Supporting Information for

Insights into the deviation from piecewise linearity in transition metal complexes from supervised machine learning models

Yael Cytter [1], Chenru Duan[1,2], Aditya Nandy[1,2], and Heather J. Kulik[1,*]

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*Corresponding author email: hjkulik@mit.edu

Contents

Table S1. Metal centers, oxidation, and spin states in dataset. Spin states are described by spin multiplicity, defined as 2*S*+1 where *S* is the total spin angular momentum.

metal	Ox	Spin multiplicity
Co	+2	2
		4
	+3	1
		5
Cr	+2	1
		5
	+3	2
		4
Fe	+2	1
		5
	+3	2
		6
Mn	+2	2
		6
	+3	1
		5

Table S2. Design space ligands with the net charge (charge), ligand denticity (dent), number of atoms (natoms), type of atom connected to the metal (CA) and SMILES string.

name	charge	dent.	natoms	CA	SMILES
acac	-1	2	14	O	O=C(C)C=[CH-](O)C
aceticacidbipyridine	0	2	32	N	n1ccc(cc1c1nccc(c1)CC(=O)O)CC(=O)O
acetonitrile	0	1	6	N	N#CC
ammonia	0	1	4	N	N
benzisc	0	1	16	C	[C-]#[N+]Cc1ccccc
bipy	0	2	20	N	n1ccccc1c1ncccc1
carbonyl	0	1	2	C	C#[O]
cyanide	-1	1	2	C	[C-]#N
cyanoaceticporphyrin	-2	4	52	N	N1=C2C=[CH2-][CH-]1=C(c1[nH]c(cc1)/C=C/1\N=C(/C(=c/3\[nH]/c(=C\2)/cc3)/C=C(/C(=O)O)\C#N)C=C1)/C=C(/C(=O)O)\C#N
cyanopyridine	0	1	12	N	C1(=CCNC=C1)C#N
en	0	2	12	N	NCCN
formaldehyde	0	1	4	O	C=O
furan	0	1	9	O	o1cccc1
isothiocyanate	-1	1	3	N	[N-]=C=S
mebipyridine	0	2	26	N	n1ccc(cc1c1nccc(c1)C)C
mec	-2	2	15	O	[O-]c1c(cc(cc1)C)[O-]
methylamine	0	1	7	N	CN
misc	0	1	6	C	[C-]#[N+]C
ox	-2	2	6	O	C(=O)(C(=O)[O-])[O-]
phen	0	2	22	N	c1cc2ccc3cccnc3c2nc1
phenisc	0	1	13	C	[C-]#[N+]c1ccccc1
pisc	0	1	25	C	[C-]#[N+]c1ccc(C(C)(C)C)cc1
porphyrin	-2	4	36	N	N1=C2C=[CH2-][CH-]1=Cc1[nH]c(cc1)/C=C/1\N=C(/C=c/3\[nH]/c(=C\2)/cc3)C=C1
pph3	0	1	34	P	c1c(P(c2ccccc2)c2ccccc2)cccc1
py	0	1	11	N	C1=CCNC=C1
tbuc	-2	2	24	O	[O-]c1c(cc(C(C)(C)C)cc1)[O-]
thiopyridine	0	1	12	N	C1(=CCNC=C1)S
water	0	1	3	O	O
fluoride ion	-1	1	1	F	[F-]
iodide ion	-1	1	1	I	[I-]
[O-][O-]	-2	1	2	O	[O-][O-]
hydroxide	-1	1	2	O	[OH-]
phosphine	0	1	4	P	[PH3]
sulfide	-2	1	1	S	[S--]
hydrogen sulfide	0	1	3	S	[SH2]
cyanate	-1	1	3	N	N#C[O-]

Table S3. Summary of 23 DFAs used in this work, as motivated in Duan *et al.*[1], including their rung on "Jacob's ladder" of DFT, HF exchange fraction, LRC range-separation parameter ($bohr^{-1}$), MP2 correlation fraction, and whether empirical (i.e., D3) dispersion correction is included.

DFA	type	exchange type	HF exchange percentage	LRC RS parameter ($bohr^{-1}$)	MP2 correlation	D3 dispersion
BP86[2,3]	GGA	GGA	--	--	--	no
BLYP[4,5]	GGA	GGA	--	--	--	no
PBE[6]	GGA	GGA	--	--	--	no
TPSS[7]	meta-GGA	meta-GGA	--	--	--	no
SCAN[8]	meta-GGA	meta-GGA	--	--	--	no
M06-L[9]	meta-GGA	meta-GGA	--	--	--	no
MN15-L[10]	meta-GGA	meta-GGA	--	--	--	no
B3LYP[11-13]	GGA hybrid	GGA	0.200	--	--	no
B3P86[2,11]	GGA hybrid	GGA	0.200	--	--	no
B3PW91[11,14]	GGA hybrid	GGA	0.200	--	--	no
PBE0[15]	GGA hybrid	GGA	0.250	--	--	no
ωB97X[16]	RS hybrid	GGA	0.158	0.300	--	no
LRC-ωPBEh[17]	RS hybrid	GGA	0.200	0.200	--	no
TPSSh[7]	meta-GGA hybrid	meta-GGA	0.100	--	--	no
SCAN0[18]	meta-GGA hybrid	meta-GGA	0.250	--	--	no
M06[19]	meta-GGA hybrid	meta-GGA	0.270	--	--	no
M06-2X[19]	meta-GGA hybrid	meta-GGA	0.540	--	--	no
MN15[20]	meta-GGA hybrid	meta-GGA	0.440	--	--	no
B2GP-PLYP[21]	double hybrid	GGA	0.650	--	0.360	no
PBE0-DH[22]	double hybrid	GGA	0.500	--	0.125	no
DSD-BLYP-D3BJ[23]	double hybrid	GGA	0.710	--	1.000	yes
DSD-PBEB95-D3BJ[23]	double hybrid	GGA	0.660	--	1.000	yes
DSD-PBEP6-D3BJ[23]	double hybrid	GGA	0.690	--	1.000	yes

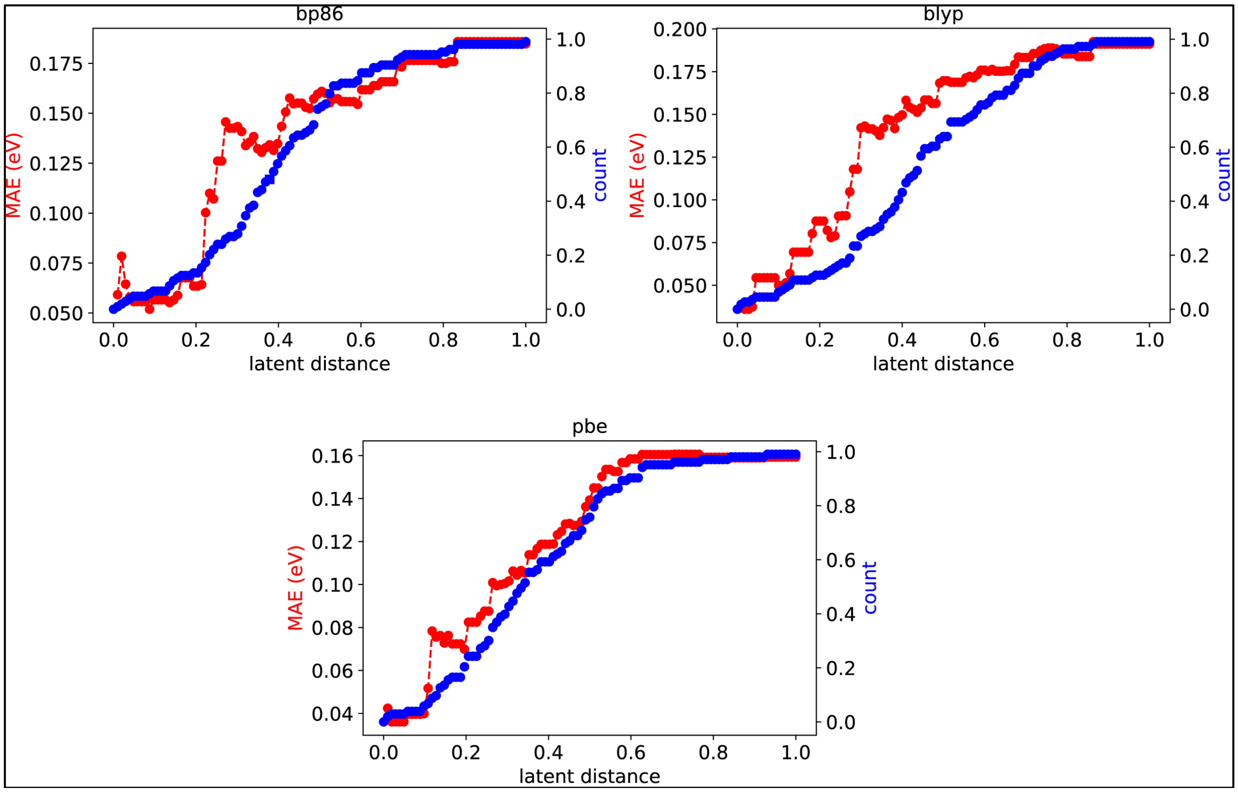

Figure S1. Mean absolute error (red markers) and fraction of included data (count, blue markers) as a function of latent distance for three average curvature models, each corresponding to a different GGA functional.

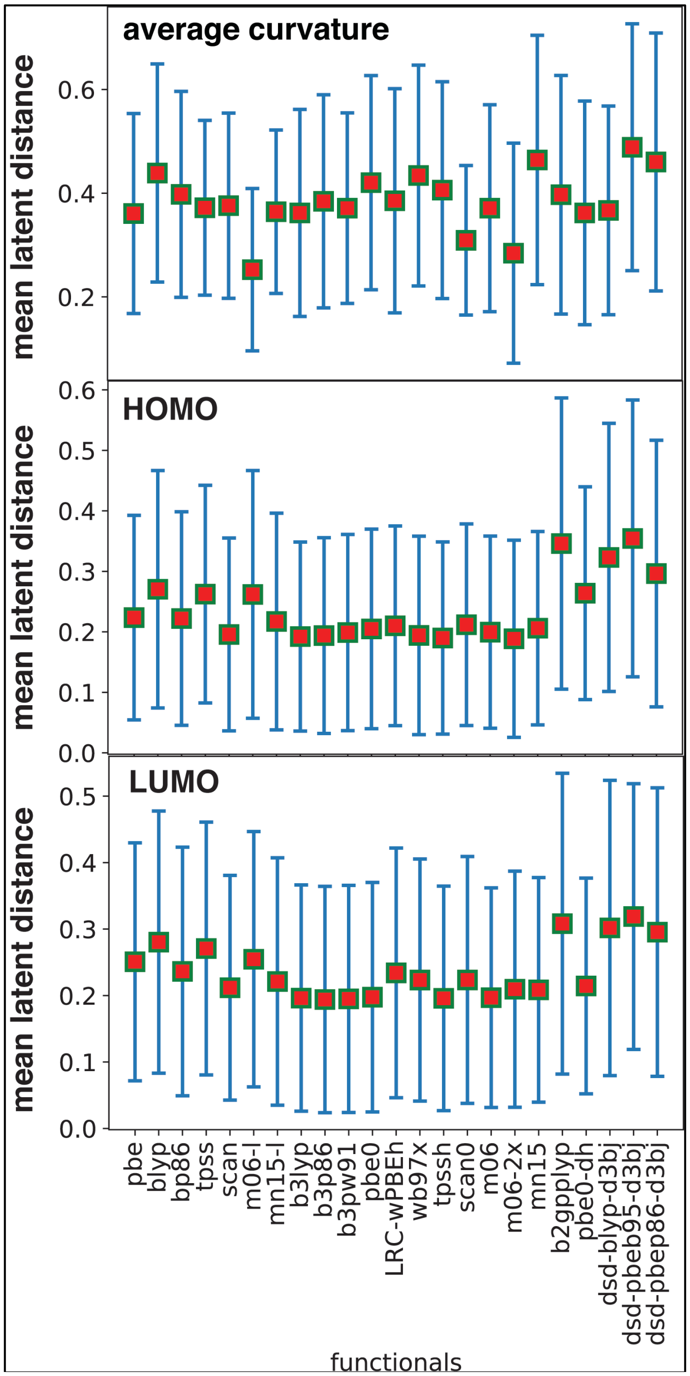

Figure S2. The latent distance obtained and averaged over the 10 nearest neighbors for each of the test set transition metal complexes to available training data evaluated for the ANN models trained to predict the average curvature (top panel), the HOMO energy of the N electron system (middle panel) and the LUMO energy of the system with *N*-1 electrons (lower panel), for each of the 23 functionals included in this work. The error bars correspond to one standard deviation of the 10-NN-averaged distance in latent space for each functional.

Table S4. The mean values (mean) and standard deviations (STD) of the curvature distributions of several functionals with different Hartree–Fock exchange fractions (HF fraction). The different gray scales represent functional families with similar correlation functional and different HF fraction (from left to right, PBE, M06 and dispersion-corrected families).

Functionals	PBE	PBE0	PBE0-DH	M06-L	M06	M06-2X	DSD-BLYP-D3BJ,	DSD-PBEB95-D3BJ	DSD-PBEP86-D3BJ
HF fraction	0.00	0.25	0.50	0.00	0.27	0.54	0.66	0.69	0.71
Mean (eV)	4.54	2.87	1.22	4.40	2.92	1.31	-0.55	-1.20	-1.55
STD (eV)	0.96	0.74	0.92	0.96	0.72	0.76	1.64	1.77	1.80

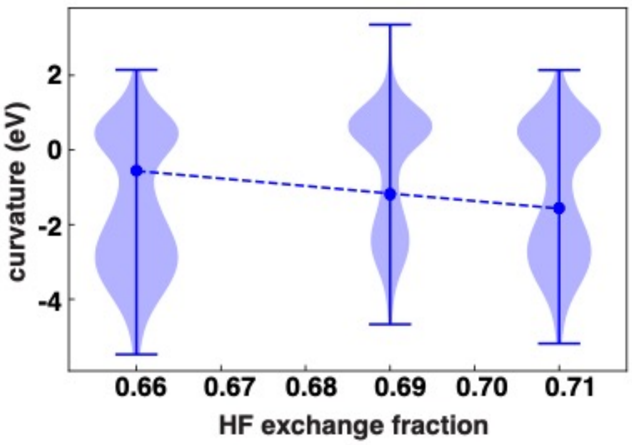

Figure S3. The curvature distribution (light blue violin plots) and its mean value (blue markers) for dispersion-corrected functionals as a function of the Hartree–Fock (HF) exchange fraction in their double hybrid xc functionals. Functionals from left to right (low to higher HF values) correspond to DSD-PBEB95-D3BJ, DSD-PBEP86-D3BJ and DSD-BLYP-D3BJ. The dashed line corresponds to a linear fit to the mean values, with a slope of -20.15 eV/HFX and R^2 of 0.999. The vertical bar indicates the full range of the distribution for each functional.

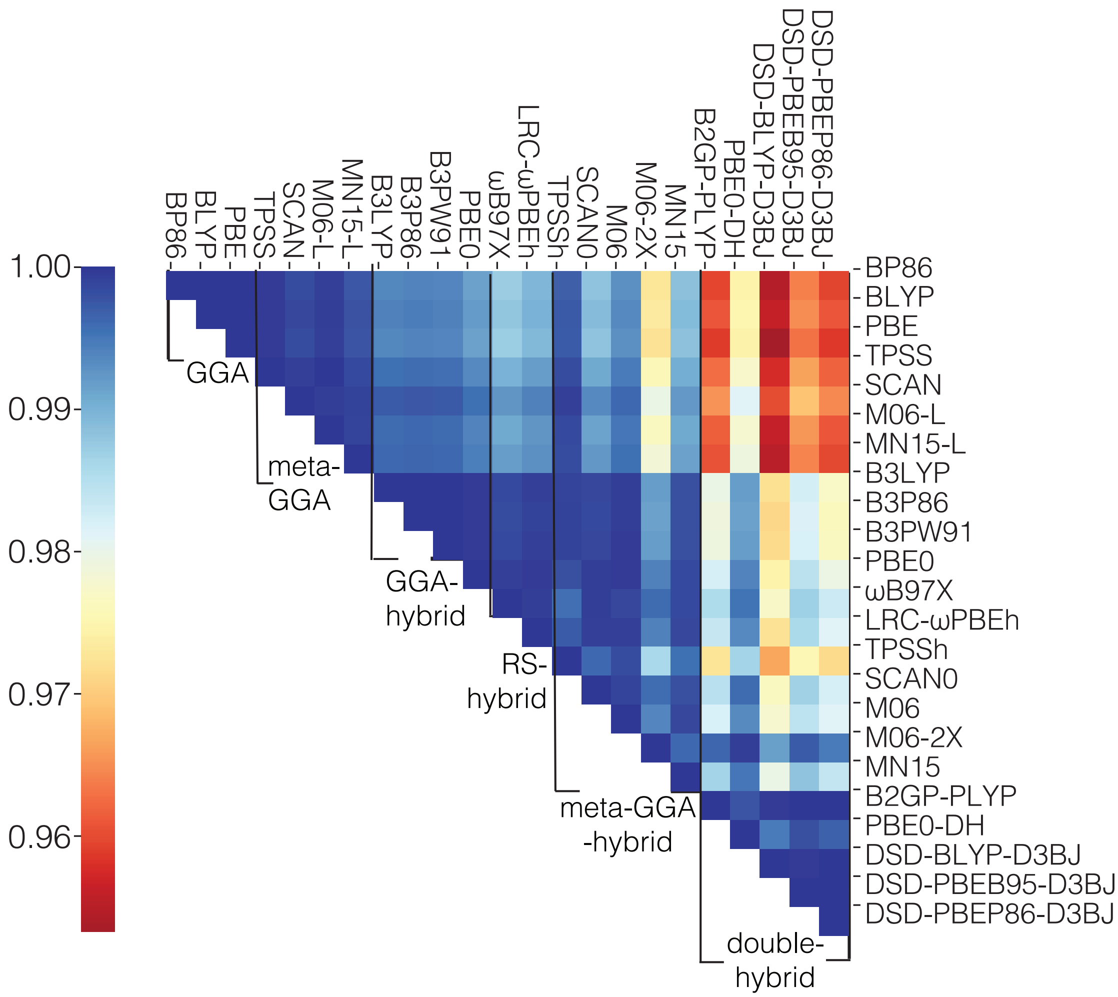

Figure S4. An upper triangular matrix colored by Pearson's r for pairs of 23 functionals for the HOMO energy of the N-electron system computed over a set of mononuclear octahedral transition metal complexes with Cr, Mn, Fe, or Co centers. The correlations are grouped by functional family from top to bottom or left to right: GGA, meta-GGA, GGA-hybrid, range-separated (RS) hybrid, meta-GGA hybrid, and double hybrid. The colorbar range (0.96-1.00) is much smaller for the HOMO energy than for the average curvature values in main text Figure 2.

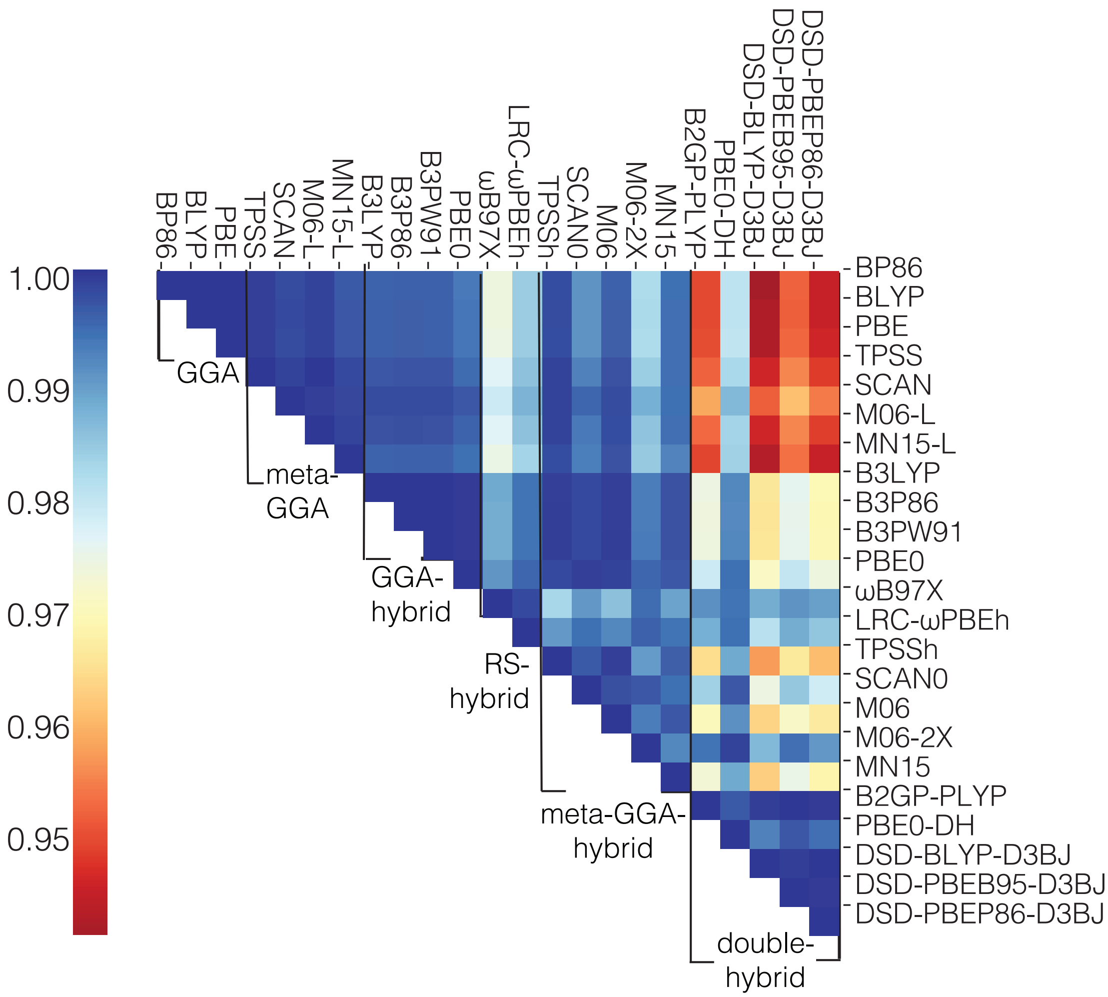

Figure S5. An upper triangular matrix colored by Pearson's r for pairs of 23 functionals for the LUMO energy of the *N-1*-electron system computed over a set of mononuclear octahedral transition metal complexes with Cr, Mn, Fe, or Co centers. The correlations are grouped by functional family from top to bottom or left to right: GGA, meta-GGA, GGA-hybrid, range-separated (RS) hybrid, meta-GGA hybrid, and double hybrid. The colorbar range (0.96-1.00) is much smaller for the HOMO energy than for the average curvature values in main text Figure 2.

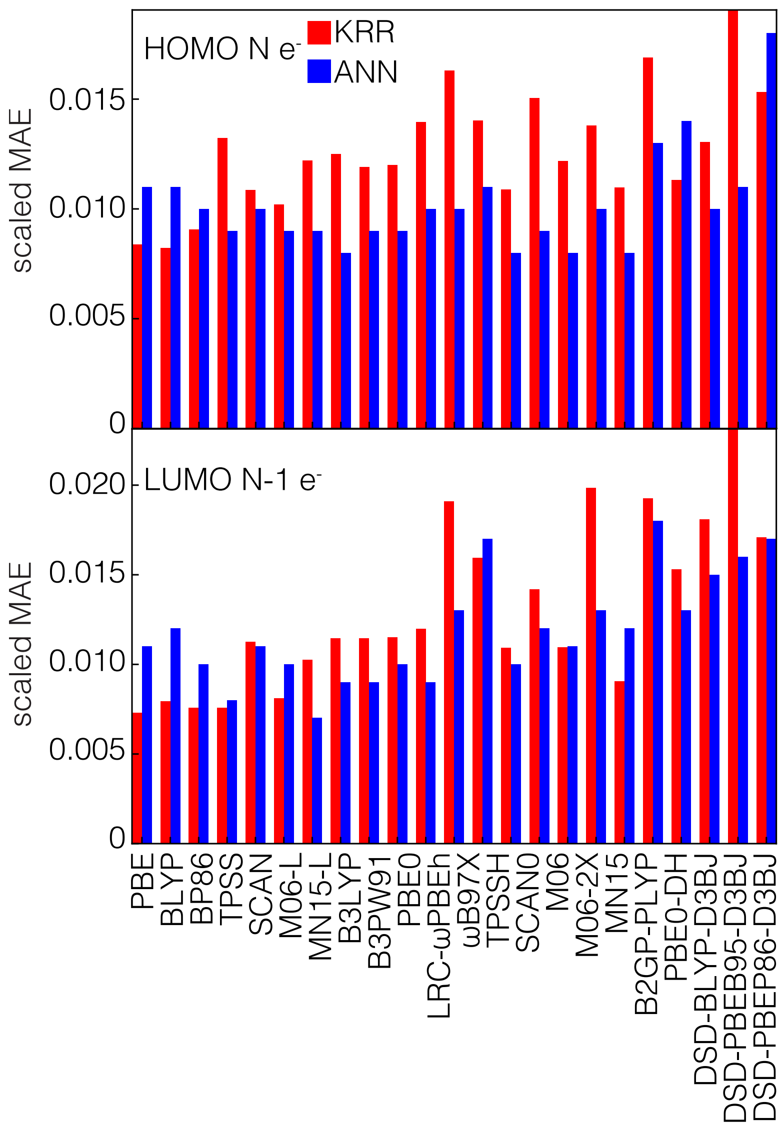

Figure S6. The scaled MAE of the predictions of HOMO of the *N*-electron system (top panel) and the LUMO of the *N*-1 electron system for KRR (red) and ANN (blue) models. The functionals are grouped by functional family on the x-axis from left to right: GGA (PBE, BLYP, BP86), meta-GGA (TPSS, SCAN, M06-L, MN15-L), GGA hybrid (B3LYP, B3P86, B3PW91, PBE0), range-separated hybrid (LRC-ωPBEH, ωB97x), meta-GGA hybrid (TPSSh, SCAN0, M06, M06-2X, MN15), and double hybrid (B2GP-PLYP, PBE0-DH, DSD-BLYP-D3BJ, DSD-PBEB95-D3BJ, DSD-PBEP86-D3BJ).

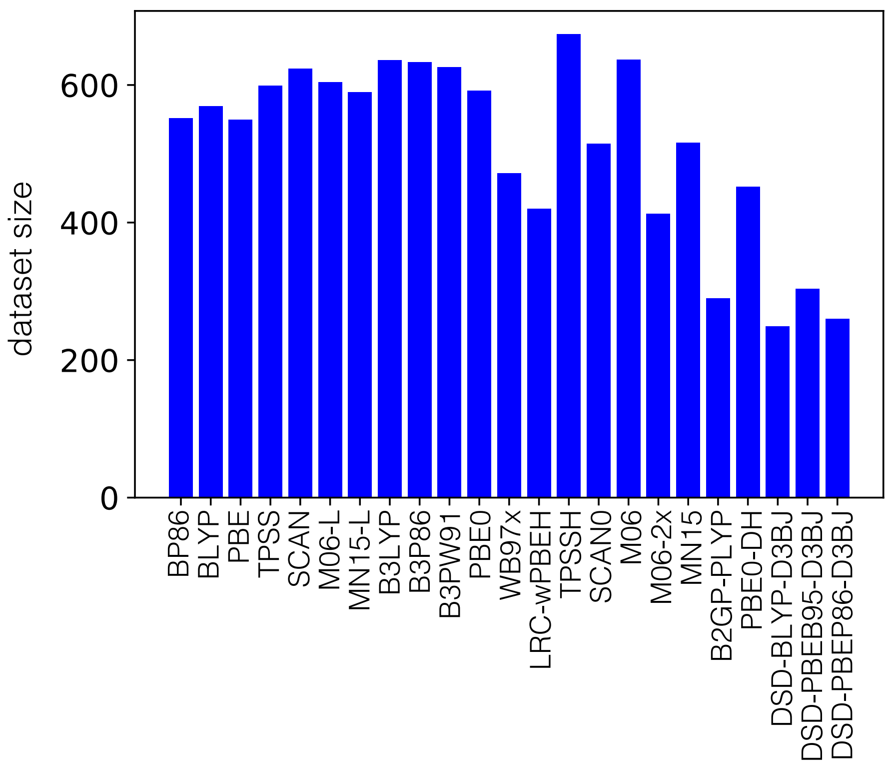

Figure S7. The total dataset sizes (including both the training and test set) of the different functionals. The theoretical maximum of the dataset size (train and test) is 948 complexes, but not all complexes converged for all functionals. Complexes are pruned when the HOMO and LUMO error are of opposite sign, leading to inconclusive predictions of curvature from the difference of the HOMO and LUMO energies. This occurs most frequently for hybrid functionals.

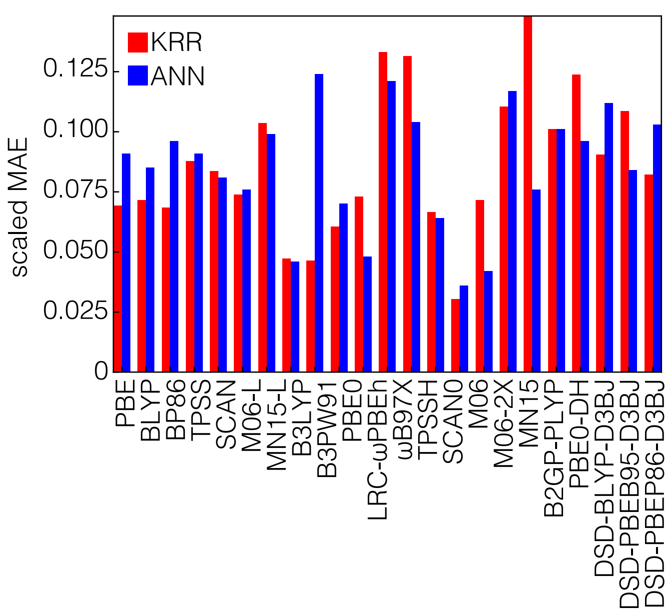

Figure S8. The scaled MAE of the predicted curvature using KRR (red) and ANN (blue) models, performed on a dataset of 64 complexes with valid curvature values for all functionals. The functionals are grouped by functional family on the x-axis from left to right: GGA (PBE, BLYP, BP86), meta-GGA (TPSS, SCAN, M06-L, MN15-L), GGA hybrid (B3LYP, B3P86, B3PW91, PBE0), range-separated hybrid (LRC-ωPBEH, ωB97x), meta-GGA hybrid (TPSSh, SCAN0, M06, M06-2X, MN15), and double hybrid (B2GP-PLYP, PBE0-DH, DSD-BLYP-D3BJ, DSD-PBEB95-D3BJ, DSD-PBEP86-D3BJ).

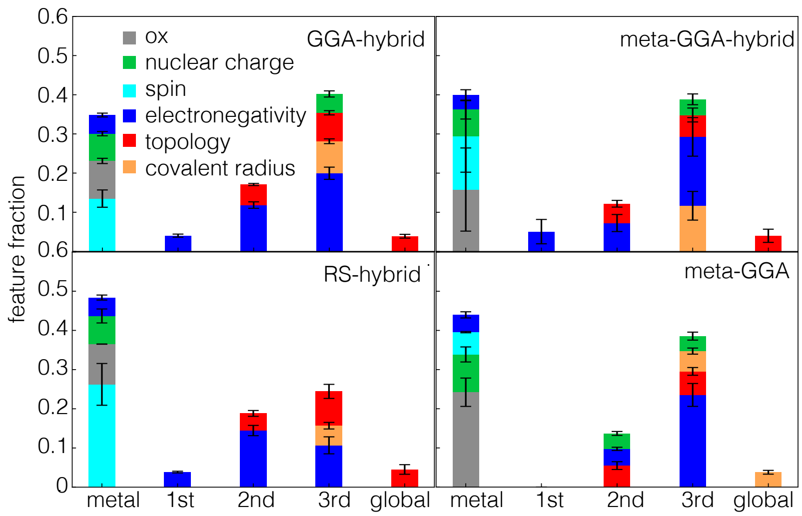

Figure S9. Stacked bar plot of the fractional weight of 15 features with the highest SHAP values in a curvature prediction model, as a function of the most metal-distal atoms for the GGA-hybrid (top left panel), meta-GGA (top-right panel), RS-hybrid (bottom-left panel) and meta-GGA functional families (bottom-right panel). Error bars reflect the standard deviation across the set of DFAs within each functional family.

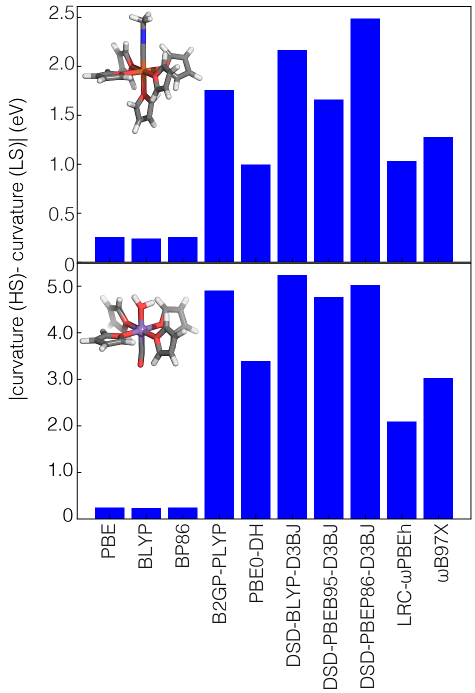

Figure S10. The absolute value of the difference between the DFT-calculated curvature in the HS state and the curvature in LS state of Fe(III)(furan)$_5$(methylisocyanide) (top panel) and Mn(II)(furan)$_4$(H_2O)(CO) (bottom panel) for GGA, double hybrid, and RS-hybrid functionals labeled from left to right.

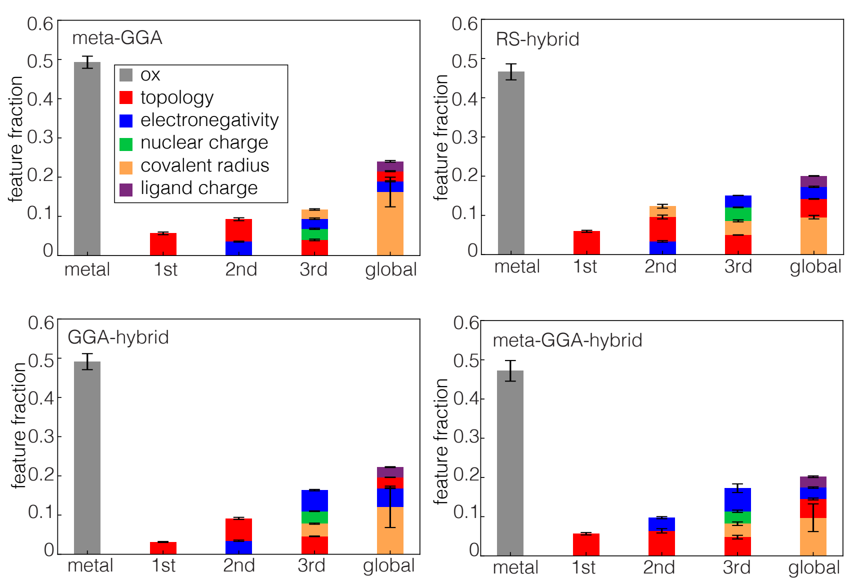

Figure S11. Stacked bar plot of the fractional weight of 15 features with the highest SHAP values in a HOMO energy prediction model of the *N*-electron system, as a function of the most metal-distal atoms for GGA-hybrid (top left panel), meta-GGA (top-right panel), RS-hybrid (bottom-left panel) and the meta-GGA functional family (bottom-right panel). Error bars are computed from the standard deviation across the DFAs for each functional family.

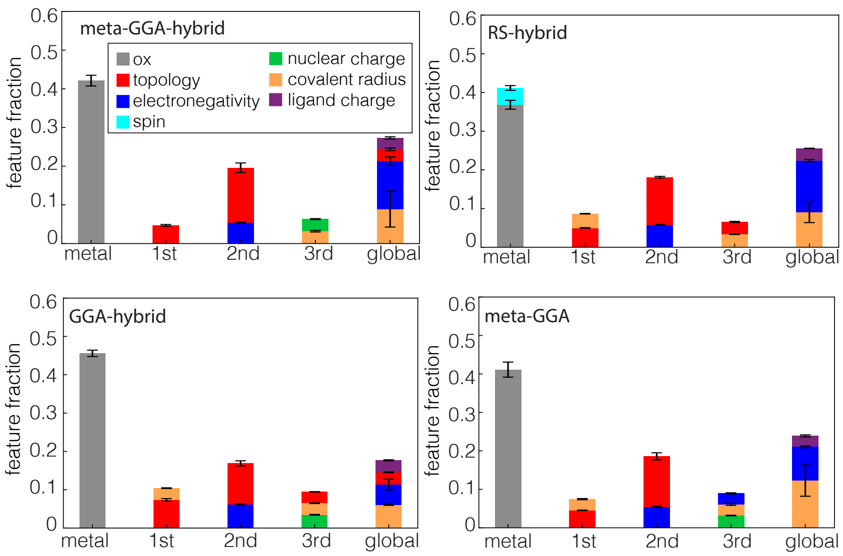

Figure S12. Stacked bar plot of the fractional weight of 15 features with the highest SHAP values in a LUMO energy prediction model of the *N*-1 electron system, as a function of the most metal-distal atoms for GGA-hybrid (top left panel), meta-GGA (top-right panel), RS-hybrid (bottom-left panel) and the meta-GGA functional family (bottom-right panel). Error bars are computed from the standard deviation across the DFAs for each functional family.

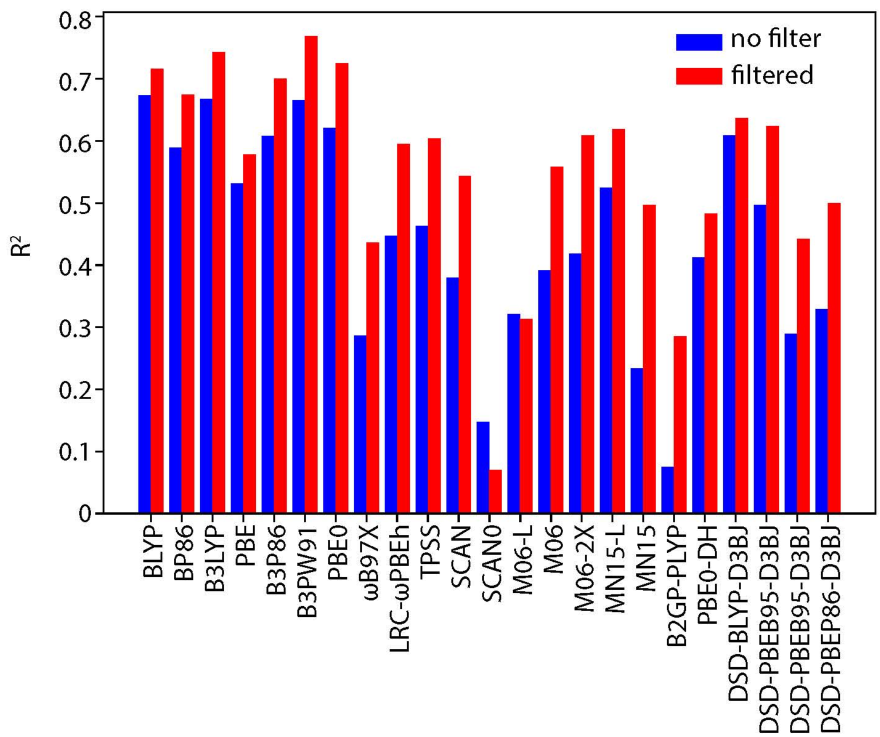

Figure S13. The R^2 between direct and indirect curvature predictions from ANN models for different functionals before applying the latent distance criteria (blue, no filter) and after (red, filtered).

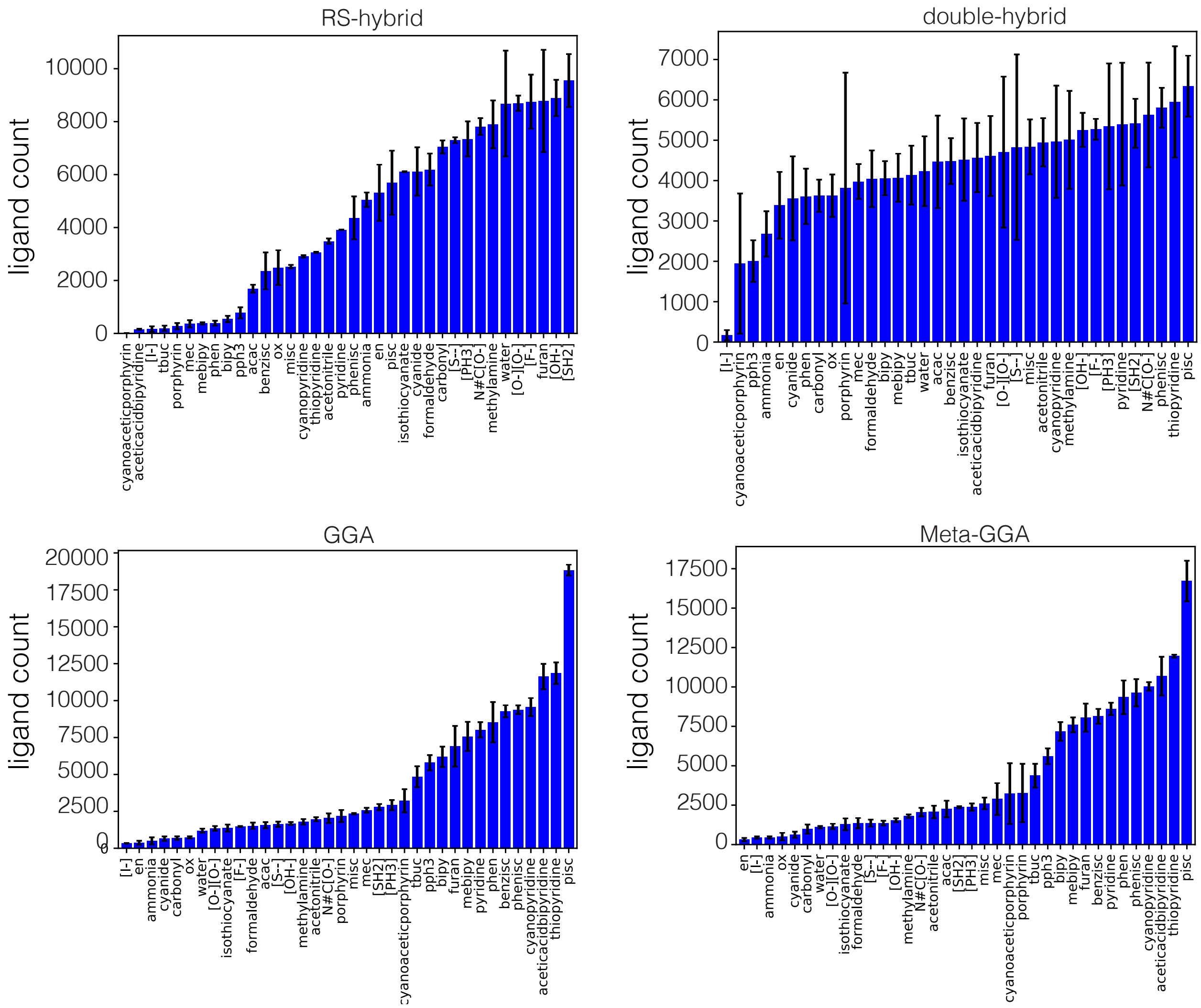

Figure S14. The ligand distribution of the 20% of the design space complexes with the lowest curvatures for RS-hybrid (top left panel) double-hybrid (top right panel), GGA (bottom left panel) and meta-GGA (bottom right panel) families ordered by increasing count from left to right for each functional family. The counts correspond to each occurrence of the ligand in a complex in the 20% of the design space with the lowest curvatures. Error bars correspond to the standard deviation between the different functionals in the same family.

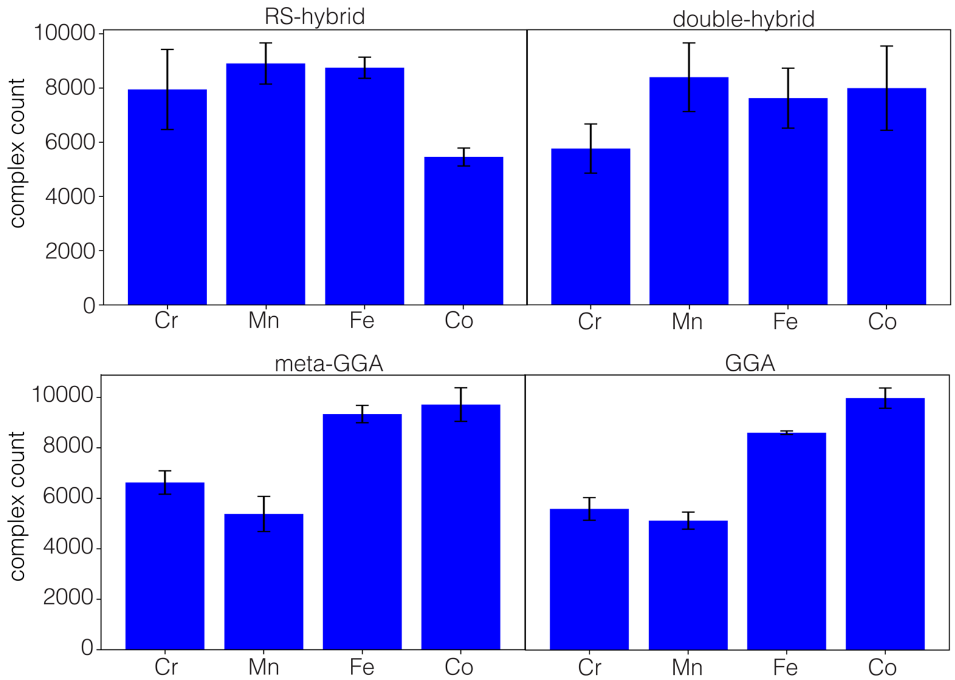

Figure S15. The metal distribution of the 20% of the design space complexes with the lowest curvatures for RS-hybrid (top left panel) double-hybrid (top right panel), GGA (bottom left panel) and meta-GGA (bottom right panel) families. The counts correspond to each occurrence of a complex in the 20% of the design space with the lowest curvatures. Error bars correspond to the standard deviation between the different functionals in the same family.

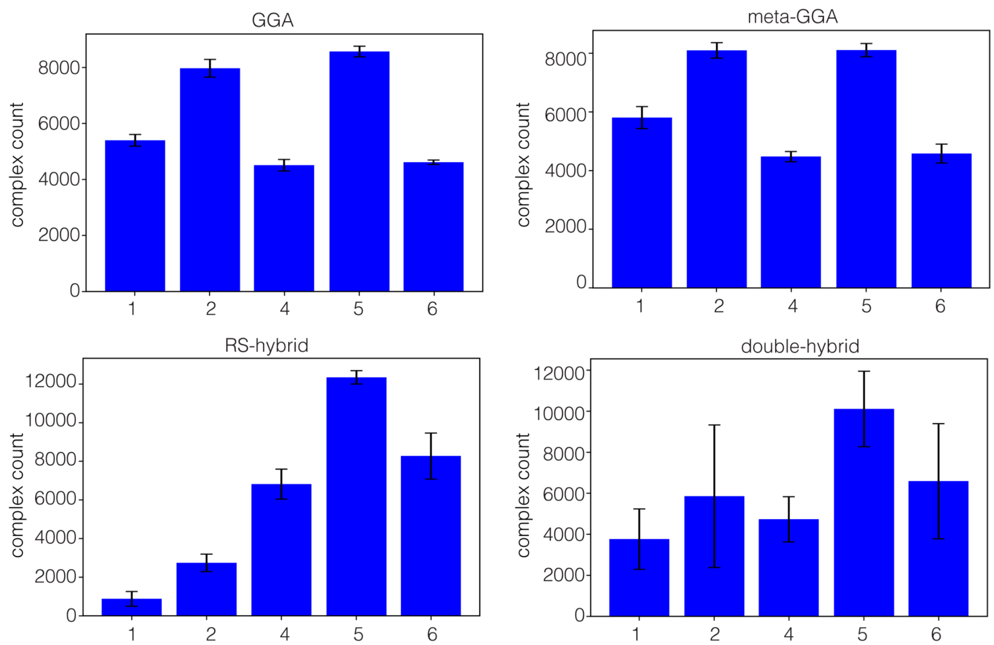

Figure S16. The spin multiplicity distribution of 20% of the design space complexes with the lowest curvatures for GGA (top left panel) meta-GGA (top right panel), RS-hybrid (bottom left panel) and double-hybrid (bottom right panel) families ordered by increasing count from left to right for each functional family. The counts correspond to each occurrence of a complex in the 20% of the design space with the lowest curvatures. Error bars correspond to the standard deviation between the different functionals in the same family.

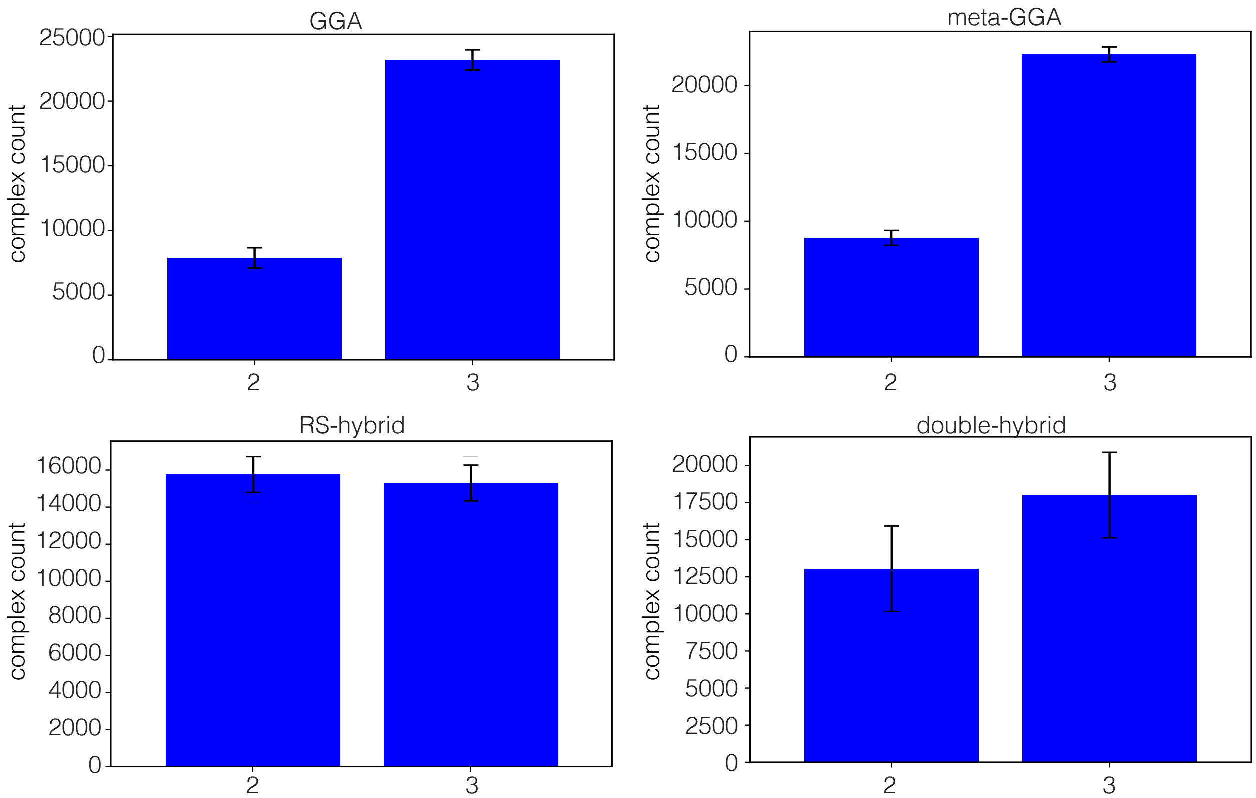

Figure S17. The oxidation state multiplicity distribution of the 20% of the design space complexes with the lowest curvatures for GGA (top left panel) meta-GGA (top right panel), RS-hybrid (bottom left panel) and double-hybrid (bottom right panel). The counts correspond to each occurrence of a complex in the 20% of the design space with the lowest curvatures. Error bars correspond to the standard deviation between the different functionals.

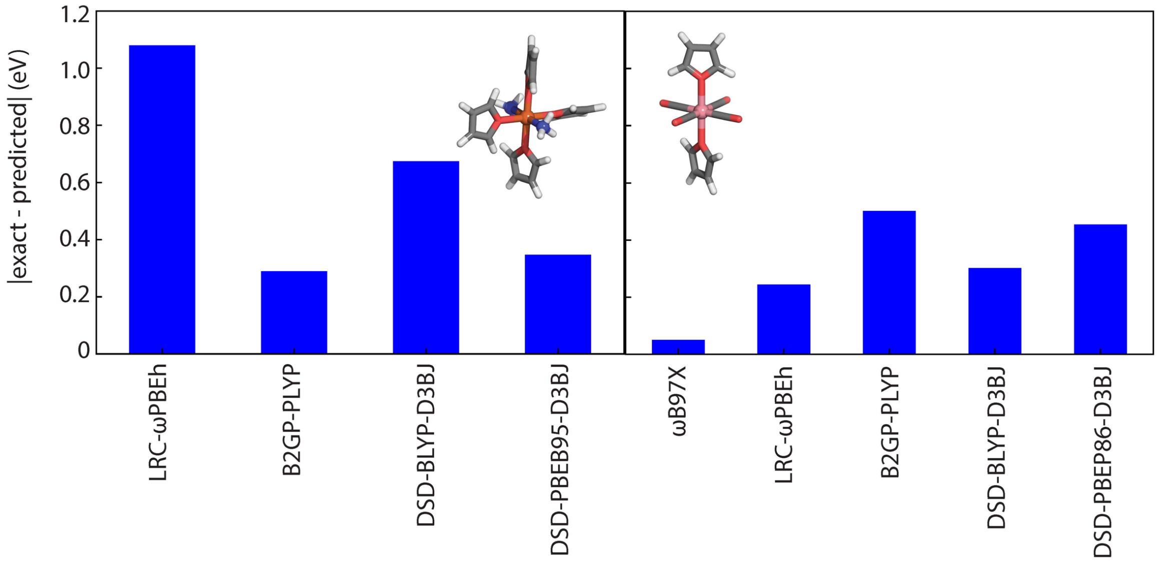

Figure S18. The absolute value of the difference between the ANN model prediction and the calculated curvature for a series of functionals that were predicted to produce curvatures between 0 and 0.4 eV for two representative complexes: (left panel) Fe(III)(furan)$_4$(ammonia)$_2$ in the HS state and (right panel) Co(II)(carbonyl)$_4$(furan)$_2$ in the LS state.

References

(1) Duan, C.; Chen, S.; Taylor, M. G.; Liu, F.; Kulik, H. J. Machine Learning to Tame Divergent Density Functional Approximations: A New Path to Consensus Materials Design Principles. *Chem Sci* **2021,** *12*, 13021-13036.

(2) Perdew, J. P. Density-Functional Approximation for the Correlation-Energy of the Inhomogeneous Electron-Gas. *Physical Review B* **1986,** *33*, 8822-8824.

(3) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior. *Physical Review A* **1988,** *38*, 3098-3100.

(4) Devlin, F. J.; Finley, J. W.; Stephens, P. J.; Frisch, M. J. Ab-Initio Calculation of Vibrational Absorption and Circular-Dichroism Spectra Using Density-Functional Force-Fields - a Comparison of Local, Nonlocal, and Hybrid Density Functionals. *Journal of Physical Chemistry* **1995,** *99*, 16883-16902.

(5) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results Obtained with the Correlation-Energy Density Functionals of Becke and Lee, Yang and Parr. *Chemical Physics Letters* **1989,** *157*, 200-206.

(6) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **1996,** *77*, 3865-3868.

(7) Tao, J. M.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-Generalized Gradient Approximation Designed for Molecules and Solids. *Physical Review Letters* **2003,** *91*, 146401.

(8) Sun, J. W.; Ruzsinszky, A.; Perdew, J. P. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Physical Review Letters* **2015,** *115*, 036402.

(9) Zhao, Y.; Truhlar, D. G. A New Local Density Functional for Main-Group Thermochemistry, Transition Metal Bonding, Thermochemical Kinetics, and Noncovalent Interactions. *Journal of Chemical Physics* **2006,** *125*, 194101.
(10) Yu, H. S.; He, X.; Truhlar, D. G. Mn15-L: A New Local Exchange-Correlation Functional for Kohn-Sham Density Functional Theory with Broad Accuracy for Atoms, Molecules, and Solids. *Journal of Chemical Theory and Computation* **2016,** *12*, 1280-1293.
(11) Becke, A. D. Density-Functional Thermochemistry. Iii. The Role of Exact Exchange. *Journal of Chemical Physics* **1993,** *98*, 5648-5652.
(12) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Physical Review B* **1988,** *37*, 785--789.
(13) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *The Journal of Physical Chemistry* **1994,** *98*, 11623-11627.
(14) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces - Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Physical Review B* **1992,** *46*, 6671-6687.
(15) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The Pbe0 Model. *Journal of Chemical Physics* **1999,** *110*, 6158-6170.
(16) Chai, J. D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *Journal of Chemical Physics* **2008,** *128*, 084106.
(17) Rohrdanz, M. A.; Martins, K. M.; Herbert, J. M. A Long-Range-Corrected Density Functional That Performs Well for Both Ground-State Properties and Time-Dependent Density Functional Theory Excitation Energies, Including Charge-Transfer Excited States. *Journal of Chemical Physics* **2009,** *130*, 054112.
(18) Hui, K.; Chai, J. D. Scan-Based Hybrid and Double-Hybrid Density Functionals from Models without Fitted Parameters. *Journal of Chemical Physics* **2016,** *144*, 044114.
(19) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theoretical Chemistry Accounts* **2008,** *120*, 215-241.
(20) Yu, H. Y. S.; He, X.; Li, S. H. L.; Truhlar, D. G. Mn15: A Kohn-Sham Global-Hybrid Exchange-Correlation Density Functional with Broad Accuracy for Multi-Reference and Single-Reference Systems and Noncovalent Interactions. *Chemical Science* **2016,** *7*, 6278-6279.
(21) Karton, A.; Tarnopolsky, A.; Lamere, J. F.; Schatz, G. C.; Martin, J. M. L. Highly Accurate First-Principles Benchmark Data Sets for the Parametrization and Validation of Density Functional and Other Approximate Methods. Derivation of a Robust, Generally Applicable, Double-Hybrid Functional for Thermochemistry and Thermochemical Kinetics. *Journal of Physical Chemistry A* **2008,** *112*, 12868-12886.
(22) Bremond, E.; Adamo, C. Seeking for Parameter-Free Double-Hybrid Functionals: The Pbe0-Dh Model. *Journal of Chemical Physics* **2011,** *135*, 024106.

(23) Kozuch, S.; Martin, J. M. L. Spin-Component-Scaled Double Hybrids: An Extensive Search for the Best Fifth-Rung Functionals Blending Dft and Perturbation Theory. *Journal of Computational Chemistry* **2013,** *34*, 2327-2344.