A quantitative uncertainty metric controls error in neural network-driven chemical discovery†

Jon Paul Janet,†a Chenru Duan,†ab Tzuhsiung Yang,a Aditya Nandyb and Heather J. Kulika,b

Machine learning (ML) models, such as artificial neural networks, have emerged as a complement to high-throughput screening, enabling characterization of new compounds in seconds instead of hours. The promise of ML models to enable large-scale chemical space exploration can only be realized if it is straightforward to identify when molecules and materials are outside the model’s domain of applicability. Established uncertainty metrics for neural network models are either costly to obtain (e.g., ensemble models) or rely on feature engineering (e.g., feature space distances), and each has limitations in estimating prediction errors for chemical space exploration. We introduce the distance to available data in the latent space of a neural network ML model as a low-cost, quantitative uncertainty metric that works for both inorganic and organic chemistry. The calibrated performance of this approach exceeds widely used uncertainty metrics and is readily applied to models of increasing complexity at no additional cost. Tightening latent distance cutoffs systematically drives down predicted model errors below training errors, thus enabling predictive error control in chemical discovery or identification of useful data points for active learning.

1. Introduction

Machine learning (ML) models for property prediction have emerged as powerful complements to high-throughput computation and experiment enabling the prediction of properties in seconds rather than the hours to days that direct observations would require. Using large data sets, trained interpolative potentials and property prediction models have achieved chemical accuracy with respect to the underlying data. Predictive models hold great promise in the discovery of new catalysts and materials enabling researchers to overcome combinatorial challenges in chemical space exploration. While application of ML to chemical space exploration is increasingly becoming a reality, a key outstanding challenge remains in knowing in which regions of chemical space a trained ML model may be confidently applied.

While trained ML models are fast to deploy to large compound spaces, many models (e.g., artificial neural networks or ANNs) are typically trained only after acquisition of thousands to millions of data points. Quantitative uncertainty metrics are most critical in applications of active learning where the model is improved by acquisition of selected data. Although some models (e.g., Gaussian process regression) inherently provide estimates of model uncertainty, uncertainty quantification for models suited to handle large data sets (e.g., ANNs) remains an active area of research.

One approach to estimating model uncertainty is to train an ensemble of identical architecture models on distinct partitions of training data to provide both a mean prediction and associated variance (Fig. 1). While widely employed in the chemistry community, ensembles increase the model training effort in proportion to the number of models used (typically an order of magnitude, ESI Text S1†). Although this additional effort...
ensemble uncertainty estimates have been most frequently embedding map their feature space distance (metrics considered in this work. Two points are compared in terms of those using many convolutional models that are increasingly used in chemical discovery, such as iterative retraining for active learning or for more complex discovery, we found that dropout-generated credible intervals added cost of model re-training. In transition metal complex modest cost of running the model multiple times rather than the ff predictions provides an e

... from the model (ESI Text S1 run repeatedly with varied dropout masks, randomly eliminating nodes from the model (ESI Text S1). The standard ANN architecture (middle) is compared at bottom for Monte-Carlo dropout (i.e., zeroed out nodes) and ensemble models (i.e., varied model weights) at bottom left and right.

Fig. 1 Schematic of an ANN annotated with the four uncertainty metrics considered in this work. Two points are compared in terms of their feature space distance (i.e., the difference between two points in the molecular representation) on a t-distributed stochastic neighbor embedding map of data in the input layer (top, left, annotations in orange) and the latent space distance (i.e., the difference between two points in the final layer latent space) on a t-SNE of the data in the last layer (top, right, annotations in green). The standard ANN architecture (middle) is compared at bottom for Monte-Carlo dropout (i.e., zeroed out nodes) and ensemble models (i.e., varied model weights) at bottom left and right.

may be practical for some models (e.g., networks with only a few layers), the training effort becomes cost-prohibitive during iterative retraining for active learning or for more complex models that are increasingly used in chemical discovery, such as those using many convolutional or recurrent layers. Thus, ensemble uncertainty estimates have been most frequently applied in the context of simpler networks, especially in neural network potentials that are trained in a one-shot manner. A key failing of ensemble metrics is that with sufficient model damping (e.g., by L2 regularization), variance over models can approach zero for compounds very distant from training data, leading to over-confidence in model predictions.

Another approach to obtain model-derived variances in dropout-regularized neural networks is Monte Carlo dropout (mc-dropout) (Fig. 1). In mc-dropout, a single trained model is run repeatedly with varied dropout masks, randomly eliminating nodes from the model (ESI Text S1). The variance over these predictions provides an effective credible interval with the modest cost of running the model multiple times rather than the added cost of model re-training. In transition metal complex discovery, we found that dropout-generated credible intervals provided a good estimate of errors on a set aside test partition but were over-confident when applied to more diverse transition metal complexes. Consistent with the ensembles and mc-dropout estimates, uncertainty in ANNs can be interpreted by taking a Bayesian view of weight uncertainty where a prior is assumed over the distribution of weights of the ANN and then updated upon observing data, giving a distribution over possible models. However, if the distribution of the new test data is distinct from training data, as is expected in chemical discovery, this viewpoint on model uncertainty may be incomplete.

A final class of widely applied uncertainty metrics employs distances in feature space of the test molecule to available training data to provide an estimate of molecular similarity and thus model applicability. The advantages of feature space distances are that they are easily interpreted, may be rapidly computed, and are readily applied regardless of the regression model (Fig. 1). We used high feature space distances to successfully reduce model prediction errors on retained points while still discovering new transition metal complexes. Limitations of this approach are that the molecular representation must be carefully engineered such that distance in feature space is representative of distance in property space, the relationship between distance cutoff and high property uncertainty must be manually chosen, and this metric cannot be applied to message-passing models that learn representations.

A chief advantage of multi-layer neural network models over simpler ML models is that successive layers act to automatically engineer features, limiting the effect of weakly-informative features that otherwise distort distances in the feature space (Fig. 1). Thus, for multi-layer ANNs, feature-based proximity can be very different from the intrinsic relationship between points in the model. Such ideas have been explored in generative modeling where distances in auto-encoded latent representations have informed chemical diversity and in anomaly detection with separate models (e.g., autoencoders or nearest-neighbor classifiers) have enabled identification of ‘poisoned’ input data. However, the relationship between latent space properties and feature space properties has not been exploited or understood in the context of error estimation for property prediction (i.e., regression) ML models.

In this work, we propose the distance in latent space, i.e., the distance of a test point to the closest training set point or points in the final layer latent space, as a new uncertainty metric (Fig. 1). The advantages of this approach are that (i) it introduces no overhead into model training or evaluation, (ii) it can work just as easily with both simple and complex ANN models that have been used for chemical property prediction (e.g., hierarchical, recurrent, or convolutional), and (iii) it naturally ignores distances corresponding to features to which the model prediction is insensitive, obviating the need for feature engineering to develop an estimate of test point proximity to prior training data. We show that these attributes yield superior performance over other metrics in chemical discovery.

2. Results & discussion

To demonstrate the advantages of the latent space distance metric in a quantitative fashion, we compare to three
established uncertainty metrics. This assessment is particularly motivated by the nature of chemical discovery applications,\(^9\) where data set sizes are often smaller and have more broadly varying chemistry than typical applications in neural network potentials\(^{19,40}\) or in quantitative structure–property relationships in cheminformatics.\(^{41,52}\) To mimic chemical discovery efforts, we train neural networks to predict transition metal complex spin state energetics\(^7\) and test them on diverse transition metal complexes from experimental databases. To confirm the generality of our observations, we also compare uncertainty estimates for neural network models trained on a very small subset (\(i.e., 5\%) of QM9,\(^{31}\) a widely used\(^{22,65,70-75}\) data set in organic chemistry ML.

For open shell transition metal chemistry, we use 1901 equilibrium high (H)/low (L) spin splitting energies (\(i.e., \Delta E_{H-L}\)) for octahedral first-row transition metal (\(i.e., M(n)\) or \(M(n)\)) complexes generated in prior work\(^7,8\) using density functional theory (DFT). We use the previously introduced\(^7\) full set of revised autocorrelation (RACs) descriptors (\(i.e., \text{RAC}-155\)) to train a fully connected ANN with three 200-node hidden layers (see Computational Details and ESI Text S2, Table S1, and Fig. S1\(^\dagger\)). RACs have been demonstrated for training predictive models of transition metal complex properties,\(^7,8,25,26\) including spin splitting, metal–ligand bond length, redox and ionization potentials, and likelihood of simulation success.

To mimic chemical discovery application of this model, we extracted a set of 116 octahedral, first-row transition metal complexes that have been characterized experimentally (\(i.e., \text{from the Cambridge Structural Database or CSD}^{77}\)) as an out-of-sample test set (Fig. 2, ESI Text S2 and Fig. S2–S5\(^\dagger\)). We selected these CSD complexes to be intentionally distinct from training data, as is apparent from principal component analysis (PCA) in the RAC-155 (ref. 7) representation (Fig. 2). Several complexes in the CSD test set fall outside the convex hull of the training data in the first two principal components (\(ca. 50\%\) of the variance) and are distant from training data, as judged by the Euclidean distance in the full RAC-155 feature space (Fig. 2 and ESI Fig. S6\(^\dagger\)). High distances are observed for complexes containing elements rarely present (\(e.g., \text{an S/N macrocycle for a Co(n)}\) complex, CSD ID: FATJIT) or completely absent from our training data (\(e.g., B\) in boronated dipyrrole ligands of the Fe(n) complex CSD ID: ECODIM and as in thioarsenite ligands in an Mn(n) complex, CSD ID: CEDTAJ) as well as ligand topologies (\(e.g., \text{acrylamide axial ligands in an Mn(n) complex, CSD ID: EYUSUO}\)) not present in training data (Fig. 2).

Due to the distinct nature of the CSD test set from the original training data, the 8.6 kcal mol\(^{-1}\) mean absolute error (MAE) of the RAC-155 ANN on the CSD data set is much larger than the 1.5 kcal mol\(^{-1}\) training set MAE (Fig. 2 and ESI Table S2\(^\dagger\)). Use of ensemble- or mc-dropout-averaged predictions unexpectedly\(^7\) worsens or does not improve test MAEs (ensemble: 9.0 kcal mol\(^{-1}\); mc-dropout: 8.5 kcal mol\(^{-1}\)), which we attribute to noise in averaging due to the relatively heterogeneous training data (ESI Fig. S7–S9\(^\dagger\)). The relative error increase on diverse data is consistent with our prior work where we achieved low errors on test set partitions of 1–3 kcal mol\(^{-1}\) (ref. 7) that increased\(^7\) to around 10 kcal mol\(^{-1}\) on sets of diverse molecules (\(e.g., 35\) molecules from a prior curation\(^7\) of the CSD\(^7\)). These observations held across feature sets (\(e.g., \text{MCDL-25 vs. RAC-155}\) and model architectures\(^7,8\) (\(e.g., \text{kernel ridge regression vs. ANNs}\)) for \(\Delta E_{H-L}\) property prediction.

Despite the increase in MAE, errors are not uniformly high across the 116 molecules in our new CSD data set (Fig. 2). A significant number (24 or 21\%) of the complexes have errors within the 1.5 kcal mol\(^{-1}\) training MAE, a substantial fraction are within the 3 kcal mol\(^{-1}\) test set error described in prior work\(^7\) (41 or 35\%), and a majority (61 or 53\%) have errors

---

**Fig. 2** (left) Comparison of inorganic training and CSD test data in the dominant two principal components of the RAC-155 representation of the training data set. The density of training data is shown as gray squares shaded as indicated in inset count colorbar. CSD test data points are shown as circles colored by the 10-nearest-neighbor-averaged Euclidean distance in RAC-155 space, as shown in dist. inset color bar. Four representative high-distance structures are shown in circle insets in ball and stick representations: (top left inset, CSD ID: FATJIT) a Co(n) complex with S/N macrocycle and axial Br\(^-\) ligands, (top right inset, CSD ID: EYUSUO) Mn(n) tetra-chlorophenyl-porphyrin with acrylamide axial ligands, (bottom left inset, CSD ID: CEDTAJ) a Mn(n) complex with thioarsenite ligands, and (bottom right inset, CSD ID: ECODIM) an Fe(n) complex with boronated dipyrrole and thiolated phenanthrene ligands. (right) Distribution of absolute CSD test set model errors for \(\Delta E_{H-L}\) (in kcal mol\(^{-1}\), bins: 2.5 kcal mol\(^{-1}\)) with the MAE annotated as a green vertical bar and the cumulative count shown in blue according to the axis on the right.
5 kcal mol$^{-1}$ or below (Fig. 2 and ESI†). At the same time, a number of outlier compounds have very large absolute errors with 31 (27%) above 10 kcal mol$^{-1}$ and 12 (10%) above 20 kcal mol$^{-1}$ (Fig. 2 and ESI†). Large errors are due to both underestimation of $\Delta E_{\text{H-L}}$ by the ANN (e.g., Fe(n) complex CSD ID: CEYSAA, $\Delta E_{\text{H-L,ANN}} = -23.8$ kcal mol$^{-1}$, $\Delta E_{\text{H-L,DFT}} = 26.6$ kcal mol$^{-1}$) and overestimation (CSD ID: Mn[m] complex CSD ID: EYUSUO, $\Delta E_{\text{H-L,ANN}} = 5.7$ kcal mol$^{-1}$, $\Delta E_{\text{H-L,DFT}} = -46.4$ kcal mol$^{-1}$, see Fig. 2). Given the heterogeneity of observed errors, we apply uncertainty metrics to this data set with the aim to (i) systematically drive down error on predicted data points by only making predictions within the model’s domain of applicability and (ii) identify data points that should be characterized and incorporated into the model training set in an active learning setting.

For heavily engineered feature sets (i.e., MCDL-25 (ref. 7)), we showed the Euclidean norm feature space distance to the closest training point could be used to control ANN errors in inorganic complex discovery,7* typically limiting discovery MAEs to only slightly larger (i.e., 4–5 kcal mol$^{-1}$) than the original test MAE. This approach required that we select a cutoff over which distances were deemed too high, a quantity that can be sensitive to the nature of the feature set and the number of nearest neighbors used in the average (ESI Fig. S10 and S11†). Averaging Euclidean norm distances in RAC-155 (ref. 7) or a feature-selected subset7+ over the nearest (i.e., 1–10) neighbors in the training data and only predicting on points sufficiently close to training data systematically eliminates the highest error points (ESI Fig. S11†). Consistent with prior work,7,8 this approach allows us to achieve sub-6 kcal mol$^{-1}$ MAE on over half (64 of 116) points in the CSD set, but further improvement of predicted-data MAEs below 5 kcal mol$^{-1}$ is not possible (ESI Fig. S11†).

In the large, non-engineered feature spaces typically used as input to neural networks, feature space distances may be insufficient for identifying when predictions lack support by data in the model. Thus, we turn to the latent space distance evaluated at the final hidden layer (Fig. 1). Using high distances in latent space as the criterion for prediction uncertainty, we drive down MAEs on predicted data nearly monotonically, well below the 5 kcal mol$^{-1}$ MAE that could be achieved using feature space distances (ESI Fig. S11†). This difference in performance is motivated by the distinct, higher effective dimensionality of the principal components in the latent space over the feature space (ESI Fig. S6†). With the distance in latent space as our guide, 76 points can be identified as falling within model domain of applicability (i.e., sub-6 kcal mol$^{-1}$ MAE), and 3 kcal mol$^{-1}$ MAE can be achieved on over 25% of the data (ca. 30 points), indicating a close relationship between high latent space distance and model error (ESI Fig. S11–S13†). The distance in latent space has the added advantage of being less sensitive to the number of nearest neighbors over which the distance evaluation is carried out than feature space distances (ESI Fig. S11†). Our approach is general and not restricted to the distance in the latent space described here. In future work, we could move beyond potential ambiguities79 in measuring high-dimensional similarity with Euclidean distances and compare to alternatives, including averaged properties55 or those that incorporate other geometric features of the latent data distribution.

Having confirmed that distances in latent space provide significant advantages over feature space distances at no additional cost, we also would like to consider the performance with respect to mc-dropout and ensemble-based uncertainty metrics (ESI Fig. S14 and S15†). To do so, we overcome the key inconvenience that the distance measure itself does not provide an error estimate in the units of the property being predicted. After model training, we calibrate the error estimate by fitting the predictive variance to a simple conditional Gaussian distribution of the error, $\epsilon$, for a point at latent space distance, $d$:

$$\epsilon(d) \sim N(0, \sigma_1^2 + d\sigma_2^2)$$

where the error is assumed to be normal with a baseline $\sigma_1^2$ term and a growing term $d\sigma_2^2$. Selection of $\sigma_1$ and $\sigma_2$ using a simple maximum likelihood estimator on a small subset (ca. 20 points) of the CSD test set is robust, leading to property-derived uncertainties (Fig. 3, ESI Fig. S16, Tables S3 and S4†). Over the 116-complex CSD test set, this latent space-derived metric spans a large 8–24 kcal mol$^{-1}$ range and correlates to absolute model errors as strongly as ensemble and mc-dropout standard deviation (std. dev.) metrics (ESI Fig. S13†).

Although not unique and dependent on the training process of the model, the distance in latent space-derived energetic uncertainties provide a superior bound on high error points (Fig. 3). Observed errors reside within one std. dev. in the majority (77%) of cases, and only a small fraction (8%) exceed two std. dev. ranges (Fig. 3). In comparison, less than half of errors are within one std. dev. evaluated from the ensemble (44%) or mc-dropout (37%), and a significant fraction of errors exceed two std. dev. (23% and 34%, respectively, Fig. 3). When the ensemble or mc-dropout uncertainty metrics are used as cutoffs to decide if predictions should be made, model over-confidence leads to inclusion of more high error (i.e., >12 kcal mol$^{-1}$) points than when using the latent distance (ESI Fig. S17†). The ability to smoothly transition between high cutoffs where more points are characterized with the ML model (e.g., to achieve 8 kcal mol$^{-1}$ MAE) vs. conservative where the error is small (e.g., 2 kcal mol$^{-1}$) but only a minority of predictions are made is important for predictive control; here, the latent distance provides the more robust separation between these two regimes, thus enabling greater distinction between the two (ESI Fig. S15†).

There are numerous cases where both ensemble and mc-dropout are relatively confident on very high error points in comparison to latent distance. For example, an Fe(n) complex with ethanimine and alkanamine ligands (CSD ID: DOQRAC) is predicted erroneously by the model to be strongly high spin ($\Delta E_{\text{H-L,ANN}} = -34.7$ kcal mol$^{-1}$ vs. $\Delta E_{\text{H-L,DFT}} = -1.4$ kcal mol$^{-1}$), but this point has a low std. dev. from the ensemble (4.3 kcal mol$^{-1}$) in comparison to a relatively high 17.2 kcal mol$^{-1}$ std. dev. from the latent space distance. Conversely, there are no cases where the latent distance uncertainty is uniquely over-confident, but there are cases...
where all metrics are overconfident. For example, an Mn(n) complex with four equatorial water ligands and two axial, oxygen-coordinating 4-pyridinone ligands is expected by all metrics to be reasonably well predicted (std. dev. ensemble = 2.5 kcal mol$^{-1}$, mc-dropout = 2.7 kcal mol$^{-1}$, and latent space = 9.4 kcal mol$^{-1}$), but the DFT preference for the high-spin state is underestimated by the ANN ($\Delta E_{H-L,ANN} = -45.5$ kcal mol$^{-1}$ vs. $\Delta E_{H-L,DFT} = -77.4$ kcal mol$^{-1}$). Although the latent distance error estimate does not bound all high error points predicted by the model, it provides a high fidelity, no cost uncertainty estimate for $>90\%$ of the data.

To assess the generality of our observations on inorganic complexes for other chemical data sets, we briefly consider the approach applied to atomization energies computed with hybrid DFT (i.e., B3LYP$^{79-82}$/6-31G$^*$) for a set of organic (i.e., C, H, N, O, and F-containing) small molecules. The QM9 data set$^{79}$ consists of 134k organic molecules with up to 9 heavy atoms and has been widely used as a benchmark for atomistic machine learning model development,$^{22,76-78}$ with the best models in the literature reporting MAEs well below 1 kcal mol$^{-1}$. As in previous work,$^7$ we employ standard autocorrelations (ACs)$^{44}$ that encode heuristic features$^{46}$ on the molecular graph and perform well (ca. 6 kcal mol$^{-1}$ MAE) even on small ($\sim$10%) training set partitions for QM9 atomization energies,$^7$ exceeding prior performance from other connectivity-only featureizations.$^{79}$ For this work, we trained a two-hidden layer residual ANN using AC features and passing the input layer forward in a ResNet-like architecture$^{86}$ to improve performance over a fully-connected architecture (Computational Details and ESI Fig. S18, Tables S5 and S6$^{†}$). We use only 5% (6614) of the data points for training, reserving the remaining 127k molecules for our test set to mimic chemical discovery in a single random partition, the choice of which does not influence overall performance (ESI Table S7$^{†}$).

Baseline model performance for QM9 atomization energies with the ANN is improved over our prior work for both train (4.6 kcal mol$^{-1}$) and test (6.8 kcal mol$^{-1}$) MAE, with some further improvement of test MAE with an ensemble model (6.1 kcal mol$^{-1}$, see ESI Tables S7 and S8$^{†}$). A wide distribution of errors is observed with some outlier points such as hexafluoropropane (error = 120 kcal mol$^{-1}$) having very large errors for both the single and ensemble models (ESI Fig. S19$^{†}$). For the residual ANN, the mc-dropout uncertainty has not been derived, and so we compare only the other three uncertainty metrics. We observe ensemble and latent space distance uncertainty metrics to have similar correlations to model errors and both to outperform feature space distance in this regard (ESI Fig. S20$^{†}$). Selecting either the distance in latent space or ensemble uncertainty as a cutoff, we systematically drive down MAEs on the predicted data fraction, and latent distance again provides superior control when error tolerance is low (ESI Fig. S21$^{†}$). For example, setting a tolerance of 3.3 kcal mol$^{-1}$ for the MAE leads to a pool of over 4200 points retained with the latent space distance metric vs. few points (74) for the ensemble std. dev. (ESI Fig. S21$^{†}$).

We again observe that the AC feature space distance is a poor indicator of increasing model errors, with as many high error points occurring at low distances as at high distances (Fig. 4). In contrast to feature space distance, ensemble std. dev. and latent distance both grow with increasing error (Fig. 4). Calibration of the latent space distance to the output property enables direct comparison to ensemble uncertainties (ESI Table S9$^{†}$). As in the inorganic data set, the ensemble std. dev. values are overconfident, capturing a smaller amount (44%) of the errors within a single std. dev. in comparison to the distance in latent space (77%) metric (Fig. 4 and ESI Fig. S22$^{†}$). For the ensemble uncertainty, a significant fraction (28%) of points have errors larger than twice the std. dev., whereas only a small fraction (5%) do so for the distance in latent space (Fig. 4 and ESI Fig. S22$^{†}$).

For both the CSD test set and the QM9 set, a systematic reduction in baseline error can be observed in a practical use case where the user adjusts the applied uncertainty metric to become more conservative (Fig. 5). Smooth reductions in MAE on data inside the uncertainty cutoffs can be achieved across a wide range of latent distance cutoffs, with errors nearly
monotonically approaching the training data MAE, which may be recognized as a qualitative lower bound on our test set error (Fig. 5). Combining all error metrics to choose the most conservative result does not improve upon the single latent space distance metric (ESI Fig. S23†). PCA or uniform manifold approximation and projection (UMAP)87 analysis of the latent space distance indicates that a large number of the latent space dimensions are needed for error estimation (ESI Fig. S24 and Table S10†). For either data set, at the point on which half of all possible predictions are made, predicted data MAE is less than half of that for the excluded points (Fig. 5).

The latent distance also shows promise for application in active learning, where a model is trained iteratively by acquiring data in regions of high model uncertainty. To mimic such an application in the context of inorganic chemistry, we returned to the CSD data set and identified the 10 least confident points based on the distance in latent space, retrained the ANN using the same protocol, and re-evaluated model MAE (ESI Table S11†). Incorporating these data points during retraining reduced model errors from 8.6 to 7.1 kcal mol\(^{-1}\), whereas simply removing these points only reduced model MAE to 7.7 kcal mol\(^{-1}\) (ESI Table S11†). This effect is particularly significant considering the relatively small change in the number of data points (i.e., 10 added to 1901 or 0.5%) and an even larger reduction in root mean square error is observed (ESI Table S11†). When compared to an ensemble or MC-dropout cutoff, selection of retraining points based on latent space distance results in the largest reduction in model MAE while also simultaneously only requiring a single model retraining (ESI Table S11†).

Although we have focused on applications in chemical discovery with fully connected neural networks, application to other network architectures is straightforward. We trained convolutional neural networks for image classification tasks on two standard benchmarks, MNIST88 and Fashion-MNIST.89 Incorrectly classified images are observed at higher latent distances in both cases (ESI Text S3, Table S12, and Fig. S25†).

3. Conclusions

We have demonstrated on two diverse chemical data sets that the distance in the latent space of a neural network model provides a measure of model confidence that out-performs the best established metrics (i.e., ensembles) at no additional cost beyond single model training. The distance in latent space
provides an improved approach to separating low- and high-confidence points, maximizing the number of retained points for prediction at low error to enable extrapolative application of machine learning models. We introduced a technique to calibrate latent distances that required only a small fraction of out-of-sample data, enabling conversion of this distance-based metric to error estimates in the units of the predicted property. In doing so, >90% of model errors were bounded within 2 std. dev. of latent distance estimates, in significant improvement beyond typically over-confident ensemble estimates. Like ensembles or mc-dropout, the latent space distance could still be challenged by unstable models, such as those trained on highly discontinuous properties. The latent space distance metric is general beyond the examples demonstrated here and is expected to be particularly useful in complex architectures that are normally time-consuming and difficult to train or in active learning approaches where rapid, iterative model retraining may be needed.

4. Computational Details

Neural networks were trained for this work with hyperparameters selected using Hyperopt\textsuperscript{89} followed by manual fine-tuning in Keras\textsuperscript{8} with the Tensorflow\textsuperscript{8} backend (ESI Fig. S17, Tables S5 and S13). Model weights are provided in the ESI.\textsuperscript{1} The ΔE_{0.\textit{H}} energy evaluation protocol for inorganic chemistry training data and the curated CSD\textsuperscript{37} test set used molSimplify\textsuperscript{40,41} to automate hybrid (\textit{i.e.}, B3LYP\textsuperscript{80,82} \textit{a}\textsubscript{u}) DFT calculations, with more details provided in ESI Text S2.\textsuperscript{1} For the organic chemistry test, the QM9 atomization energy data set was obtained from the literature.\textsuperscript{33} In all cases, we normalized the representations and properties to make the training data have zero mean and unit variance. For calculating ensemble properties, we employed 10 sub-models trained on 10-fold cross-validation splits of the training data. For mc-dropout, we used the same 8.25% dropout as in training with 100 realizations, and we employed maximum likelihood to optimize the baseline uncertainty parameter, \(\tau\) (ESI Text S1 and Table S2).\textsuperscript{1} We did not apply mc-dropout to the organic test case because it has not been developed for residual-connectivity networks. For feature space distance, we measured Euclidean distance in the normalized feature space (\textit{e.g.}, RAC-155 (ref. 7)) directly. Features of relevant complexes are provided in the ESI.\textsuperscript{1} For latent distances, we used the latent space after the last hidden layer, which has the dimensionality of the model (\textit{i.e.}, 200 for spin splitting, 120 for the organic model).

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

The authors acknowledge primary support by DARPA grant D18AP00039 for the generation of the latent space uncertainty metric (for C. D., A. N., and J. P. J). Inorganic complex data set construction (for T. Y.) was supported by the Office of Naval Research under grant number N00014-18-1-2434. This work was also supported in part by an AAAS Marion Milligan Mason Award. H. J. K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. The authors thank Adam H. Steeves for providing a critical reading of the manuscript.

References


