GDB-9 & Small Molecule Universes

Stefan O. Gugler

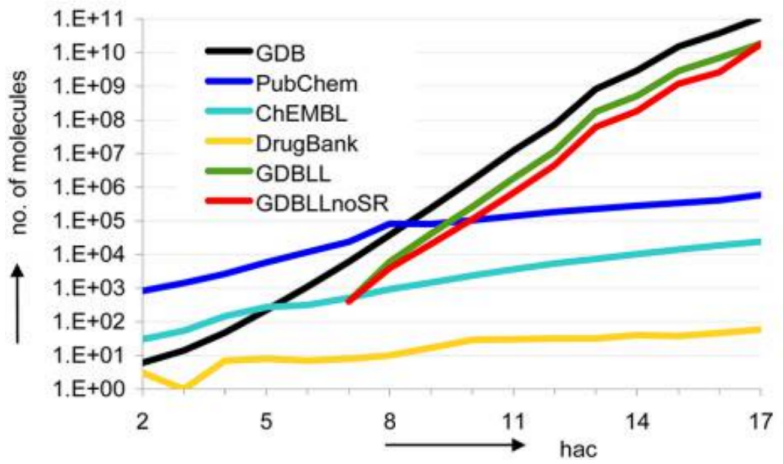
Massachusetts Institute of Technology
Department of Chemical Engineering

May 31, 2018

1. GDB-9 dataset
2. Small Ligand Universe
   - General
   - Reduction Rules
   - SLU Analysis
3. Assembly of complexes
   - General

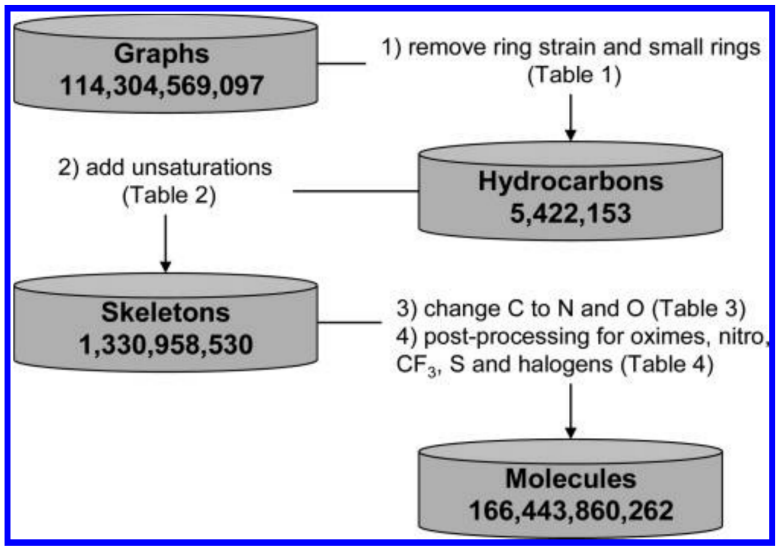
GDB-17

- Up to 17 atoms with CNOS and halogens
- Results in 166 billion molecules
- Contains more nonaromatic heterocycles, quarternary centers, stereoisomers than PubChem



GDB-17

- H4: C15 & C16 graphs are allowed max one 3- or 4-membered ring.
- S1: No allenes (C=C=C).
- F2: Maximum one N or O in small rings.
- P4: No aliphatic thiols or thioethers.



Examples of up to 4 atoms

1: C, N, O
2: C#C, C#N, C=O, CC, CO
3: CC#C, CC#N, CC=O, NC=O, CCC, CCO, COC
4: C1CC1, C1CO1, CC(=O)C, CC(=O)N, NC(=O)N, CC(C)C, CC(C)O, C(#C)C#C, C(#C)C#N, N#CC#N, O=CC#C, O=CC#N, O=CC=O, CC#CC, CCC#C, CCC#N, NCC#N, OCC#C, OCC#N, CCC=O, CNC=O, COC=O, OCC=O, CCCC, CCCO, CCOC, OCCO

GDB-9 and QM-9

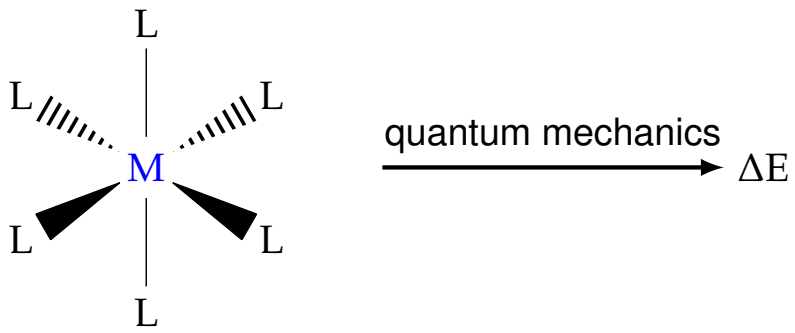
- 134,000 molecules up to 9 CHNOF atoms from GDB-17
- Minimum energies, harmonic frequencies, dipole moments, polarizabilities, enthalpies, free energies of atomization
- B3LYP/6-31G(2df,p)

Small Ligand Universe: Motivation

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

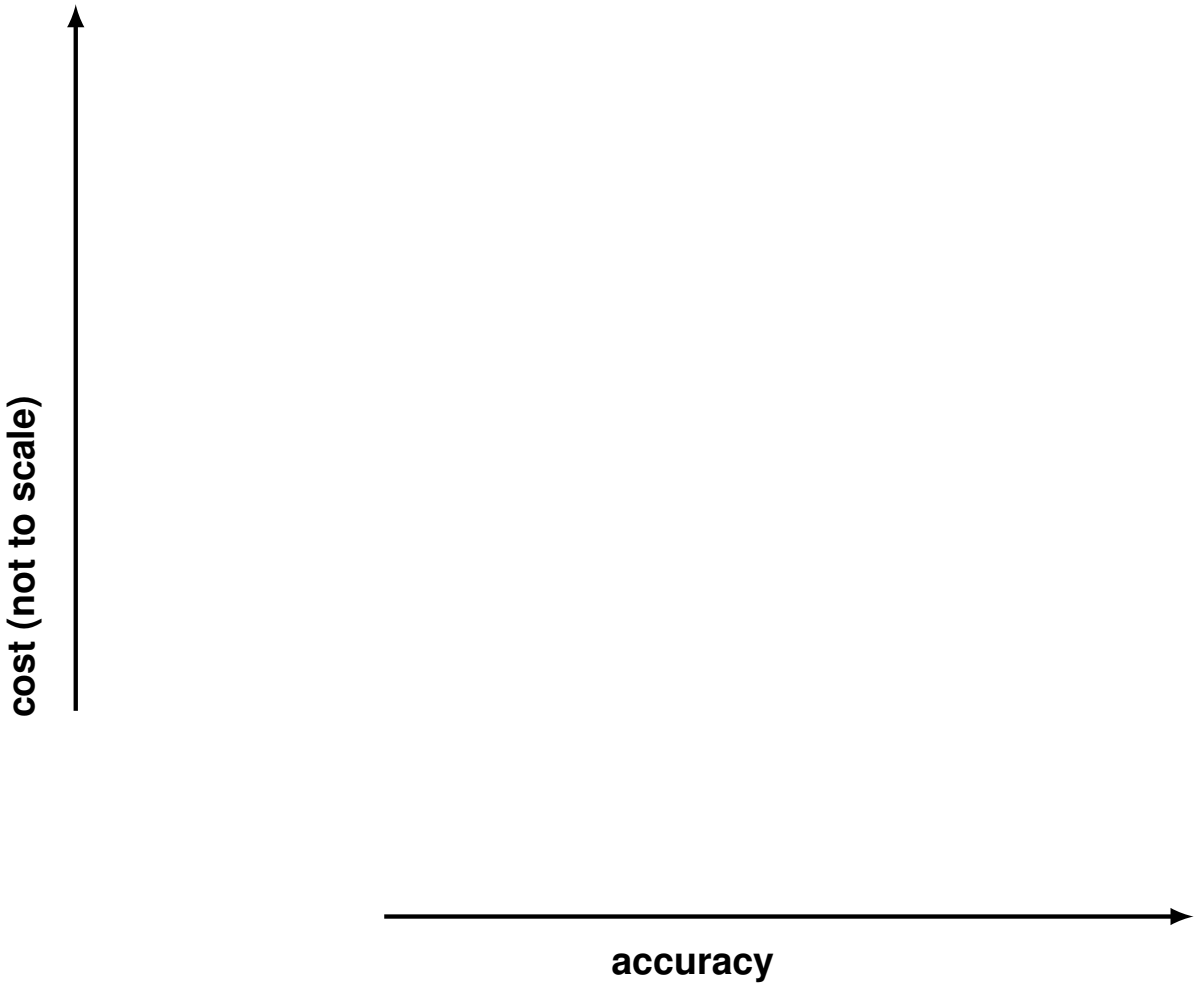
Scaling of different methods

Where do we find DLPNO?



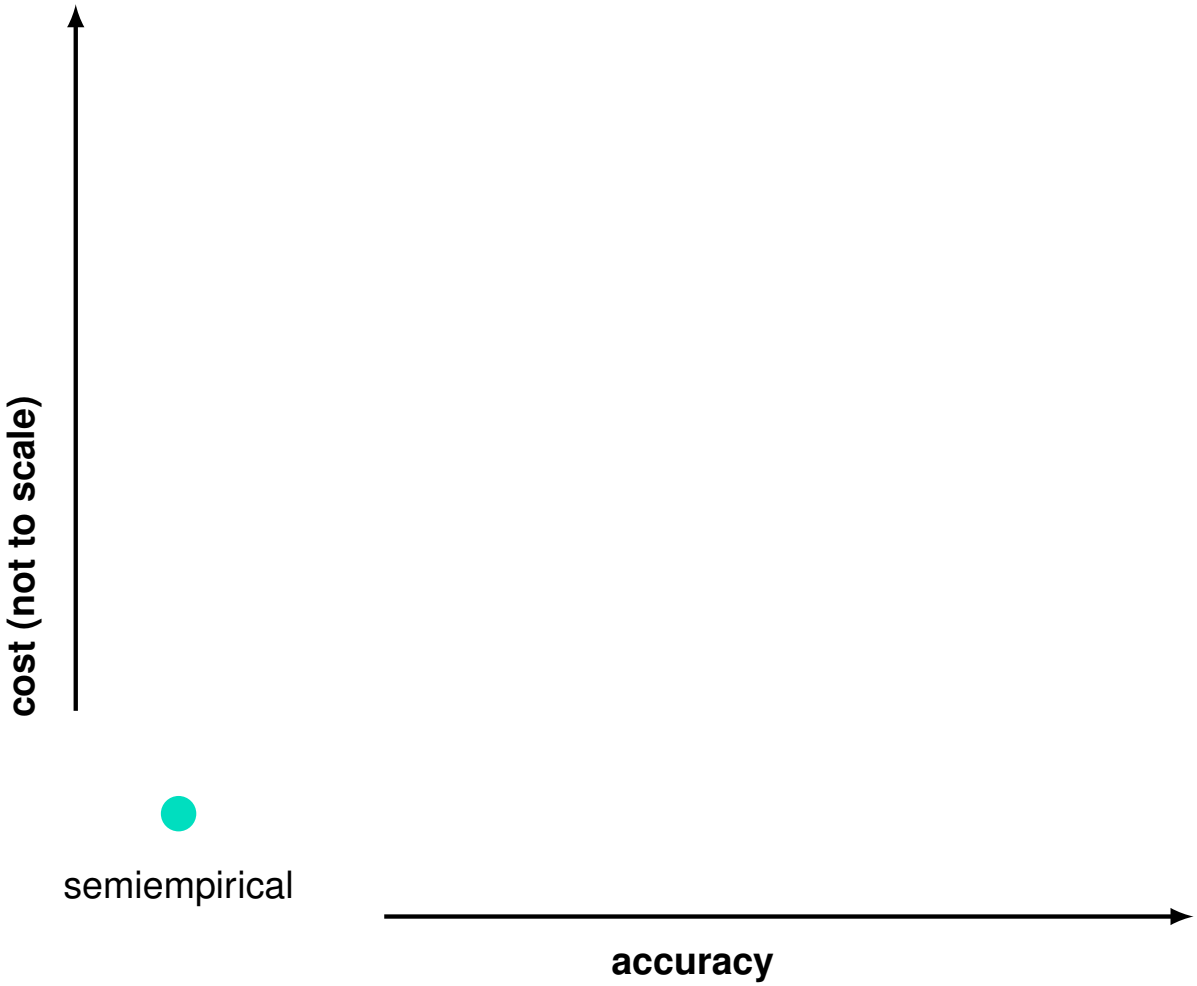
Scaling of different methods

Where do we find DLPNO?



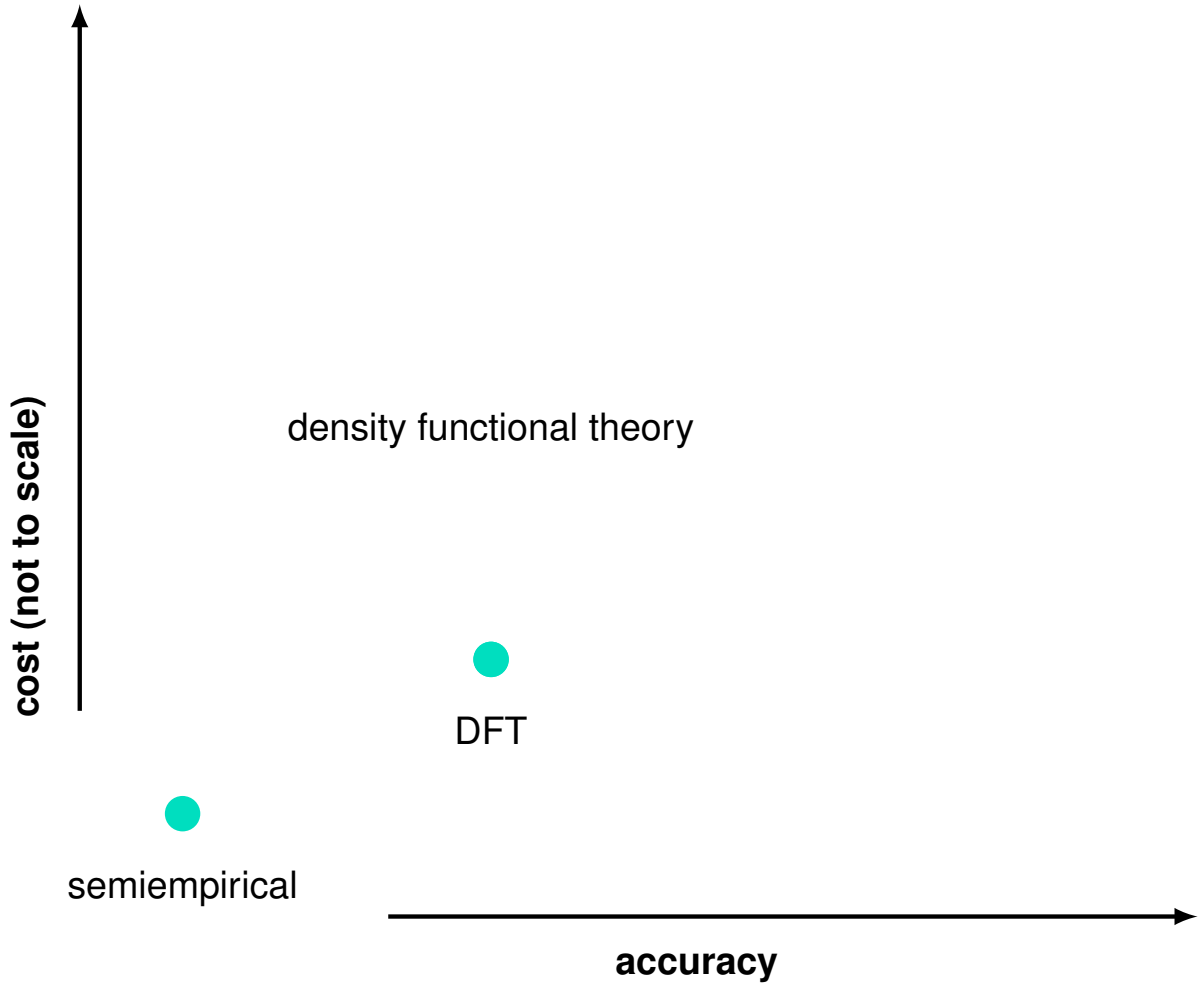
Scaling of different methods

Where do we find DLPNO?



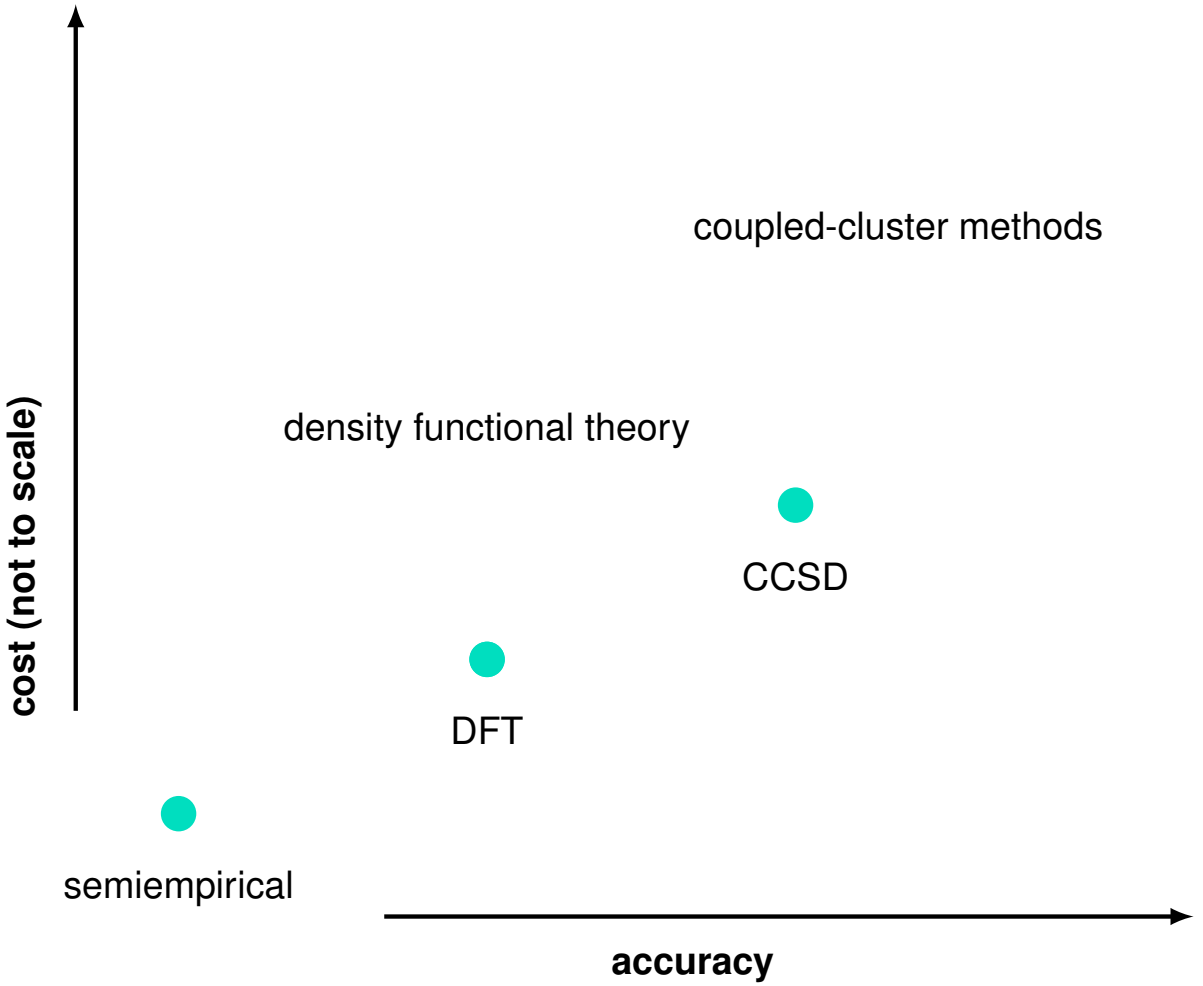
Scaling of different methods

Where do we find DLPNO?



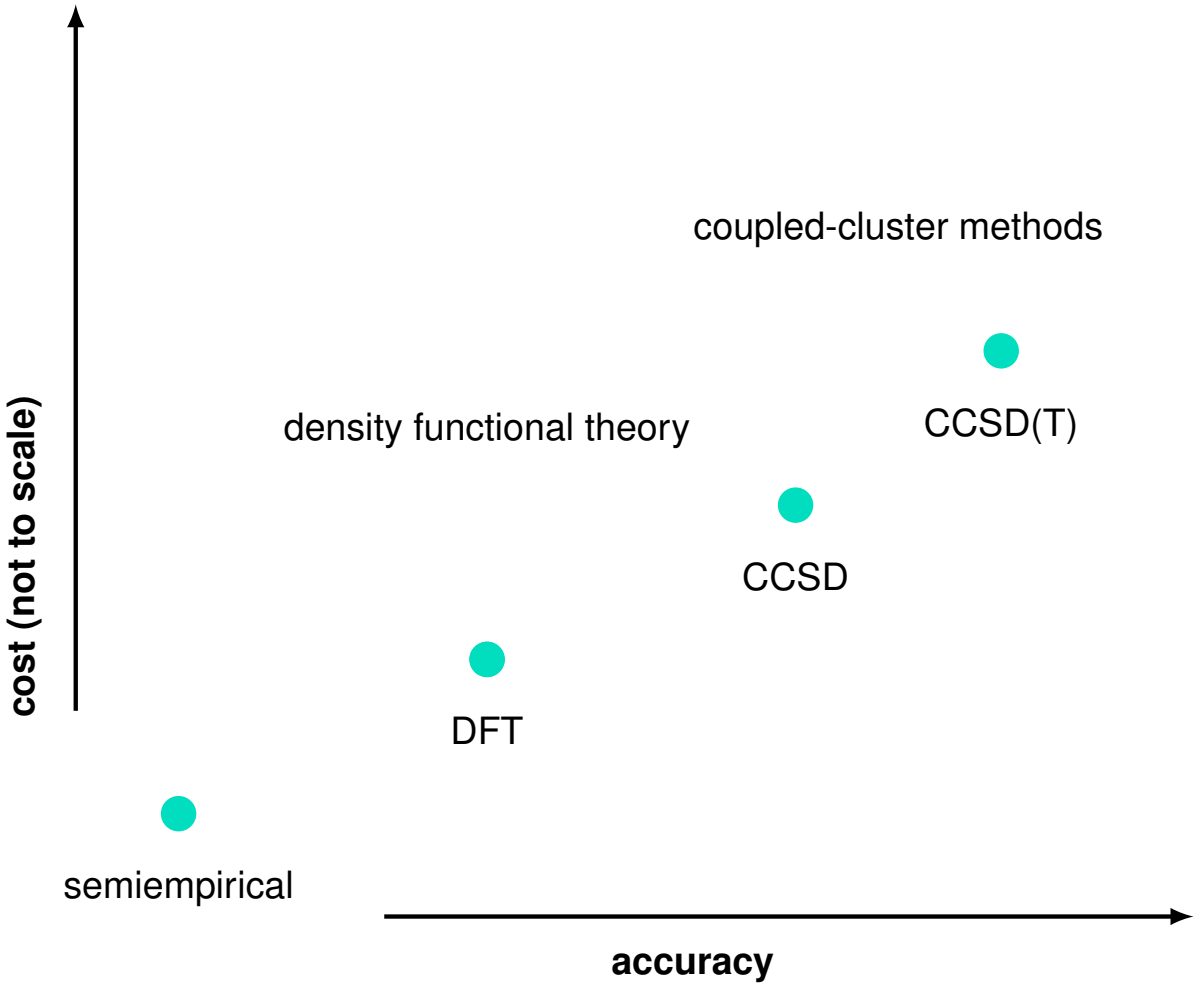
Scaling of different methods

Where do we find DLPNO?



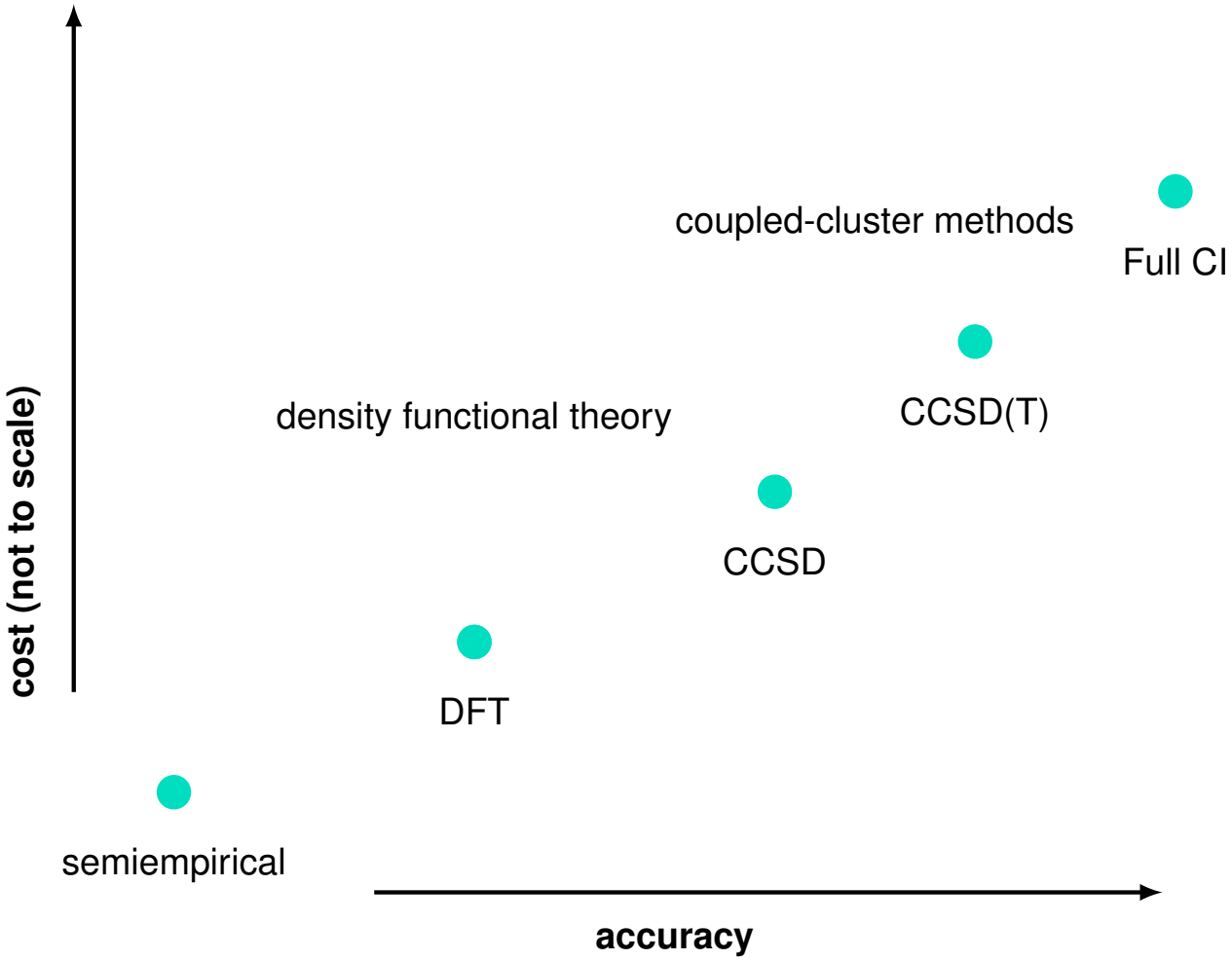
Scaling of different methods

Where do we find DLPNO?



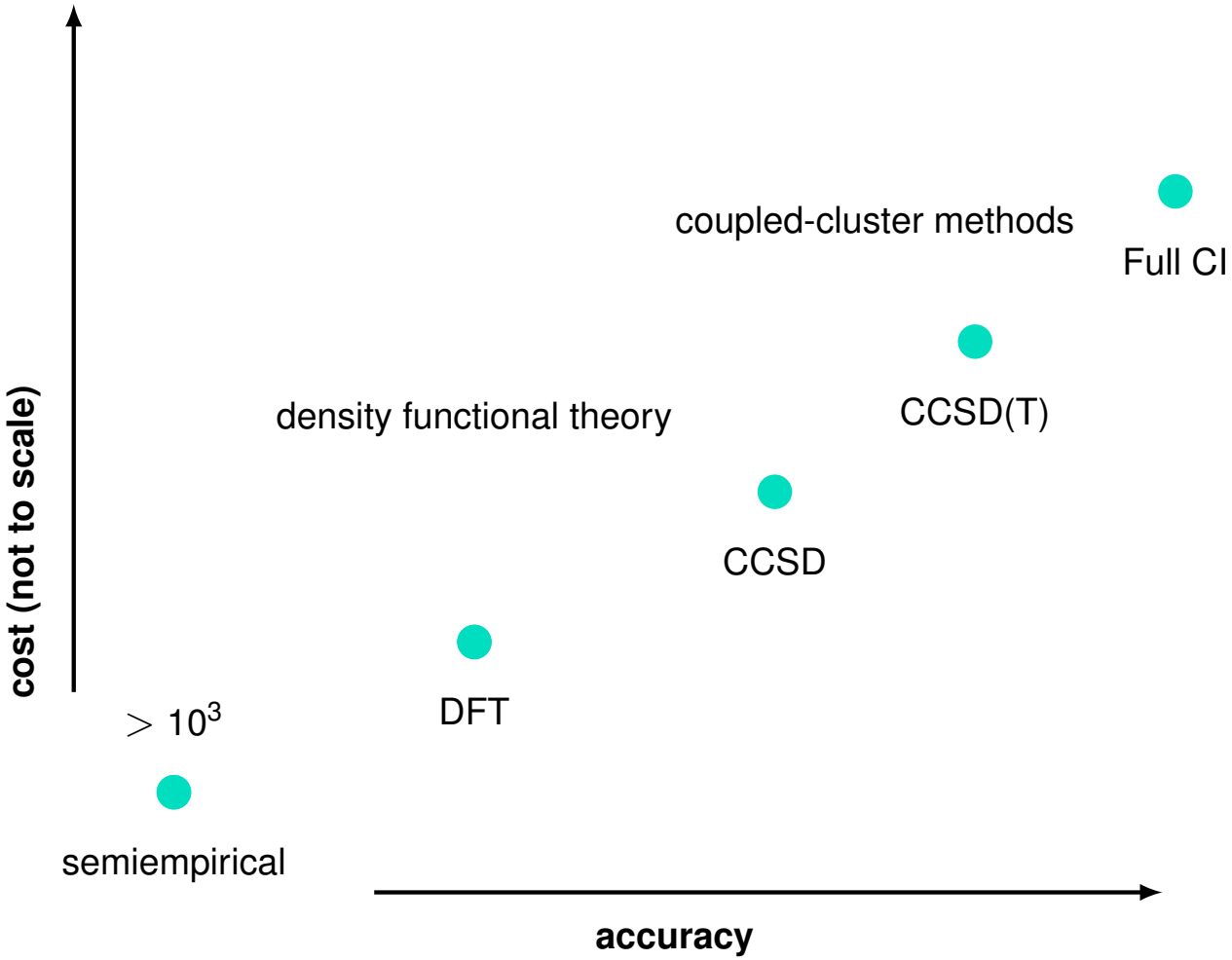
Scaling of different methods

Where do we find DLPNO?



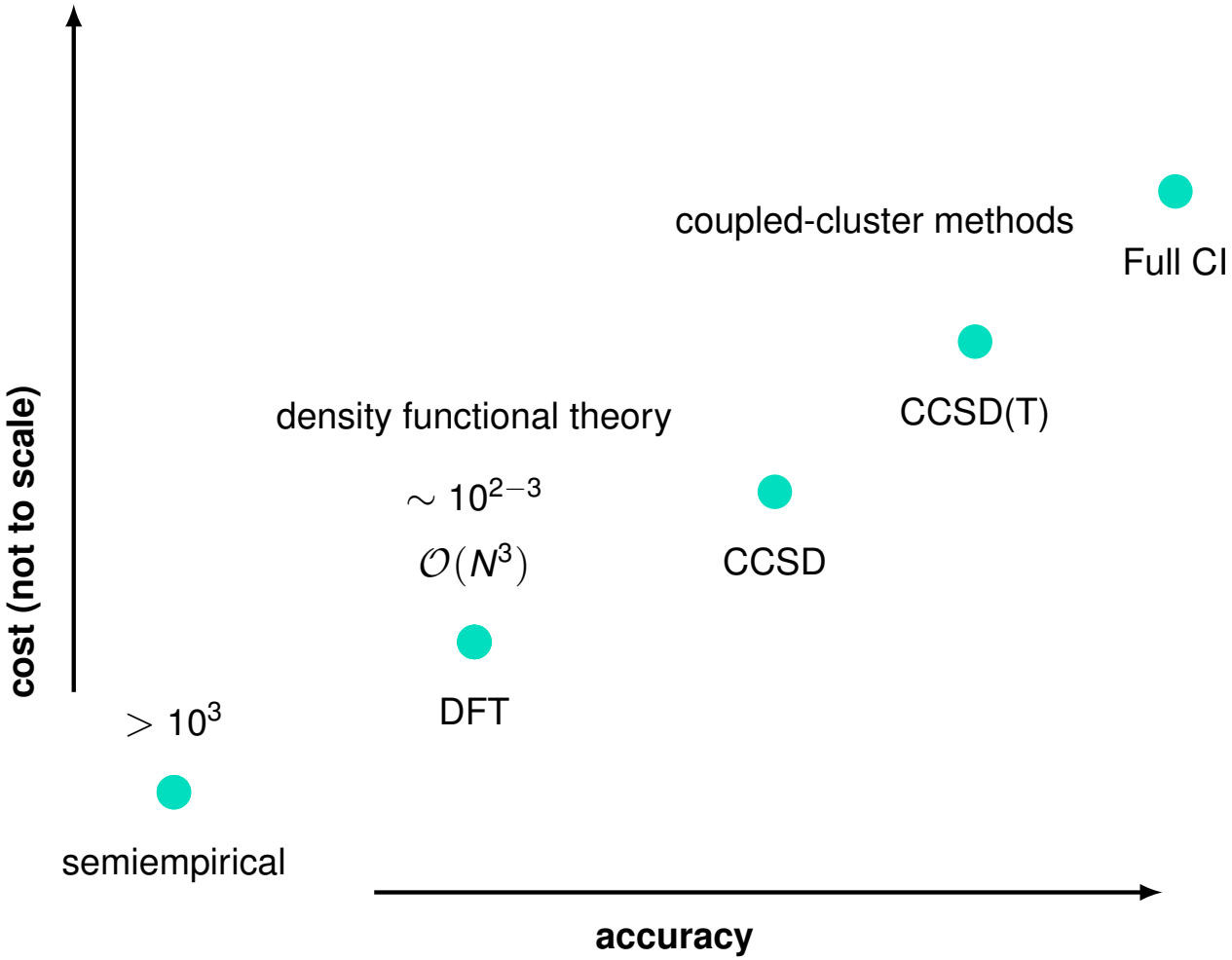
Scaling of different methods

Where do we find DLPNO?



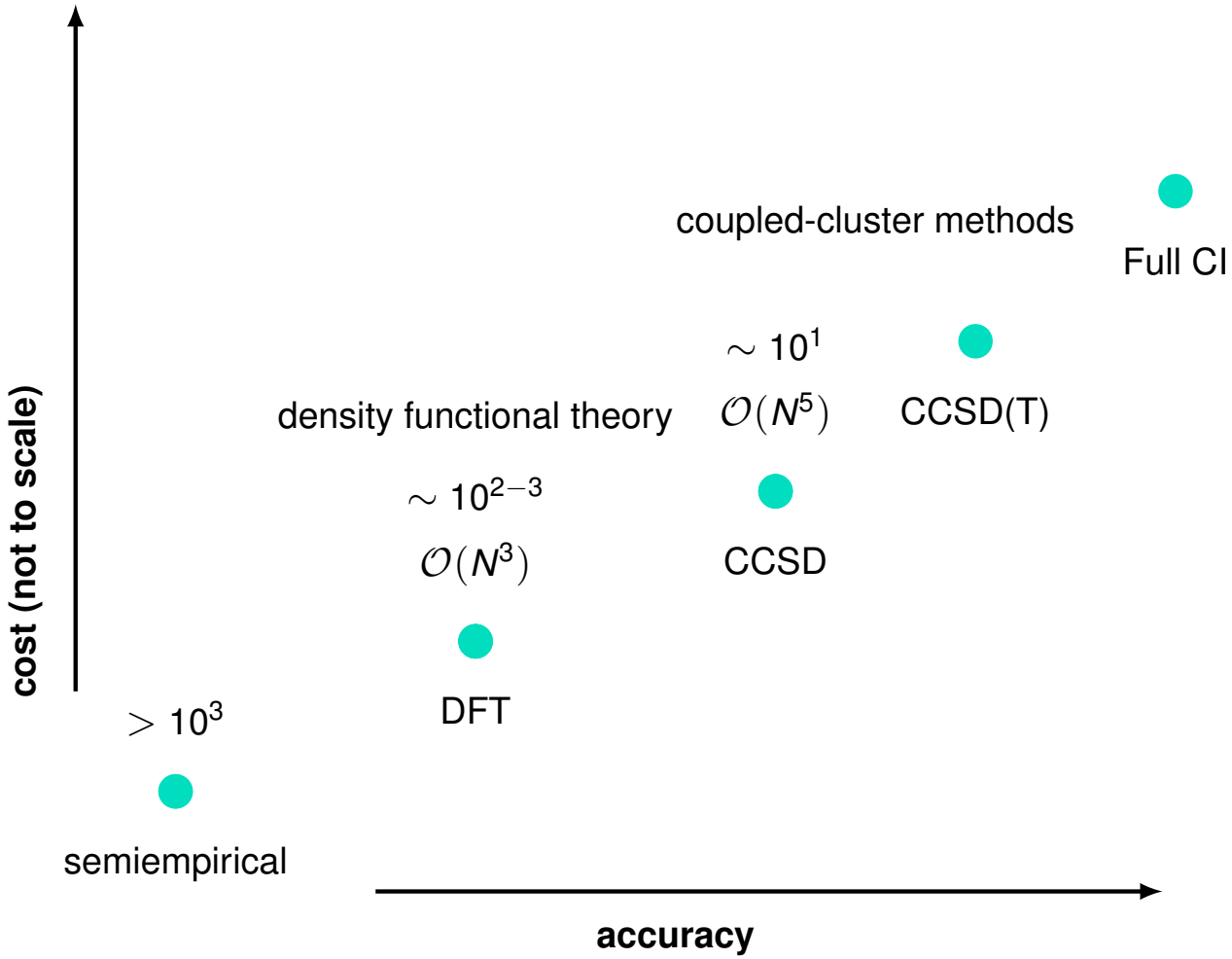
Scaling of different methods

Where do we find DLPNO?



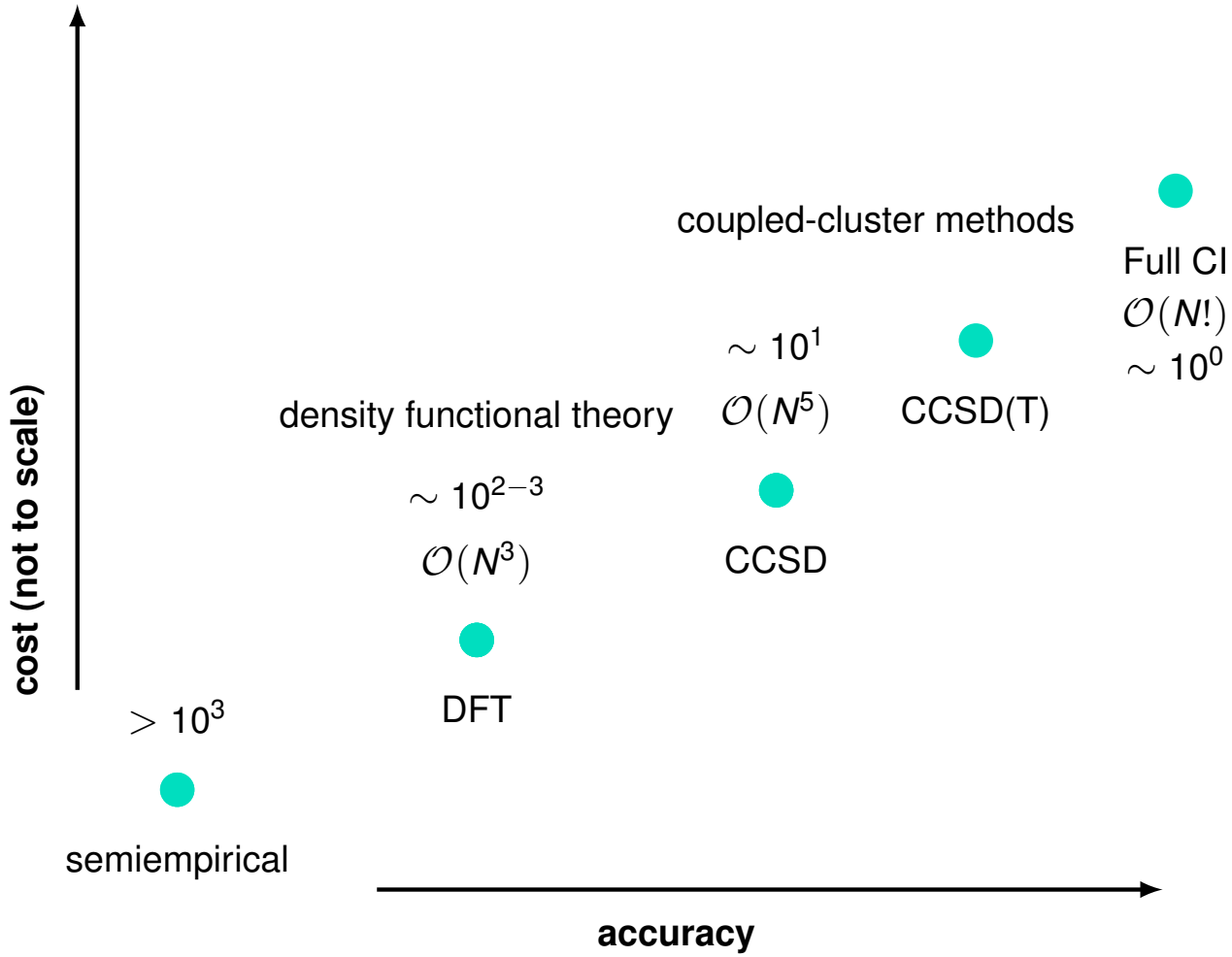
Scaling of different methods

Where do we find DLPNO?



Scaling of different methods

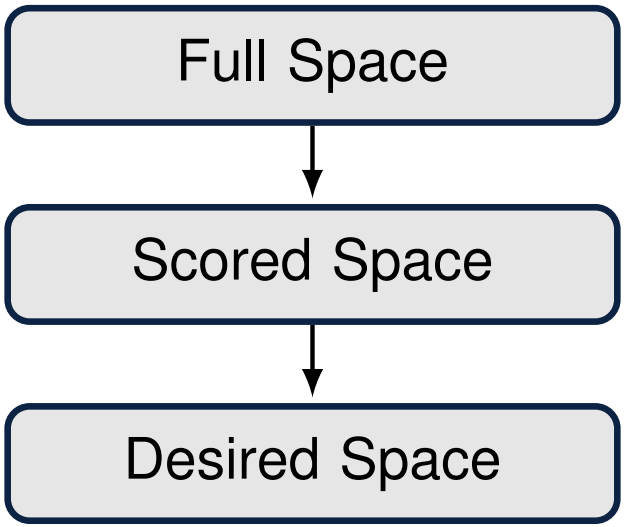
Where do we find DLPNO?



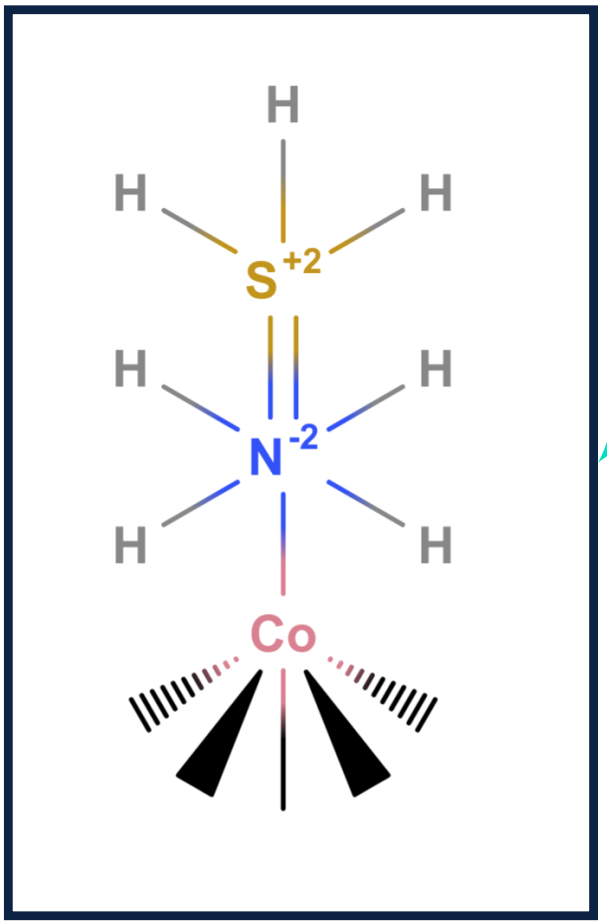
Procedure

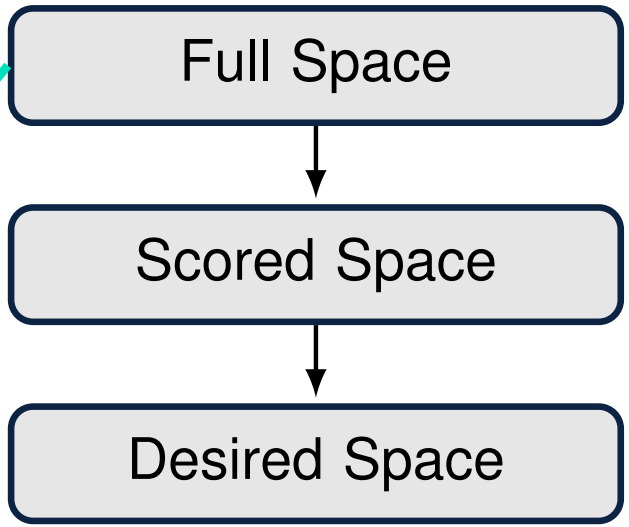
1. Generate small ligand universe (SLU)
2. Assemble complexes with different symmetries from truncated SLU
3. Analysis of symmetry classes

Reduction of the full space

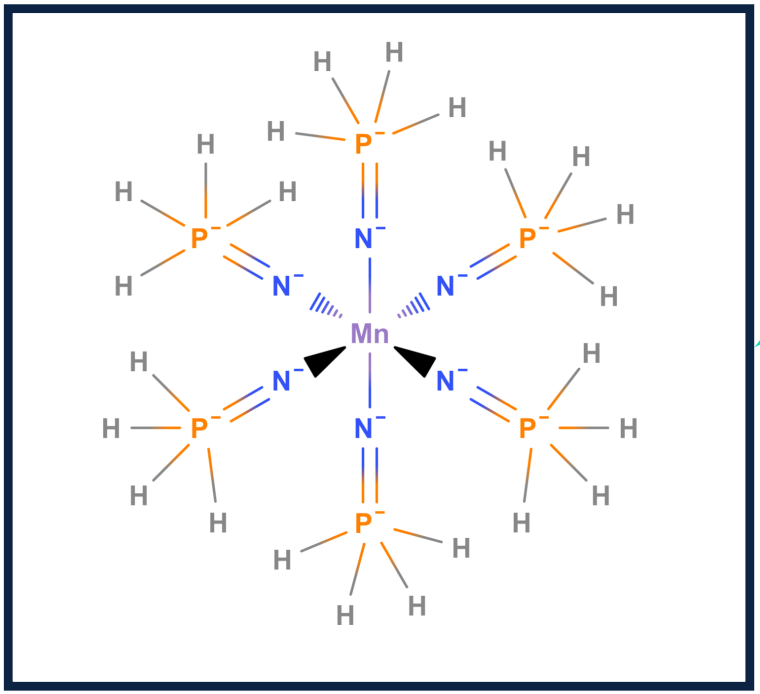


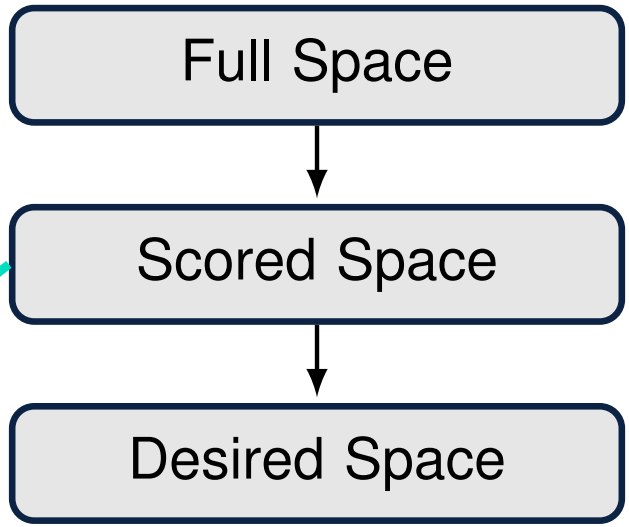
Reduction of the full space



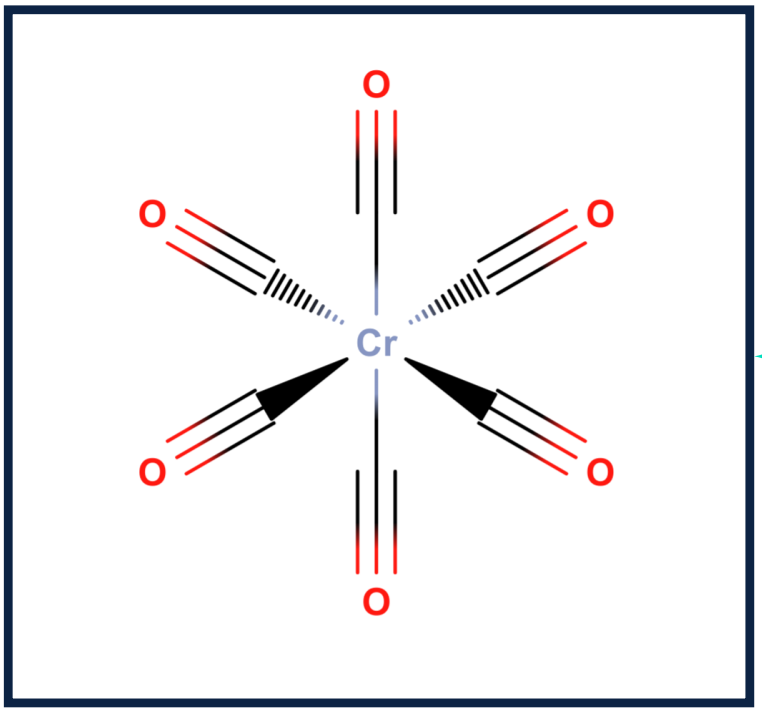


Reduction of the full space





Reduction of the full space



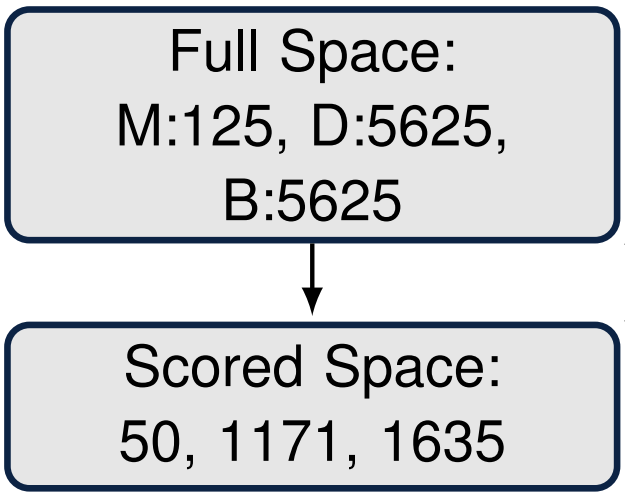
Full Space

Scored Space

Desired Space

Full Space $\rightarrow$ Scored Space

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



} charge, sterics, octet, shell, bond order

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)

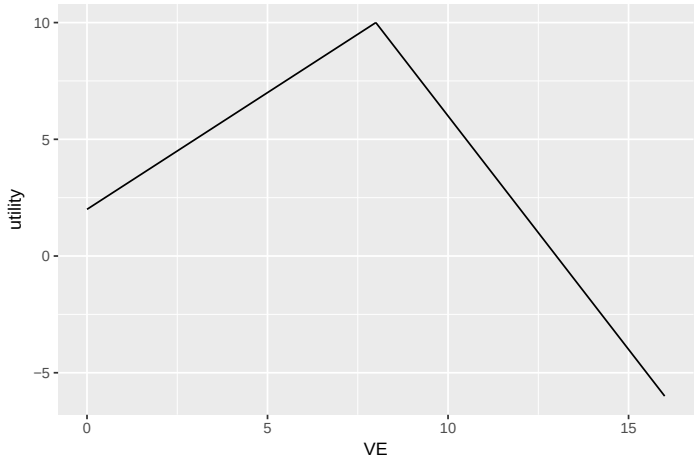
$\rightarrow$ Reduces 5625 ligands to 1171.

Octet Rule

Molecules that fulfill the octet rule are more stable.

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

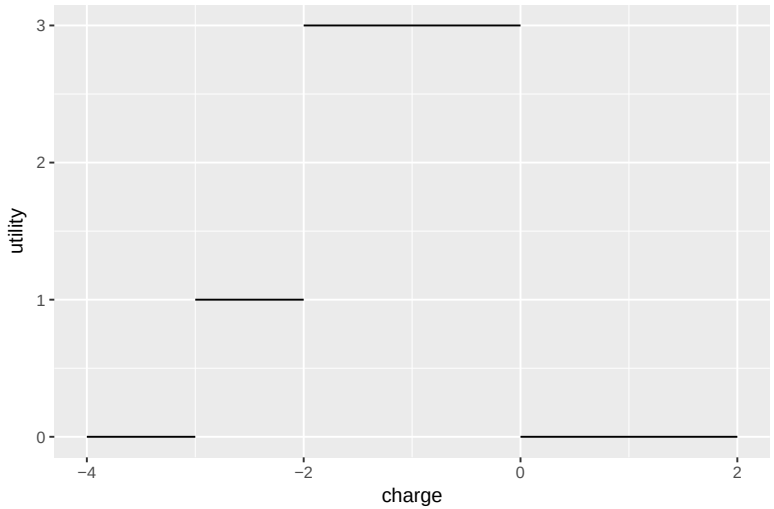
$$u_{\text{octet}} = \frac{1}{2} \sum_{i}^{2} u_{\text{octet,1}} + u_{\text{octet,2}} \tag{2}$$



Charge Constraints

We reward only mildly negatively charged compounds.

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



VSEPR Constraints

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

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

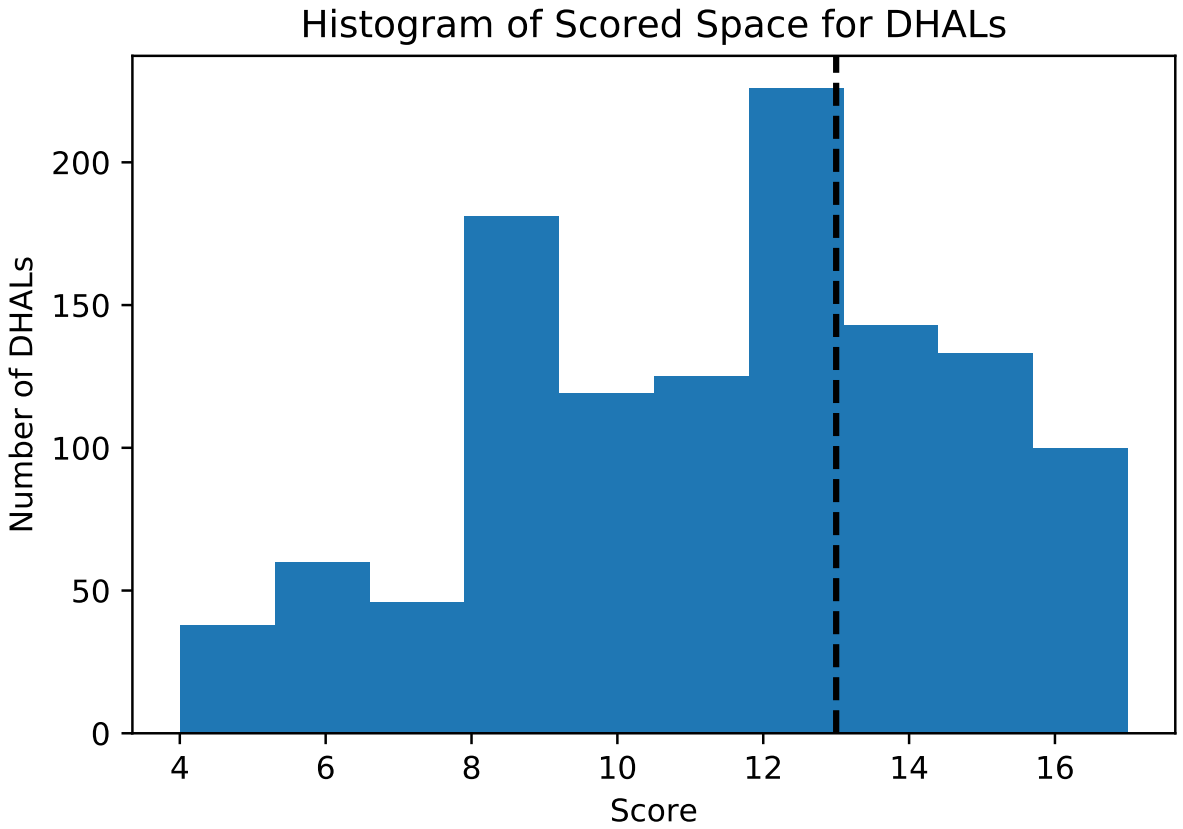
Sterics

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

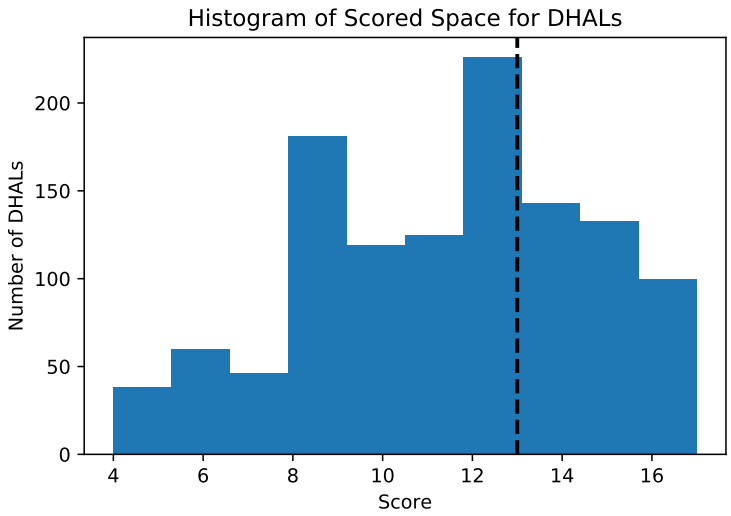
$$u_{\mathrm{CA}} = \begin{cases} 2 & \text{if } h_1 = 3 \\ 3 & \text{if } h_1 < 3 \end{cases} \tag{5}$$

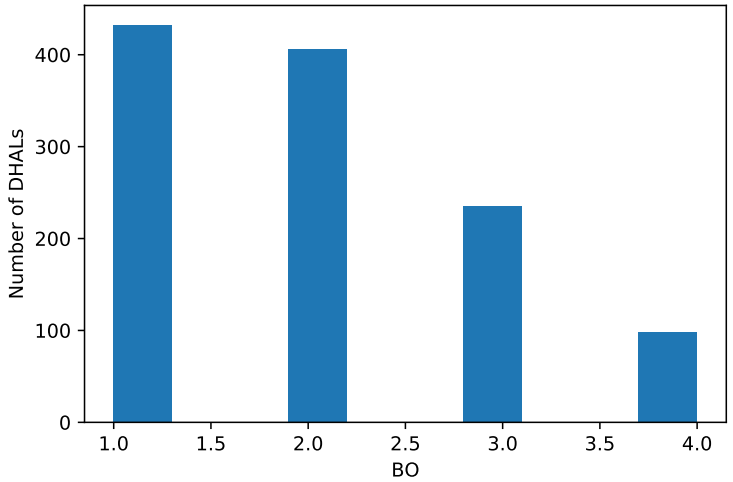
Global utility function

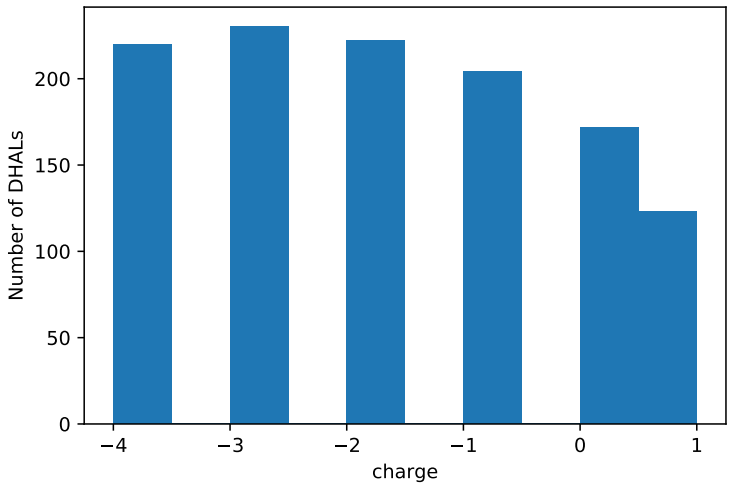
$$u_{\text{total}} = u_{\text{octet}} + u_{\text{charge}} + u_{\text{VSEPR}} + u_{\text{CA}} \tag{6}$$

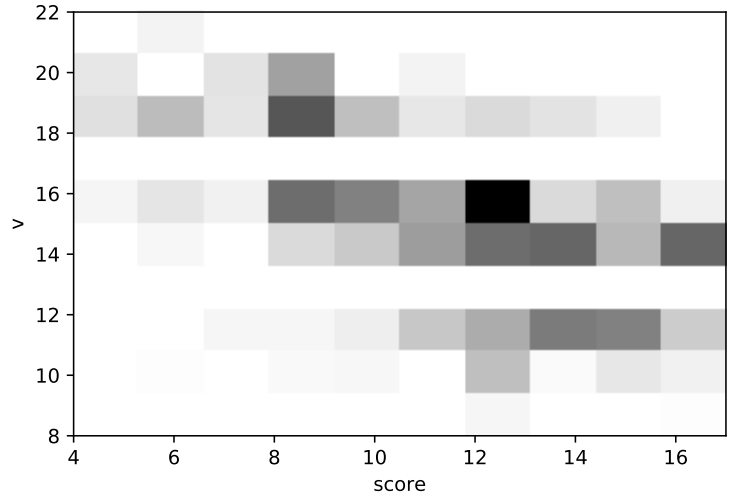


Analysis: Score, bond order, charge, valence electrons









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

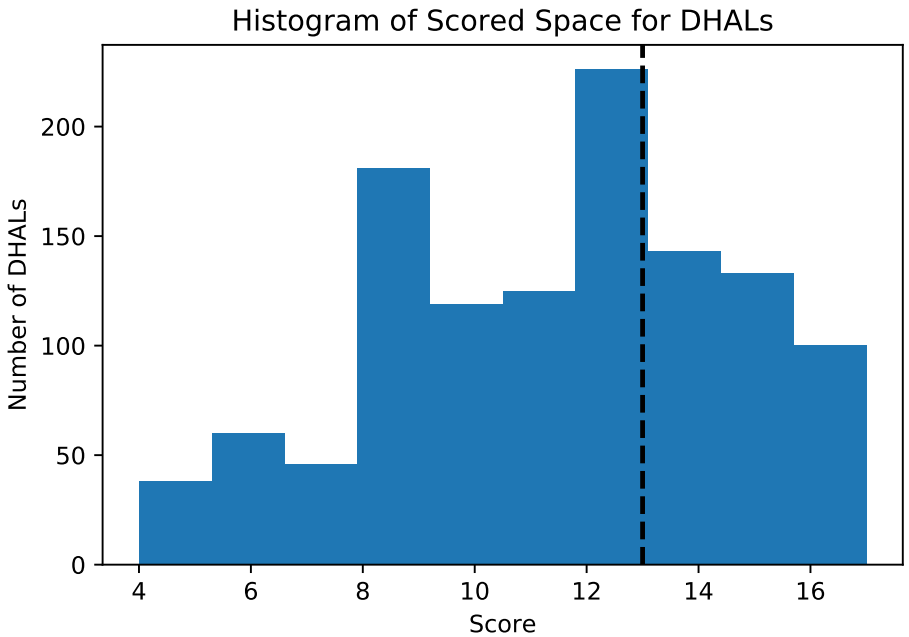
We include all ligands with a score at least as high as the lowest scoring spectrochemical series ligand ($s = 13$)

Scored Space:
50, 1171, 1635

↓

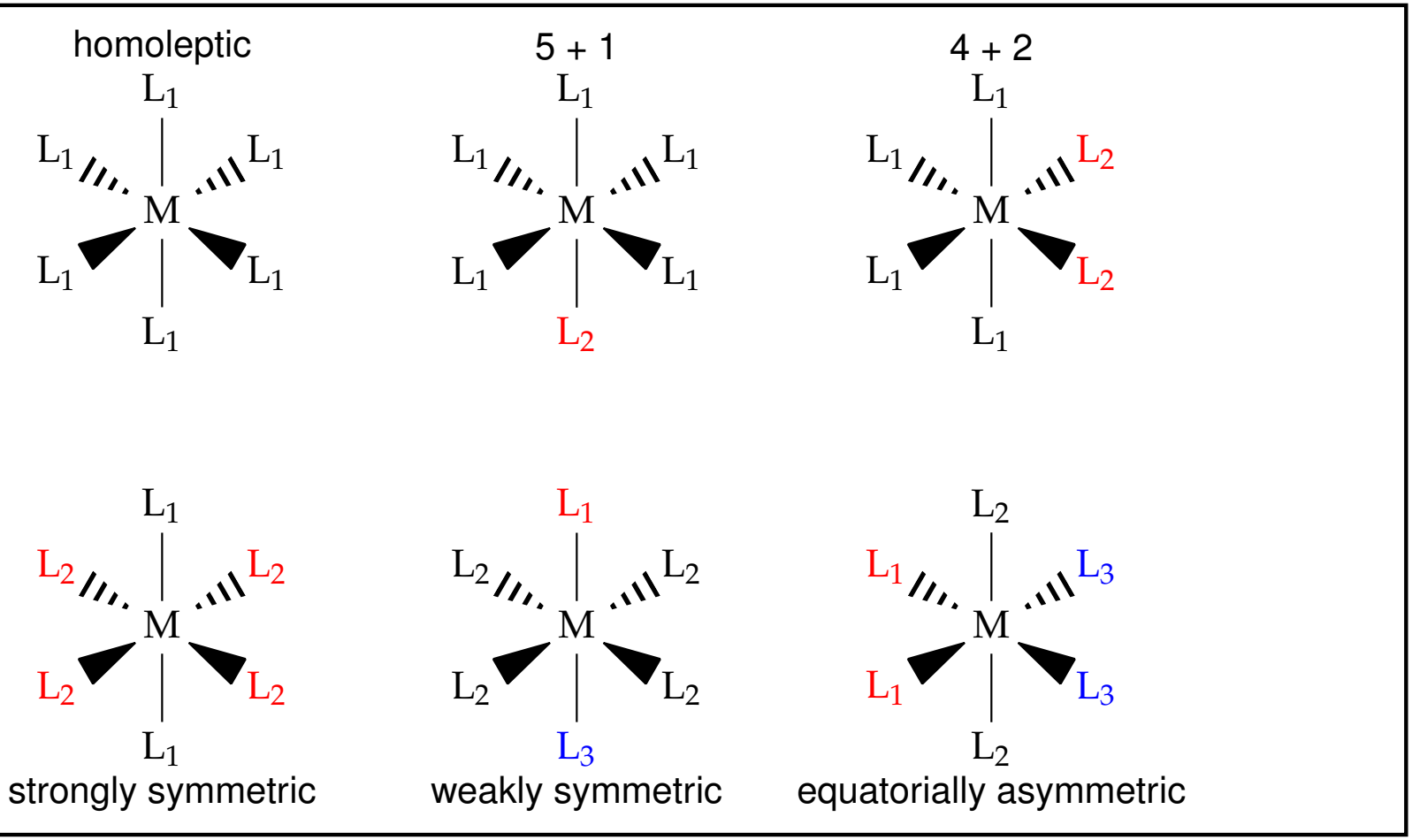
Desired Space:
29, 374, 148

} score $\geq s$



This gives us 553 ligands in total.

Overview of considered symmetry classes

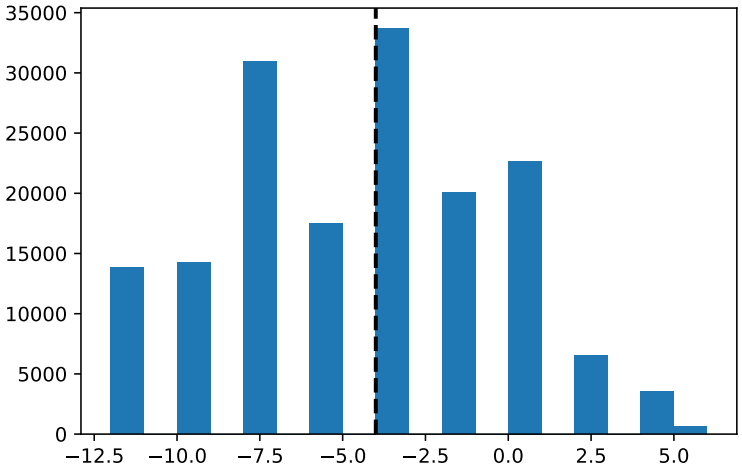


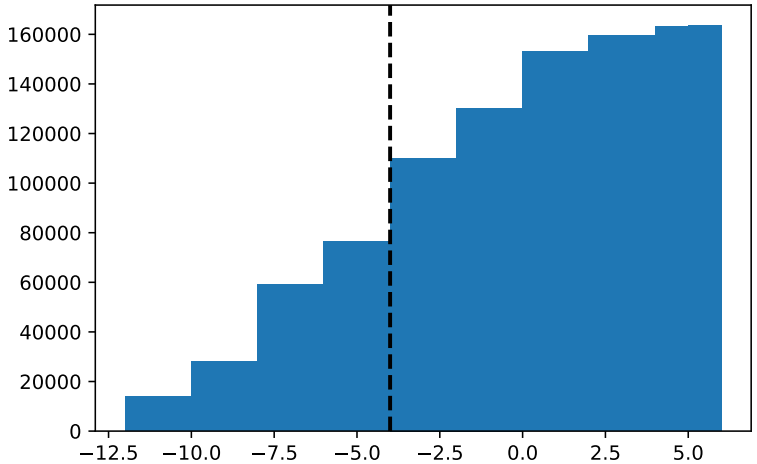
Subsets of octahedral space

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

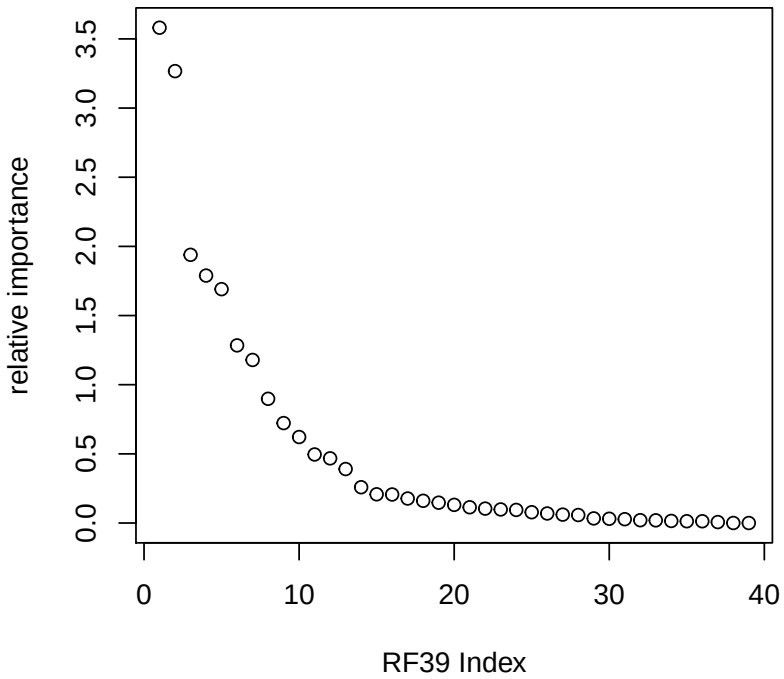
- Reduce space to facilitate sampling from non-homoleptics
- Example: strongly symmetric, monodentate ligand fields (163,620)
- Exclude all with charge smaller than -4, which results in 87,150 ligand fields (53 %).

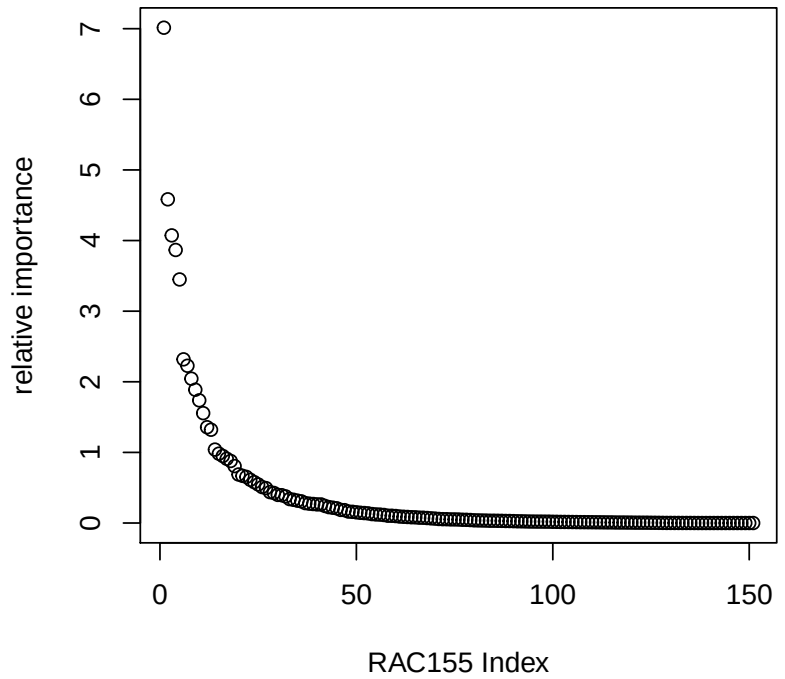




Principal Component Analysis

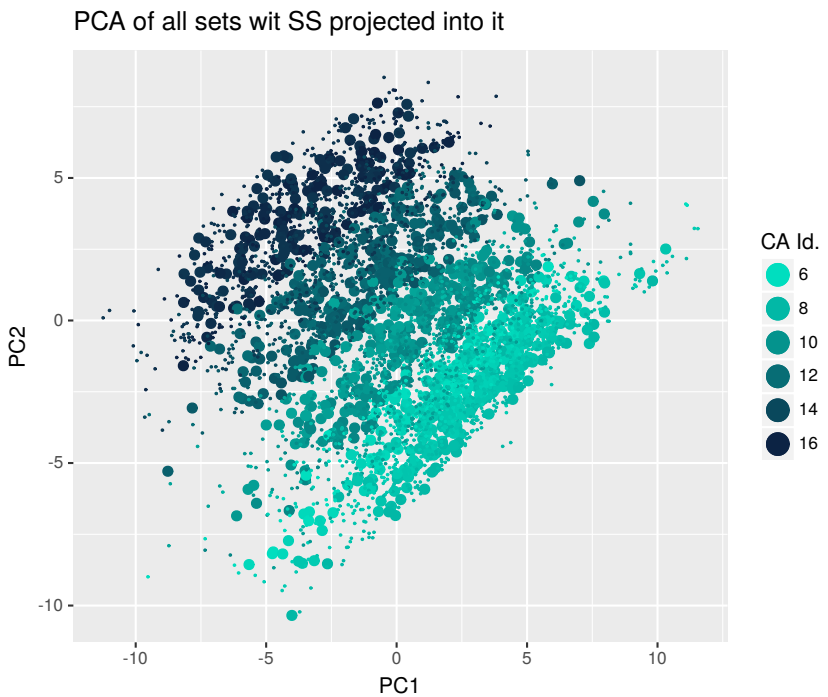
- Use RACs to characterize complexes (JP in JPCA 2017).
- PCA shows highest variance
- Compare eigendecay on RAC155 and RAC RF39.

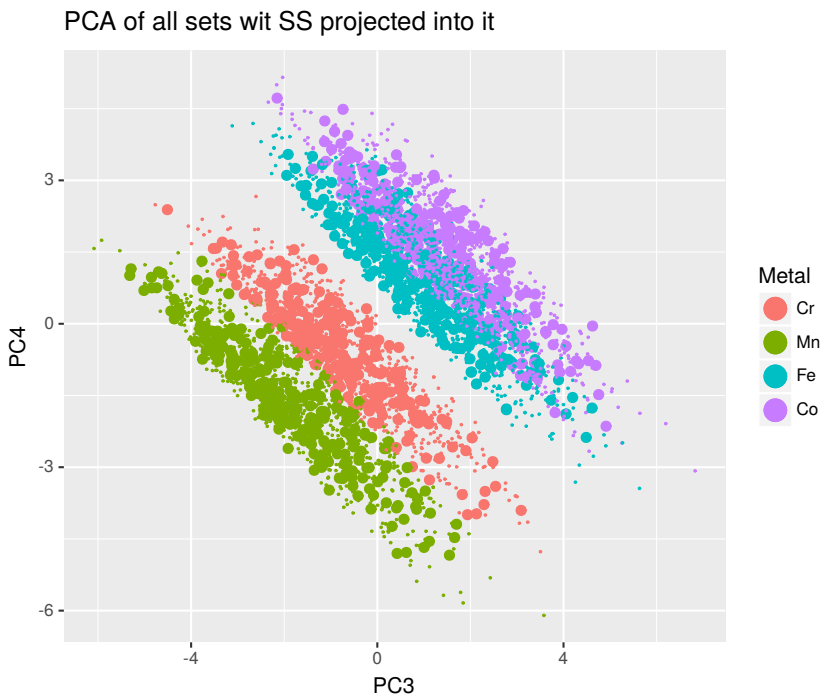




PCA: Strongly symmetric complexes

- Only RF39.
- Colored by connecting atom type (PC1 and PC2)
- Colored by metal type (PC3 and PC4)





PCA: Homoleptics span the space

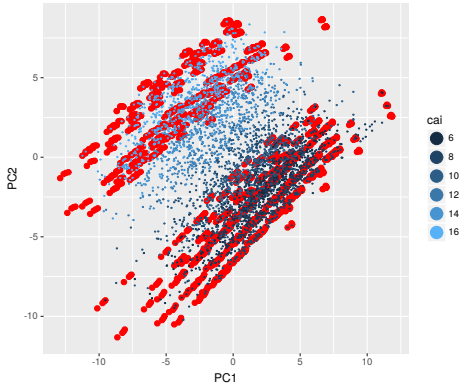


Figure : Homoleptics

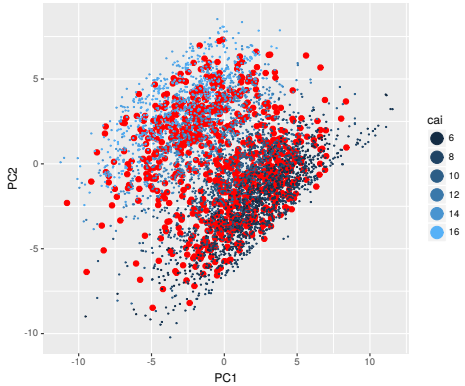


Figure : 5+1

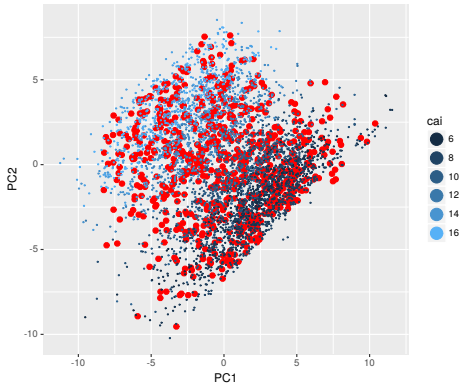


Figure : 4+2

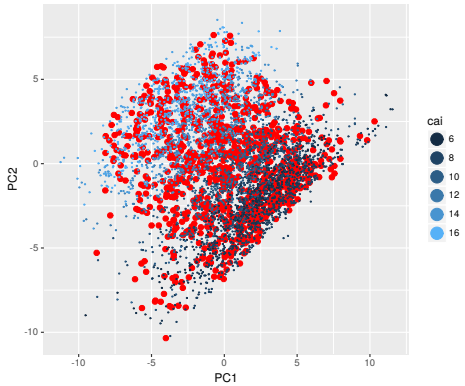


Figure : Strongly Symmetric

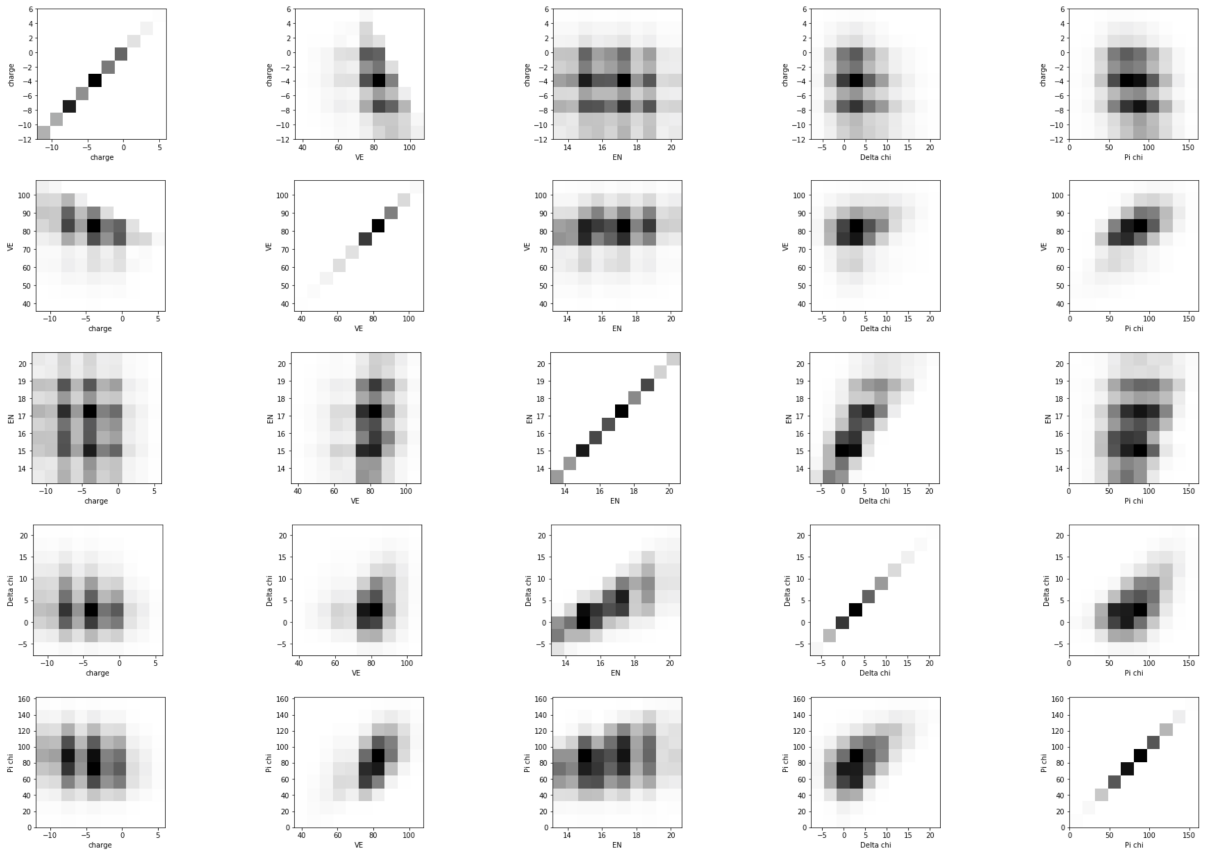
Footprint

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
- ${}^{\mathrm{lc}}_{\mathrm{ax,eq}}\chi_1 = \sum EN_{\mathrm{CA}} \cdot EN_i$
- ${}^{\mathrm{lc}}_{\mathrm{ax,eq}}\chi'_1 = \sum EN_{\mathrm{CA}} - EN_i$

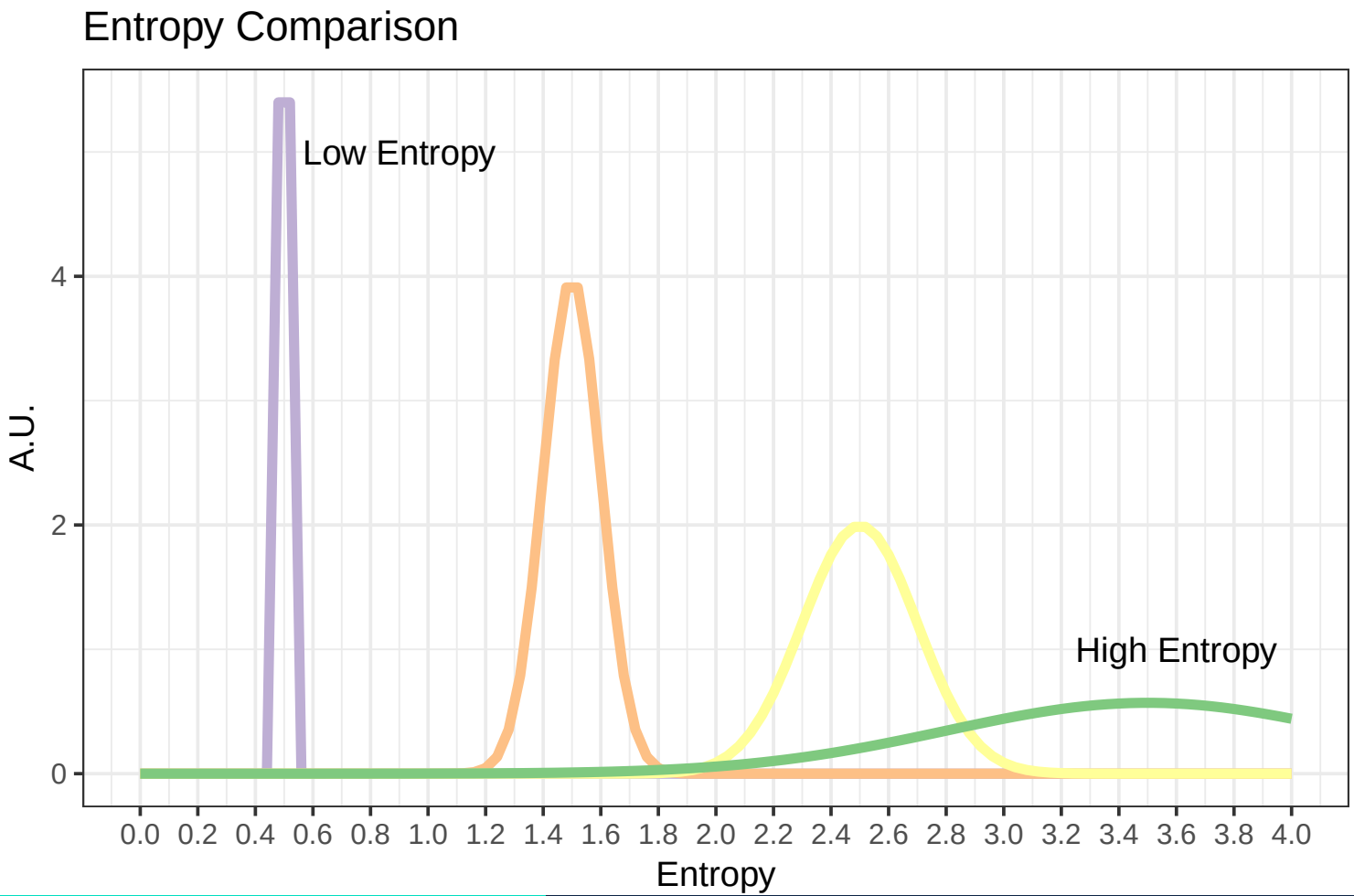
Correlations for strongly symmetric monodentates

The less peaky the better.



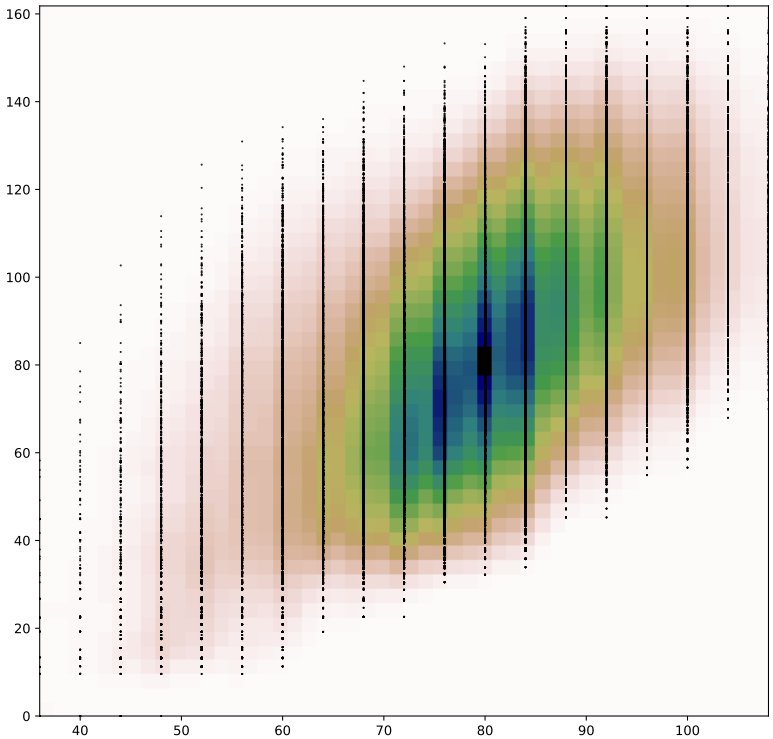
Entropy and KDE

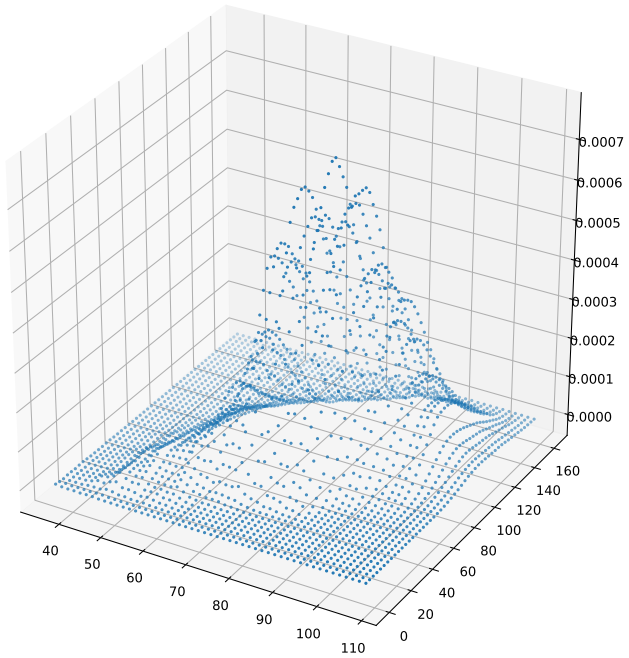
We then calculate the entropy, H_{KDE}, of the Kernel Density Estimated distrbution. We want it to be uniform (high entropy) not to oversample.



Example of KDE slice

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



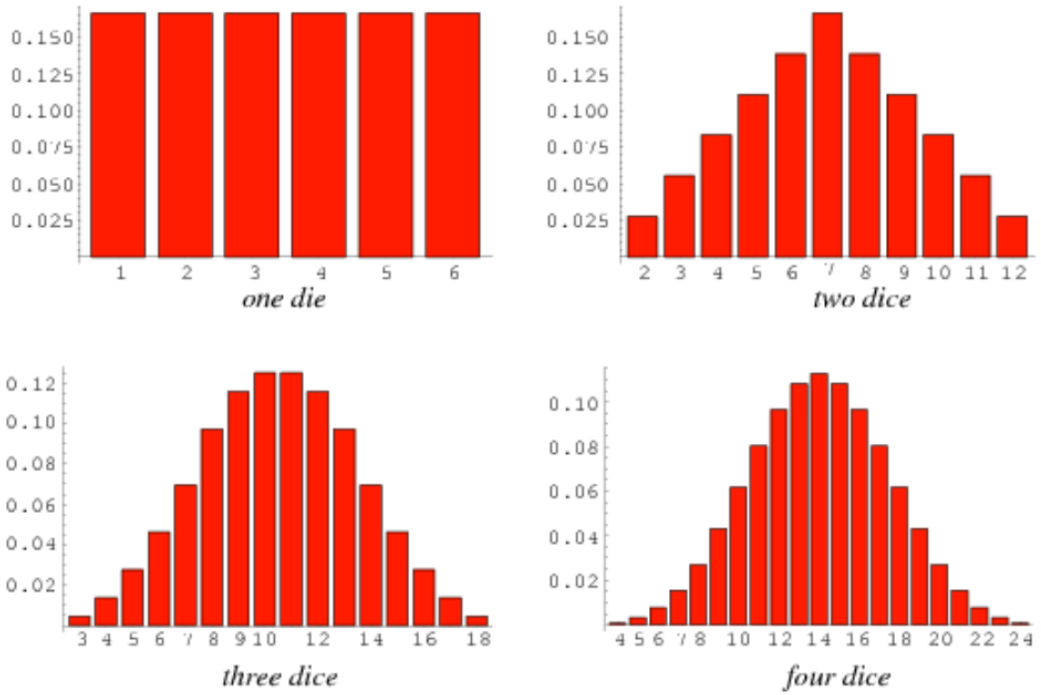


Entropies of various symmetry classes

Table : Entropic footprint

Set	H_{KDE}
Homoleptics	20.18
Strongly symmetric	14.04
”4+2” symmetric	13.68
”5+1” symmetric	13.65
Equatorially asymmetric	10.57
Weakly symmetric	8.86

Multinomial peaking



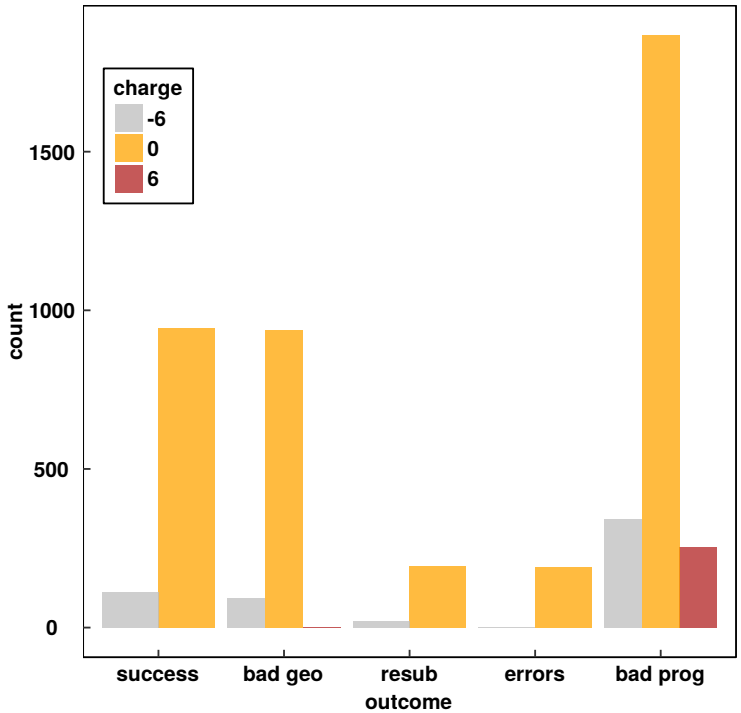
- The more dice, the more the distribution peaks which results in low entropy.
- Sample from low entropy uniformly and get similar molecules.

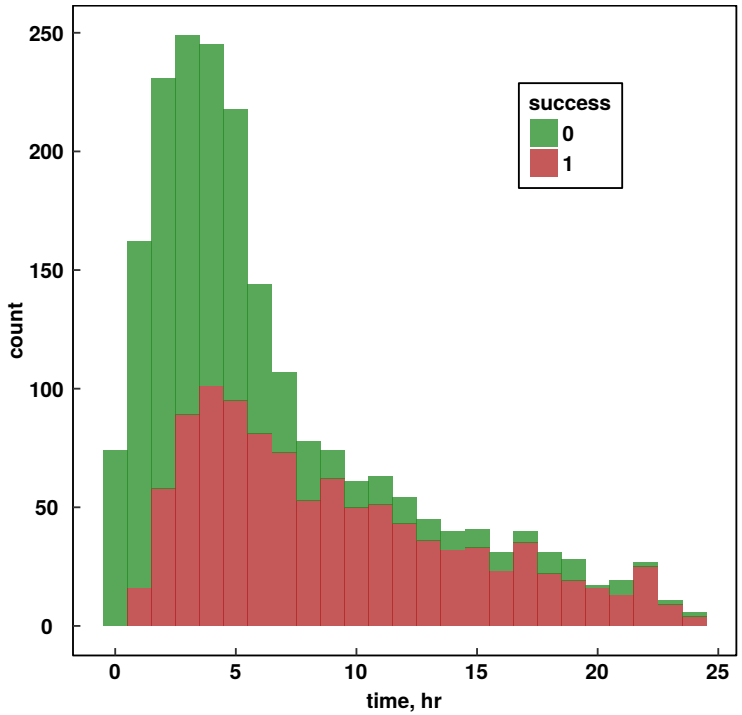
DFT calculations

We calculated the homoleptic sets (405 ligands).

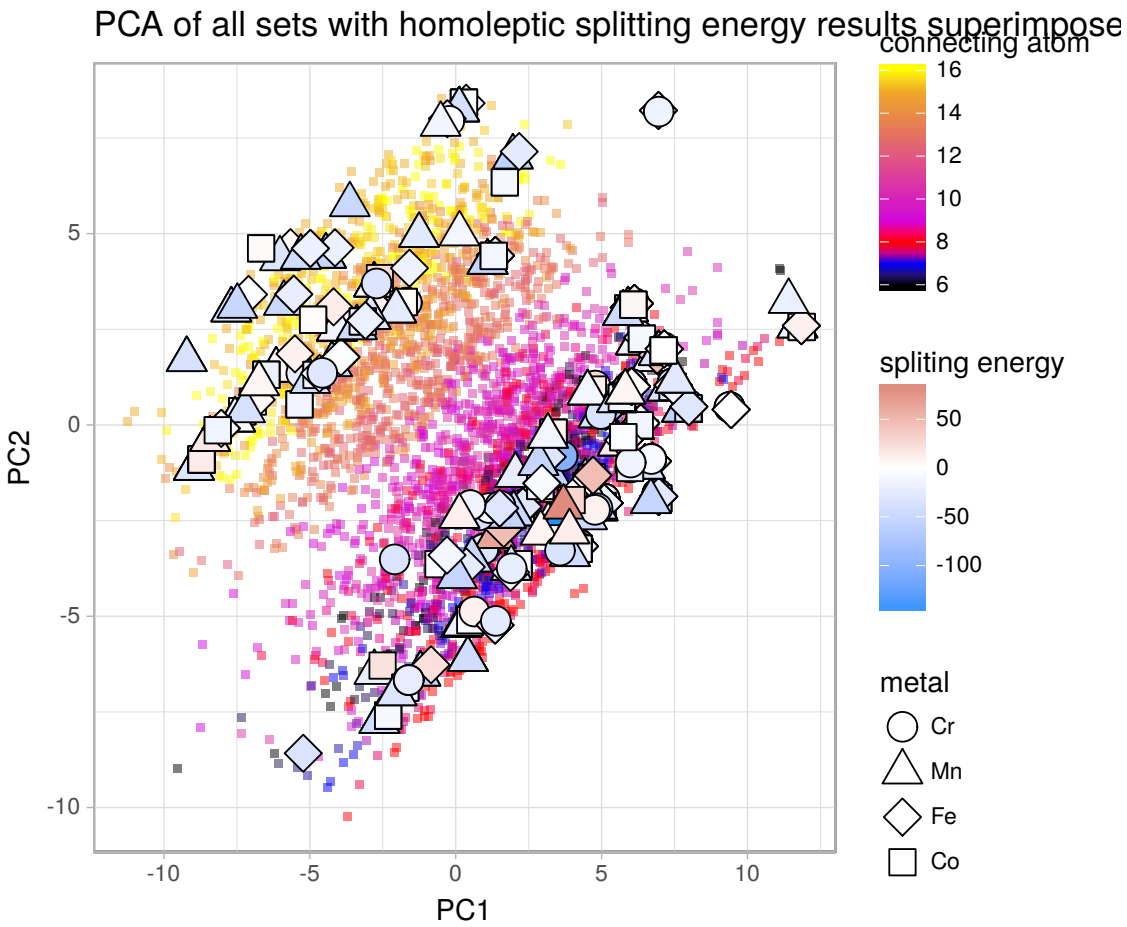
- Basis set: LACVPS (6-31G*)
- Effective core potential: LANL2DZ
- Functional: B3LYP

DFT calculations analysis





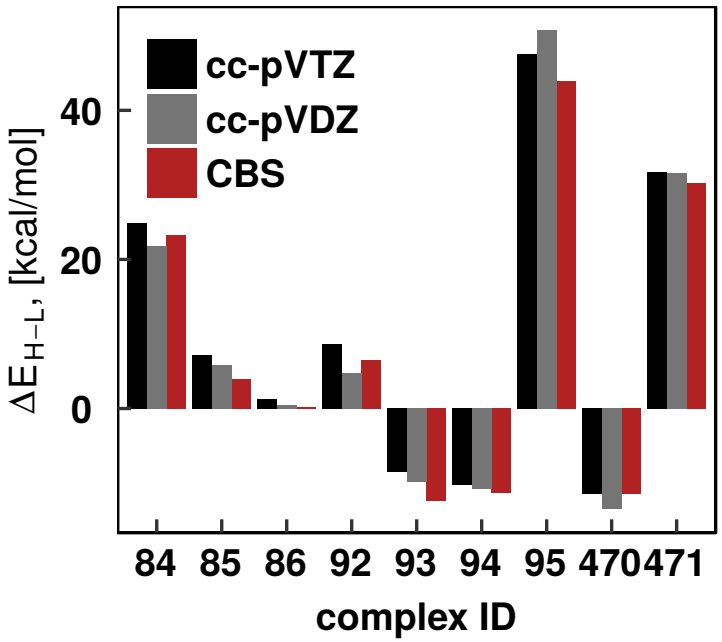
DFT calculations analysis

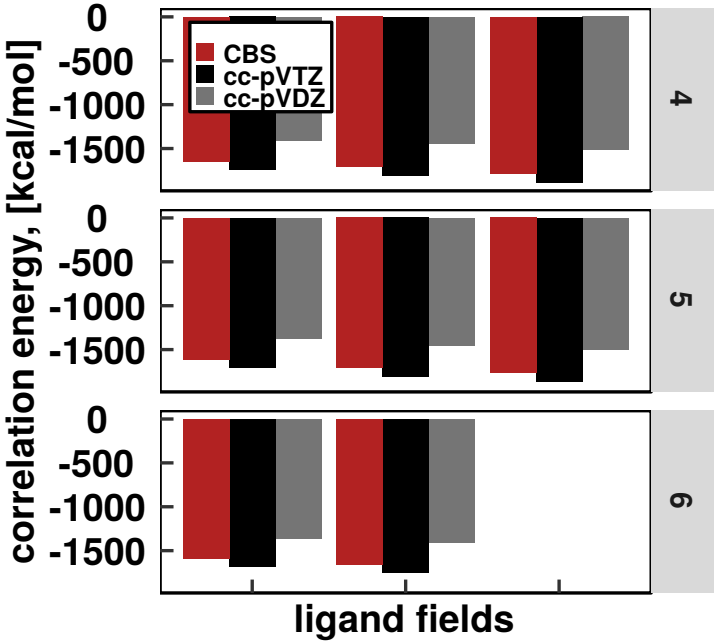


DLPNO-CCSD(T) calculations

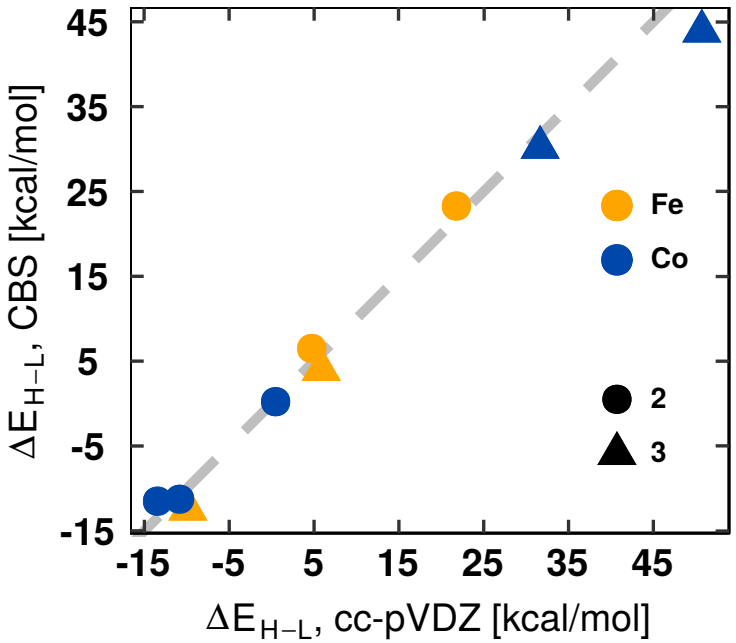
- Basis set: CC-pV(D/T)Z (also tried def2)
- Auxiliary Basis: AutoAux
- RIJCOSX approximation for speed-up
- Increased maximum iterations for SCF to 500
- NormalPNO vs. TightPNO

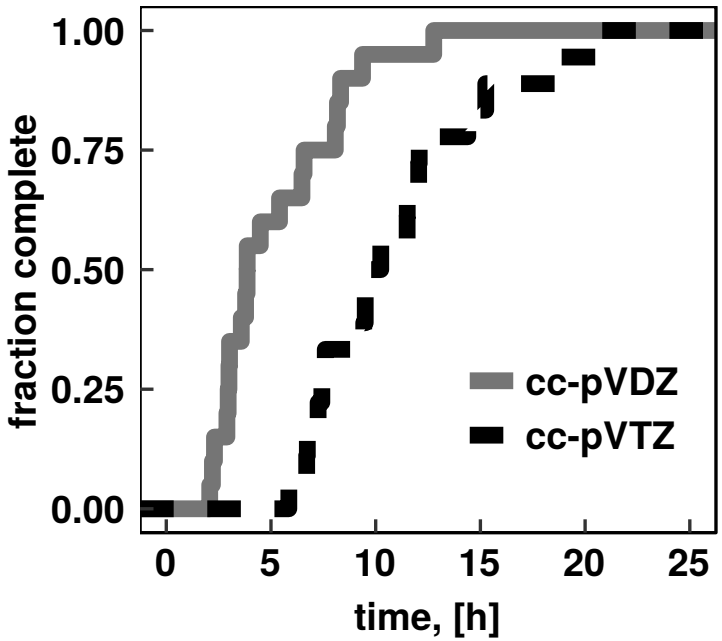
Basis set comparison



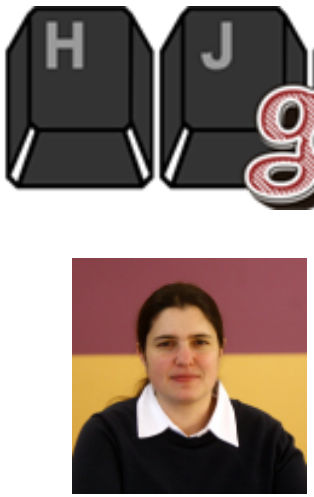


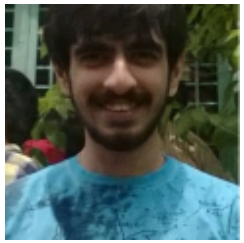
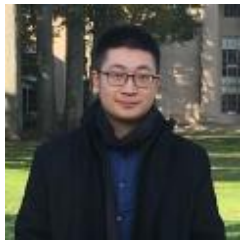
CBS comparison and time analysis





Thanks for the opportunity to work in this lab!








ECCS, CBET
CAMD
CTC/CCS
BURROUGHS WELLCOME FUND
BOSCH
NIH
MIT CEHS