Electronic Supplementary Material (ESI) for Molecular Systems Design & Engineering.

Electronic Supplementary Information for Enumeration of *de novo* inorganic complexes for chemical discovery and machine learning

Stefan Gugler[1], Jon Paul Janet[1], and Heather J. Kulik [*1]

[1] *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*

[*]Corresponding Author: hjkulik@mit.edu

Text S0: supporting files index

In addition to results reported here, additional files are provided as follows:

provided files:

- **ligand-scores.zip**
  - *readme.txt*
  - *scores-m1.csv*
  - *scores-m2.csv*
  - *scores-b4.csv*
- **DFT-results.zip**
  - *readme.txt*
  - *CSD-results.csv*
  - *train-results.csv*
  - *OHLDB-results.csv*
- **predictions.zip**
  - *readme.txt*
  - *old-csd-predictions.csv*
  - *retrained-csd-predictions.csv*
  - *ohldb-predictions.csv*
- **geometries.zip**
  - *readme.txt*
  - **CSD**
    - *CODE_[spin].xyz*
  - **inorganic_training**
    - *[metal]_[oxidation]_[equatorial_ligand]_[axial_ligand_1]_[axial_ligand_2]_[%HFX]_[spin].xyz*
  - **OHLDB**
    - *[metal]_[oxidation]_[ligand]_[spin].xyz*
- **models.zip**
  - *readme.txt*
  - **old-model**
    - *split_vars.csv*
    - *split_mean_x.csv/u0_mean_y.csv*
    - *split_var_x.csv/u0_var_y.csv*
    - *split_model.json*
    - *split_model.h5*
  - **retrained-model**
    - *split_vars.csv*
    - *split_mean_x.csv/u0_mean_y.csv*
    - *split_var_x.csv/u0_var_y.csv*
    - *split_model.json*
    - *split_model.h5*

Table S1: Component and total scores along with SMILES string representations for 29 (of 125 theoretical) retained M1 ligands. The ligands selected have an $s_{\text{tot}} \geq 8$. The theoretical maximum s_{tot} is 10, and the variations in overall score can be due to variations in s_{charge}, the score for overall ligand charge, s_{octet}, the score for violating the octet rule, and s_{sterics}, the score for how much steric repulsion occurs at the metal-coordinating heavy atom. M1 ligands present in the spectrochemical series include NH_3, O^{2-}, OH_2, OH^-, and S^{2-}.

SMILES	s_{charge}	s_{octet}	s_{sterics}	s_{tot}
[CH2- -]	3	4	3	10
[CH3-]	3	4	3	10
[NH- -]	3	4	3	10
[NH2-]	3	4	3	10
[NH3]	3	4	3	10
[O- -]	3	4	3	10
[OH-]	3	4	3	10
[OH2]	3	4	3	10
[PH- -]	3	4	3	10
[PH2-]	3	4	3	10
[PH3]	3	4	3	10
[S- -]	3	4	3	10
[SH-]	3	4	3	10
[SH2]	3	4	3	10
[C- -]	3	2	3	8
[CH-]	3	2	3	8
[CH2]	3	2	3	8
[NH3- -]	3	2	3	8
[N-]	3	2	3	8
[NH]	3	2	3	8
[OH2- -]	3	2	3	8
[OH3-]	3	2	3	8
[O]	3	2	3	8
[PH3- -]	3	2	3	8
[P-]	3	2	3	8
[PH]	3	2	3	8
[SH2- -]	3	2	3	8
[SH3-]	3	2	3	8
[S]	3	2	3	8

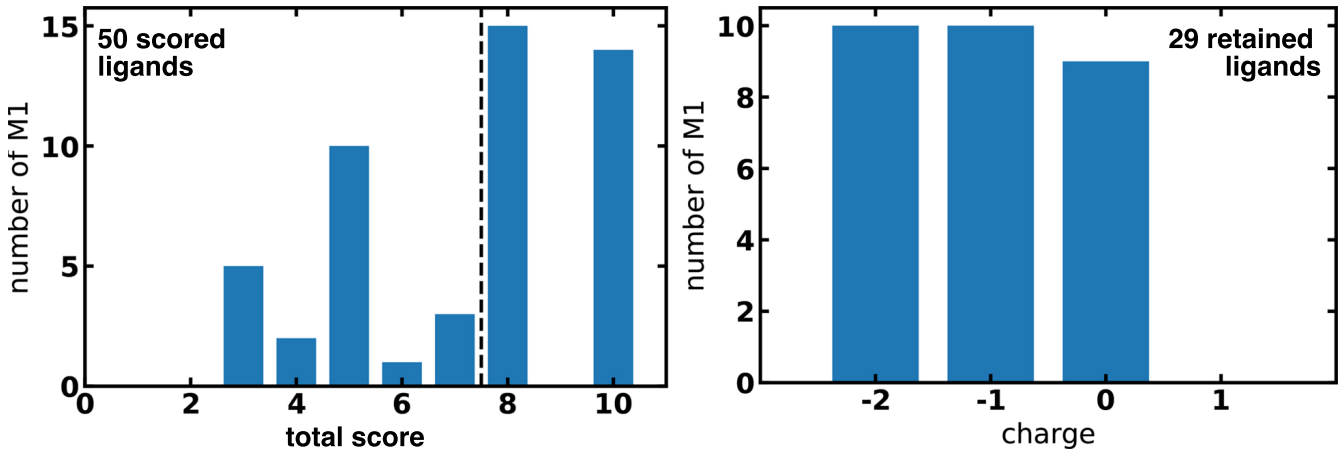

Figure S1: Left: Histogram of total score for 50 scored M1 ligands with the cutoff for retained ligands (i.e., $s_{\text{tot}} \geq 8$) indicated as a vertical dashed line. Right: Charge distribution of the 29 retained M1 ligands. No $+1$ charged ligands are retained because they can score at most an $s_{tot} = 7$ due to the large penalty for positively charged ligands in s_{charge}.

Algorithm S1: Determination of charge and bond order for M2 ligands for a given total charge, c_{tot}. The individual atom 1 and 2 charges, c_1 and c_2 are varied while satisfying the relationship between the valence, v, and the number of lone pairs l, standard number of valence electrons ve, number of hydrogen atoms h. The bond order, b is then set.

Require: $c_{\text{tot}} \in [-4, 4]$
  for $c_1 = -2$ **to** 2 **do**
    for $c_2 = -2$ **to** 2 **do**
      if $c_{\text{tot}} == c_1 + c_2$ **then**
        $v_1 \leftarrow ve_1 - c_1 - 2 \cdot l_1 - h_1$
        $v_2 \leftarrow ve_2 - c_2 - 2 \cdot l_2 - h_2$
        $\delta \leftarrow |v_1 - v_2|$
        $b \leftarrow \min\{v_1, v_2\}$
      end if
    end for
  end for
  if $b \in (0, 4]$ **then**
    return b, c_1, c_2 at smallest δ
  else
    $c_1 \leftarrow c_{\text{tot}}$
    $b \leftarrow 0$
    return b, c_1, c_2
  end if

Table S2: The total valence score v assigned is the net of ve-c-2l-h. The c and h values depend on the enumeration, whereas ve and l are assigned as follows below for neutral atoms.

elem.	l	ve
C	0	4
N	1	5
O	2	6
P	1	5
S	2	6

Table S3: Component and total scores along with SMILES string representations for 494 (of 5,625 theoretical) retained M2 ligands. The ligands selected have an $s_{\text{tot}} > 13$. The theoretical maximum s_{tot} is 17, and the variations in overall score can be due to variations in any of five component scores: s_{pol}, the score for how much charge separation there is in the two heavy atoms in the M2 ligand, s_{bond}, the score for the bond order between the two heavy atoms, s_{charge}, the score for overall ligand charge, s_{octet}, the score for violating the octet rule, and s_{sterics}, the score for how much steric repulsion occurs at the metal-coordinating heavy atom. M2 scored ligands present in the spectrochemical series include O_2^{2-}, CN^-, and CO.

SMILES	s_{pol}	s_{bond}	s_{charge}	s_{octet}	s_{sterics}	s_{tot}
[CH2+]-[CH4-]	3	3	5	3	3	17
[CH2+]-[NH3-]	3	3	5	3	3	17
[CH2+]-[OH2-]	3	3	5	3	3	17
[CH2+]-[PH3-]	3	3	5	3	3	17
[CH2+]-[SH2-]	3	3	5	3	3	17
[NH+]-[CH4-]	3	3	5	3	3	17
[NH2]-[CH3]	3	3	5	3	3	17
[NH+]-[NH3-]	3	3	5	3	3	17
[NH2]-[NH2]	3	3	5	3	3	17
[NH+]-[OH2-]	3	3	5	3	3	17
[NH2]-[OH]	3	3	5	3	3	17
[NH+]-[PH3-]	3	3	5	3	3	17
[NH2]-[PH2]	3	3	5	3	3	17
[NH+]-[SH2-]	3	3	5	3	3	17
[NH2]-[SH]	3	3	5	3	3	17
[O+]-[CH4-]	3	3	5	3	3	17
[OH]-[CH3]	3	3	5	3	3	17
[OH2-]-[CH2+]	3	3	5	3	3	17
[O+]-[NH3-]	3	3	5	3	3	17
[OH]-[NH2]	3	3	5	3	3	17
[OH2-]-[NH+]	3	3	5	3	3	17
[O+]-[OH2-]	3	3	5	3	3	17
[OH]-[OH]	3	3	5	3	3	17
[OH2-]-[O+]	3	3	5	3	3	17
[O+]-[PH3-]	3	3	5	3	3	17
[OH]-[PH2]	3	3	5	3	3	17
[OH2-]-[PH+]	3	3	5	3	3	17
[O+]-[SH2-]	3	3	5	3	3	17
[OH]-[SH]	3	3	5	3	3	17
[OH2-]-[S+]	3	3	5	3	3	17
[PH+]-[CH4-]	3	3	5	3	3	17
[PH2]-[CH3]	3	3	5	3	3	17
[PH+]-[NH3-]	3	3	5	3	3	17
[PH2]-[NH2]	3	3	5	3	3	17
[PH+]-[OH2-]	3	3	5	3	3	17

[PH2]-[OH]	3	3	5	3	3	17
[PH+]-[PH3-]	3	3	5	3	3	17
[PH2]-[PH2]	3	3	5	3	3	17
[PH+]-[SH2-]	3	3	5	3	3	17
[PH2]-[SH]	3	3	5	3	3	17
[S+]-[CH4-]	3	3	5	3	3	17
[SH]-[CH3]	3	3	5	3	3	17
[SH2-]-[CH2+]	3	3	5	3	3	17
[S+]-[NH3-]	3	3	5	3	3	17
[SH]-[NH2]	3	3	5	3	3	17
[SH2-]-[NH+]	3	3	5	3	3	17
[S+]-[OH2-]	3	3	5	3	3	17
[SH]-[OH]	3	3	5	3	3	17
[SH2-]-[O+]	3	3	5	3	3	17
[S+]-[PH3-]	3	3	5	3	3	17
[SH]-[PH2]	3	3	5	3	3	17
[SH2-]-[PH+]	3	3	5	3	3	17
[S+]-[SH2-]	3	3	5	3	3	17
[SH]-[SH]	3	3	5	3	3	17
[SH2-]-[S+]	3	3	5	3	3	17
[C]#4[C]	3	3	5	2	3	16
[C+]#[CH2-]	3	3	5	2	3	16
[CH]#[CH]	3	3	5	2	3	16
[CH+]=[CH3-]	3	3	5	2	3	16
[CH2-]#[C+]	3	3	5	2	3	16
[CH2]=[CH2]	3	3	5	2	3	16
[C+]#[NH-]	3	3	5	2	3	16
[CH]#[N]	3	3	5	2	3	16
[CH+]=[NH2-]	3	3	5	2	3	16
[CH2]=[NH]	3	3	5	2	3	16
[C+]#[O-]	3	3	5	2	3	16
[CH+]=[OH-]	3	3	5	2	3	16
[CH2]=[O]	3	3	5	2	3	16
[C+]#[PH-]	3	3	5	2	3	16
[CH]#[P]	3	3	5	2	3	16
[CH+]=[PH2-]	3	3	5	2	3	16
[CH2]=[PH]	3	3	5	2	3	16
[C+]#[S-]	3	3	5	2	3	16
[CH+]=[SH-]	3	3	5	2	3	16
[CH2]=[S]	3	3	5	2	3	16
[NH2]-[CH4-]	2	3	5	3	3	16
[N]#[CH]	3	3	5	2	3	16
[N+]=[CH3-]	3	3	5	2	3	16
[NH-]#[C+]	3	3	5	2	3	16
[NH]=[CH2]	3	3	5	2	3	16
[NH2-]=[CH+]	3	3	5	2	3	16
[NH2]-[NH4- -]	2	3	5	3	3	16

[NH2]-[NH3-]	2	3	5	3	3	16
[N]#[N]	3	3	5	2	3	16
[N+]=[NH2-]	3	3	5	2	3	16
[NH]=[NH]	3	3	5	2	3	16
[NH2-]=[N+]	3	3	5	2	3	16
[NH2]-[OH3- -]	2	3	5	3	3	16
[NH2]-[OH2-]	2	3	5	3	3	16
[N+]=[OH-]	3	3	5	2	3	16
[NH]=[O]	3	3	5	2	3	16
[NH2]-[PH4- -]	2	3	5	3	3	16
[NH2]-[PH3-]	2	3	5	3	3	16
[N]#[P]	3	3	5	2	3	16
[N+]=[PH2-]	3	3	5	2	3	16
[NH]=[PH]	3	3	5	2	3	16
[NH2-]=[P+]	3	3	5	2	3	16
[NH2]-[SH3- -]	2	3	5	3	3	16
[NH2]-[SH2-]	2	3	5	3	3	16
[N+]=[SH-]	3	3	5	2	3	16
[NH]=[S]	3	3	5	2	3	16
[OH2-]-[CH4-]	2	3	5	3	3	16
[OH]-[CH4-]	2	3	5	3	3	16
[OH2-]-[CH3]	2	3	5	3	3	16
[O-]#[C+]	3	3	5	2	3	16
[O]=[CH2]	3	3	5	2	3	16
[OH-]=[CH+]	3	3	5	2	3	16
[OH]-[NH4- -]	2	3	5	3	3	16
[OH2-]-[NH3-]	2	3	5	3	3	16
[OH]-[NH3-]	2	3	5	3	3	16
[OH2-]-[NH2]	2	3	5	3	3	16
[O]=[NH]	3	3	5	2	3	16
[OH-]=[N+]	3	3	5	2	3	16
[OH]-[OH3- -]	2	3	5	3	3	16
[OH2-]-[OH2-]	2	3	5	3	3	16
[OH]-[OH2-]	2	3	5	3	3	16
[OH2-]-[OH]	2	3	5	3	3	16
[O]=[O]	3	3	5	2	3	16
[OH]-[PH4- -]	2	3	5	3	3	16
[OH2-]-[PH3-]	2	3	5	3	3	16
[OH]-[PH3-]	2	3	5	3	3	16
[OH2-]-[PH2]	2	3	5	3	3	16
[O]=[PH]	3	3	5	2	3	16
[OH-]=[P+]	3	3	5	2	3	16
[OH]-[SH3- -]	2	3	5	3	3	16
[OH2-]-[SH2-]	2	3	5	3	3	16
[OH]-[SH2-]	2	3	5	3	3	16
[OH2-]-[SH]	2	3	5	3	3	16
[O]=[S]	3	3	5	2	3	16

[PH2]-[CH4-]	2	3	5	3	3	16
[P]#[CH]	3	3	5	2	3	16
[P+]=[CH3-]	3	3	5	2	3	16
[PH-]#[C+]	3	3	5	2	3	16
[PH]=[CH2]	3	3	5	2	3	16
[PH2-]=[CH+]	3	3	5	2	3	16
[PH2]-[NH4- -]	2	3	5	3	3	16
[PH2]-[NH3-]	2	3	5	3	3	16
[P]#[N]	3	3	5	2	3	16
[P+]=[NH2-]	3	3	5	2	3	16
[PH]=[NH]	3	3	5	2	3	16
[PH2-]=[N+]	3	3	5	2	3	16
[PH2]-[OH3- -]	2	3	5	3	3	16
[PH2]-[OH2-]	2	3	5	3	3	16
[P+]=[OH-]	3	3	5	2	3	16
[PH]=[O]	3	3	5	2	3	16
[PH2]-[PH4- -]	2	3	5	3	3	16
[PH2]-[PH3-]	2	3	5	3	3	16
[P]#[P]	3	3	5	2	3	16
[P+]=[PH2-]	3	3	5	2	3	16
[PH]=[PH]	3	3	5	2	3	16
[PH2-]=[P+]	3	3	5	2	3	16
[PH2]-[SH3- -]	2	3	5	3	3	16
[PH2]-[SH2-]	2	3	5	3	3	16
[P+]=[SH-]	3	3	5	2	3	16
[PH]=[S]	3	3	5	2	3	16
[SH2-]-[CH4-]	2	3	5	3	3	16
[SH]-[CH4-]	2	3	5	3	3	16
[SH2-]-[CH3]	2	3	5	3	3	16
[S-]#[C+]	3	3	5	2	3	16
[S]=[CH2]	3	3	5	2	3	16
[SH-]=[CH+]	3	3	5	2	3	16
[SH]-[NH4- -]	2	3	5	3	3	16
[SH2-]-[NH3-]	2	3	5	3	3	16
[SH]-[NH3-]	2	3	5	3	3	16
[SH2-]-[NH2]	2	3	5	3	3	16
[S]=[NH]	3	3	5	2	3	16
[SH-]=[N+]	3	3	5	2	3	16
[SH]-[OH3- -]	2	3	5	3	3	16
[SH2-]-[OH2-]	2	3	5	3	3	16
[SH]-[OH2-]	2	3	5	3	3	16
[SH2-]-[OH]	2	3	5	3	3	16
[S]=[O]	3	3	5	2	3	16
[SH]-[PH4- -]	2	3	5	3	3	16
[SH2-]-[PH3-]	2	3	5	3	3	16
[SH]-[PH3-]	2	3	5	3	3	16
[SH2-]-[PH2]	2	3	5	3	3	16

[S]=[PH]	3	3	5	2	3	16
[SH-]=[P+]	3	3	5	2	3	16
[SH]-[SH3- -]	2	3	5	3	3	16
[SH2-]-[SH2-]	2	3	5	3	3	16
[SH]-[SH2-]	2	3	5	3	3	16
[SH2-]-[SH]	2	3	5	3	3	16
[S]=[S]	3	3	5	2	3	16
[C]#4[CH2- -]	2	3	5	2	3	15
[CH-]#4[CH-]	2	3	5	2	3	15
[CH]#[CH3- -]	2	3	5	2	3	15
[CH2- -]#4[C]	2	3	5	2	3	15
[CH2-]#[CH2-]	2	3	5	2	3	15
[CH2]=[CH4- -]	2	3	5	2	3	15
[C]#4[CH-]	2	3	5	2	3	15
[CH-]#4[C]	2	3	5	2	3	15
[CH]#[CH2-]	2	3	5	2	3	15
[CH2-]#[CH]	2	3	5	2	3	15
[CH2]=[CH3-]	2	3	5	2	3	15
[C]#4[NH- -]	2	3	5	2	3	15
[CH-]#4[N-]	2	3	5	2	3	15
[CH]#[NH2- -]	2	3	5	2	3	15
[CH2-]#[NH-]	2	3	5	2	3	15
[CH2]=[NH3- -]	2	3	5	2	3	15
[C]#4[N-]	2	3	5	2	3	15
[CH]#[NH-]	2	3	5	2	3	15
[CH2-]#[N]	2	3	5	2	3	15
[CH2]=[NH2-]	2	3	5	2	3	15
[C]#4[O- -]	2	3	5	2	3	15
[CH]#[OH- -]	2	3	5	2	3	15
[CH2-]#[O-]	2	3	5	2	3	15
[CH2]=[OH2- -]	2	3	5	2	3	15
[CH]#[O-]	2	3	5	2	3	15
[CH2]=[OH-]	2	3	5	2	3	15
[C]#4[PH- -]	2	3	5	2	3	15
[CH-]#4[P-]	2	3	5	2	3	15
[CH]#[PH2- -]	2	3	5	2	3	15
[CH2-]#[PH-]	2	3	5	2	3	15
[CH2]=[PH3- -]	2	3	5	2	3	15
[C]#4[P-]	2	3	5	2	3	15
[CH]#[PH-]	2	3	5	2	3	15
[CH2-]#[P]	2	3	5	2	3	15
[CH2]=[PH2-]	2	3	5	2	3	15
[C]#4[S- -]	2	3	5	2	3	15
[CH]#[SH- -]	2	3	5	2	3	15
[CH2-]#[S-]	2	3	5	2	3	15
[CH2]=[SH2- -]	2	3	5	2	3	15
[CH]#[S-]	2	3	5	2	3	15

[CH2]=[SH-]	2	3	5	2	3	15
[N-]#4[CH-]	2	3	5	2	3	15
[N]#[CH3- -]	2	3	5	2	3	15
[NH- -]#4[C]	2	3	5	2	3	15
[NH-]#[CH2-]	2	3	5	2	3	15
[NH]=[CH4- -]	2	3	5	2	3	15
[NH2- -]#[CH]	2	3	5	2	3	15
[NH2-]=[CH3-]	2	3	5	2	3	15
[N-]#4[C]	2	3	5	2	3	15
[N]#[CH2-]	2	3	5	2	3	15
[NH-]#[CH]	2	3	5	2	3	15
[NH]=[CH3-]	2	3	5	2	3	15
[NH2-]=[CH2]	2	3	5	2	3	15
[N-]#4[N-]	2	3	5	2	3	15
[N]#[NH2- -]	2	3	5	2	3	15
[NH-]#[NH-]	2	3	5	2	3	15
[NH]=[NH3- -]	2	3	5	2	3	15
[NH2- -]#[N]	2	3	5	2	3	15
[NH2-]=[NH2-]	2	3	5	2	3	15
[N]#[NH-]	2	3	5	2	3	15
[NH-]#[N]	2	3	5	2	3	15
[NH]=[NH2-]	2	3	5	2	3	15
[NH2-]=[NH]	2	3	5	2	3	15
[N]#[OH- -]	2	3	5	2	3	15
[NH-]#[O-]	2	3	5	2	3	15
[NH]=[OH2- -]	2	3	5	2	3	15
[NH2-]=[OH-]	2	3	5	2	3	15
[N]#[O-]	2	3	5	2	3	15
[NH]=[OH-]	2	3	5	2	3	15
[NH2-]=[O]	2	3	5	2	3	15
[N-]#4[P-]	2	3	5	2	3	15
[N]#[PH2- -]	2	3	5	2	3	15
[NH-]#[PH-]	2	3	5	2	3	15
[NH]=[PH3- -]	2	3	5	2	3	15
[NH2- -]#[P]	2	3	5	2	3	15
[NH2-]=[PH2-]	2	3	5	2	3	15
[N]#[PH-]	2	3	5	2	3	15
[NH-]#[P]	2	3	5	2	3	15
[NH]=[PH2-]	2	3	5	2	3	15
[NH2-]=[PH]	2	3	5	2	3	15
[N]#[SH- -]	2	3	5	2	3	15
[NH-]#[S-]	2	3	5	2	3	15
[NH]=[SH2- -]	2	3	5	2	3	15
[NH2-]=[SH-]	2	3	5	2	3	15
[N]#[S-]	2	3	5	2	3	15
[NH]=[SH-]	2	3	5	2	3	15
[NH2-]=[S]	2	3	5	2	3	15

[O- -]#4[C]	2	3	5	2	3	15
[O-]#[CH2-]	2	3	5	2	3	15
[O]=[CH4- -]	2	3	5	2	3	15
[OH- -]#[CH]	2	3	5	2	3	15
[OH-]=[CH3-]	2	3	5	2	3	15
[OH2- -]=[CH2]	2	3	5	2	3	15
[O-]#[CH]	2	3	5	2	3	15
[O]=[CH3-]	2	3	5	2	3	15
[OH-]=[CH2]	2	3	5	2	3	15
[O-]#[NH-]	2	3	5	2	3	15
[O]=[NH3- -]	2	3	5	2	3	15
[OH- -]#[N]	2	3	5	2	3	15
[OH-]=[NH2-]	2	3	5	2	3	15
[OH2- -]=[NH]	2	3	5	2	3	15
[O-]#[N]	2	3	5	2	3	15
[O]=[NH2-]	2	3	5	2	3	15
[OH-]=[NH]	2	3	5	2	3	15
[O-]#[O-]	2	3	5	2	3	15
[O]=[OH2- -]	2	3	5	2	3	15
[OH-]=[OH-]	2	3	5	2	3	15
[OH2- -]=[O]	2	3	5	2	3	15
[O]=[OH-]	2	3	5	2	3	15
[OH-]=[O]	2	3	5	2	3	15
[O-]#[PH-]	2	3	5	2	3	15
[O]=[PH3- -]	2	3	5	2	3	15
[OH- -]#[P]	2	3	5	2	3	15
[OH-]=[PH2-]	2	3	5	2	3	15
[OH2- -]=[PH]	2	3	5	2	3	15
[O-]#[P]	2	3	5	2	3	15
[O]=[PH2-]	2	3	5	2	3	15
[OH-]=[PH]	2	3	5	2	3	15
[O-]#[S-]	2	3	5	2	3	15
[O]=[SH2- -]	2	3	5	2	3	15
[OH-]=[SH-]	2	3	5	2	3	15
[OH2- -]=[S]	2	3	5	2	3	15
[O]=[SH-]	2	3	5	2	3	15
[OH-]=[S]	2	3	5	2	3	15
[P-]#4[CH-]	2	3	5	2	3	15
[P]#[CH3- -]	2	3	5	2	3	15
[PH- -]#4[C]	2	3	5	2	3	15
[PH-]#[CH2-]	2	3	5	2	3	15
[PH]=[CH4- -]	2	3	5	2	3	15
[PH2- -]#[CH]	2	3	5	2	3	15
[PH2-]=[CH3-]	2	3	5	2	3	15
[P-]#4[C]	2	3	5	2	3	15
[P]#[CH2-]	2	3	5	2	3	15
[PH-]#[CH]	2	3	5	2	3	15

[PH]=[CH3-]	2	3	5	2	3	15
[PH2-]=[CH2]	2	3	5	2	3	15
[P-]#4[N-]	2	3	5	2	3	15
[P]#[NH2- -]	2	3	5	2	3	15
[PH-]#[NH-]	2	3	5	2	3	15
[PH]=[NH3- -]	2	3	5	2	3	15
[PH2- -]#[N]	2	3	5	2	3	15
[PH2-]=[NH2-]	2	3	5	2	3	15
[P]#[NH-]	2	3	5	2	3	15
[PH-]#[N]	2	3	5	2	3	15
[PH]=[NH2-]	2	3	5	2	3	15
[PH2-]=[NH]	2	3	5	2	3	15
[P]#[OH- -]	2	3	5	2	3	15
[PH-]#[O-]	2	3	5	2	3	15
[PH]=[OH2- -]	2	3	5	2	3	15
[PH2-]=[OH-]	2	3	5	2	3	15
[P]#[O-]	2	3	5	2	3	15
[PH]=[OH-]	2	3	5	2	3	15
[PH2-]=[O]	2	3	5	2	3	15
[P-]#4[P-]	2	3	5	2	3	15
[P]#[PH2- -]	2	3	5	2	3	15
[PH-]#[PH-]	2	3	5	2	3	15
[PH]=[PH3- -]	2	3	5	2	3	15
[PH2- -]#[P]	2	3	5	2	3	15
[PH2-]=[PH2-]	2	3	5	2	3	15
[P]#[PH-]	2	3	5	2	3	15
[PH-]#[P]	2	3	5	2	3	15
[PH]=[PH2-]	2	3	5	2	3	15
[PH2-]=[PH]	2	3	5	2	3	15
[P]#[SH- -]	2	3	5	2	3	15
[PH-]#[S-]	2	3	5	2	3	15
[PH]=[SH2- -]	2	3	5	2	3	15
[PH2-]=[SH-]	2	3	5	2	3	15
[P]#[S-]	2	3	5	2	3	15
[PH]=[SH-]	2	3	5	2	3	15
[PH2-]=[S]	2	3	5	2	3	15
[S- -]#4[C]	2	3	5	2	3	15
[S-]#[CH2-]	2	3	5	2	3	15
[S]=[CH4- -]	2	3	5	2	3	15
[SH- -]#[CH]	2	3	5	2	3	15
[SH-]=[CH3-]	2	3	5	2	3	15
[SH2- -]=[CH2]	2	3	5	2	3	15
[S-]#[CH]	2	3	5	2	3	15
[S]=[CH3-]	2	3	5	2	3	15
[SH-]=[CH2]	2	3	5	2	3	15
[S-]#[NH-]	2	3	5	2	3	15
[S]=[NH3- -]	2	3	5	2	3	15

[SH- -]#[N]	2	3	5	2	3	15
[SH-]=[NH2-]	2	3	5	2	3	15
[SH2- -]=[NH]	2	3	5	2	3	15
[S-]#[N]	2	3	5	2	3	15
[S]=[NH2-]	2	3	5	2	3	15
[SH-]=[NH]	2	3	5	2	3	15
[S-]#[O-]	2	3	5	2	3	15
[S]=[OH2- -]	2	3	5	2	3	15
[SH-]=[OH-]	2	3	5	2	3	15
[SH2- -]=[O]	2	3	5	2	3	15
[S]=[OH-]	2	3	5	2	3	15
[SH-]=[O]	2	3	5	2	3	15
[S-]#[PH-]	2	3	5	2	3	15
[S]=[PH3- -]	2	3	5	2	3	15
[SH- -]#[P]	2	3	5	2	3	15
[SH-]=[PH2-]	2	3	5	2	3	15
[SH2- -]=[PH]	2	3	5	2	3	15
[S-]#[P]	2	3	5	2	3	15
[S]=[PH2-]	2	3	5	2	3	15
[SH-]=[PH]	2	3	5	2	3	15
[S-]#[S-]	2	3	5	2	3	15
[S]=[SH2- -]	2	3	5	2	3	15
[SH-]=[SH-]	2	3	5	2	3	15
[SH2- -]=[S]	2	3	5	2	3	15
[S]=[SH-]	2	3	5	2	3	15
[SH-]=[S]	2	3	5	2	3	15
[CH3]-[CH3]	3	0	5	3	3	14
[CH2+]-[CH3]	0	3	5	3	3	14
[CH]-[NH4- -]	2	3	3	3	3	14
[CH2+]-[NH4- -]	2	3	5	3	1	14
[CH++]-[NH4- -]	3	3	5	3	0	14
[CH3]-[NH2]	3	0	5	3	3	14
[CH2+]-[NH2]	0	3	5	3	3	14
[CH]-[OH3- -]	2	3	3	3	3	14
[CH2+]-[OH3- -]	2	3	5	3	1	14
[CH++]-[OH3- -]	3	3	5	3	0	14
[CH3]-[OH]	3	0	5	3	3	14
[CH2+]-[OH]	0	3	5	3	3	14
[CH]-[PH4- -]	2	3	3	3	3	14
[CH2+]-[PH4- -]	2	3	5	3	1	14
[CH++]-[PH4- -]	3	3	5	3	0	14
[CH3]-[PH2]	3	0	5	3	3	14
[CH2+]-[PH2]	0	3	5	3	3	14
[CH]-[SH3- -]	2	3	3	3	3	14
[CH2+]-[SH3- -]	2	3	5	3	1	14
[CH++]-[SH3- -]	3	3	5	3	0	14
[CH3]-[SH]	3	0	5	3	3	14

[CH2+]-[SH]	0	3	5	3	3	14
[NH3-]-[CH2+]	3	0	5	3	3	14
[NH+]-[CH3]	0	3	5	3	3	14
[NH2]-[CH2+]	0	3	5	3	3	14
[N]-[NH4- -]	2	3	3	3	3	14
[NH+]-[NH4- -]	2	3	5	3	1	14
[N++]-[NH4- -]	3	3	5	3	0	14
[NH3-]-[NH+]	3	0	5	3	3	14
[NH+]-[NH2]	0	3	5	3	3	14
[NH2]-[NH+]	0	3	5	3	3	14
[N]-[OH3- -]	2	3	3	3	3	14
[NH+]-[OH3- -]	2	3	5	3	1	14
[N++]-[OH3- -]	3	3	5	3	0	14
[NH3-]-[O+]	3	0	5	3	3	14
[NH+]-[OH]	0	3	5	3	3	14
[NH2]-[O+]	0	3	5	3	3	14
[N]-[PH4- -]	2	3	3	3	3	14
[NH+]-[PH4- -]	2	3	5	3	1	14
[N++]-[PH4- -]	3	3	5	3	0	14
[NH3-]-[PH+]	3	0	5	3	3	14
[NH+]-[PH2]	0	3	5	3	3	14
[NH2]-[PH+]	0	3	5	3	3	14
[N]-[SH3- -]	2	3	3	3	3	14
[NH+]-[SH3- -]	2	3	5	3	1	14
[N++]-[SH3- -]	3	3	5	3	0	14
[NH3-]-[S+]	3	0	5	3	3	14
[NH+]-[SH]	0	3	5	3	3	14
[NH2]-[S+]	0	3	5	3	3	14
[O+]-[CH3]	0	3	5	3	3	14
[OH]-[CH2+]	0	3	5	3	3	14
[O+]-[NH4- -]	2	3	5	3	1	14
[O+]-[NH2]	0	3	5	3	3	14
[OH]-[NH+]	0	3	5	3	3	14
[O+]-[OH3- -]	2	3	5	3	1	14
[O+]-[OH]	0	3	5	3	3	14
[OH]-[O+]	0	3	5	3	3	14
[O+]-[PH4- -]	2	3	5	3	1	14
[O+]-[PH2]	0	3	5	3	3	14
[OH]-[PH+]	0	3	5	3	3	14
[O+]-[SH3- -]	2	3	5	3	1	14
[O+]-[SH]	0	3	5	3	3	14
[OH]-[S+]	0	3	5	3	3	14
[PH3-]-[CH2+]	3	0	5	3	3	14
[PH+]-[CH3]	0	3	5	3	3	14
[PH2]-[CH2+]	0	3	5	3	3	14
[P]-[NH4- -]	2	3	3	3	3	14
[PH+]-[NH4- -]	2	3	5	3	1	14

[P++]-[NH4- -]	3	3	5	3	0	14
[PH3-]-[NH+]	3	0	5	3	3	14
[PH+]-[NH2]	0	3	5	3	3	14
[PH2]-[NH+]	0	3	5	3	3	14
[P]-[OH3- -]	2	3	3	3	3	14
[PH+]-[OH3- -]	2	3	5	3	1	14
[P++]-[OH3- -]	3	3	5	3	0	14
[PH3-]-[O+]	3	0	5	3	3	14
[PH+]-[OH]	0	3	5	3	3	14
[PH2]-[O+]	0	3	5	3	3	14
[P]-[PH4- -]	2	3	3	3	3	14
[PH+]-[PH4- -]	2	3	5	3	1	14
[P++]-[PH4- -]	3	3	5	3	0	14
[PH3-]-[PH+]	3	0	5	3	3	14
[PH+]-[PH2]	0	3	5	3	3	14
[PH2]-[PH+]	0	3	5	3	3	14
[P]-[SH3- -]	2	3	3	3	3	14
[PH+]-[SH3- -]	2	3	5	3	1	14
[P++]-[SH3- -]	3	3	5	3	0	14
[PH3-]-[S+]	3	0	5	3	3	14
[PH+]-[SH]	0	3	5	3	3	14
[PH2]-[S+]	0	3	5	3	3	14
[S+]-[CH3]	0	3	5	3	3	14
[SH]-[CH2+]	0	3	5	3	3	14
[S+]-[NH4- -]	2	3	5	3	1	14
[S+]-[NH2]	0	3	5	3	3	14
[SH]-[NH+]	0	3	5	3	3	14
[S+]-[OH3- -]	2	3	5	3	1	14
[S+]-[OH]	0	3	5	3	3	14
[SH]-[O+]	0	3	5	3	3	14
[S+]-[PH4- -]	2	3	5	3	1	14
[S+]-[PH2]	0	3	5	3	3	14
[SH]-[PH+]	0	3	5	3	3	14
[S+]-[SH3- -]	2	3	5	3	1	14
[S+]-[SH]	0	3	5	3	3	14
[SH]-[S+]	0	3	5	3	3	14

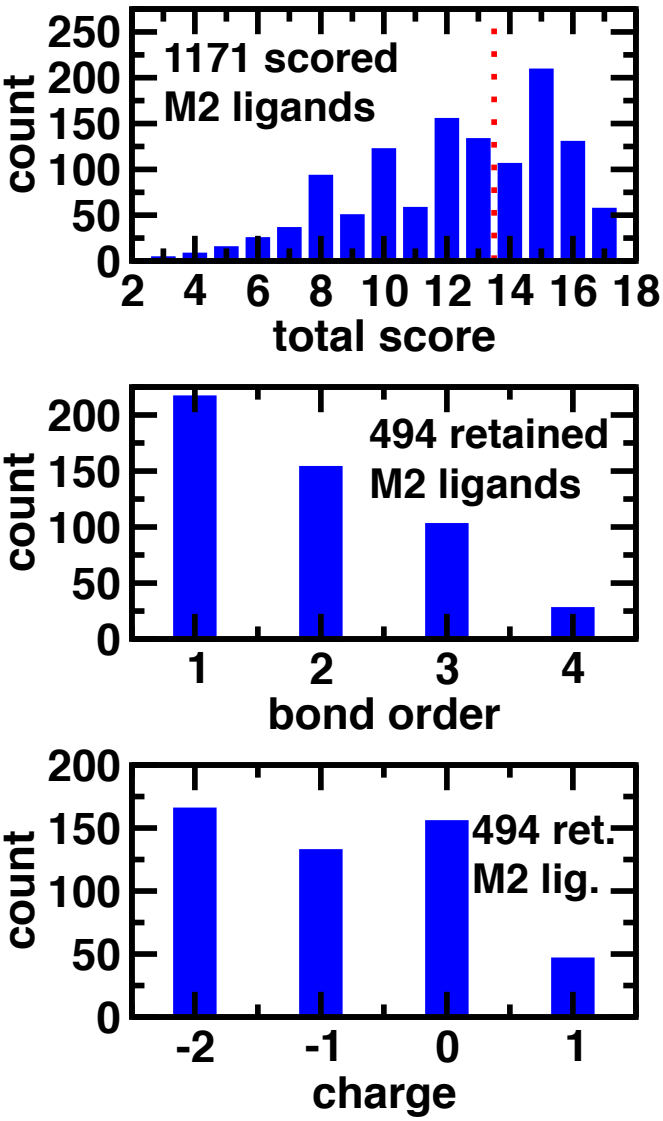

Figure S2: Top: Histogram of total score for 1,171 scored M2 ligands with the cutoff for retained ligands (i.e., $s_{\text{tot}} \geq 14$) indicated as a red vertical dotted line. Middle: heavy atom bond order distribution of the 494 retained M2 ligands. Bottom: ligand charge distribution of the 494 retained M2 ligands. The high penalty on positive charge ligands means relatively few ligands carry a net positive charge.

Table S4: Component and total scores along with SMILES string representations for 47 (of 5,625 theoretical) retained B4 ligands. The ligands selected have an $s_{\text{tot}} > 13$. The theoretical maximum s_{tot} is 16, and the variations in overall score can be due to variations in any of five component scores: s_{pol}, the score for how much charge separation there is in the two heavy atoms 1 and 2 in either of the symmetric units of the B4 ligand, s_{bond}, the score for the bond order between the two fragments, s_{charge}, the score for overall ligand charge, as evaluated over half of the ligand, s_{octet}, the score for violating the octet rule, and s_{sterics}, the score for how much steric repulsion occurs at either symmetric copy of the metal-coordinating heavy atom. The only B4 scored ligand present in the spectrochemical series is ethylenediamine ($C_2H_8N_2$).

SMILES	s_{pol}	s_{charge}	s_{sterics}	s_{bond}	s_{octet}	s_{tot}
[NH2]-[CH]=[CH]-[NH2]	2	3	3	3	5	16
[NH2]-[CH2]-[CH2]-[NH2]	2	3	3	3	5	16
[NH2]-[N]=[N]-[NH2]	2	3	3	3	5	16
[NH2]-[NH]-[NH]-[NH2]	2	3	3	3	5	16
[NH2]-[O]-[O]-[NH2]	2	3	3	3	5	16
[NH2]-[P]=[P]-[NH2]	2	3	3	3	5	16
[NH2]-[PH]-[PH]-[NH2]	2	3	3	3	5	16
[NH2]-[S]-[S]-[NH2]	2	3	3	3	5	16
[OH]-[CH]=[CH]-[OH]	2	3	3	3	5	16
[OH]-[CH2]-[CH2]-[OH]	2	3	3	3	5	16
[OH]-[N]=[N]-[OH]	2	3	3	3	5	16
[OH]-[NH]-[NH]-[OH]	2	3	3	3	5	16
[OH]-[O]-[O]-[OH]	2	3	3	3	5	16
[OH]-[P]=[P]-[OH]	2	3	3	3	5	16
[OH]-[PH]-[PH]-[OH]	2	3	3	3	5	16
[OH]-[S]-[S]-[OH]	2	3	3	3	5	16
[PH2]-[CH]=[CH]-[PH2]	2	3	3	3	5	16
[PH2]-[CH2]-[CH2]-[PH2]	2	3	3	3	5	16
[PH2]-[N]=[N]-[PH2]	2	3	3	3	5	16
[PH2]-[NH]-[NH]-[PH2]	2	3	3	3	5	16
[PH2]-[O]-[O]-[PH2]	2	3	3	3	5	16
[PH2]-[P]=[P]-[PH2]	2	3	3	3	5	16
[PH2]-[PH]-[PH]-[PH2]	2	3	3	3	5	16
[PH2]-[S]-[S]-[PH2]	2	3	3	3	5	16
[SH]-[CH]=[CH]-[SH]	2	3	3	3	5	16
[SH]-[CH2]-[CH2]-[SH]	2	3	3	3	5	16
[SH]-[N]=[N]-[SH]	2	3	3	3	5	16
[SH]-[NH]-[NH]-[SH]	2	3	3	3	5	16
[SH]-[O]-[O]-[SH]	2	3	3	3	5	16
[SH]-[P]=[P]-[SH]	2	3	3	3	5	16
[SH]-[PH]-[PH]-[SH]	2	3	3	3	5	16
[SH]-[S]-[S]-[SH]	2	3	3	3	5	16
[CH2]=[CH]-[CH]=[CH2]	2	3	3	1	5	14
[CH2]=[N]-[N]=[CH2]	2	3	3	1	5	14

[CH2]=[P]-[P]=[CH2]	2	3	3	1	5	14
[NH]=[CH]-[CH]=[NH]	2	3	3	1	5	14
[NH]=[N]-[N]=[NH]	2	3	3	1	5	14
[NH]=[P]-[P]=[NH]	2	3	3	1	5	14
[O]=[CH]-[CH]=[O]	2	3	3	1	5	14
[O]=[N]-[N]=[O]	2	3	3	1	5	14
[O]=[P]-[P]=[O]	2	3	3	1	5	14
[PH]=[CH]-[CH]=[PH]	2	3	3	1	5	14
[PH]=[N]-[N]=[PH]	2	3	3	1	5	14
[PH]=[P]-[P]=[PH]	2	3	3	1	5	14
[S]=[CH]-[CH]=[S]	2	3	3	1	5	14
[S]=[N]-[N]=[S]	2	3	3	1	5	14
[S]=[P]-[P]=[S]	2	3	3	1	5	14

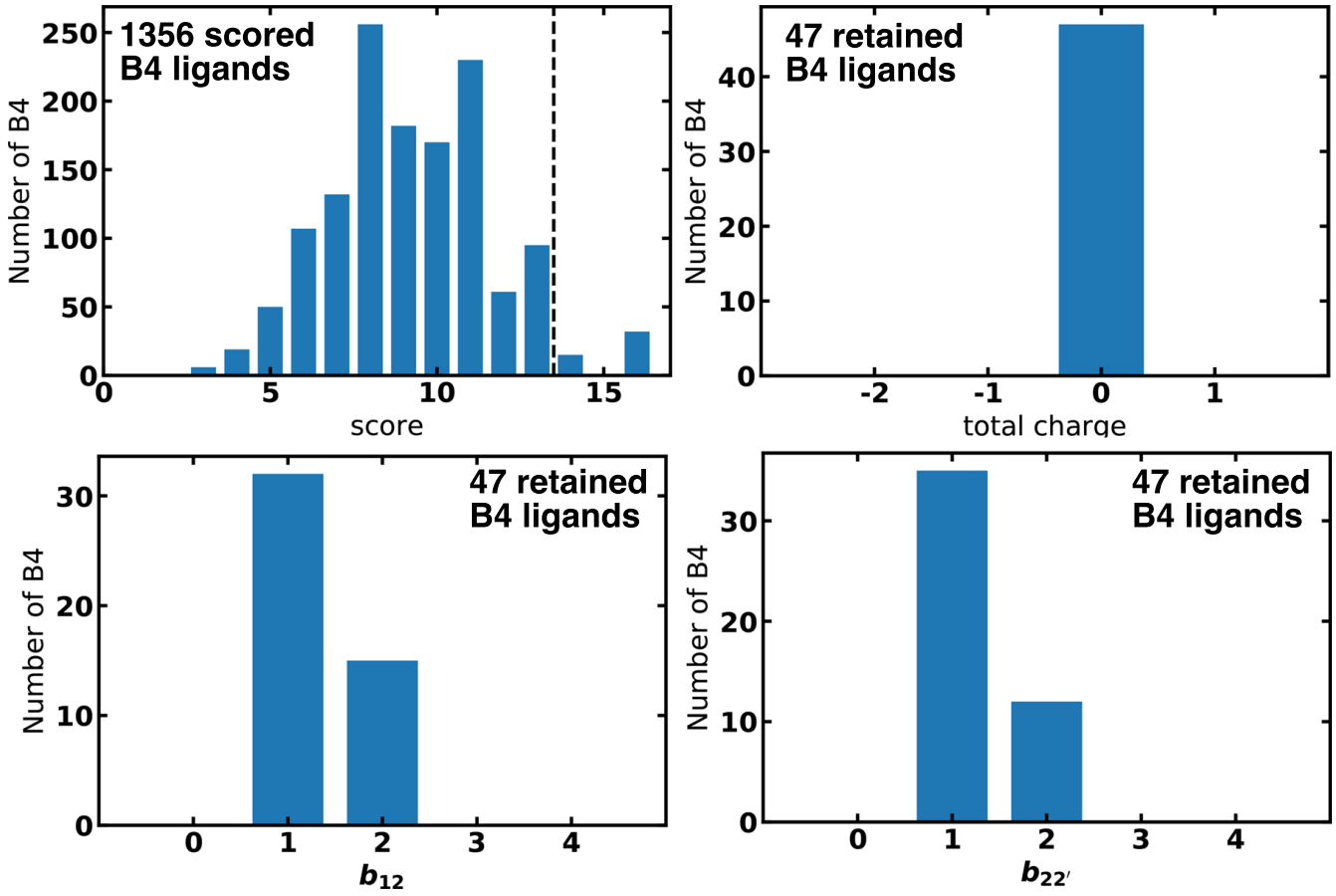

Figure S3: Top left: The distribution of the 1,356 scored B4 ligands. The vertical dashed line indicates the retained score threshold $s_{\text{tot}} \geq 14$. Top right: Total charge distribution of the 47 selected B4 ligands with a score $s_{tot} \geq 14$. The strict score threshold (i.e., no more than 2 less than the maximum value) leads to retention of only neutral B4 ligands. Bottom: Bond order (b_{12}, left, $b_{22'}$, right) distribution of the 47 scored B4 ligands with a score $s_{tot} \geq 14$, that result from only a combination of single and double bonds, which are necessary for bidentate-chelating geometries.

Algorithm S2: Determination of charge and bond order for B4 ligands for a given total charge, c_{tot}. The individual atom 1 and 2 charges, c_1 and c_2 are varied while satisfying the relationship between the valence, v, and the number of lone pairs l, standard number of valence electrons ve, number of hydrogen atoms h. The bond order, b is then set differently from how the individual M2 ligands were set by redistributing electrons into the 2-2′ bond.

Require: $c_{\text{tot}} \in [-4, 4]$
  for $c_1 = -2$ **to** 2 **do**
    for $c_2 = -2$ **to** 2 **do**
      if $c_{\text{tot}} == c_1 + c_2$ **then**
        $v_1 \leftarrow ve_1 - c_1 - 2 \cdot l_1 - h_1$
        $v_2 \leftarrow ve_2 - c_2 - 2 \cdot l_2 - h_2$
        if $v_1 > 0$ and $v_2 > 1$ **then**
          for $b_{12} = 1$ **to** 3 **do**
            for $b_{22'} = 1$ **to** 3 **do**
              $\delta = |v_1 - b_{12}| + |v_2 - b_{12} - b_{22'}|$
              $\alpha = |c_1| + |c_2|$
              $\min\{\delta\}$
              if $\exists!\delta$ **then**
                return $b, c_1, c_2, b_{12}, b_{22'}$
              else
                return $b, c_1, c_2, b_{12}, b_{22'}$ at smallest α
              end if
            end for
          end for
        end if
      else
        $b_{12} \leftarrow 0$
        $b_{22'} \leftarrow 0$
        $c_1 \leftarrow c_{\text{tot}}$
        $c_2 \leftarrow 0$
        return $b, c_1, c_2, b_{12}, b_{22'}$
      end if
    end for
  end for

Table S5: From 2,577 M1, M2, and B4 enumerated ligands, 71 ligands occur at least once in ChEMBL, DiRef, or GDB-9. Of these 71, 18 score below the retention threshold for their respective ligand grouping (3 B4, 13 M2, and 2 M1). Of these 71 ligands, 55 correspond to the M2 type, and 46 of those form pairs of identical ligands distinguishable only by their metal-coordinating atom. For most M2 ligands, this leads to two entries below that are identical except for the order of the atoms in the SMILES string. However, for CH_3NH_2 and CH_3OH, the score differs depending on the metal-coordinating atom due to differences in sterics. Overall, counting only unique chemical compositions, 36 of 48 ligands (75%) are above threshold. Including duplicate molecules that are unique ligands, 53 of 71 ligands (75%) are above threshold. All scores listed are the scores adjusted to span a 0 to 100 range. Presence in each set is indicated by a ✓or absence by --. Ligands are grouped first by whether they were above threshold (yes or no, last column), then by their classification (B4, M2, M1), and finally sorted by score within these groupings, excluding only duplicate molecules distinguished only by their coordinating atoms with distinct scores. These results were extracted from the downloadable databases (i.e., for GDB-9 and ChEMBL) or through manual search (i.e., DiRef M2 ligands).

SMILES	score	set	DiRef	GDB-9	ChEMBL	above threshold
[NH2]-[CH2]-[CH2]-[NH2]	100	B4	--	--	✓	yes
[OH]-[CH2]-[CH2]-[OH]	100	B4	--	✓	✓	yes
[O]=[CH]-[CH]=[O]	86	B4	--	✓	✓	yes
[CH2]=[CH]-[CH]=[CH2]	86	B4	--	--	✓	yes
[NH2]-[CH3]	100	M2	--	--	✓	yes
[CH3]-[NH2]	79	M2	--	--	✓	yes
[NH2]-[NH2]	100	M2	--	--	✓	yes
[OH]-[CH3]	100	M2	--	✓	✓	yes
[CH3]-[OH]	79	M2	--	✓	✓	yes
[NH2]-[OH]	100	M2	--	--	✓	yes
[OH]-[NH2]	100	M2	--	--	✓	yes
[OH]-[OH]	100	M2	--	--	✓	yes
[C]#4[C]	93	M2	✓	--	--	yes
[CH]#[CH]	93	M2	--	✓	✓	yes
[CH2]=[CH2]	93	M2	--	✓	✓	yes
[CH]#[N]	93	M2	--	✓	✓	yes
[N]#[CH]	93	M2	--	✓	✓	yes
[CH2]=[O]	93	M2	--	✓	✓	yes
[O]=[CH2]	93	M2	--	✓	✓	yes
[N]#[N]	93	M2	✓	--	✓	yes
[N]#[P]	93	M2	✓	--	--	yes
[P]#[N]	93	M2	✓	--	--	yes
[O]=[O]	93	M2	✓	--	✓	yes
[O]=[S]	93	M2	✓	--	--	yes
[S]=[O]	93	M2	✓	--	--	yes
[P]#[P]	93	M2	✓	--	--	yes
[S]=[S]	93	M2	✓	--	--	yes
[C+]#[O-]	93	M2	✓	--	--	yes

[O-]#[C+]	93	M2	✓	--	--	yes
[C+]#[S-]	93	M2	✓	--	--	yes
[S-]#[C+]	93	M2	✓	--	--	yes
[N-]#4[C]	86	M2	✓	--	--	yes
[C]#4[N-]	86	M2	✓	--	--	yes
[O-]#[N]	86	M2	✓	--	--	yes
[N]#[O-]	86	M2	✓	--	--	yes
[O-]#[P]	86	M2	✓	--	--	yes
[P]#[O-]	86	M2	✓	--	--	yes
[P-]#4[C]	86	M2	✓	--	--	yes
[C]#4[P-]	86	M2	✓	--	--	yes
[S-]#[N]	86	M2	✓	--	--	yes
[N]#[S-]	86	M2	✓	--	--	yes
[S-]#[P]	86	M2	✓	--	--	yes
[P]#[S-]	86	M2	✓	--	--	yes
[O-]#[O-]	86	M2	✓	--	--	yes
[S-]#[S-]	86	M2	✓	--	--	yes
[CH3]-[CH3]	79	M2	--	✓	✓	yes
[NH3]	100	M1	--	✓	✓	yes
[OH-]	100	M1	--	--	✓	yes
[OH2]	100	M1	--	✓	✓	yes
[PH3]	100	M1	--	--	✓	yes
[SH-]	100	M1	--	--	✓	yes
[SH2]	100	M1	--	--	✓	yes
[S]	71	M1	--	--	✓	yes
[CH3]-[CH2]-[CH2]-[CH3]	79	B4	--	✓	✓	no
[CH3]-[NH]-[NH]-[CH3]	79	B4	--	--	✓	no
[CH3]-[O]-[O]-[CH3]	79	B4	--	--	✓	no
[C]#4[C- -]	71	M2	✓	--	--	no
[C+]#[N]	71	M2	✓	--	--	no
[N]#[C+]	71	M2	✓	--	--	no
[C+]#[P]	71	M2	✓	--	--	no
[P]#[C+]	71	M2	✓	--	--	no
[N+]=[O]	71	M2	✓	--	--	no
[O]=[N+]	71	M2	✓	--	--	no
[N+]=[S]	71	M2	✓	--	--	no
[S]=[N+]	71	M2	✓	--	--	no
[P+]=[O]	71	M2	✓	--	--	no
[O]=[P+]	71	M2	✓	--	--	no
[P+]=[S]	71	M2	✓	--	--	no
[S]=[P+]	71	M2	✓	--	--	no
[CH4]	57	M1	--	✓	✓	no
[C]	43	M1	--	--	✓	no

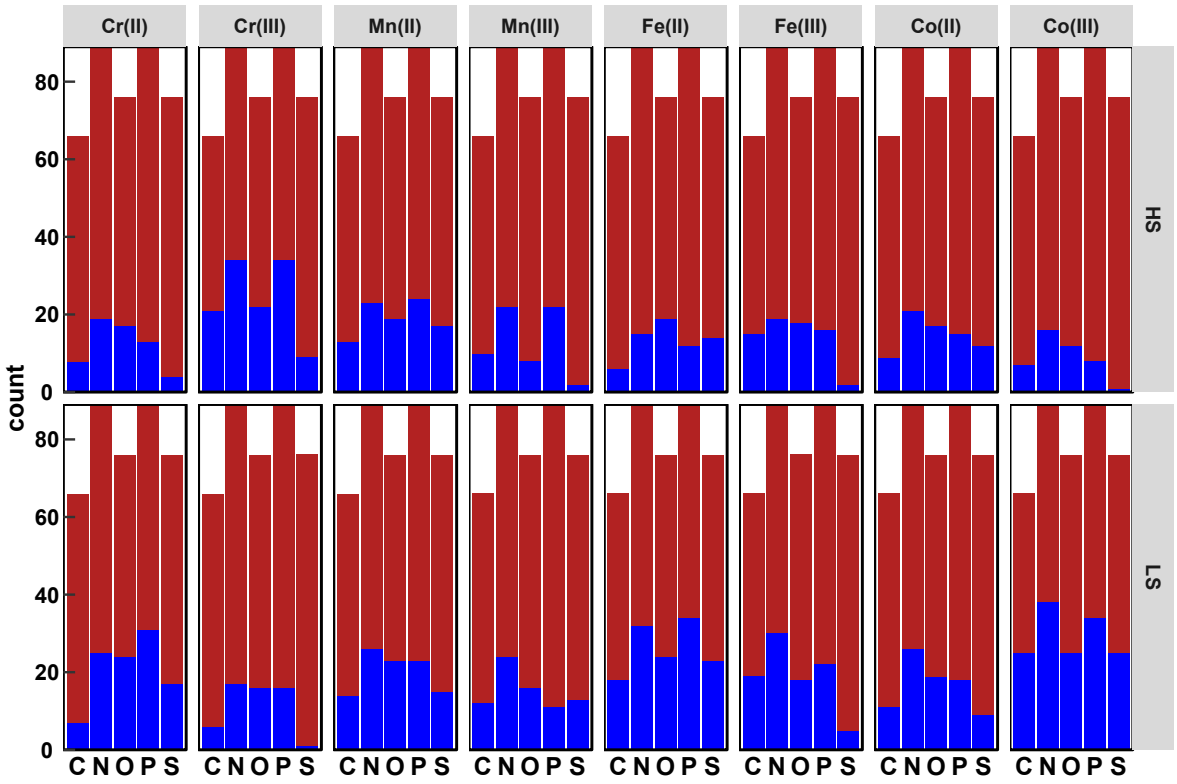

Figure S4: Success/failure stacked bar plot for DFT optimization of complexes with the potential ligands divided by metal, oxidation and spin states, and the atomic identity of the atom coordinating to the metal center. The successful calculations are shown in blue and the remaining unsuccessful calculations are shown in brick red. The top plot corresponds to the high-spin state and the bottom corresponds to the low-spin state.

Text S1: Description of RAC features and RAC-155

Revised Autocorrelations (RACs) are a set of molecular fingerprints based on the the molecular graph that were developed[2] to represent transition metal complexes. RACs are derived by applying a function, either the product or difference, to atomic properties of atoms separated by a fixed number of bonds, d:

$$P_{\mathrm{d}} = \sum_i \sum_j P_i P_j \delta(d_{ij})$$

Here, d_{ij} is the number of bonds between atoms i and j and $\delta(d_{ij}) = 1 \iff d_{ij} = d$ and is otherwise zero. We consider the following atomic properties, P: nuclear charge (Z), Pauling electronegativity (χ), covalent radius (S), the number of bonds the atom has (T), and a counter that is 1 for all atoms (I). The results of these functions are summed up over a scope, which could be the whole molecule or a subgraph. Original autocorrelations[1] were based on products only and are summed over the entire molecular graph. We supplement these with these sums that are restricted to a single atomic atomic site, being either the metal or ligand connecting atom, and by restricting the scope to the ligand only. These additional features enrich the representation of the local metal center, which is known to important for transition metal chemistry. We consider a maximum depth (number of bonds) of 3, which we previously identified was the optimal empirically. RACs are related to graph convolutions but have fixed kernels that are not updated during training time and consider information exchange across 1–3 bonds simultaneously. RAC-155[2] is a set of 155 RACs that were previously shown to have good performance in predicting spin splitting energy.

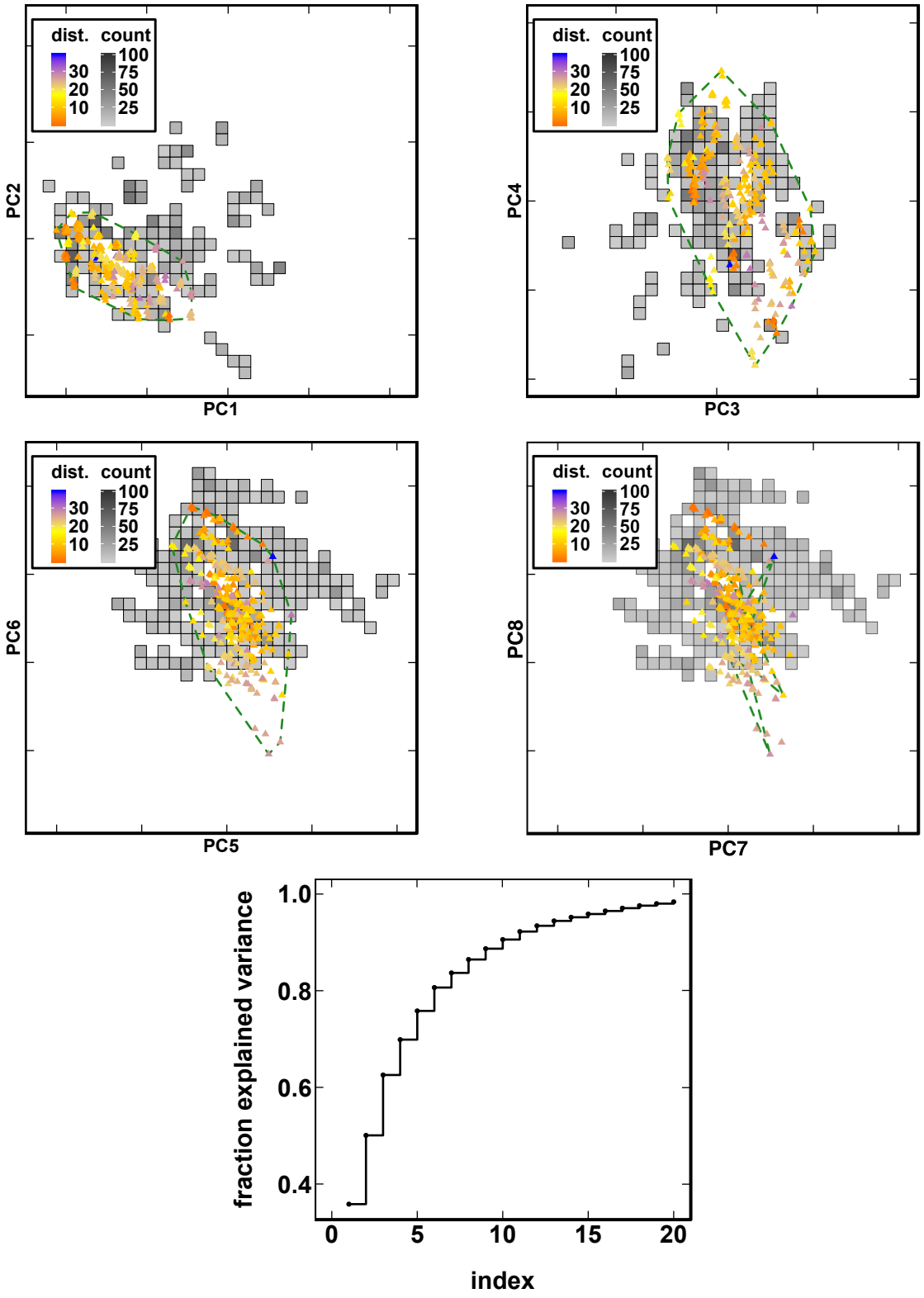

Figure S5: Principal component (PC) histograms based on RAC-155 representation for 1901 preexisting data points, shown as gray squares shaded by population. The new data obtained in this work plotted on as triangles, colored by average distance to 10-nearest preexisting points. The convex hull of the new data is shown with a green dashed line. The different plots show different pairs of PCs. The final frame shows the percentage of explained variance in the first 20 PCs, with PCs 1-8 accounting for 89% of total variance. The PCs are computed based on preexisting data only.

Table S6: Hyperparameters and topology for inorganic spin splitting ANN showing original parameters used for training and then parameters for retraining on combined data.

parameter	orignal training	retraining
layer 1 size	200	
layer 2 size	200	
layer 3 size	200	
activation function	relu	
learning rate	0.00163	0.001
optimizer	sgd	adam
momentum	0.998	$\beta_1 = 0.9$, $\beta_2 = 0.999$
decay	0.0015719	0
Nesterov acceleration	yes	NA
dropout (all hidden)	0.0825	
batch size	128	100
epochs	2000	2000
L^2 regularization	7.101148E-14	
semibatch normalization	yes	
early stopping	none	

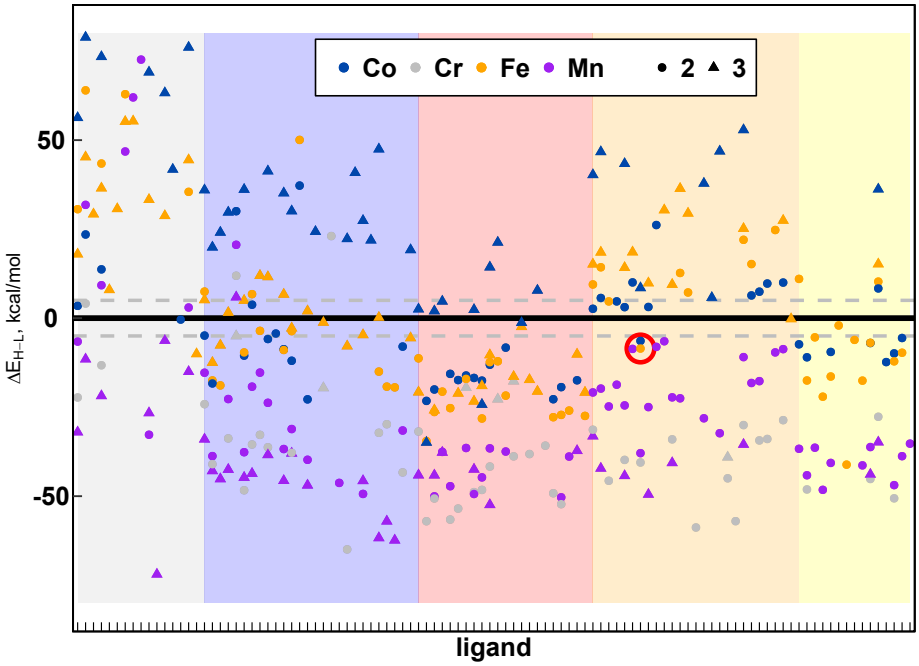

Figure S6: Spin splitting energies for 343 new DFT calculations in kcal/mol, plotted for each ligand colored by connecting atom. The region with 5 kcal/mol of zero (uncertain spin assignment) is indicated by gray dashed lines. Colors for atom type are gray = carbon, blue = nitrogen, orange = phosphorus, sulfur = yellow. An unusual high spin iron(II) phosphorus complex with [P]#[P] ligands is circled in red

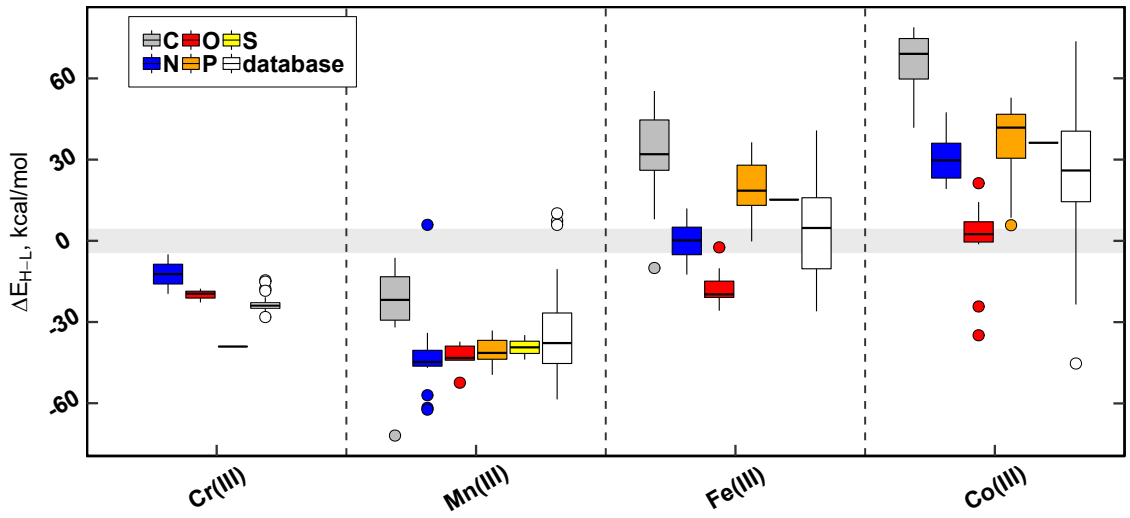

Figure S7: Box-and-whisker diagram of spin splitting energies for new DFT calculations in kcal/mol, colored by the connecting atom identity, plotted by metals in oxidation state 3+. Ranges for a previous database of DFT calculations are shown in white. Each box shows the interquartile range (IQR) of the data with the median indicated by a horizontal line. Whiskers indicate 1.5 the IQR and data outside this range are plotted with circles. The region within 5 kcal/mol of zero (uncertain spin assignment) is indicated by a shaded gray rectangle. Colors for atom type are gray = carbon, blue = nitrogen, orange = phosphorus, sulfur = yellow. A similar plot for oxidation state II is provided in the Main text, Figure 7.

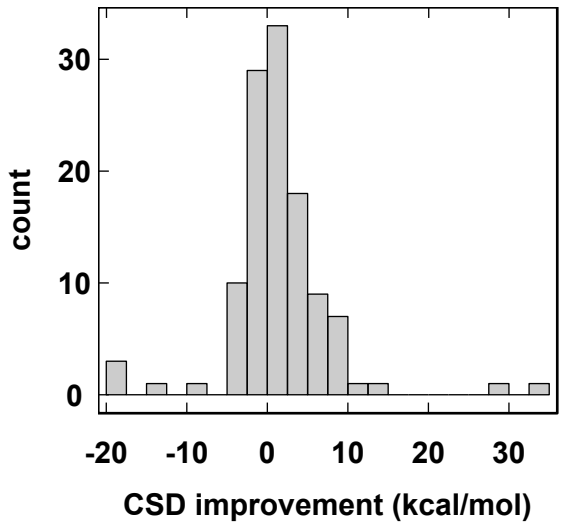

Figure S8: Histogram of changes in CSD prediction errors after adding new data.

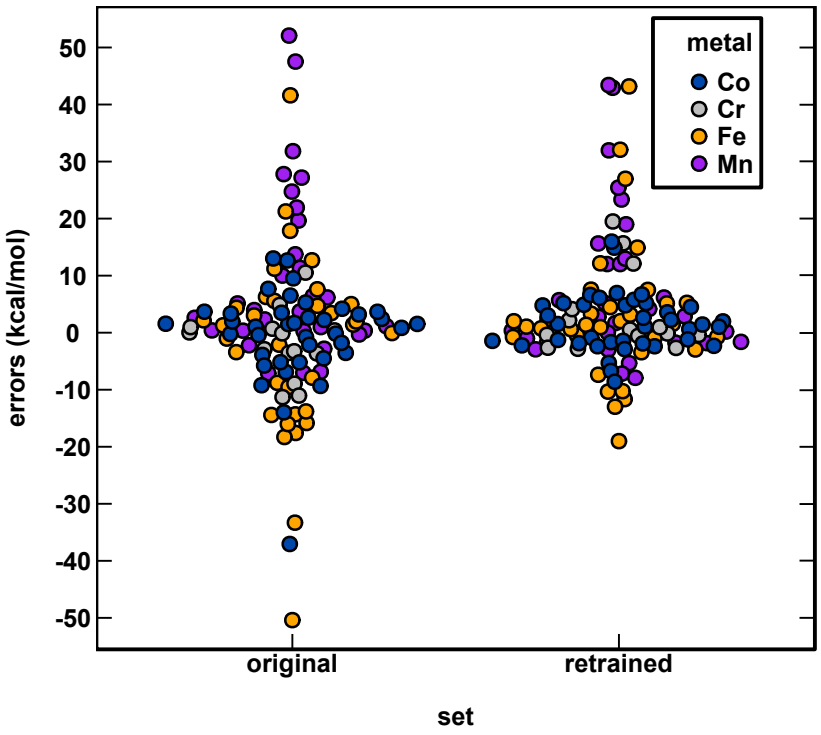

Figure S9: Swarm plot of ANN errors on CSD data before and after adding the new 343 points on out-of-distribution CSD test case. Points are colored by metal.

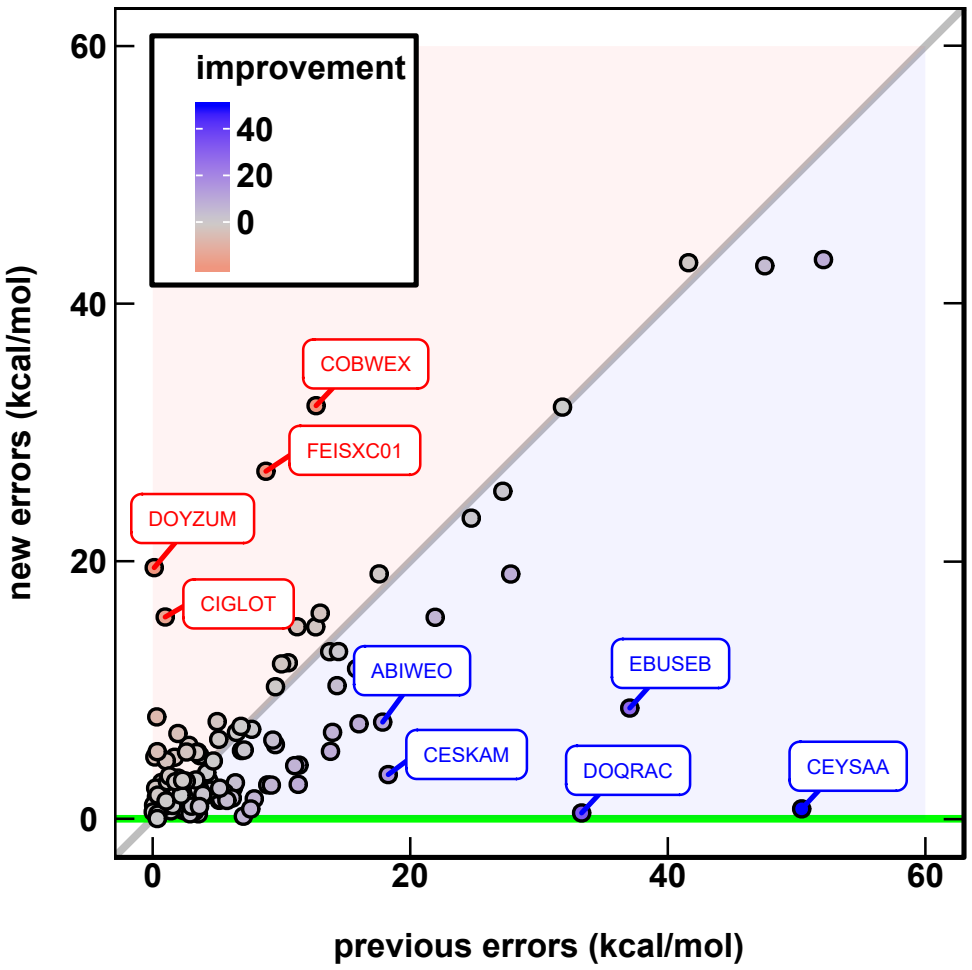

Figure S10: Parity plot of old (before adding new data) and new ANN predictions for the out-of-distribution CSD test set. Points in the blue region represent improvement (72 points, mean improvement ~ 4.97 kcal/mol), while points in the red region represent degraded performance (44 points, mean degradation ~ -3.36 kcal/mol). Points with $\geq |10|$ kcal/mol improvement or worsening after model retraining are labeled with CSD access codes. The gray line indicates parity, while the green line represents zero case (no error).

References

1. P Broto, G Moreau, and C Vandycke. Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *Eur. J. Med. Chem.*, **1984**, 19 (1), 71–78.

2. JP Janet and HJ Kulik. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A*, **2017**, 121 (46), 8939–8954.