Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2019, **10**, 7913

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 11th May 2019
Accepted 11th July 2019

DOI: 10.1039/c9sc02298h

rsc.li/chemical-science

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

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

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[1–8] as powerful complements to high-throughput computation[8–13] and experiment,[14–16] 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[17–21] and property prediction models[1–8] have achieved chemical accuracy with respect to the underlying data.[22] Predictive models hold great promise in the discovery of new catalysts[5,6,23,24] and materials[8,25–31] by 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.[32]

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[33] to millions[17,34] of data points. Quantitative uncertainty metrics are most critical in applications of active learning[35,36] where the model is improved by acquisition of selected data. Although some models (*e.g.*, Gaussian process regression) inherently provide estimates of model uncertainty,[37,38] uncertainty quantification for models suited to handle large data sets (*e.g.*, ANNs) remains an active area of research.[39–41]

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,[19,39,40,42,43] 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

[a]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: hjkulik@mit.edu; Tel: +1-617-253-4584*

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

† Electronic supplementary information (ESI) available: Information about ensembles and mc-dropout procedure, information about DFT methods, information about training ligands and structures, information about CSD test cases, PCA decay plots and model performance information for inorganic dataset, comparison between single ANN and ensemble predictions and distribution of CSD errors, error distributions with different distance metrics and numbers of neighbors, correlation between errors and uncertainty metrics, maximum and average retained inorganic errors as function of uncertain metrics, CSD codes used to calibrate latent distance model, variation in uncertainty model calibration parameters for CSD data, type I error rates and with retained errors, architecture and hyperparameters used for QM9 prediction task, performance results for different architectures on QM9 prediction task, variation in QM9 performance with test/train split, performance and error distribution for QM9 for single ANN and ensembles, correlation between QM9 errors and uncertainty metrics, retained mean QM9 errors with different uncertainty metrics, variation in uncertainty model calibration parameters for QM9 data, distribution of predicted and actual errors for QM9 data, results of active learning experiment, hyperparameters for inorganic ANN (PDF). Summary DFT results and information for training and CSD data; model predictions, errors and uncertainties for all tests; DFT-optimized geometries of training and CSD data; model weights, architectures and scaling data for inorganic and QM9 ANNs (ZIP). See DOI: 10.1039/c9sc02298h

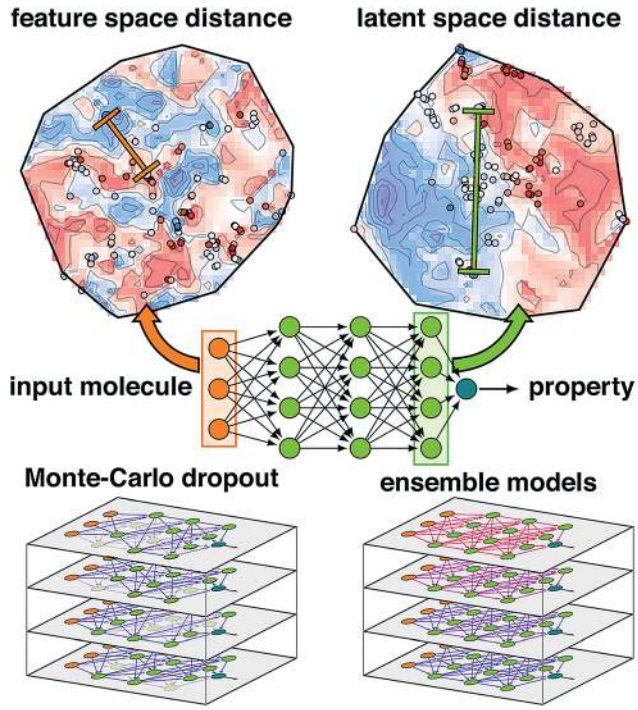

Fig. 1 Schematic of an ANN annotated with the four uncertainty metrics considered in this work. Two points are compared in terms of their feature space distance (*i.e.*, the difference between two points in the molecular representation) on a t-distributed stochastic neighbor embedding map[49] (t-SNE) 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[44] during iterative retraining for active learning or for more complex models that are increasingly used in chemical discovery, such as those using many convolutional[45,46] or recurrent[47,48] layers. Thus, ensemble uncertainty estimates have been most frequently applied[19,40] 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[41] 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)[50] (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.[7,8] 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.[51] 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[7,8,41,52] (Fig. 1). We used[7,8] 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.[53,54]

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[55,56] and in anomaly detection with separate models[57,58] (*e.g.*, autoencoders[59–61] or nearest-neighbor classifiers[62,63]) have enabled identification of 'poisoned' input data.[64] 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,[65] recurrent,[47,48] or convolutional[46,66–69]), 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,[8] 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,[33] 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.*, $\Delta E_{H\text{-}L}$) for octahedral first-row transition metal (*i.e.*, M(II) or M(III) where M = Cr, Mn, Fe, or Co) 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.*, 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†). RACs have been demonstrated for training predictive models of transition metal complex properties,[7,8,25,76] 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.*, from the Cambridge Structural Database or CSD[77]) as an out-of-sample test set (Fig. 2, ESI Text S2 and Fig. S2–S5†). 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†). High distances are observed for complexes containing elements rarely present (*e.g.*, an S/N macrocycle for a Co(II) complex, CSD ID: FATJIT) or completely absent from our training data (*e.g.*, B in boronated dipyrazole ligands of the Fe(II) complex CSD ID: ECODIM and as in thioarsenite ligands in an Mn(II) complex, CSD ID: CEDTAJ) as well as ligand topologies (*e.g.*, acrylamide axial ligands in an Mn(II) 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†). Use of ensemble- or mc-dropout-averaged predictions unexpectedly[78] 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†). 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[77]). These observations held across feature sets[7] (*e.g.*, MCDL-25 *vs.* RAC-155) and model architectures[7,8] (*e.g.*, kernel ridge regression *vs.* ANNs) for $\Delta E_{H\text{-}L}$ 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

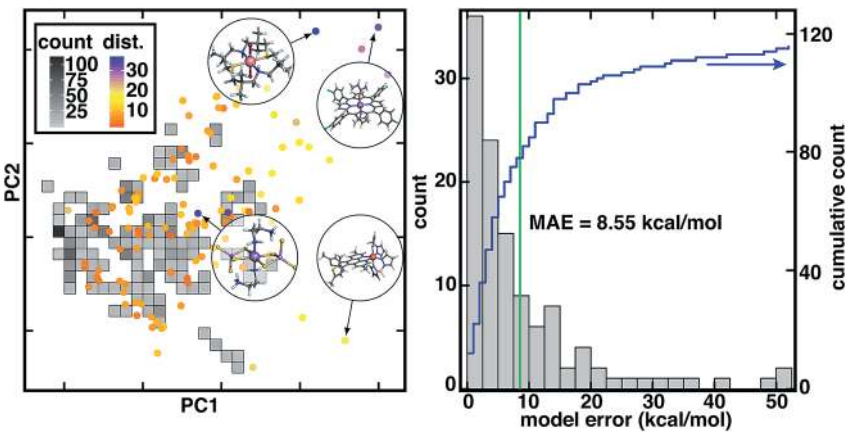

Fig. 2 (left) Comparison of inorganic training and CSD test data in the dominant two principal components of the RAC-155 representation of the training data set. The density of training data is shown as gray squares shaded as indicated in inset count colorbar. CSD test data points are shown as circles 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(II) complex with S/N macrocycle and axial Br^- ligands, (top right inset, CSD ID: EYUSUO) Mn(II) tetra-chlorophenyl-porphyrin with acrylamide axial ligands, (bottom left inset, CSD ID: CEDTAJ) a Mn(II) complex with thioarsenite ligands, and (bottom right inset, CSD ID: ECODIM) an Fe(II) complex with boronated dipyrazole and thiolated phenanthrene ligands. (right) Distribution of absolute CSD test set model errors for $\Delta E_{H\text{-}L}$ (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(II) 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(III) 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,[7,8] 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 subset[7,25] 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 ambiguities[79] in measuring high-dimensional similarity with Euclidean distances and compare to alternatives, including averaged properties[55] 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, ε, for a point at latent space distance, d:

$$\varepsilon(d) \sim \mathcal{N}\left(0,\ \sigma_1{}^2 + d\sigma_2{}^2\right) \quad (1)$$

where the error is assumed to be normal with a baseline $\sigma_1{}^2$ term and a growing term $\sigma_2{}^2$. Selection of σ_1 and σ_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(II) 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

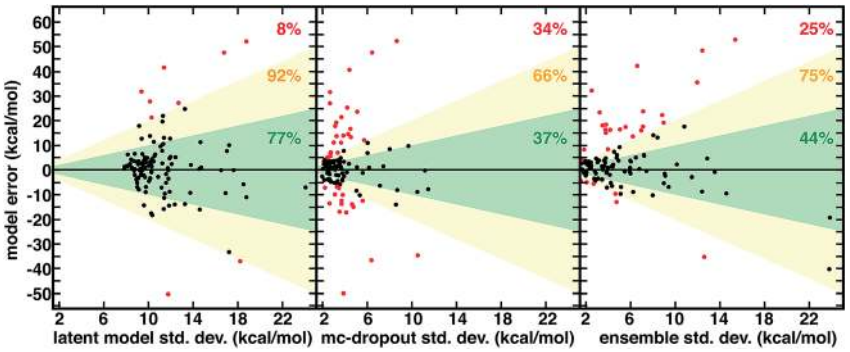

Fig. 3 Relationship between spin-splitting ANN model errors (in kcal mol^{-1}) on a 116 molecule CSD set and three uncertainty metrics all in kcal mol^{-1}: latent model energetic, calibrated std. dev. (left), mc-dropout std. dev. (middle), and 10-model ensemble std. dev. (right). The translucent green 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, and the percentage of points in this group is annotated in each graph in red. Three points are omitted from the ensemble plot to allow for a consistent *x*-axis range.

where all metrics are overconfident. For example, an Mn(II) 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 ($\Delta E_{\text{H-L,ANN}} = -45.5$ kcal mol^{-1} *vs.* $\Delta E_{\text{H-L,DFT}} = -77.4$ 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[80–82]/6-31G[83]) for a set of organic (*i.e.*, C, H, N, O, and F-containing) small molecules. The QM9 data set[33] 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,[22,70–72] with the best models in the literature reporting MAEs well below 1 kcal mol^{-1}.[22,65,70,73–75] As in previous work,[7] we employ standard autocorrelations (ACs)[84] that encode heuristic features[85] on the molecular graph and perform well (*ca.* 6 kcal mol^{-1} MAE) even on small (<10%) training set partitions for QM9 atomization energies,[7] exceeding prior performance from other connectivity-only featurizations.[70] 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 3.5 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

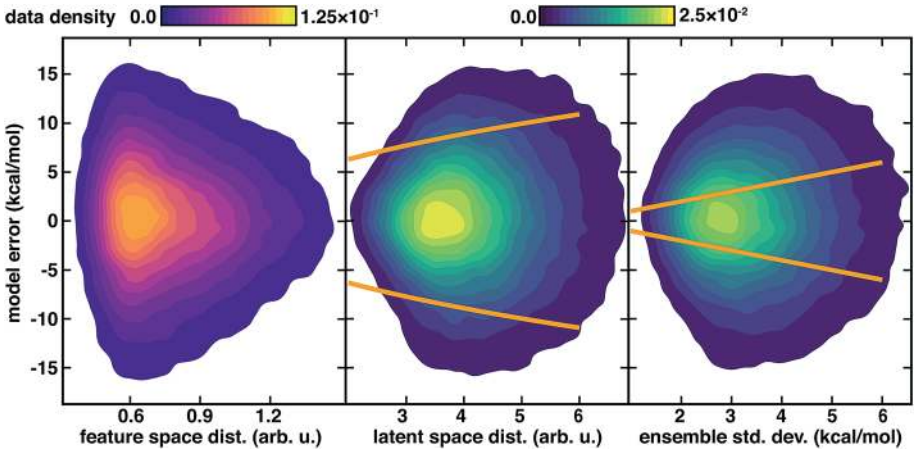

Fig. 4 Model errors (in kcal mol^{-1}) for 127k QM9 atomization energy test points shown as contours as a function of uncertainty metrics. The three uncertainty metrics compared are: feature space distance (in arb. u., left, with top left color bar), latent space distance (in arb. u., middle, with top right color bar), and 10-model ensemble std. dev. (in kcal mol^{-1}, with top right color bar). One standard deviation cutoffs are shown as orange lines for the latent space distance from the calibrated error model (center) and directly from the ensemble (right).

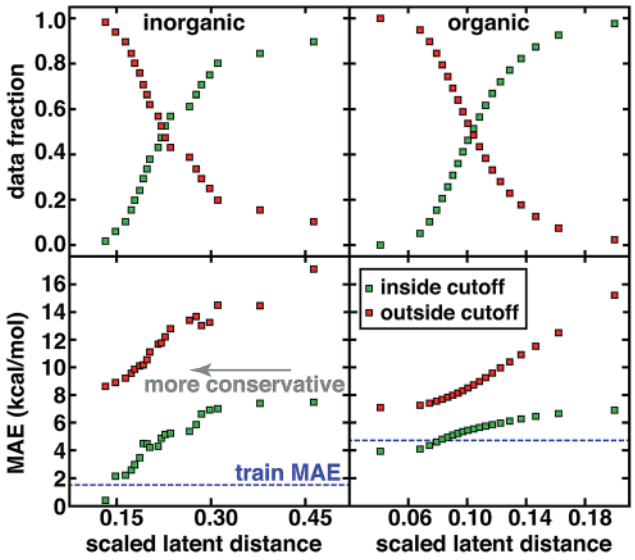

Fig. 5 MAE for predicted points (inside cutoff, green squares) and those not predicted (outside cutoff, orange squares) compared to the training data MAE (blue horizontal dashed line) along with data fraction in each set for the inorganic CSD test set (left) and organic QM9 set (right). The most distant point in the test set is scaled to have a latent distance of 1.0 for comparison across data sets but the x-axis range is then truncated to focus on the range of latent distance cutoffs that affect most of the data.

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, MNIST[88] 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[90] followed by manual fine-tuning in Keras[91] with the Tensorflow[92] backend (ESI Fig. S17, Tables S5 and S13†). Model weights are provided in the ESI.† The $\Delta E_{H\text{-}L}$ energy evaluation protocol for inorganic chemistry training data and the curated CSD[77] test set used molSimplify[8,11] to automate hybrid (*i.e.*, B3LYP[80–82]) DFT calculations, with more details provided in ESI Text S2.† For the organic chemistry test, the QM9 atomization energy data set was obtained from the literature.[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, τ (ESI Text S1 and Table S2†). 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 (*e.g.*, RAC-155 (ref. 7)) directly. Featurizations of relevant complexes are provided in the ESI.† For latent distances, we used the latent space after the last hidden layer, which has the dimensionality of the model (*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

1 Y. Zhuo, A. Mansouri Tehrani and J. Brgoch, Predicting the Band Gaps of Inorganic Solids by Machine Learning, *J. Phys. Chem. Lett.*, 2018, **9**, 1668–1673.
2 S. De, A. P. Bartok, G. Csanyi and M. Ceriotti, Comparing Molecules and Solids across Structural and Alchemical Space, *Phys. Chem. Chem. Phys.*, 2016, **18**, 13754–13769.
3 L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials, *npj Comput. Mater.*, 2016, **2**, 16028.
4 G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, Accelerating Materials Property Predictions Using Machine Learning, *Sci. Rep.*, 2013, **3**, 2810.
5 B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld and C. Corminboeuf, Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts, *Chem. Sci.*, 2018, **9**, 7069–7077.
6 X. Ma, Z. Li, L. E. K. Achenie and H. Xin, Machine-Learning-Augmented Chemisorption Model for CO_2 Electroreduction Catalyst Screening, *J. Phys. Chem. Lett.*, 2015, **6**, 3528–3533.
7 J. P. Janet and H. J. Kulik, Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships, *J. Phys. Chem. A*, 2017, **121**, 8939–8954.
8 A. Nandy, C. Duan, J. P. Janet, S. Gugler and H. J. Kulik, Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry, *Ind. Eng. Chem. Res.*, 2018, **57**, 13973–13986.
9 S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang and O. Levy, AFLOW: An Automatic Framework for High-Throughput Materials Discovery, *Comput. Mater. Sci.*, 2012, **58**, 218–226.
10 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
11 E. I. Ioannidis, T. Z. H. Gani and H. J. Kulik, molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry, *J. Comput. Chem.*, 2016, **37**, 2106–2117.
12 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, Open Babel: An Open Chemical Toolbox, *J. Cheminf.*, 2011, **3**, 33.
13 T. J. Martínez, *Ab Initio* Reactive Computer Aided Molecular Design, *Acc. Chem. Res.*, 2017, **50**, 652–656.
14 J. Caruthers, J. A. Lauterbach, K. Thomson, V. Venkatasubramanian, C. Snively, A. Bhan, S. Katare and G. Oskarsdottir, Catalyst Design: Knowledge Extraction

from High-Throughput Experimentation, *J. Catal.*, 2003, **216**, 98–109.

15 S. Katare, J. M. Caruthers, W. N. Delgass and V. Venkatasubramanian, An Intelligent System for Reaction Kinetic Modeling and Catalyst Design, *Ind. Eng. Chem. Res.*, 2004, **43**, 3484–3512.

16 A. Corma, M. J. Díaz-Cabanas, M. Moliner and C. Martínez, Discovery of a New Catalytically Active and Selective Zeolite (ITQ-30) by High-Throughput Synthesis Techniques, *J. Catal.*, 2006, **241**, 312–318.

17 K. Yao, J. E. Herr, D. W. Toth, R. Mckintyre and J. Parkhill, The Tensormol-0.1 Model Chemistry: A Neural Network Augmented with Long-Range Physics, *Chem. Sci.*, 2018, **9**, 2261–2269.

18 J. Behler, Perspective: Machine Learning Potentials for Atomistic Simulations, *J. Chem. Phys.*, 2016, **145**, 170901.

19 J. S. Smith, O. Isayev and A. E. Roitberg, ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost, *Chem. Sci.*, 2017, **8**, 3192–3203.

20 L. Zhang, J. Han, H. Wang, R. Car and E. Weinan, Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics, *Phys. Rev. Lett.*, 2018, **120**, 143001.

21 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt and K.-R. Müller, Machine Learning of Accurate Energy-Conserving Molecular Force Fields, *Sci. Adv.*, 2017, **3**, e1603015.

22 F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. Von Lilienfeld, Prediction Errors of Molecular Machine Learning Models Lower Than Hybrid DFT Error, *J. Chem. Theory Comput.*, 2017, **13**, 5255–5264.

23 B. R. Goldsmith, J. Esterhuizen, J. X. Liu, C. J. Bartel and C. Sutton, Machine Learning for Heterogeneous Catalyst Design and Discovery, *AIChE J.*, 2018, **64**, 2311–2323.

24 J. R. Kitchin, Machine Learning in Catalysis, *Nat. Catal.*, 2018, **1**, 230.

25 J. P. Janet, F. Liu, A. Nandy, C. Duan, T. Yang, S. Lin and H. J. Kulik, *Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry*, Inorganic Chemistry, 2019, ASAP.

26 S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li and J. Wang, Accelerated Discovery of Stable Lead-Free Hybrid Organic–Inorganic Perovskites *via* Machine Learning, *Nat. Commun.*, 2018, **9**, 3405.

27 R. Yuan, Z. Liu, P. V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue and T. Lookman, Accelerated Discovery of Large Electrostrains in $BaTiO_3$-Based Piezoelectrics Using Active Learning, *Adv. Mater.*, 2018, **30**, 1702884.

28 B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons and J. Hattrick-Simpers, Can Machine Learning Identify the Next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery, *Mol. Syst. Des. Eng.*, 2018, **3**, 819–825.

29 F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers and A. Mehta, Accelerated Discovery of Metallic Glasses through Iteration of Machine Learning and High-Throughput Experiments, *Sci. Adv.*, 2018, **4**, eaaq1566.

30 B. Sanchez-Lengeling and A. Aspuru-Guzik, Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering, *Science*, 2018, **361**, 360.

31 Y. He, E. D. Cubuk, M. D. Allendorf and E. J. Reed, Metallic Metal–Organic Frameworks Predicted by the Combination of Machine Learning Methods and *Ab Initio* Calculations, *J. Phys. Chem. Lett.*, 2018, **9**, 4562–4569.

32 B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski and T. Yong-Jin Han, *Reliable and Explainable Machine Learning Methods for Accelerated Material Discovery*, arXiv:1901.02717, 2019.

33 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, Quantum Chemistry Structures and Properties of 134 Kilo Molecules, *Sci. Data*, 2014, **1**, 140022.

34 J. S. Smith, O. Isayev and A. E. Roitberg, ANI-1, a Data Set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules, *Sci. Data*, 2017, **4**, 170193.

35 J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, Less Is More: Sampling Chemical Space with Active Learning, *J. Chem. Phys.*, 2018, **148**, 241733.

36 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Machine Learning for Molecular and Materials Science, *Nature*, 2018, **559**, 547.

37 G. N. Simm and M. Reiher, Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes, *J. Chem. Theory Comput.*, 2018, **14**, 5238–5248.

38 Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations, *Nat. Commun.*, 2017, **8**, 14621.

39 F. Musil, M. J. Willatt, M. A. Langovoy and M. Ceriotti, Fast and Accurate Uncertainty Estimation in Chemical Machine Learning, *J. Chem. Theory Comput.*, 2019, **15**, 906–915.

40 A. A. Peterson, R. Christensen and A. Khorshidi, Addressing Uncertainty in Atomistic Machine Learning, *Phys. Chem. Chem. Phys.*, 2017, **19**, 10978–10985.

41 R. Liu and A. Wallqvist, Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies out-of-Domain Compounds, *J. Chem. Inf. Model.*, 2019, **59**, 181–189.

42 I. Cortés-Ciriano and A. Bender, Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks, *J. Chem. Inf. Model.*, 2018, **59**, 1269–1281.

43 C. L. M. Morais, K. M. G. Lima and F. L. Martin, Uncertainty Estimation and Misclassification Probability for Classification Models Based on Discriminant Analysis and Support Vector Machines, *Anal. Chim. Acta*, 2018, **1063**, 40–46.

44 G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft and K. Q. Weinberger, *Snapshot Ensembles: Train 1, Get M for Free*, eprint arXiv:1704.00109, 2017.

45 K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko and K.-R. Müller, SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions, in *Advances in Neural Information Processing Systems*, 2017, pp. 991–1001.

46 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, SchNet–a Deep Learning Architecture for Molecules and Materials, *J. Chem. Phys.*, 2018, **148**, 241722.

47 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, Quantum-Chemical Insights from Deep Tensor Neural Networks, *Nat. Commun.*, 2017, **8**, 13890.

48 M. H. Segler, T. Kogej, C. Tyrchan and M. P. Waller, Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks, *ACS Cent. Sci.*, 2017, **4**, 120–131.

49 L. van der Maaten and G. Hinton, Visualizing Data Using t-SNE, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.

50 Y. Gal and Z. Ghahramani, in *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, international conference on machine learning*, 2016, pp. 1050–1059.

51 R. M. Neal, *Bayesian Learning for Neural Networks*, Springer Science & Business Media, 2012, vol. 118.

52 R. Liu, K. P. Glover, M. G. Feasel and A. Wallqvist, General Approach to Estimate Error Bars for Quantitative Structure–Activity Relationship Predictions of Molecular Activity, *J. Chem. Inf. Model.*, 2018, **58**, 1561–1575.

53 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Gómez-Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, Convolutional Networks on Graphs for Learning Molecular Fingerprints, *Adv. Neural Inf. Process. Syst.*, 2015, 2215–2223.

54 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, *Neural Message Passing for Quantum Chemistry*, arXiv preprint arXiv:1704.01212, 2017.

55 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, *ACS Cent. Sci.*, 2018, **4**, 268–276.

56 N. C. Iovanac and B. M. Savoie, Improved Chemical Prediction from Scarce Data Sets *via* Latent Space Enrichment, *J. Phys. Chem. A*, 2019, **123**, 4295–4302.

57 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, The Atomic Simulation Environment—a Python Library for Working with Atoms, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.

58 J. H. Metzen, T. Genewein, V. Fischer and B. Bischoff, On Detecting Adversarial Perturbations, in *5th International Conference on Learning Representations*, ICLR, 2017.

59 S. Gu and L. Rigazio, *Towards Deep Neural Network Architectures Robust to Adversarial Examples*, eprint arXiv:1412.5068, 2014.

60 C. Zhou and R. C. Paffenroth, Anomaly Detection with Robust Deep Autoencoders, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*, Halifax, NS, Canada, 2017, pp. 665–674.

61 T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth and G. Langs, in *Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery, Information Processing in Medical Imaging*, Springer International Publishing, 2017, pp. 146–157.

62 H. Jiang, B. Kim, M. Y. Guan and M. R. Gupta, *To Trust or Not to Trust a Classifier*, 2018, pp. 5546–5557, arxiv:1805.11783.

63 N. Papernot and P. D. McDaniel, Deep K-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning, arXiv:1803.04765.

64 B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy and B. Srivastava, *Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering*, eprint arXiv:1811.03728, 2018.

65 N. Lubbers, J. S. Smith and K. Barros, Hierarchical Modeling of Molecular Energies Using a Deep Neural Network, *J. Chem. Phys.*, 2018, **148**, 241715.

66 J. Gomes, B. Ramsundar, E. N. Feinberg and V. S. Pande, *Atomic convolutional networks for predicting protein-ligand binding affinity*, 2017, arXiv preprint arXiv:1703.10603.

67 Z. Q. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, MoleculeNet: A Benchmark for Molecular Machine Learning, *Chem. Sci.*, 2018, **9**, 513–530.

68 C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola and K. F. Jensen, Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction, *J. Chem. Inf. Model.*, 2017, **57**, 1757–1772.

69 T. Xie and J. C. Grossman, Hierarchical Visualization of Materials Space with Graph Convolutional Neural Networks, *J. Chem. Phys.*, 2018, **149**, 174111.

70 C. R. Collins, G. J. Gordon, O. A. von Lilienfeld and D. J. Yaron, Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties, *J. Chem. Phys.*, 2018, **148**, 241718.

71 B. Huang and O. A. von Lilienfeld, Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity, *J. Chem. Phys.*, 2016, **145**, 161102.

72 K. Yao, J. E. Herr, S. N. Brown and J. Parkhill, Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network, *J. Phys. Chem. Lett.*, 2017, **8**, 2689–2694.

73 K. Hansen, F. Biegler, R. Ramakrishnan and W. Pronobis, Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.

74 K. Gubaev, E. V. Podryabinkin and A. V. Shapeev, Machine Learning of Molecular Properties: Locality and Active Learning, *J. Chem. Phys.*, 2018, **148**, 241727.

75 P. Bjørn Jørgensen, K. Wedel Jacobsen and M. N. Schmidt, *Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials*, arXiv e-prints, 2018.
76 C. Duan, J. P. Janet, F. Liu, A. Nandy and H. J. Kulik, Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models, *J. Chem. Theory Comput.*, 2019, **15**, 2331–2345.
77 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, The Cambridge Structural Database, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
78 L. Breiman, Bagging Predictors, *Mach. Learn.*, 1996, **24**, 123–140.
79 C. C. Aggarwal, A. Hinneburg and D. A. Keim, in *On the Surprising Behavior of Distance Metrics in High Dimensional Space, Database Theory—ICDT 2001*, ed. J. Van den Bussche and V. Vianu, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 420–434.
80 C. Lee, W. Yang and R. G. Parr, Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
81 A. D. Becke, Density-Functional Thermochemistry. III. The Role of Exact Exchange, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
82 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *Ab Initio* Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
83 R. Ditchfield, W. J. Hehre and J. A. Pople, Self-Consistent Molecular Orbital Methods 9. Extended Gaussian-Type Basis for Molecular Orbital Studies of Organic Molecules, *J. Chem. Phys.*, 1971, **54**, 724.
84 P. Broto, G. Moreau and C. Vandycke, Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies: System of Atomic Contributions for the Calculation of the N-Octanol/Water Partition Coefficients, *Eur. J. Med. Chem.*, 1984, **19**, 71–78.
85 A. M. Virshup, J. Contreras-García, P. Wipf, W. Yang and D. N. Beratan, Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds, *J. Am. Chem. Soc.*, 2013, **135**, 7296–7303.
86 K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
87 L. McInnes and J. Healy, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426.
88 Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, Gradient-Based Learning Applied to Document Recognition, *Proc. IEEE*, 1998, **86**, 2278–2324.
89 H. Xiao, K. Rasul and R. Vollgraf, Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms, arXiv:1708.07747.
90 J. C. Bergstra, D. Yamins and D. D. Cox, Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms, *Proceedings of the 12th Python in science conference*, 2013, pp. 13–20.
91 Keras, https://keras.io/, accessed Jan 17, 2019.
92 Tensorflow, https://www.tensorflow.org, accessed Jan 17, 2019.