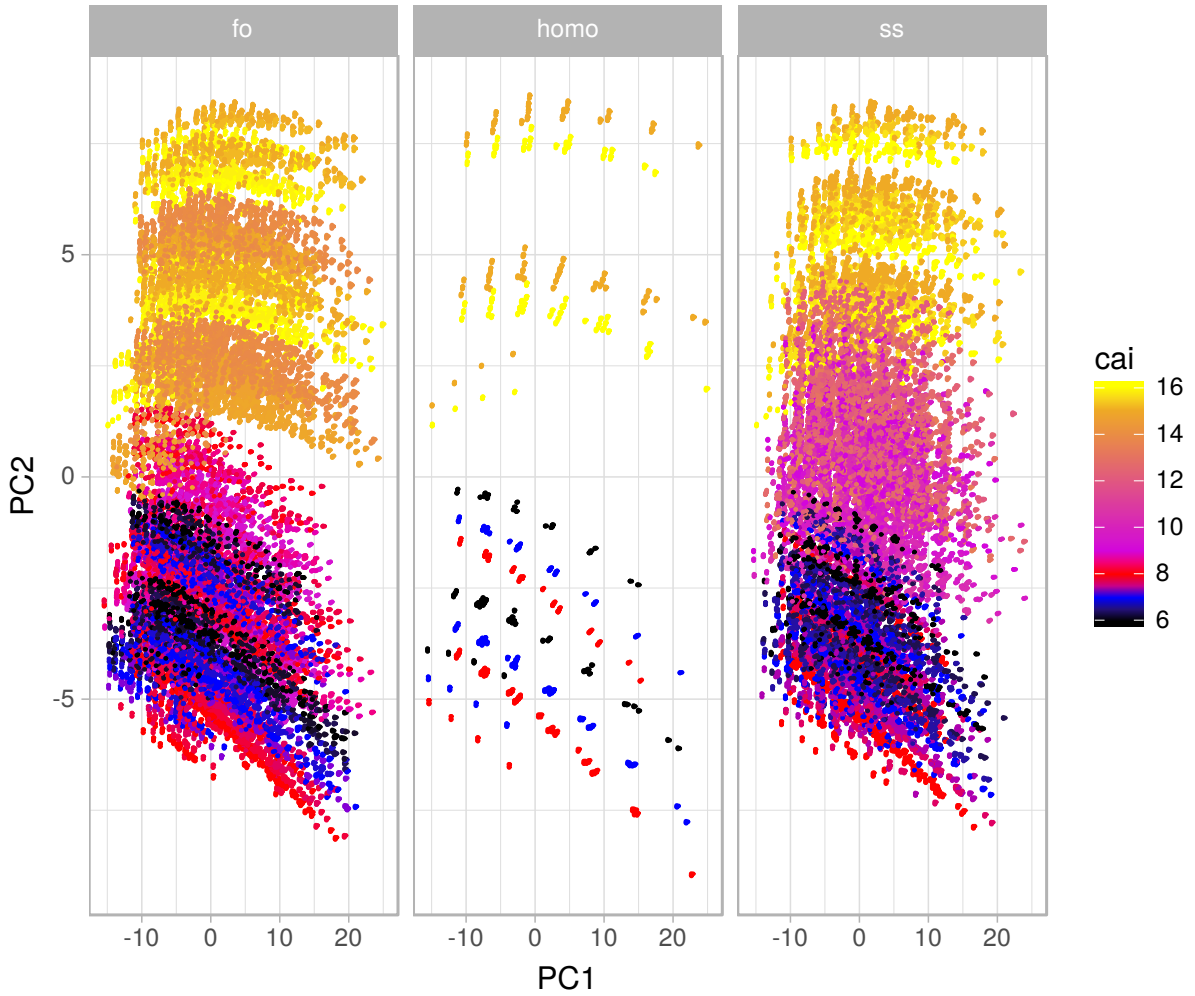
- Reduce space to facilitate sampling from non-homoleptics
- Example: strongly symmetric, monodentate ligand fields (163,620)
- Exclude all with charge smaller than -4, which results in 87,150 ligand fields (53 %).





Principal Component Analysis

The homoleptics (ho) span the strong symmetry (ss) and ”5+1” (fo) set.



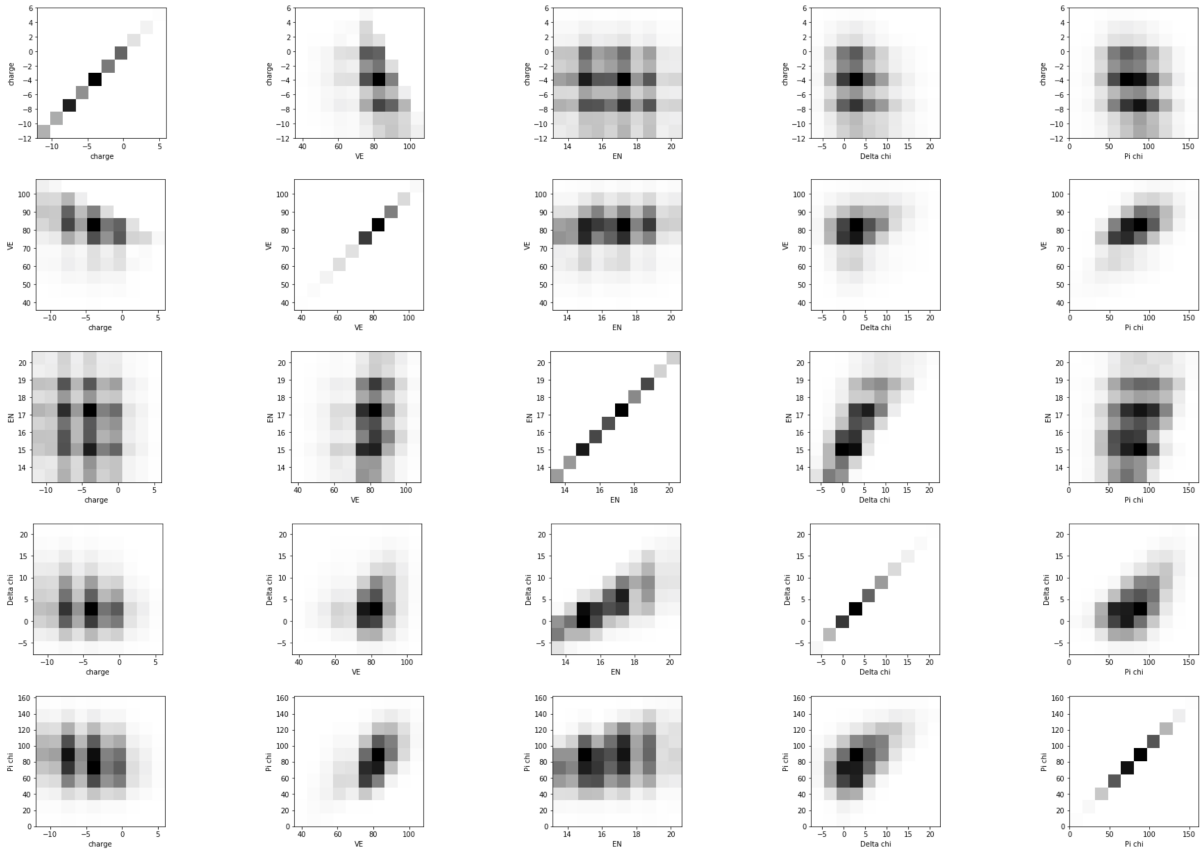
Footprint

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

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

Correlation analysis for strongly symmetric monodentates

best case: uniform gray. correl are EN and stuff obv. we dont want very peaky comps.

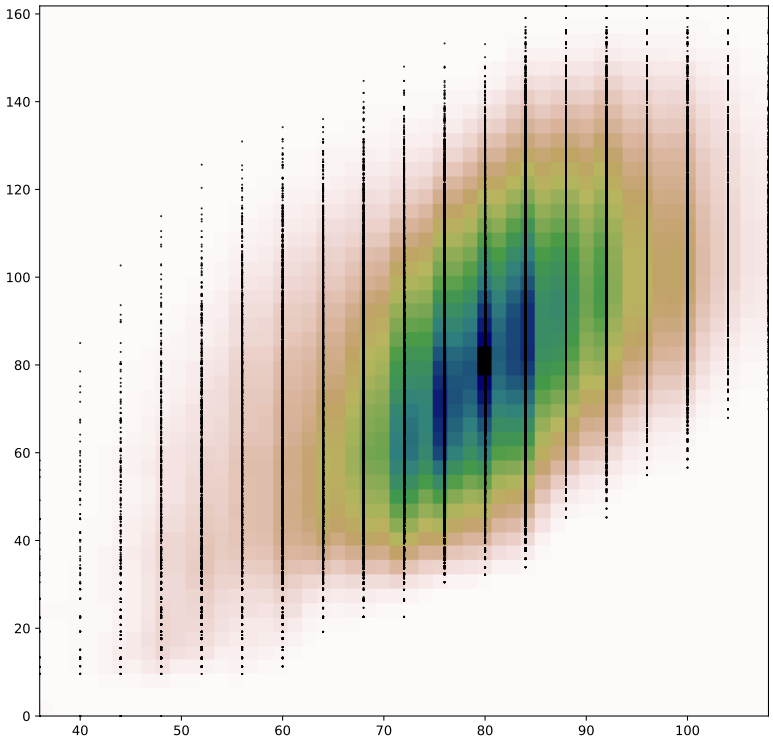


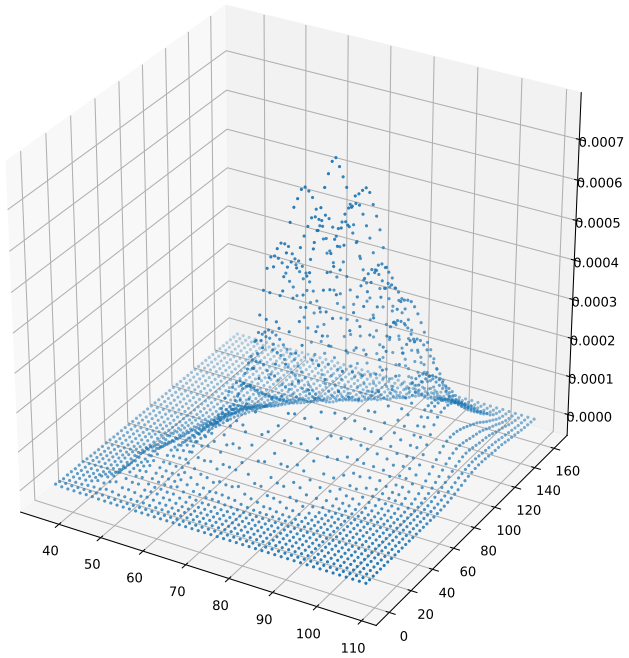
Entropy and KDE

We then calculate the entropy, H_{KDE}, of the Kernel Density Estimated distrbution. We want it to be uniform not to oversample.
...VIZ of what high and low entropy means...

Example of KDE slice

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



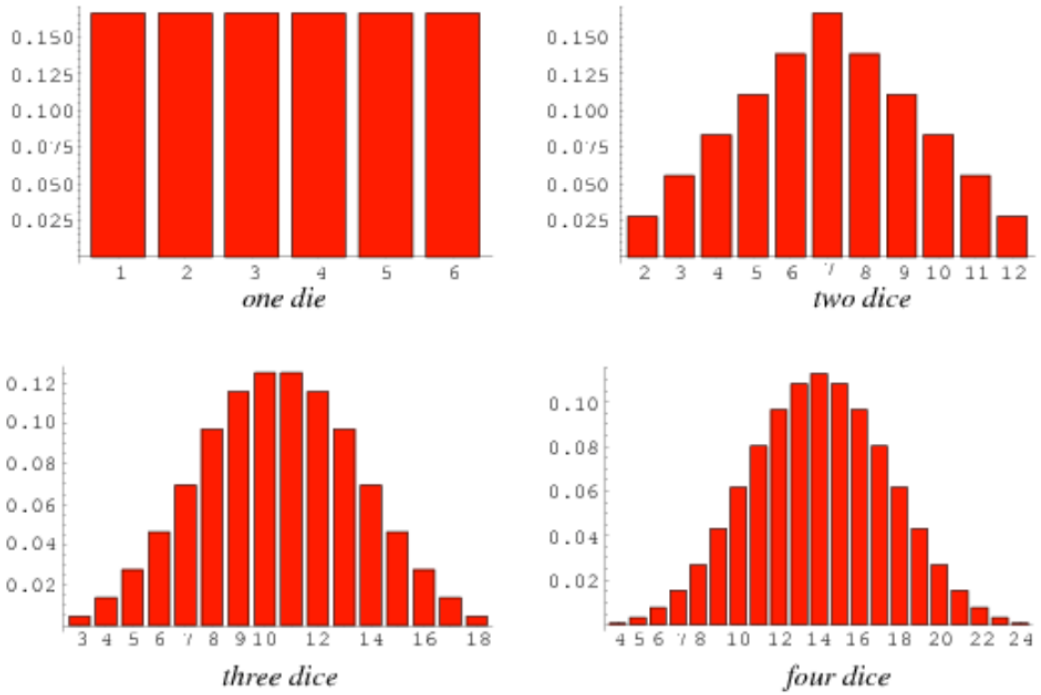


Monodentate Footprints

Table : Entropic footprint

Set	$H_{\mathrm{KDE}}^{\mathrm{monodent}}$	$H_{\mathrm{KDE}}^{\mathrm{bident}}$
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

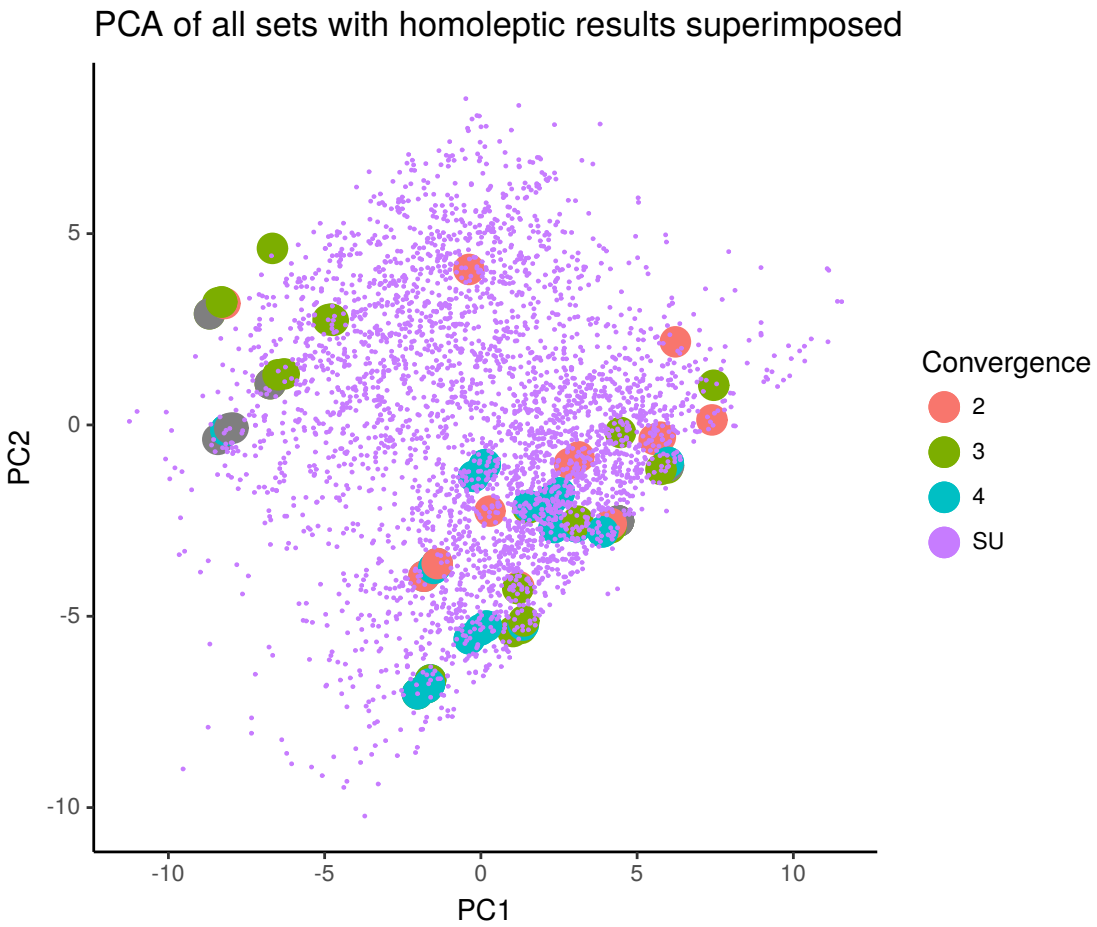


- The more dice, the more the distribution peaks which results in low entropy.
- Sample from low entropy uniformly and get similar molecules.

DFT calculations

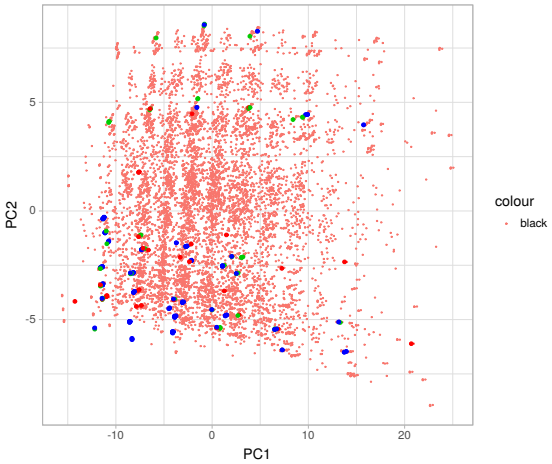
- lacvps_ecp (631G*)
- core potential: LANL2DZ
- functional: b3lyp

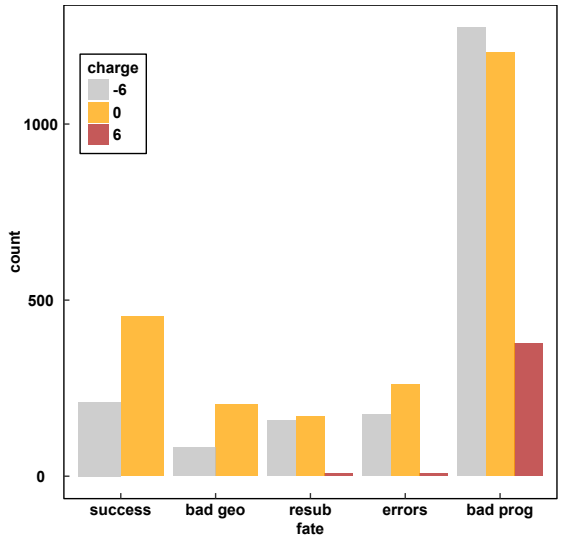
DFT calculations analysis

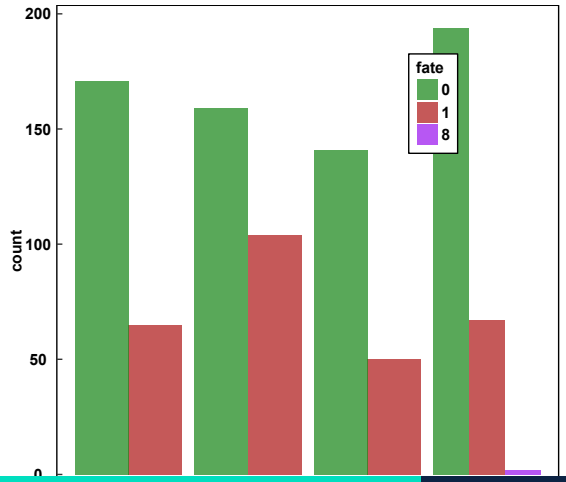


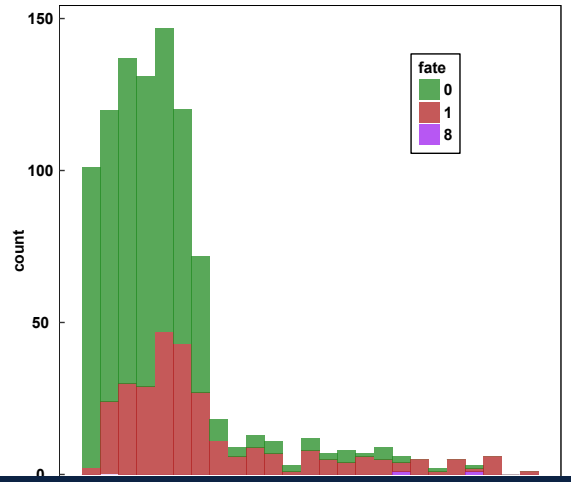
DFT calculations analysis

pull thenm apart









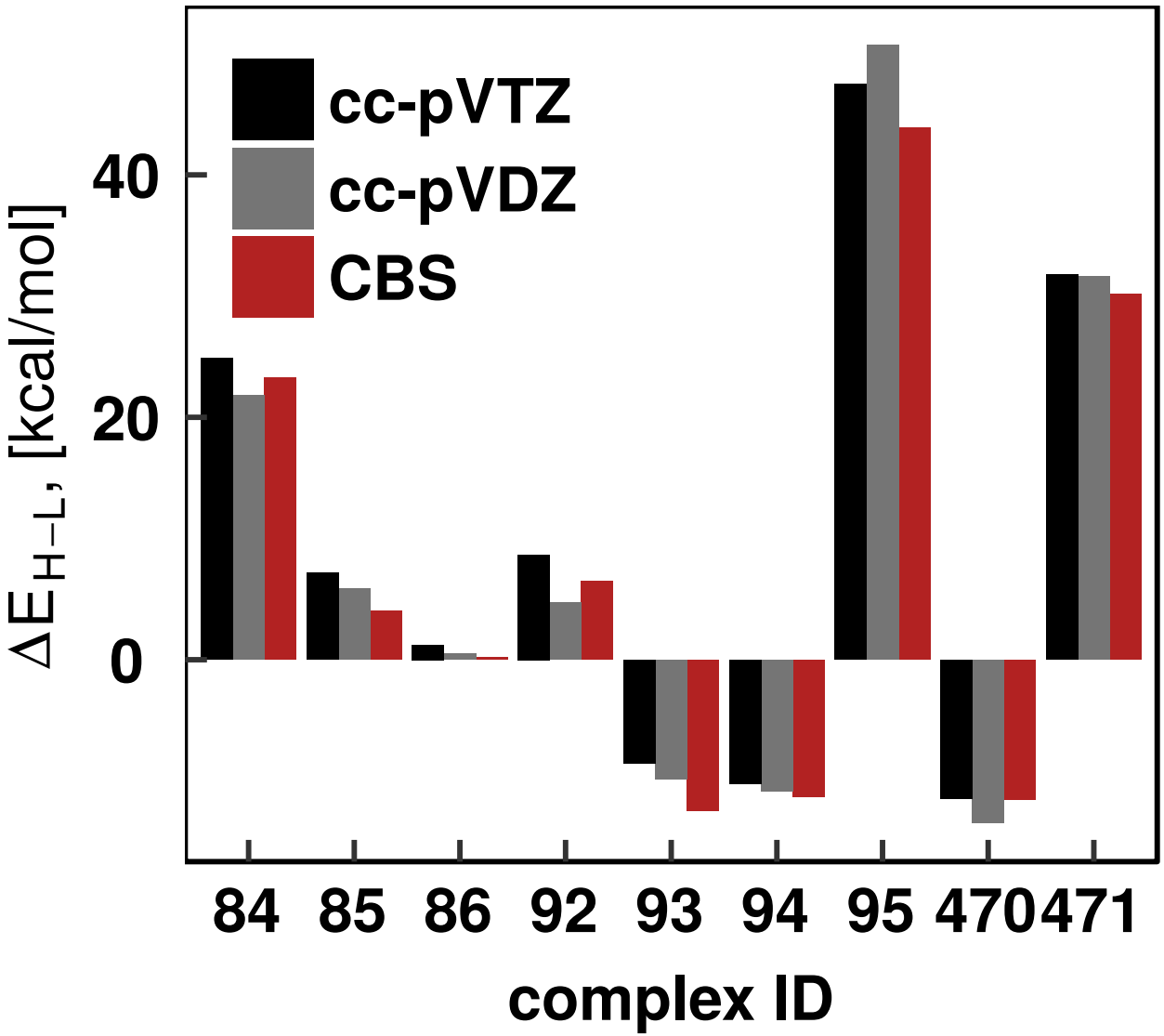
DFT calculations analysis

krr results vs DFT results in SS energy (linear model)

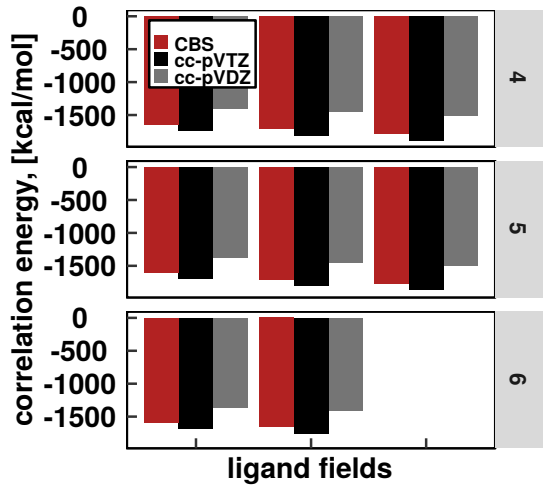
DLPNO calculations

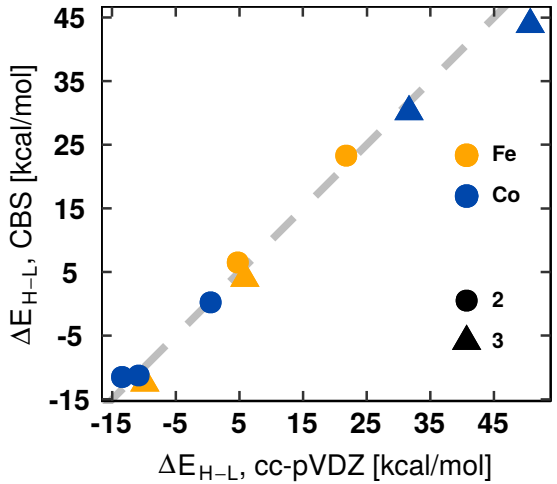
Linear model DLPNO vs DFT

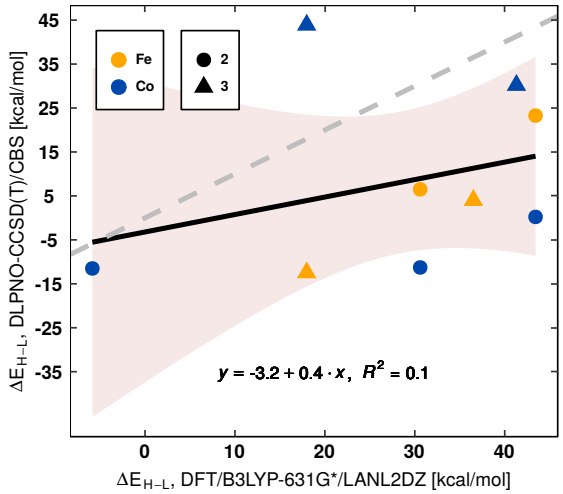
Split All

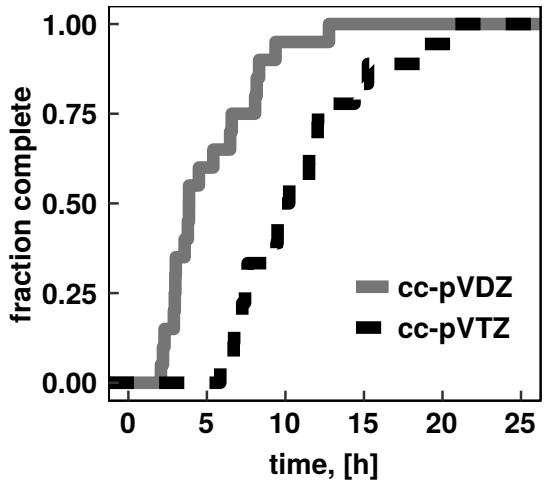

cc-pVTZ
cc-pVDZ
CBS
40
20
0
ΔE H−L, [kcal/mol]
84
85
86
92
93
94
95
470
471
complex ID

Pull apart DLPNO results









thank you slide

thanks slide pics