Electronic Supplementary Material (ESI) for Chemical Science.

Supplementary Material for A quantitative uncertainty metric controls error in neural network-driven chemical discovery

Jon Paul Janet[1], Chenru Duan[1,2], Tzuhsiung Yang[1], Aditya Nandy[1,2], and Heather J. Kulik *[1]

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

- **DFT-results.zip**
  - *readme.txt*
  - *CSD-results.csv*
  - *train-results.csv*
- **predictions.zip**
  - *readme.txt*
  - *CSD-predictions.csv*
  - *QM9-predictions.csv*
  - *QM9-test-SMILES.csv*
- **geometries.zip**
  - *readme.txt*
  - **CSD**
    - *CODE_[spin].xyz*
  - **inorganic_training**
    - *[metal]_[oxidation]_[equatorial_ligand]_[axial_ligand_1]_[axial_ligand_2]_[%HFX]_[spin].xyz*
- **models.zip**
  - *readme.txt*
  - **atomization**
    - *u0_vars.csv*
    - *u0_mean_x.csv/u0_mean_y.csv*
    - *u0_var_x.csv/u0_var_y.csv*
    - *u0_model.json*
    - *u0_model.h5*
  - **split**
    - *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*

Text S1: Details of ensemble and mc-dropout

Ensembles: One common approach to assign uncertainty estimates to predictions from data-driven models is to generate an ensemble of J different models. The mean of the predictions of these models is used as the predicted value at a new point and the variance in these predictions is used as a metric for model confidence. If x^* is a new trial point, and σ_{x^*} is the standard deviation associated with this prediction, the ensemble prediction is given as:

$$\bar{\tilde{y}}(x^*) = \frac{1}{n_{\text{ens}}} \sum_{j=1}^{n_{\text{ens}}} \hat{y}_j(x^*) \tag{S1}$$

with a variance of:

$$\sigma^2{}_{x^*} = \frac{1}{n_{\text{ens}}} \sum_{j=1}^{n_{\text{ens}}} \left(\bar{\tilde{y}}(x^*) - \hat{y}_j(x^*)\right)^2 \tag{S2}$$

The prediction mean could be expected to have lower generalization error with respect to individual models. Typically, ensembles are generated by partitioning data to generate submodels, where each is trained on distinct subsets of data. Detection of uncertain points with ensemble models relies on the submodels being incorrect in different ways (i.e., high variance), which can occur when the model is evaluated for molecules dissimilar to training examples, where the behavior is only weakly constrained.

Monte-Carlo dropout: A lower cost framework for deriving uncertainty estimates for dropout regularized neural networks has recently been suggested[1] in analogy to Gaussian processes. In practice, this entails running the model J times with the dropout mask kept on, removing random notes from the network each time. The average of these predictions are used as in the case with ensembles. The predictive uncertainty is estimated from:

$$\sigma^2{}_{x^*} = \frac{1}{J} \sum_{j=1}^{J} \left(\bar{\tilde{y}}(x^*) - \hat{y}_j(x^*)\right)^2 + \tau^{-1} I \tag{S3}$$

This expression differs from the ensemble expression by also including a learned baseline uncertainty term, τ^{-1}, which must be estimated from training data. In comparison to ensemble models, the cost of this approach is lower because the model only needs to be trained once. For mc-dropout, we determine a representative value of τ by maximizing the log predictive likelihood of the corresponding GP based on the training data. This is a measure of how likely the observed data are under the GP, and is approximated[1] by

$$\log p\left(\mathbf{y}\left(\mathbf{x}_n\right)|\mathbf{x}_n, \mathbf{X}, \mathbf{Y}\right) \approx \log\left[\sum_{j=1}^{J} e^{-\frac{1}{2}\tau\left\|\tilde{\mathbf{y}}(\mathbf{x}_n) - \tilde{\mathbf{y}}_j(\mathbf{x}_n)\right\|_2^2}\right] - \log J - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \tau^{-1} \tag{S4}$$

In the application here (i.e., for the fully connected spin splitting neural network), we have scalar output and we use the training data to optimize equation S4 with respect to τ numerically. We use $J = 100$ repeats, as in the network itself. The determined value of τ based on the training data is 3.6×10^8 in dimensionless units.

Text S2: Simulation details for inorganic complexes and CSD test set

In this work, we primarily use 1901 spin splitting energies from DFT data sets generated over several prior works[4,9,3,6] to train new machine learning models. We also generate new DFT data on a 116-molecule CSD data set. We concisely summarize some of the details of these efforts here but refer the reader to the original work for more detail. 788 of the compounds are from Ref.[4], 286 of the compounds are from Ref.[3], 19 of the compounds are from Ref.[6], 87 of the compounds had revised spin states first published in Ref.[9], and 721 of the compounds had not been previously published, including revised spins for compounds from previous sets. All energies and structures are provided in the Supporting Information zip file.

Despite originating from several original sources, a consistent workflow has been employed, with distinctions noted as follows. The molSimplify[2] toolkit was used to generate octahedral transition metal complex structures from a pool of organic ligands common in inorganic chemistry (listed in Table S1) with enforced equatorial symmetry but allowing up to two distinct axial ligands. DFT geometry optimizations were then carried out using TeraChem[11] with the B3LYP hybrid DFT functional, varying the fraction of Hartree-Fock (HF) exchange from its default 20% value in 5% increments over the range of 0-30% HF exchange. Thus, the 1901 data points corresponds to 564 unique chemical structures, with additional repeats at varied exchange fractions. The LANL2DZ effective core potential was employed for transition metals and heavy elements (i.e., Br) with the 6-31G* basis for all other atoms. The effect of using a modest basis set, which enables larger data set generation for ML models, was found to be limited in prior work on the relative energies of interest[5]. The metals studied throughout were Cr, Mn, Fe, and Co in M(II) and M(III) oxidation states. The high-spin/low-spin definitions used to calculate the adiabatic electronic energy spin splitting were: quintet-singlet for both d^4 Mn(III)/Cr(II) and d^6 Co(III)/Fe(II), sextet-doublet for d^5 Fe(III)/Mn(II), and quartet-doublet for both d^3 Cr(III) and d^7 Co(II). These spin states are a revision from initial work[4] that employed a triplet ground state for Cr(II) and Mn(III).

All open shell complexes (i.e., all non-singlets) are treated with spin-unrestricted DFT with virtual and occupied orbitals level-shifted[10] by 1.0 and 0.1 Ha. respectively, to aid convergence to an unrestricted solution. Geometry optimizations were conducted for 788 cases with DL-FIND[7] in Cartesian coordinates. The protocol was shifted to employ the TRIC (translation rotation internal coordinates)[12] optimizer for the 1113 most recent cases. Both optimizers are available in TeraChem, and the same default tolerances were employed of $4.5x10^{-4}$ hartree/bohr for the maximum gradient and $1x10^{-6}$ hartree for the change in energy between steps.

Prior to their use in model training, structures are filtered and removed if they fail metrics of quality geometries we recently introduced[9]. Specifically, these metrics include preserved coordination number of 6 with reasonable bond lengths and no ligand distortions. Additionally we removed any complexes with large (i.e., 1.0 μ_B or larger) deviation of $\langle S^2 \rangle$ from the expected value based on the assigned spin.

For the CSD data set, we searched for diverse octahedral transition metal complexes with M(II)/M(III) M = Cr, Mn, Fe, or Co transition metals. For the geometry optimization, the same method, basis, and optimization approach was followed. Geometry checks and $\langle S^2 \rangle$ deviations were used to eliminate structures. Additionally, we manually screened the collected points to exclude any that were duplicates within each other as judged through comparable connectivity but differing accession codes. We also removed those that were duplicates of data in the original data set, as judged through the assigned connectivity in RAC-155. As an additional constraint, we filtered out any complexes with evidence of ligand non-innocence. Specifically, we computed the Mulliken spin of the metal center and discarded complexes with Mulliken spin that was more than 1.0 μ_B less than the expected spin from the overall spin assigned to the complex.

Table S1: Ligand identity and occurrence among 654 unique metal-ligand combinations in the inorganic complex training set. Occurrence sums over all instances of the ligand in either axial site and the equatorial site. SMILES are given in the final column with the connection atom(s) shown in red.

	Ligand	Cumulative total	SMILES	Charge	Formula
1	misc	293	C[N]#[C]	0	C_2H_3N
2	water	292	O	0	H_2O
3	carbonyl	275	CO	0	CO
4	pyr	267	c1ccncc1	0	C_5H_5N
5	furan	168	o1cccc1	0	C_4H_4O
6	ammonia	91	N	0	NH_3
7	pisc	64	CC(C)(C)C1=CC=C(C=C1)[N]#[C]	0	$(CH_3)_3CC_6H_4NC$
8	isothiocyanate	57	[N-]=C=S	-1	NCS^-
9	cyanide	52	[C-]#N	-1	CN^-
10	en	42	NCCN	0	NCH_2CH_2N
11	acac	38	CC(=O)C=C(-O-)C	-1	$C_5H_8O_2^-$
12	chloride	36	Cl	-1	Cl^-
13	phen	35	C1=CC2=CC=C3C=CC=NC3=C2N=C1	0	$C_{12}H_8N_2$
14	ox	28	[O-]C(=O)C([O-])=O	-2	$C_2O_4^{2-}$
15	tbuc	27	CC(C)(C)C1=CC(=C([O-])C=C1)[O-]	-2	$(CH_3)_3CC_6H_3O_2^{2-}$
16	bipy	26	C1ccnc(c1)c2ccccn2	0	$C_{10}H_8N_2$
17	tbisc	22	[C]#[N]C(C)(C)C	0	$(CH_3)CCN$
18	etesacac	21	O=C(C)/C(=C(\[O])/C)/CC(=O)OCC	-1	$C_9H_{13}O_4^-$
19	cat	18	[O]c1c(cccc1)[O]	-2	$C_6H_4O_2^{2-}$
20	methylamine	18	NC	0	NH_2CH_3
21	phenacac	18	C1=CC=C(C=C1)C(=O)C C(=O)C2=CC=CC=C2	-1	$(C_6H_5CO)_2[CH]_1^-$
22	phenisc	14	[C][N]c1ccccc1	0	C_6H_5NC
23	pyrrole	12	C1=C[N]C=C1	-1	$C_4H_4N^-$
24	cyanopyr	10	c1(ccncc1)C#N	0	NCC_5H_4N
25	benzisc	8	[C][N]Cc1ccccc1	0	$C_6H_5CH_2NC$
26	mebpy	8	n1ccc(cc1c1ncccc(c1)C)C	0	$C_{12}H_{12}N_2$
27	porphyrin	7	[N-]1C2=CC3=NC(=CC4=CC=C([N-]4) C=C5C=CC(=N5)C=C1C=C2)C=C3	-2	$C_{20}H_{12}N_4^{2-}$
28	ethbpy	4	n1ccc(cc1c1ncccc(c1)CC)CC	0	$C_{14}H_{16}N_2$
29	phosacidbpy	4	n1ccc(cc1c1ncccc(c1)P(=O)(O)O)P(=O) (O)O	0	$C_{10}P_2O_6H_{10}$
30	aceticacidbpy	2	n1ccc(cc1c1ncccc(c1)CC(=O)O)CC(=O)O	0	$C_{14}H_{14}O_4N_2$
31	chloropyr	2	c1c(cncc1)Cl	0	ClC_5H_4N
32	mec	2	[O-]c1c(cc(cc1)C)[O-]	2-	$CH_3C_6H_4O_2^{2-}$
33	thiopyr	1	c1(cc—ncc1)S	0	SC_5H_4N

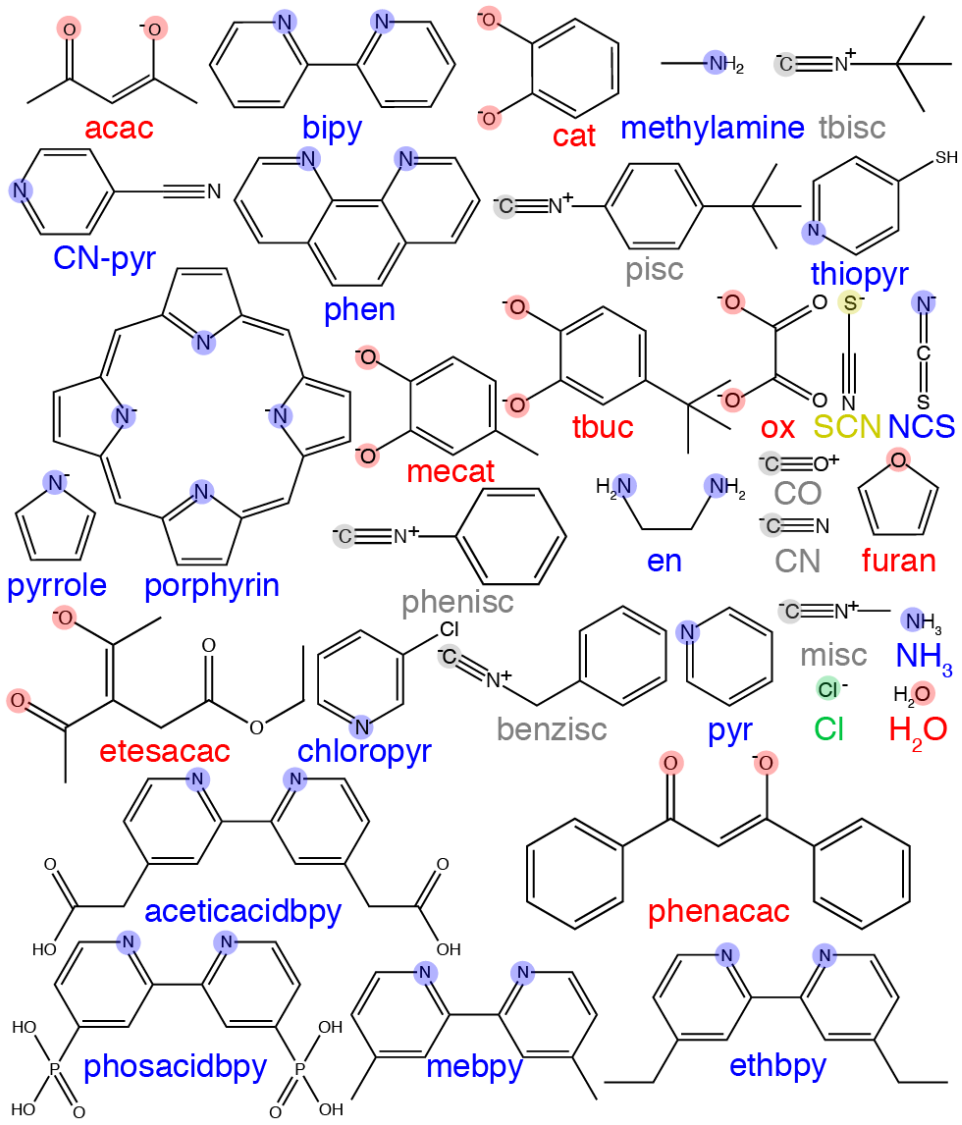

Figure S1: Ligands used to train inorganic complex spin splitting ANN, metal connection atoms highlighted, with the highlight corresponding to the element: oxygen in red, nitrogen in blue, chlorine in red, carbon in gray, and sulfur in yellow. Charges are also shown on relevant atoms.

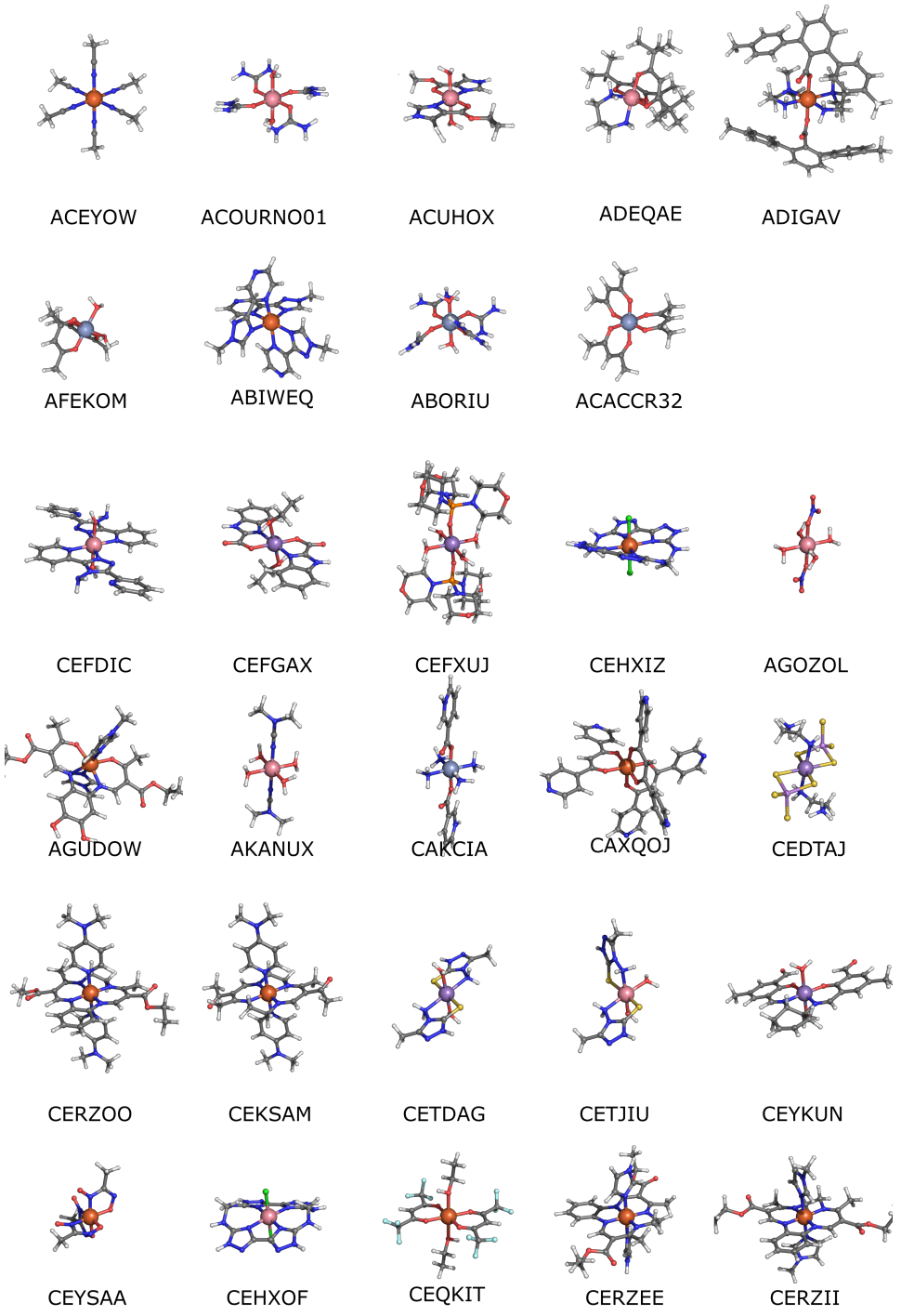

Figure S2: Visualization of CSD Structures used in this work at DFT-optimized ground spin states. CSD accession codes shown below each structure. Non-metal atoms are colored as follows: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is green, bromine is rust, fluorine is cyan, sulfur is yellow, phosphorous is orange, boron is pink and arsenic is purple. Metal centers are shown as large spheres and colored as follows: iron is orange, manganese is purple, cobalt is pink and chromium is metallic blue.

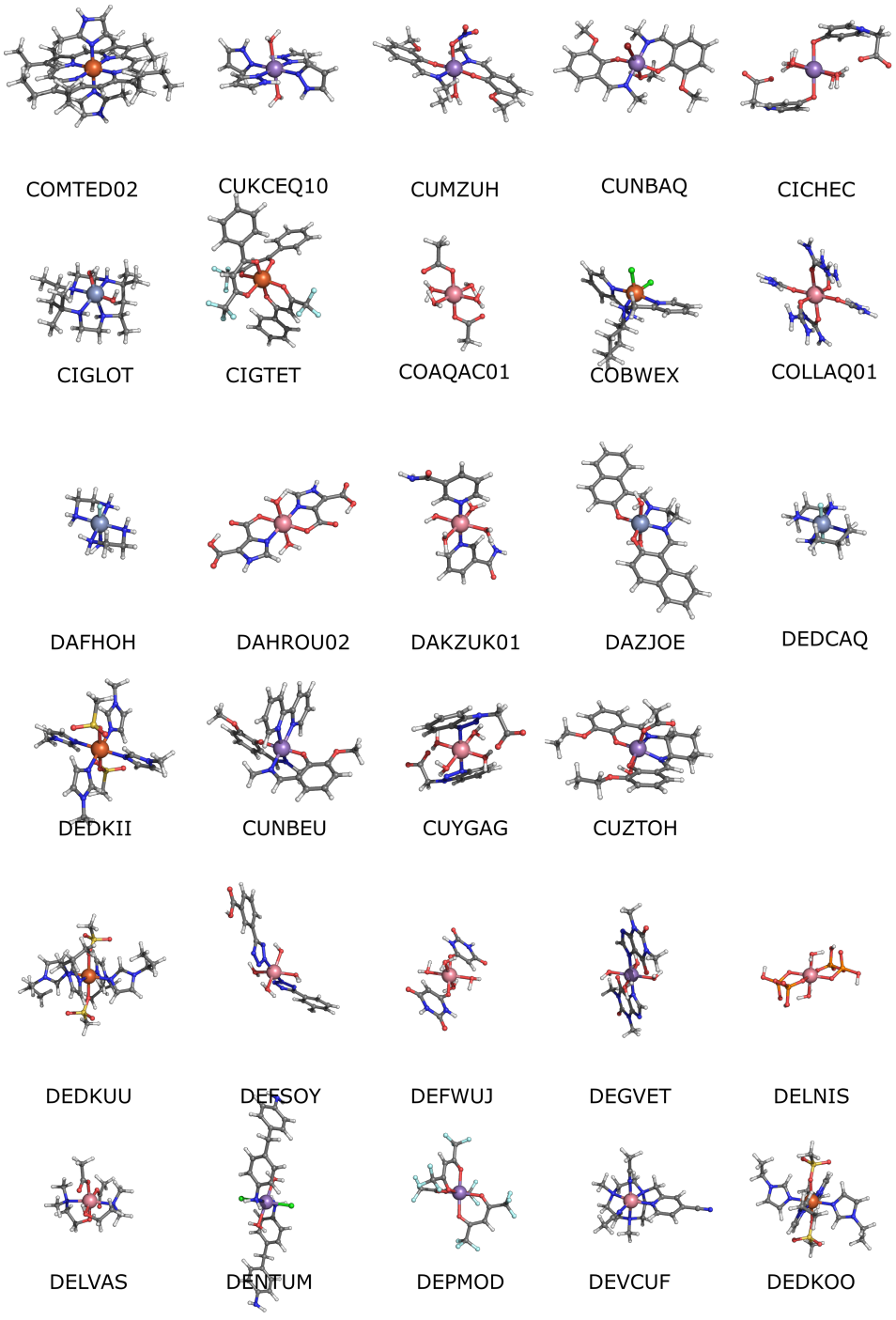

Figure S3: Visualization of CSD Structures used in this work at DFT-optimized ground spin states. CSD accession codes shown below each structure. Non-metal atoms are colored as follows: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is green, bromine is rust, fluorine is cyan, sulfur is yellow, phosphorous is orange, boron is pink and arsenic is purple. Metal centers are shown as large spheres and colored as follows: iron is orange, manganese is purple, cobalt is pink and chromium is metallic blue.

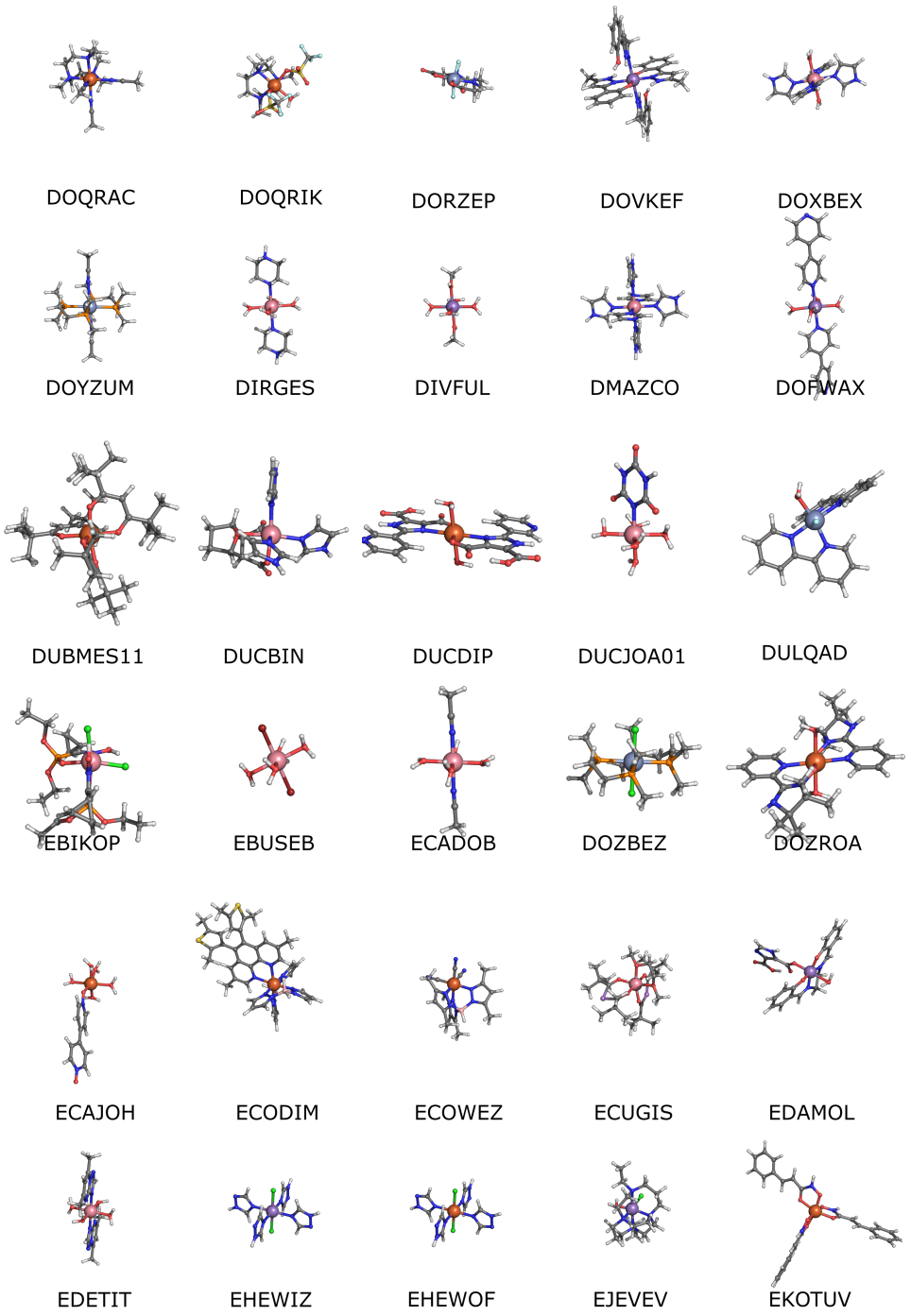

Figure S4: Visualization of CSD Structures used in this work at DFT-optimized ground spin states. CSD accession codes shown below each structure. Non-metal atoms are colored as follows: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is green, bromine is rust, fluorine is cyan, sulfur is yellow, phosphorous is orange, boron is pink and arsenic is purple. Metal centers are shown as large spheres and colored as follows: iron is orange, manganese is purple, cobalt is pink and chromium is metallic blue.

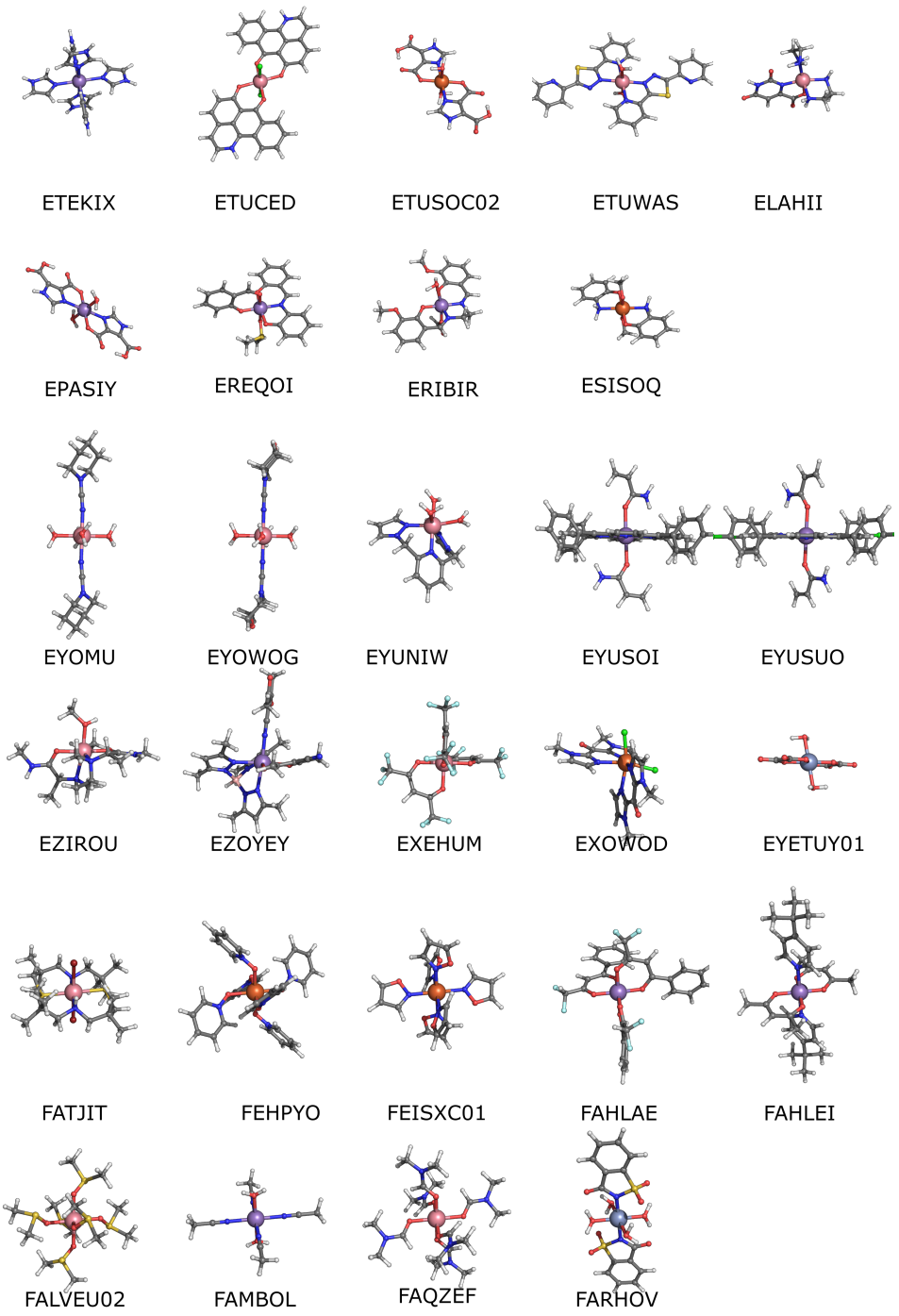

Figure S5: Visualization of CSD Structures used in this work at DFT-optimized ground spin states. CSD accession codes shown below each structure. Non-metal atoms are colored as follows: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is green, bromine is rust, fluorine is cyan, sulfur is yellow, phosphorous is orange, boron is pink and arsenic is purple. Metal centers are shown as large spheres and colored as follows: iron is orange, manganese is purple, cobalt is pink and chromium is metallic blue.

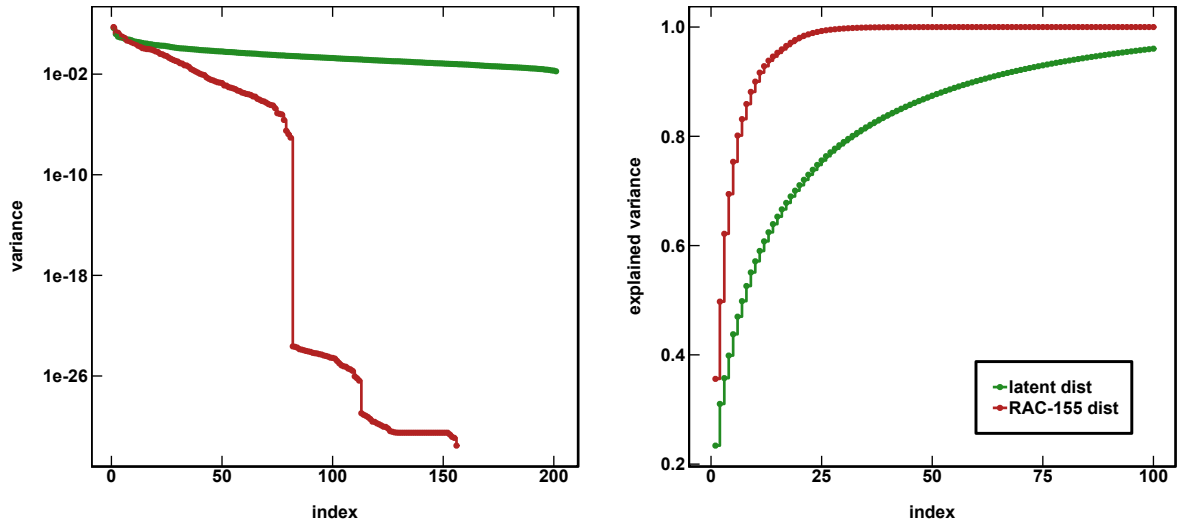

Figure S6: Decay of variance (left) and cumulative relative explained variance of dimensions from principal component analysis of 1901 inorganic training points with RAC-155 representation (red) and the final model latent space (green).

Table S2: Mean absolute error (MAE) and root-mean square error (RMSE) metrics for inorganic spin splitting ANN on training data and out of sample CSD prediction task. Errors are shown from a single model, the average of an ensemble of 10 models and the average of 100 Monte-Carlo dropout realizations of the single model. All error units are kcal/mol. This performance is comparable to a similar test in which we trained on 1400 transition metal complexes with the MCDL-25 descriptor set in a 2-hidden layer ANN. In that work[4], we studied a set of 35 CSD test structures. In those cases, we observed an increase from 2.5 kcal/mol test set MAE to 9.78 kcal/mol MAE and 13.26 kcal/mol RMSE on the 35 CSD test structures.

model	training MAE	CSD MAE (kcal/mol)	CSD RMSE
single ANN	1.52	8.55	13.61
10-model ensemble	-	8.95	14.76
100-model mc-dropout	-	8.53	13.45

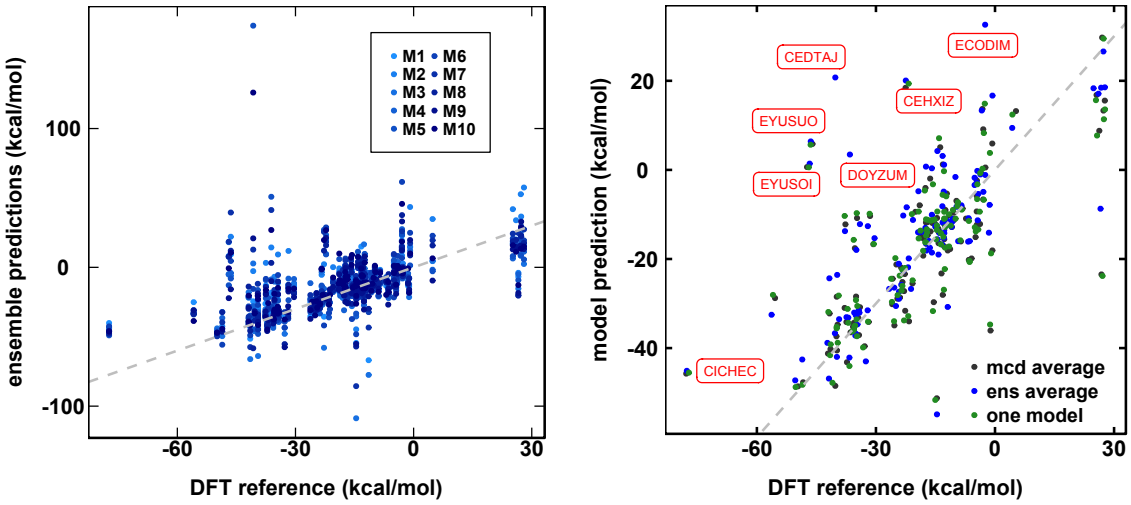

Figure S7: Parity plots of DFT-calculated splitting energy of CSD structures and predictions from a 10-model ensemble (left) and a single model (green), the average of the 10-model ensemble (blue) and the average of 100 mc-dropout realizations (charcoal) compared (right). The parity line is shown as a dashed gray line, while the CSD codes for high error (≥ 30 kcal/mol) points are shown in red. All units are kcal/mol.

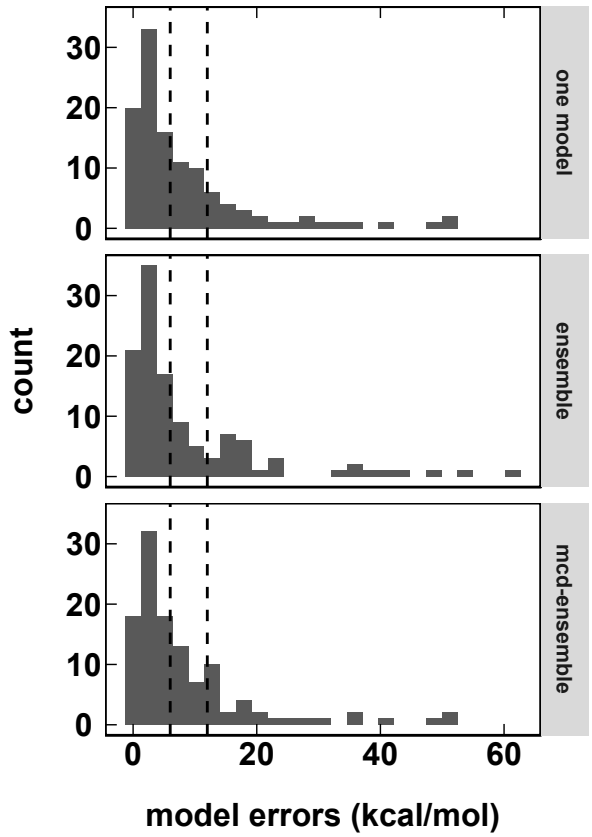

Figure S8: Distribution of a errors for CSD prediction task from a single model, the average of an ensemble of 10 models and the average of 100 Monte-Carlo dropout realizations of the single model. Dashed vertical lines show nominal tolerances of 6 and 12 kcal/mol. All error units are kcal/mol.

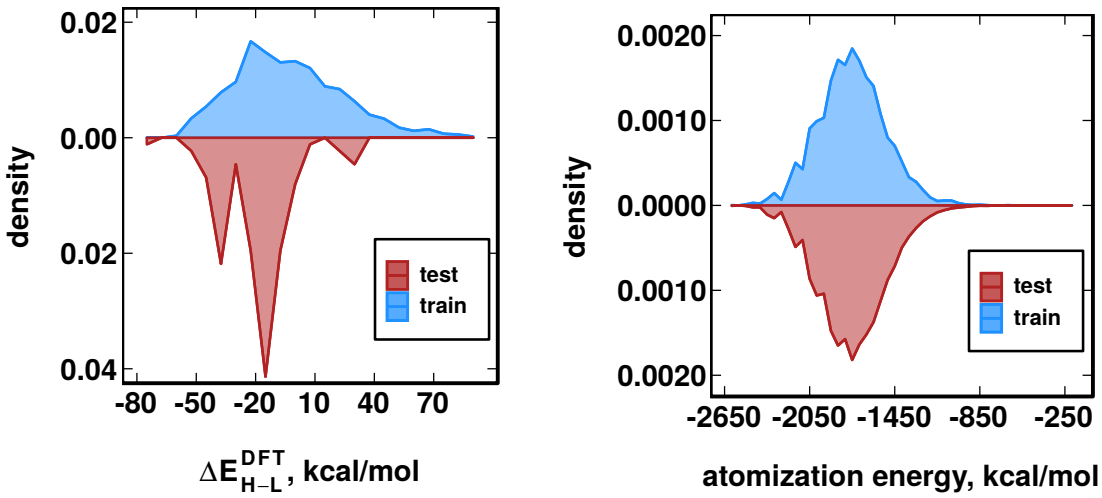

Figure S9: Comparison of train and test distributions for inorganic spin splitting task with 1900 training points and CSD test data (left) and QM9 atomization energy task with uniform random 5% training data and the remaining 95% used as test (right). All units are kcal/mol.

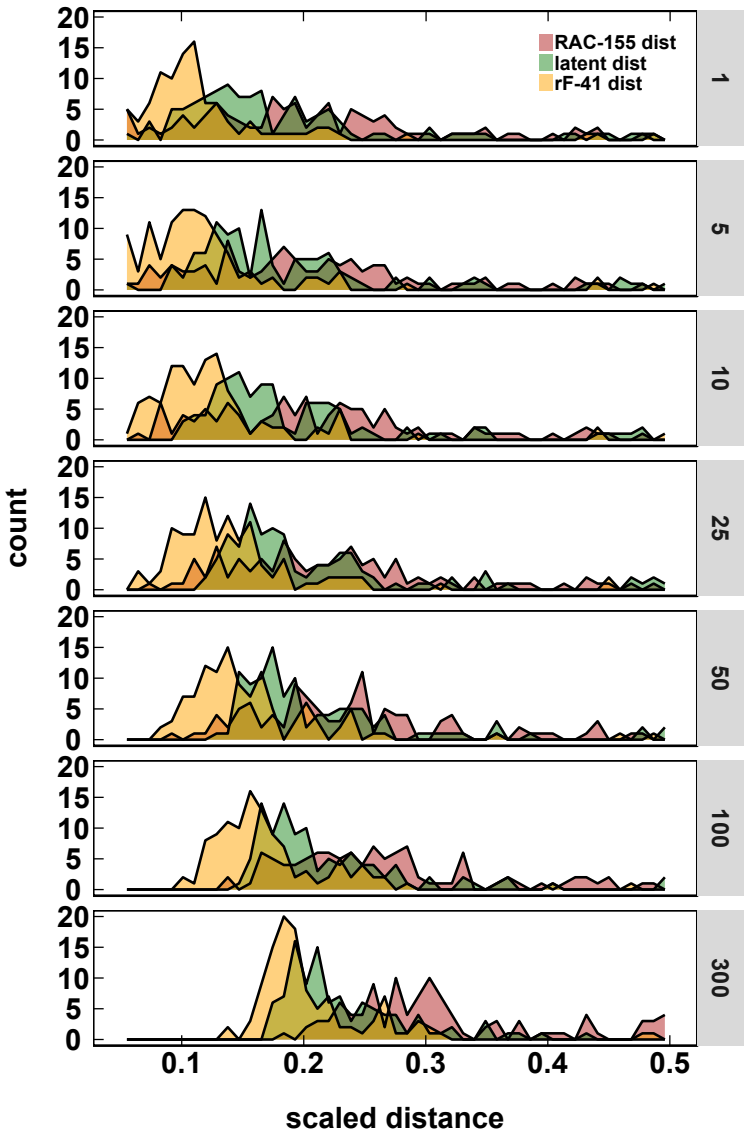

Figure S10: Distribution of average distance to nearest training data as a function of number of neighbors over which the distance is averaged for 1 to 300 neighbors (as labeled on each graph) for the CSD prediction task, showing three different distance metrics: RAC-155, random forest 41-feature subset of RAC-155 (rF-41), and latent. Distances are normalized to $[0, 1]$ for comparison and truncated to the region $[0, 0.5]$. Similarity of complexes in feature space (e.g., the simple Euclidean distance in feature space or a cheminformatic similarity metric such as the Tanimoto distance) can be measured to the nearest training point or averaged over multiple training points. Using nearest neighbor data only is likely sensitive to outlier training data, whereas using all training data will likely overestimate distances for new molecules supported by a relatively small amount of training data. Although we previously found good success in both using a single nearest neighbor or over 5-10 nearest neighbors, we now compare potential effects of nearest neighbor averaging on distance distributions. Feature space distances may not be a good proxy for chemical similarity and this approach also ignores automatic feature-engineering that occurs in complex models (e.g., multi-layer neural networks). Furthermore, high-dimensional feature spaces may contain weakly informative features that can ”pollute” isotropic distance metrics.

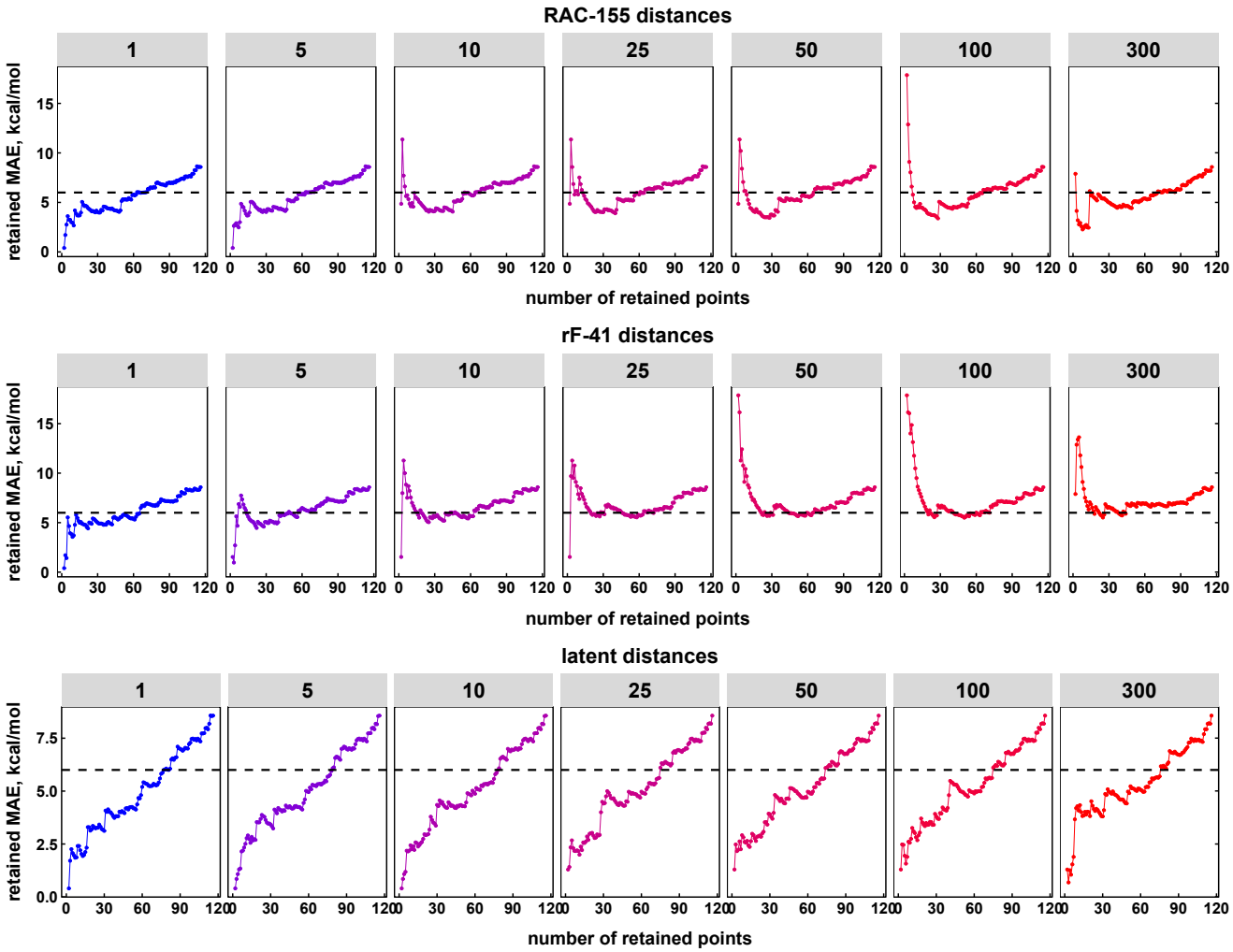

Figure S11: Mean absolute spin splitting error (MAE) as a function of number of retained points for thresholds set using different distances: RAC-155 (top),41-feature subset of RAC-155 selected with random forest (i.e., rF-41, middle), and latent (bottom), averaged over different numbers of nearest neighbors from 1 to 300 (in panels). Depending on how conservatively the boundary between trustworthy chemical space and untrustworthy chemical space is set, we include more or less test data. we therefore consider using each distance and the number of neighbors it is averaged over as a decision boundary and examine how error of retained points varies. Using feature space distances, the effect of nearest neighbors used in the average is most significant for highly conservative decisions that retain less than 20 of the 116 CSD cases. Feature space distances are generally poor at effectively classifying low error points. For intermediate data retention, feature-space-derived models are less sensitive to number of nearest neighbors and generall in agreement with each other. Latent space distance shows the least nearest neighbor dependence. Distances are normalized to $[0, 1]$ for comparison. The horizontal black line represents a nominal error tolerance of 6 kcal/mol.

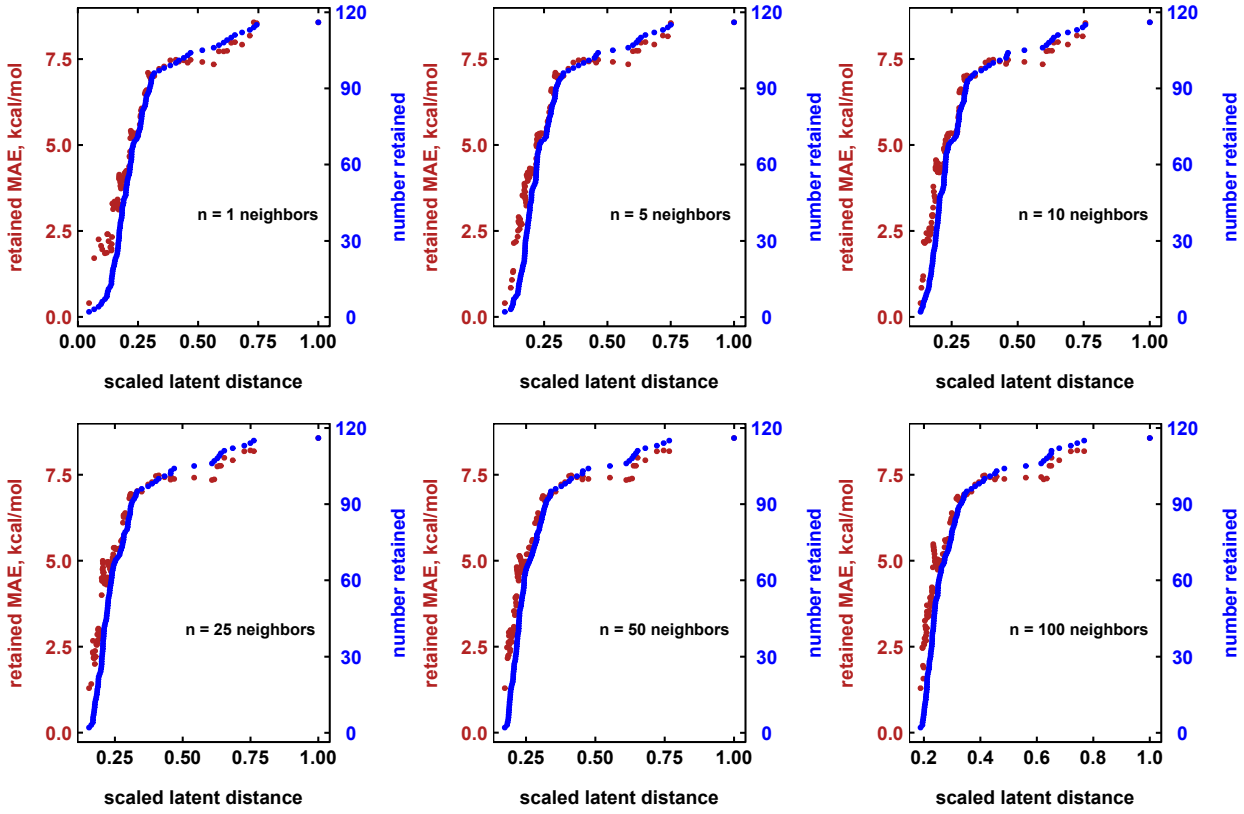

Figure S12: Mean absolute spin splitting error (MAE) on retained data and number of retained candidates as a function of threshold latent distance to nearest training points, averaged over 1 to 100 nearest neighbors. Distances are normalized to $[0, 1]$ for comparison.

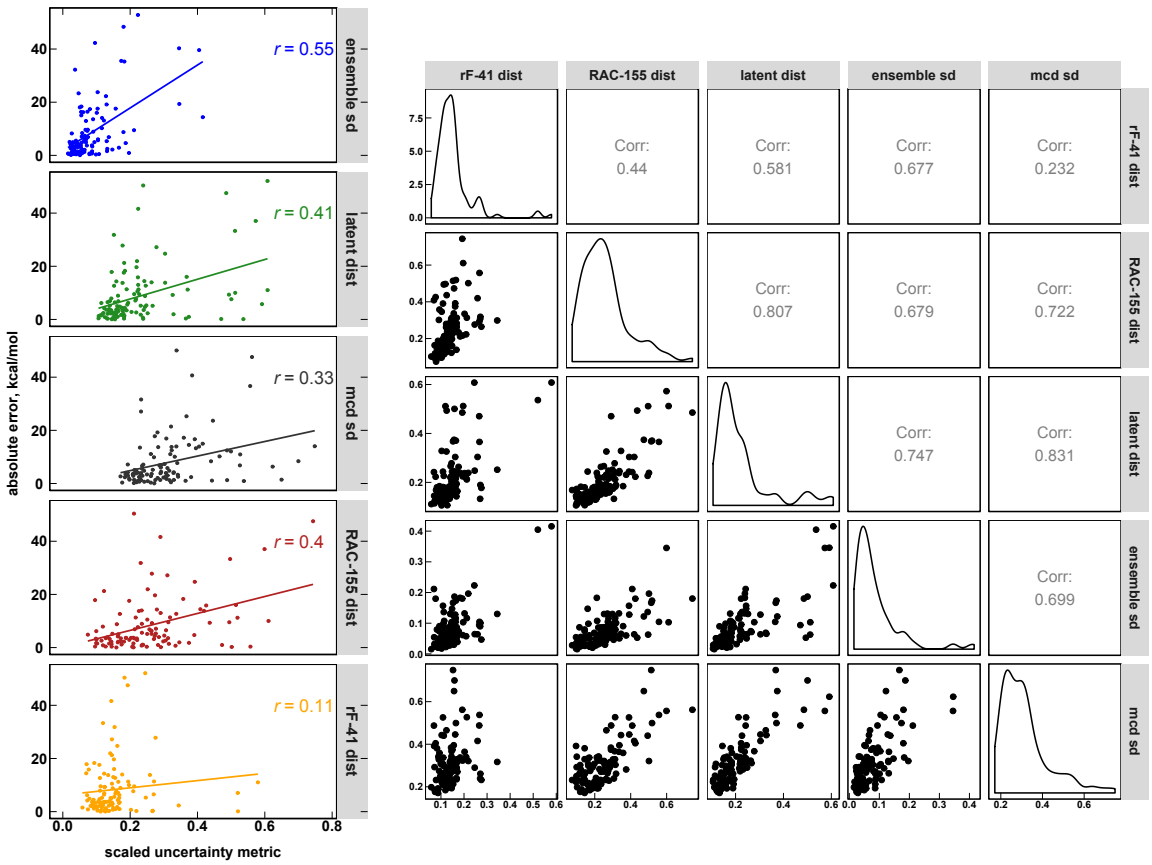

Figure S13: Correlation between different uncertainty metrics (panels) and absolute model errors on CSD data (left), showing the correlation coefficient inset along with best fit line and (right) showing all pairwise cross-correlations and distributions of uncertainty metrics. Metrics shown are the standard deviations from 10-model ensemble, 10-neighbor average latent distance, standard deviation of of 100 mc-dropout realizations, 10-neighbor average feature space distance using RAC-155 and rF-41 representations. All units are kcal/mol and all metrics are normalized to $[0, 1]$ for comparison. We truncate the plot at 0.75 to remove the few outlying points at extreme distances for clarity, excluding 1 ensemble point, 1 latent distance point, 7 mc-dropout points, 6 RAC-155 distance points, and 2 rF-41 points.

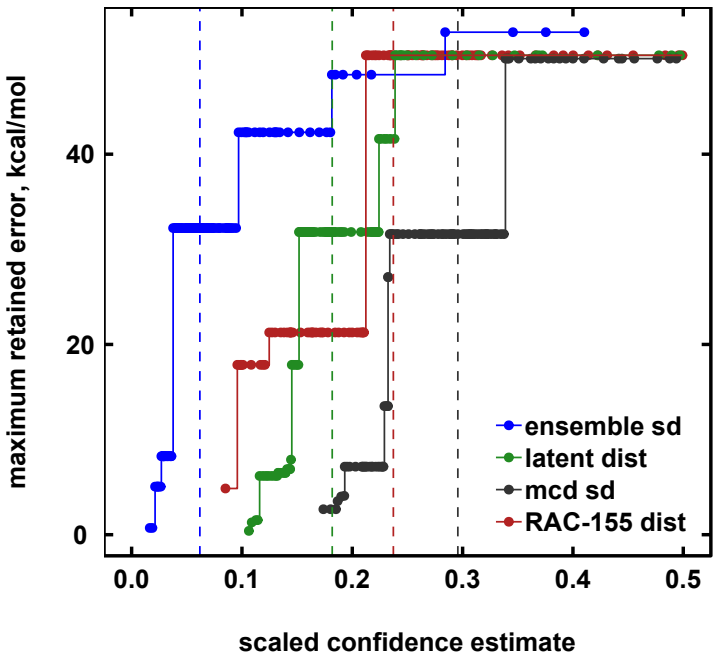

Figure S14: Variation in the maximum ANN error (in kcal/mol) for retained points on CSD data as a function of thresholds in different uncertainty metrics, showing that the largest errors can be effectively avoided by truncating with respect to latent distance and ensemble metrtics but not raw distances. Compared metrics are the 10-neighbor average distance to training data in both feature (RAC-155) and latent spaces, the standard deviation of a 10-model ensemble and the standard deviation of a 100 realizations of a mc-dropout ensemble. All metrics are normalized to $[0, 1]$ for comparison. Vertical lines indicate the median of each scaled metric.

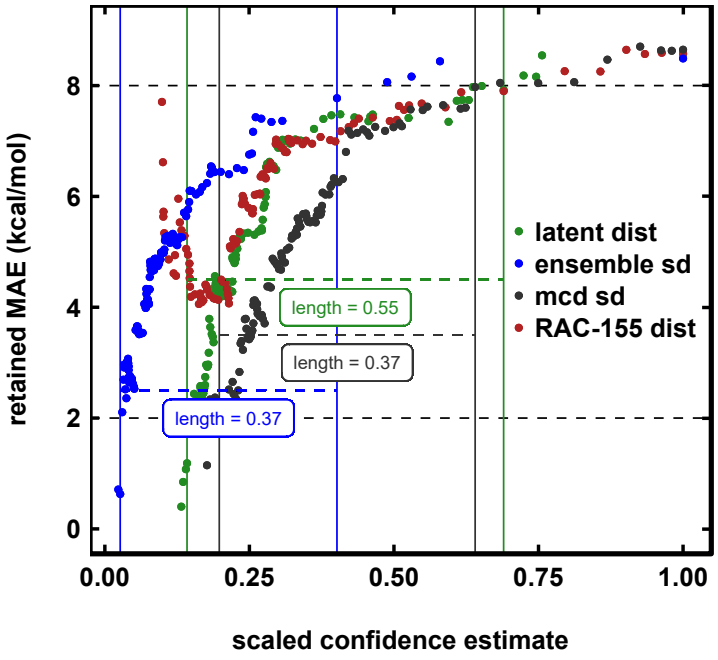

Figure S15: Variation in the mean absolute error (MAE, in kcal/mol) from ANN models for retained points on CSD data as a function of thresholds in different uncertainty metrics, showing that 1) average retained errors can be controlled with all metrics and 2) different metrics show different sharpness in response to chaging thresholds, as indicated by the annotation showing the length of the interval from MAE$= 2$ kcal/mol to MAE$= 8$ kcal/mol with a horizontal line. The interval for each model is also marked with solid vertical lines. Metrics compared are the 10-neighbor average distance in both feature (RAC-155) and latent spaces, the standard deviation of a 10-model ensemble and the standard deviation of 100 realizations of a mc-dropout ensemble. All metrics are normalized to $[0, 1]$ for comparison. Annotation is not provided for RAC-155 owing to non-monotonic behavior at low thresholds.

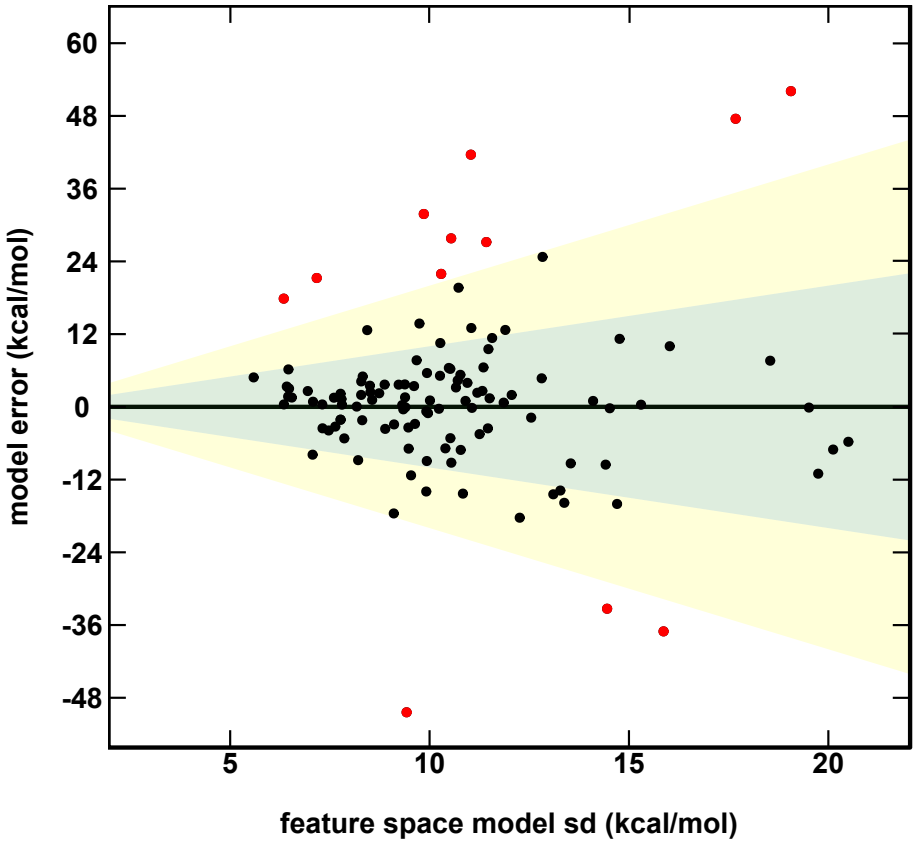

Figure S16: Relationship between spin-splitting ANN model errors (in kcal/mol) on a 116 molecule CSD set and a calibrated distance-based uncertainty model using feature space distances. The model is fit using eq. (1) from the main text with $\sigma_1 = 0$, $\sigma_2 = 3.42$. The translucent green region corresponds to one std. dev. and translucent yellow to two std. dev.. The points with model errors that lie inside either of these two bounds are shown in black, and the percentage within the green or yellow regions are annotated in each graph in green and yellow, respectively. The points outside two std. dev. are colored red.

Table S3: CSD accession codes for points used to calibrate latent-distance uncertainty model.

ABORIU	ADEQAE	AGUDOW	CAKCIA	CEFDIC
CERZEE	CICHEC	CIGTET	COAQAC01	COMTED02
DEDKII	DEFWUJ	DUCBIN	EBUSEB	ECADOB
ECOWEZ	EKOTUV	ELAHII	EZIROU	FEHPYO

Table S4: Values for σ_1 and σ_2 in latent-distance uncertainty model calibrated using maximum likelihood estimation on 5 different random samples of 20 CSD points. The bold values in the first row indicate those used in the rest of this work, corresponding to accession codes given in Table S3

repeat	σ_1 (kcal/mol)	σ_2
1	$\mathbf{4.57 \times 10^{-9}}$	**3.20**
2	2.24×10^{-8}	3.12
3	1.61×10^{-8}	2.95
4	8.93×10^{-9}	3.22
5	2.58×10^{-8}	4.16

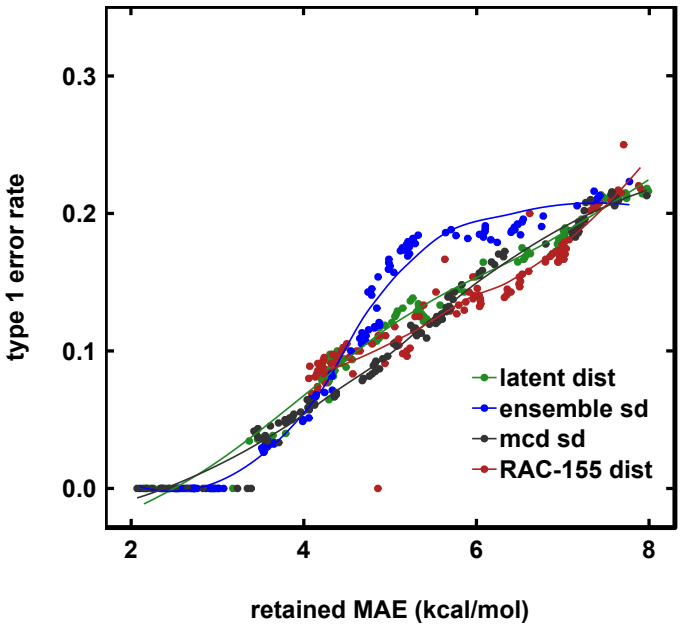

Figure S17: Comparison of type I error rate, defined as the fraction of retained points with absolute errors > 12 kcal/mol from ANN models, as a function of the mean absolute retained error when setting thresholds in different uncertainty metrics. Compared metrics are the 10-neighbor average distance to training data in both feature (RAC-155) and latent spaces, the standard deviation of a 10-model ensemble and the standard deviation of 100 realizations of mc-dropout. A smoothing spline is shown for each metric. Higher error rates are observed for 10-model ensemble for retained MAEs between 5 and 7 kcal/mol.

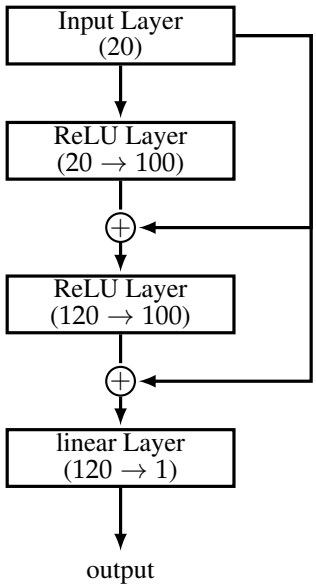

Figure S18: Neural network architecture used for QM9 prediction task, showing two fully-connected layers with input pass-through connections. The size of each mapping is shown in parentheses under the layer name. The $\oplus$ symbol represents concatenation. Dropout and batch normalization are applied to the ReLU layers.

Table S5: Hyperparameters and topology for organic atomization energy ANN on QM9 benchmark.

parameter	value
layer 1 size	100
layer 2 size	100
activation function	relu
learning rate	0.00033
optimizer	adam
β_1	0.9945
β_2	0.9936
decay	0
dropout (all hidden)	0.053
batch size	128
epochs	800
L^2 regularization	1.32818E-8
semibatch normalization	yes
early stopping	none

Table S6: Comparison of single-model performance of QM9 atomization ANNs with two hidden layers of 100 nodes, using residual archietcture (original), without residual links at the same hyperparameters and without residual links after reoptimizing hyperparameters using hyperopt. The reoptimized hyperparameters are the same as in Table S5 except for learning rate $= 0.00196$, $\beta_1 = 0.9694$, $\beta_2 = 0.9779$, decay$= 0$, L^2 regularization $= 2.33317 \times 10^{-9}$.

model	training RMSE	test MAE (kcal/mol)	test RMSE
original	6.24	6.79	9.97
no residual links	18.32	15.28	19.60
hyperparameter reoptimized	6.97	8.58	11.80

Table S7: Repetition test showing train and test mean absolute errors (MAE) for atomization energy prediction on QM9 data using 100 different 5% training data samples. In all cases, all points not in the training set are used as test. Average and standard deviations are given at the end of the table.

	train MAE (kca/mol)	test MAE (kcal/mol)		train MAE (kca/mol)	test MAE (kcal/mol)		train MAE (kca/mol)	test MAE (kcal/mol)		train MAE (kca/mol)	test MAE (kcal/mol)
0	4.84	7.06	1	4.83	7.10	2	4.60	6.95	3	4.68	6.84
4	4.37	6.73	5	4.44	6.76	6	4.44	6.89	7	4.77	7.11
8	4.17	6.75	9	4.57	6.71	10	4.55	6.9	11	4.82	6.94
12	4.44	6.80	13	4.53	6.79	14	4.59	7.04	15	4.55	6.92
16	4.80	7.15	17	4.84	7.0	18	4.32	6.72	19	4.69	7.03
20	4.39	6.77	21	5.00	7.08	22	4.98	7.0	23	5.05	6.94
24	4.52	6.93	25	4.62	6.94	26	4.63	6.79	27	4.34	6.55
28	4.54	6.96	29	4.50	6.79	30	4.99	7.03	31	4.30	6.69
32	4.56	7.02	33	4.70	6.83	34	4.78	6.87	35	4.50	6.69
36	4.59	7.02	37	4.27	6.78	38	4.43	6.96	39	4.34	6.84
40	4.59	6.96	41	4.83	6.87	42	4.46	6.68	43	4.82	7.13
44	4.59	6.99	45	4.72	6.84	46	4.38	6.7	47	4.63	7.07
48	4.52	6.97	49	4.81	6.93	50	4.49	6.95	51	4.41	6.73
52	4.38	6.81	53	5.57	6.95	54	4.30	6.67	55	5.08	7.05
56	4.35	6.81	57	4.80	7.02	58	4.65	6.96	59	4.32	6.78
60	4.41	6.87	61	4.73	7.20	62	4.76	7.21	63	4.77	7.08
64	4.23	6.63	65	4.76	6.86	66	4.42	6.82	67	4.49	6.91
68	4.77	6.96	69	4.41	6.81	70	5.06	7.19	71	4.85	7.23
72	4.54	6.78	73	4.52	6.80	74	4.57	6.96	75	4.56	7.00
76	4.46	6.77	77	4.99	7.11	78	4.63	6.79	79	4.26	6.84
80	4.63	7.01	81	4.48	6.75	82	4.55	6.86	83	4.68	6.90
84	4.31	6.60	85	4.42	6.85	86	4.53	6.86	87	4.30	6.68
88	4.34	6.84	89	4.24	6.65	90	4.33	6.58	91	4.57	6.85
92	4.34	6.77	93	4.54	7.07	94	4.36	6.8	95	4.65	7.07
96	4.56	6.79	97	4.73	6.85	98	4.66	6.89	99	5.08	7.10
average train MAE = 4.59 kcal/mol sd train MAE = 0.23 kcal/mol						average test MAE = 6.89 kcal/mol sd test MAE = 0.15 kcal/mol					

Table S8: Mean absolute error (MAE) and root-mean square error (RMSE) metrics for QM9 atomization energy ANN trained on a random 5% of data tested on the remaining 127217 points. Errors are shown from a single model and the average of an ensemble of 10 models. All error units are kcal/mol.

model	training RMSE	test MAE (kcal/mol)	test RMSE
single ANN	6.24	6.79	9.97
10-model ensemble	-	6.13	9.14

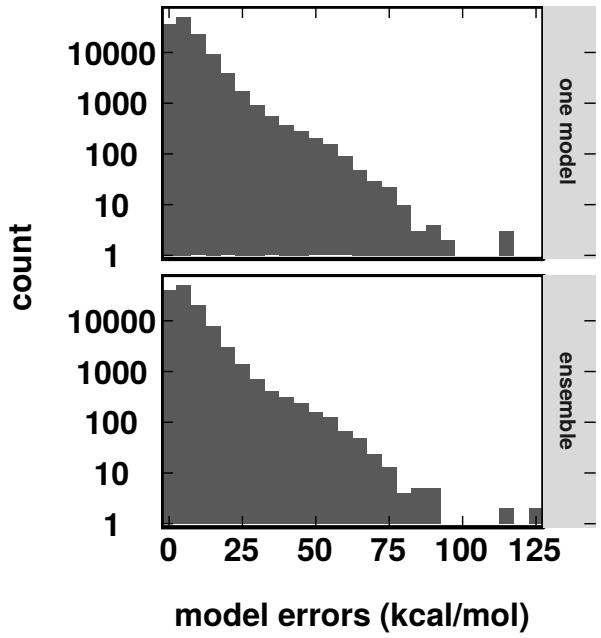

Figure S19: Distribution of a errors for QM9 atomization prediction task from a single model and the average of an ensemble of 10 models. All error units are kcal/mol and counts are shown on a log y-axis. The maximum error for a single model is 119.97 kcal/mol and 124.10 kcal/mol for the ensemble model. These large errors are observed on the SMILES strings FC(F)(F)CC(F)(F)F (hexafluoropropane) and CC1N2C3C4=CCC13C24 (a cyclic tertiary amine), respectively.

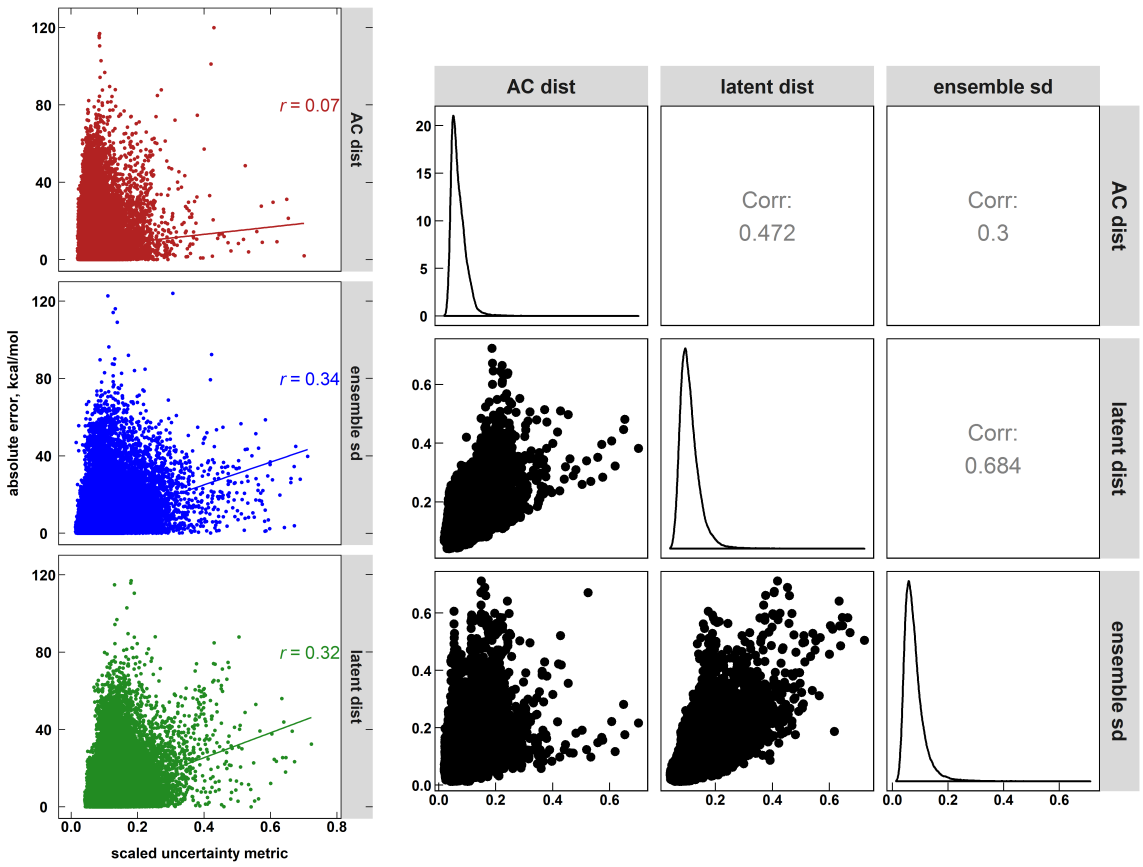

Figure S20: Correlation between different uncertainty metrics (panels) and absolute model errors on QM9 atomization energy test data (left), showing the correlation coefficient inset along with best fit line and (right) showing all pairwise cross-correlations and distributions of uncertainty metrics. Metrics shown are the standard deviations from 10-model ensemble and 10-neighbor average latent distance and 10-neighbor average feature space distance using AC representations. All units are kcal/mol and all metrics are normalized to $[0, 1]$ for comparison. We truncate the plot at 0.75 to remove the few outlying points at extreme distances for clarity, excluding 8 ensemble point, 3 latent distance point and 6 AC distance points.

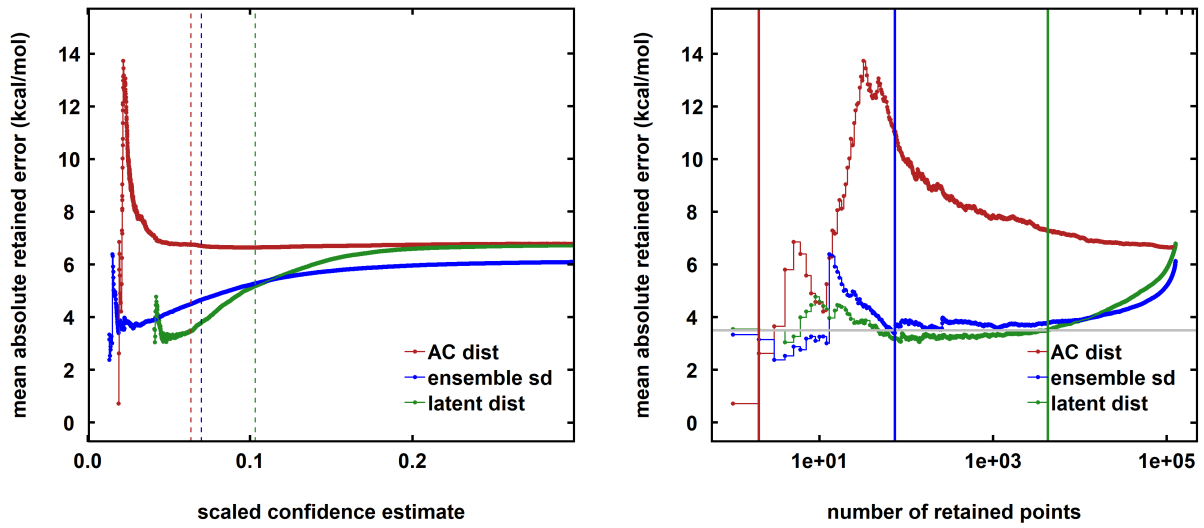

Figure S21: Variation in the mean absolute error (MAE, in kcal/mol) on retained points from ANN models on QM9 atomization energy data as a function of (left) the thresholds in different uncertainty metrics and (right) the number of points retained. The metrics compared are the 10-neighbor average distance in both feature (AC) and latent spaces and the standard deviation of a 10-model ensemble. We also plot the maximum number of retained points before the retained MAE (right) crosses a 3.5 kcal/mol threshold (horizontal gray line) with solid vertical lines at 2, 74 and 4299 points for AC distances, ensembles and latent distances respectively. All metrics are normalized to $[0, 1]$ for comparison. Dashed vertical lines show the median of each metric (left).

Table S9: Values for σ_1 and σ_2 in latent-distance uncertainty model calibrated using maximum likelihood estimation on different numbers of random samples of QM9 test points. For each number of points, we present the mean and standard deviation over 10 random samples. The bold values in the last row indicate the single sample with 500 points used in the rest of this work. Thus, the conclusion is that only 500 points from $>$ 120k are needed to calibrate parameters, indicating the proposed model learns this mapping easily from sparse data.

	σ_1		σ_2	
# of points	mean	std	mean	std
	(kcal/mol)			
100	0.325	0.0053	15.70	0.528
500	1.71×10^{-7}	3.43×10^{-7}	4.54	0.313
1000	1.39×10^{-7}	2.20×10^{-7}	4.40	0.120
5000	8.08×10^{-7}	1.28×10^{-6}	4.41	0.0787
10000	2.04×10^{-7}	3.23×10^{-7}	4.45	0.0446
25000	6.93×10^{-7}	1.49×10^{-6}	4.45	0.0284
50000	7.02×10^{-7}	1.84×10^{-6}	4.45	0.0289
100000	3.36×10^{-7}	6.61×10^{-7}	4.46	0.011
500	$\mathbf{1.79 \times 10^{-6}}$		**4.45**	

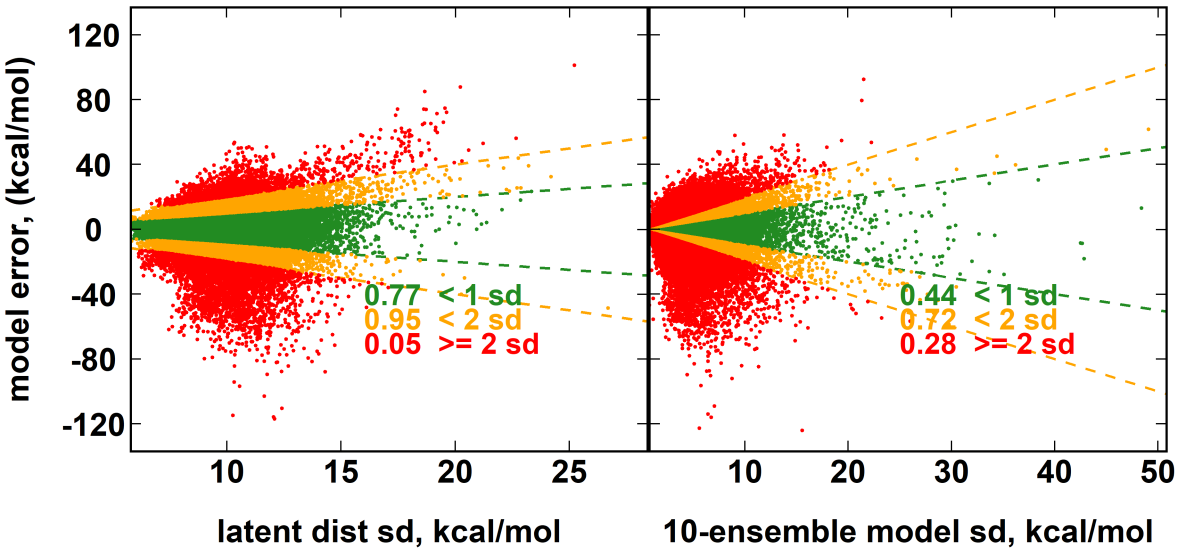

Figure S22: Relationship between model errors and different uncertainty metrics for QM9 atomization energy ANN on test-set points. Standard deviations from calibrated latent distance model using 500 points (left) and 10-model ensemble (right) are compared, with points lying in one (two) sd colored green (yellow). Points outside two sd are colored red. Dotted lines indicate one and two standard deviations. Error units are kcal/mol.

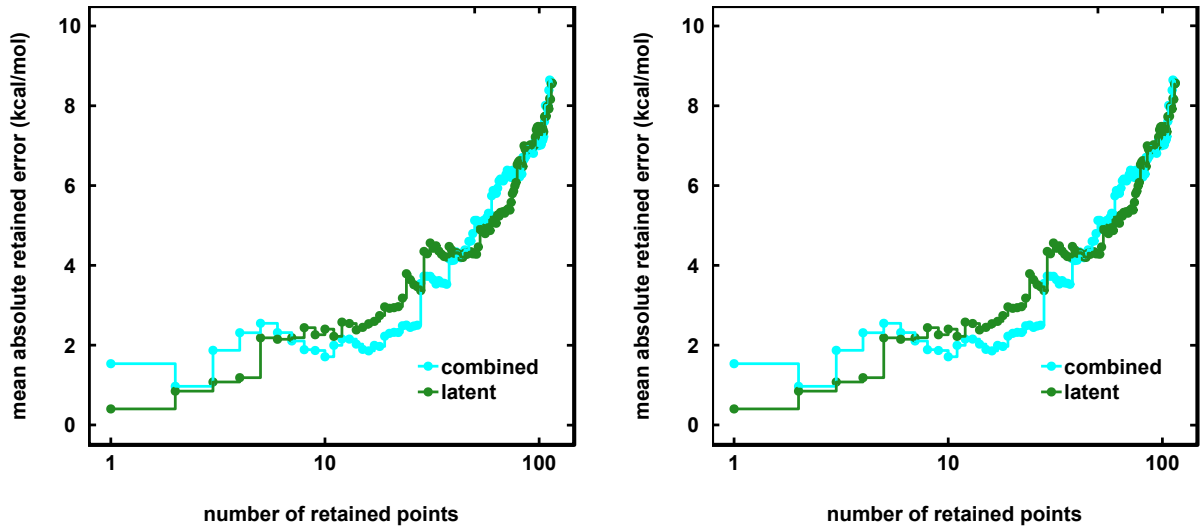

Figure S23: Variation in the mean absolute error (MAE, in kcal/mol) on retained points from ANN models on CSD splitting energy prediction task as a function of (left) the thresholds in different uncertainty metrics and (right) the number of points retained. The metrics compared are the *minimum of the combination of the* 10*-neighbor average distance in latent space and the standard deviation of a 10-model ensemble and 100 realizations mc-dropout* and as well as the 10-neighbor average distance in latent space alone. All metrics are normalized to $[0, 1]$ for comparison. Dashed vertical lines show the median of each metric (left). Errors are taken from single-ANN predictions only. It is apparent the minimum of the combined metrics can provide marginally better error control over some of the range, though latent distances alone perform better or equivalent for retained MAE values $\gtrapprox$ 4.00 kcal/mol

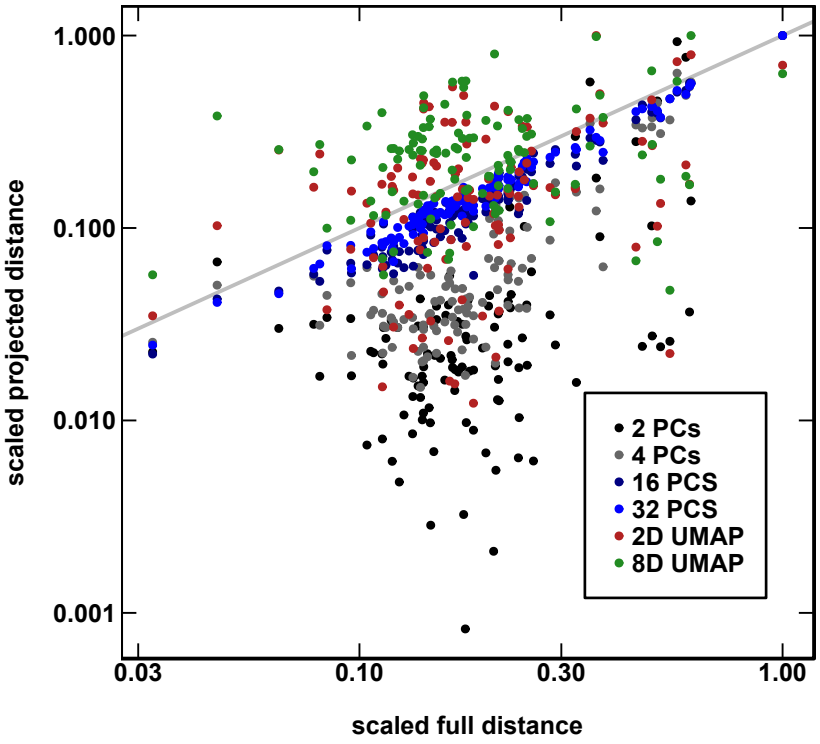

Figure S24: Distance to nearest training point for CSD points in the full spin splitting ANN latent space and low dimensional spaces from principal component analysis (PCA) and uniform manifold approximation (UMAP). In all cases the dimensionality reduction is conducted based on the training data only. All distances are normalized by the largest value and the gray line shows parity.

Table S10: Spearman (rank) correlation between distances from CSD points and nearest training data in the full spin splitting ANN latent space and low dimensional spaces from principal component analysis (PCA) and uniform manifold approximation (UMAP). In all cases the dimensionality reduction is conducted based on training data only.

method	Spearman correlation
2D PCA	0.32
4D PCA	0.68
16D PCA	0.93
32D PCA	0.97
2D UMAP	0.32
8D UMAP	0.17

Table S11: CSD active learning experiment: mean absolute error (MAE) and root-mean square error (RMSE) metrics for out-of-sample CSD prediction task with a single model for the original 116 points, after removing the 10 points with lowest model confidence determined with different metrics and then after retraining with the 10 excluded points. Uncertainty metrics compared are the latent distance of 10-model ensemble metrics. All errors are in kcal/mol.

original single ANN		data selection method	10 removed		retrained	
MAE	RMSE		MAE	RMSE	MAE	RMSE
(kcal/mol)			(kcal/mol)			
		latent distance	7.73	12.22	7.10	10.62
8.55	13.61	10-model ensemble	7.61	11.65	7.56	10.87
		mcd-ensemble	7.57	11.79	7.46	11.39

Text S3: Application to MNIST and Fashion-MNIST classification task

In order to test the application of out proposed method to other tasks, we consider two standard benchmark test image classification tasks, MNIST[8] and Fashion-MNIST[13]. Both consist of 60k training and 10k test grayscale images of size 28×28 pixels divided into 10 classes. We use a convolutional neural network (CNN) with the same hyperparameters for each task, trained with cross entropy loss and no explicit regularization (Table S12).

Training on MNIST gives a train/test accuracy (top-1) of $100.00\%/99.06\%$ ($0/94$ errors), while training on Fashion-MNIST gives a train/test accuracy (top-1) of $99.87\%/91.51\%$ ($77/849$ errors).

As before, we average the distance of each test point to the nearest 10 training points to generate a confidence metric (Figure S25). Comparison of the distribution of correctly and incorrectly classified points reveals a shift towards high distance for the incorrectly classified points, with an increase in mean distance of the incorrect points of 66.64% for MNIST and 11.90% for Fashion-MNIST. We perform a Mann–Whitney test to estimate if the difference in distances is significant and find $p = 9.3 \times 10^{-47}$ and $p = 1.12 \times 10^{-36}$ for MNIST and Fashion-MNIST respectively, although in both cases the number incorrect samples is low.

This suggests that the methods proposed could be applied to other types of the neural networks (CNNs), datasets (images) and tasks (classification).

Table S12: Hyperparameters and topology for image classification CNN

parameter	value
layer 1	64 filter 3×3 2D convolution
layer 2	32 3×3 2D convolution
layer 3	64 unit dense
layer 4	64 unit dense
layer 5	10 unit softmax
activation function	relu
learning rate	0.01
optimizer	adam
β_1	0.9
β_2	0.999
decay	0
dropout (all hidden)	none
batch size	128
epochs	50
L^2 regularization	none
semibatch normalization	no
early stopping	none

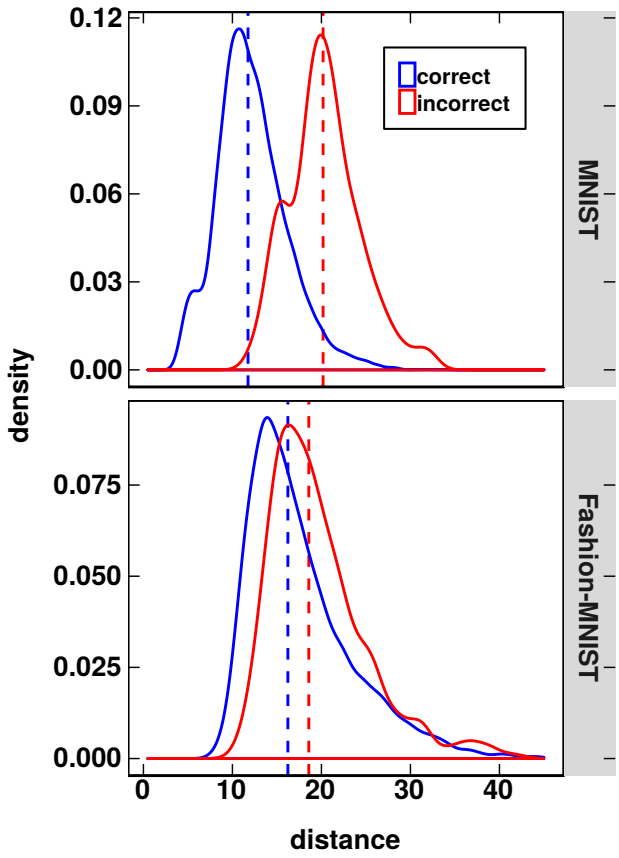

Figure S25: Comparison of kernel density estimates for 10-neighbor average latent distance to training data for image classification task using a CNN, showing correctly (blue) and incorrectly (red) classified points for MNIST (top) and Fashion-MNIST (bottom) benchmarks. The dashed vertical lines represent the median values for each curve.

Table S13: Hyperparameters and topology for inorganic spin splitting ANN.

parameter	value
layer 1 size	200
layer 2 size	200
layer 3 size	200
activation function	relu
learning rate	0.00163
optimizer	sgd
momentum	0.998
decay	0.0015719
Nesterov acceleration	yes
dropout (all hidden)	0.0825
batch size	128
epochs	2000
L^2 regularization	7.101148E-14
semibatch normalization	yes
early stopping	none

References

1. Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv e-prints*, art. arXiv:1506.02142, Jun 2015.

2. Efthymios I Ioannidis, Terry ZH Gani, and Heather J Kulik. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.*, 37(22):2106–2117, 2016.

3. Jon Paul Janet and Heather J Kulik. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

4. Jon Paul Janet and Heather J. Kulik. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.*, 8:5137–5152, 2017.

5. Jon Paul Janet, Terry Z. H. Gani, Adam H. Steeves, Efthymios I. Ioannidis, and Heather J. Kulik. Leveraging cheminformatics strategies for inorganic discovery: Application to redox potential design. *Ind. Eng. Chem. Res.*, 56(17):4898–4910, 2017.

6. Jon Paul Janet, Lydia Chan, and Heather J Kulik. Accelerating chemical discovery with machine learning: simulated evolution of spin crossover complexes with an artificial neural network. *J. Phys. Chem. Lett.*, 9(5):1064–1071, 2018.

7. Johannes Kästner, Joanne M Carr, Thomas W Keal, Walter Thiel, Adrian Wander, and Paul Sherwood. DL-FIND: an open-source geometry optimizer for atomistic simulations. *J. Phys. Chem. A*, 113(43):11856–11865, 2009.

8. Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

9. Aditya Nandy, Chenru Duan, Jon Paul Janet, Stefan Gugler, and Heather J. Kulik. Strategies and software for machine learning accelerated discovery in transition metal chemistry. *Ind. Eng. Chem. Res.*, 57(42):13973–13986, 2018.

10. V. R. Saunders and I. H. Hillier. A “Level–Shifting” method for converging closed shell Hartree–Fock wave functions. *International Journal of Quantum Chemistry*, 7 (4):699–705, 1973. doi: $10.1002/\mathrm{qua}.560070407$.

11. Ivan S Ufimtsev and Todd J Martinez. Quantum chemistry on graphical processing units. 3. analytical energy gradients, geometry optimization, and first principles molecular dynamics. *J. Chem. Theory Comput.*, 5(10):2619–2628, 2009.

12. Lee-Ping Wang and Chenchen Song. Geometry optimization made simple with translation and rotation coordinates. *J. Chem. Phys.*, 144(21):214108, 2016.

13. Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *Computing Research Repository*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.