ACS central science

http://pubs.acs.org/journal/acscii

Research Article

Evidential Deep Learning for Guided Molecular Property Prediction and Discovery

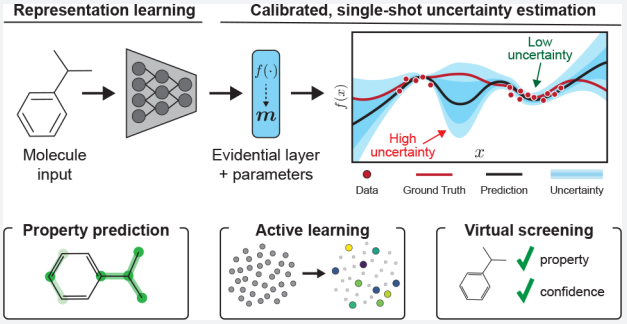
Ava P. Soleimany,[○] Alexander Amini,[○] Samuel Goldman,[○] Daniela Rus, Sangeeta N. Bhatia, and Connor W. Coley*

Cite This: *ACS Cent. Sci.* 2021, 7, 1356–1367

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: While neural networks achieve state-of-the-art performance for many molecular modeling and structure–property prediction tasks, these models can struggle with generalization to out-of-domain examples, exhibit poor sample efficiency, and produce uncalibrated predictions. In this paper, we leverage advances in evidential deep learning to demonstrate a new approach to uncertainty quantification for neural network-based molecular structure–property prediction at no additional computational cost. We develop both evidential 2D message passing neural networks and evidential 3D atomistic neural networks and apply these networks across a range of different tasks. We demonstrate that evidential uncertainties enable (1) calibrated predictions where uncertainty correlates with error, (2) sample-efficient training through uncertainty-guided active learning, and (3) improved experimental validation rates in a retrospective virtual screening campaign. Our results suggest that evidential deep learning can provide an efficient means of uncertainty quantification useful for molecular property prediction, discovery, and design tasks in the chemical and physical sciences.



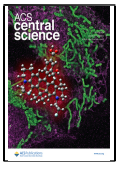
INTRODUCTION

As quantitative structure–activity relationship (QSAR) models are increasingly applied across the chemical and physical sciences to guide time- and resource-intensive experimentation, an understanding of when to trust model predictions is of critical importance.[1−3] Though neural networks have shown tremendous promise in QSAR modeling,[4,5] they remain difficult to interpret, are susceptible to pathological failures in out-of-domain regimes, and lack guarantees on their robustness. Therefore, a better understanding of predictive confidence of neural models is essential, particularly for drug discovery and virtual screening applications where model predictions can inform safety-critical experimental pipelines.

Uncertainty quantification (UQ) methods can help meet this critical need to facilitate the robust application of neural models in the chemical sciences. Indeed, significant work has been done in establishing general methods to estimate epistemic uncertainties (i.e., model uncertainty due to uncertainty in parameters and predictions) and aleatoric uncertainties (i.e., data uncertainty due to noise inherent in the observations) of neural network predictions.[6] Recent studies have demonstrated the importance of focusing explicitly on epistemic uncertainty in the contexts of property and reaction prediction in the chemical sciences,[7] discovery in the biological sciences,[8] as well as in healthcare more broadly.[9] While a plethora of distance-based and nonparametric methods for UQ has been developed,[10,11] Bayesian neural networks[12] and sampling-based approaches, such as model ensembling[13] and dropout sampling,[14] are still accepted as state of the art for epistemic UQ in neural networks, due in part to their model-agnostic nature and ease of implementation.[7,15,16]

However, these approaches only generate approximations to the underlying uncertainty functions via stochastic sampling, incurring computational costs and runtimes that are routinely an order of magnitude higher than those of single models. This poses a significant challenge to using these epistemic uncertainty models in iterative active learning procedures, scans of very large chemical libraries, and molecular dynamics simulations.[16,17] Additionally, the most recent adaptations of atomistic neural networks for prediction of potential energy surfaces and quantum mechanical properties have achieved state-of-the-art results by using more expressive, larger network architectures that sacrifice speed for predictive accuracy.[18,19] Large model sizes compound the computational expense of deploying sampling-based UQ methods and necessitate the development

Received: May 5, 2021
Published: July 27, 2021



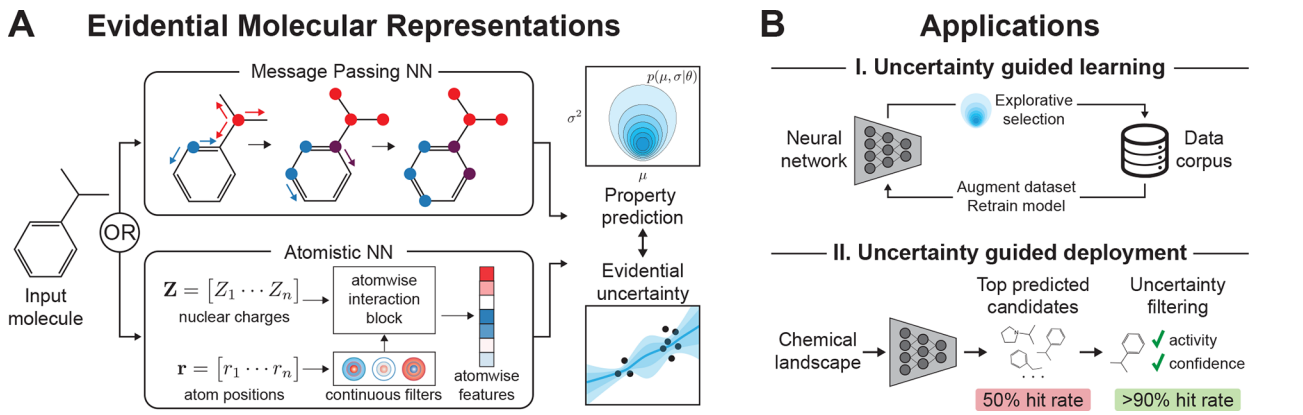


Figure 1. Evidential uncertainty for molecular prediction and discovery. (A) Evidential direct message passing or atomistic neural networks learn molecular representations, predict target properties, and infer the parameters of an underlying evidential distribution that captures the evidence in support of each prediction and enables uncertainty estimation. (B) Uncertainties are applied during learning (I) to guide sample acquisition and during deployment (II) to discover high confidence candidates with high empirical success rates.

of more efficient approaches. Though neural networks can be trained to obtain closed-form solutions of only aleatoric uncertainty without sampling, these methods fail to provide estimates of epistemic uncertainties, limiting their broad utility.[20−24] More generally, recent analyses have revealed an overwhelming lack of consensus as to the top performing UQ methods across molecular property prediction data sets.[15,25] Thus, there remains a need for fast, calibrated, and scalable UQ methods for QSAR models that provide estimates of model uncertainty and can be deployed across a range of molecular property prediction and discovery tasks.

Emerging evidential deep learning algorithms have the potential to address these limitations in their ability to directly learn grounded representations of epistemic uncertainty without the need for sampling.[26,27] Specifically, these methods formulate learning as an evidence acquisition process, wherein new training examples add support to a learned evidential distribution that parametrizes a probability distribution over the network's likelihood function. Evidential learning therefore offers the promise of efficient and calibrated uncertainty learning without the need for sampling. Furthermore, evidential neural networks can be implemented without significant architecture changes, but rather via modifications to the training loss function, and could thus enable tight integration with domain-specific architectures. However, while evidential learning formulations for both regression[27] and classification[26] have recently been presented, the utility of these methods on complex, nonuniform inputs, such as molecular graphs pervasive throughout the chemical sciences, has yet to be shown.

In this work, we establish evidential deep learning as a new approach to UQ for molecular structure−property prediction (Figure 1). Specifically, this work makes the following contributions:

1. Development of evidential message passing and atomistic networks that learn 2D or 3D molecular representations, respectively, and return well-calibrated epistemic uncertainties without any sampling;
2. Evaluation of evidential uncertainties on benchmark QSAR regression tasks against gold-standard, sampling-based UQ methods;
3. Validation of the relevance of evidential deep learning to key molecular discovery applications that require sample prioritization from a larger screening library, namely the following:
   (i) Uncertainty-guided learning for sample-efficient model training and accelerated property optimization;
   (ii) Uncertainty-guided deployment for prioritization of high confidence candidates in virtual screening.

Taken together, our experiments validate a framework to use evidential deep learning as a powerful and flexible replacement for UQ in molecular property prediction and discovery tasks across the chemical sciences.

APPROACH

Formulating Evidential Learning for Molecules. Evidential deep learning models[26,27] are a recent approach to training single networks to estimate predictive uncertainties. While neural networks have been trained to output probabilities, for example with Softmax[28] for classification or Gaussian distributions (MVE)[20] for regression, these approaches estimate the probability of an output but neglect the model's uncertainty associated with that output. Evidential deep learning extends the idea of learning the parameters of a probability distribution further to predict higher-order distributions over the original likelihood parameters themselves. These higher-order parameters define the evidential distribution and capture both the model's prediction as well as the degree of evidence associated with that prediction. These models estimate uncertainty by directly learning the parameters defining this evidential distribution and are closely related to Bayesian neural networks[14] and ensembling approaches[13] which estimate the model's uncertainty by sampling from the likelihood distribution, instead of directly learning to output it.

In the regression setting (e.g., prediction of a continuous target), we are given a data set of paired training examples $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$, for which the targets, $y_i \in \mathbb{R}$, can be assumed to be drawn i.i.d. from a Gaussian distribution with mean and variance $\boldsymbol{\theta} = \{\mu, \sigma^2\}$. In the case of MVE networks, the likelihood parameters $\boldsymbol{\theta}$ are deterministic and fixed, such that the model is optimized during training to predict these values directly, preventing estimation of model uncertainty. As an extension to this approach, evidential models assume these parameters are unknown and must instead be probabilistically estimated. This is done by placing priors over the likelihood parameters, such that

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367

the mean μ is drawn from a Gaussian distribution and the variance σ^2 is drawn from an Inverse-Gamma distribution. The resulting higher-order distribution (also referred to as the evidential distribution) thus can be represented by a Normal-Inverse-Gamma distribution, $p(\boldsymbol{\theta}|\boldsymbol{m})$. This evidential distribution is specified by four parameters $\boldsymbol{m} = \{\gamma, \lambda, \alpha, \beta\}$. For continuous targets, evidential models directly learn these parameters $\boldsymbol{m}$ which in turn define full distributions on top of the likelihood parameters $\{\mu, \sigma^2\}$, thus capturing the uncertainty in the model's prediction (Figure 2A,B). Accordingly, the model

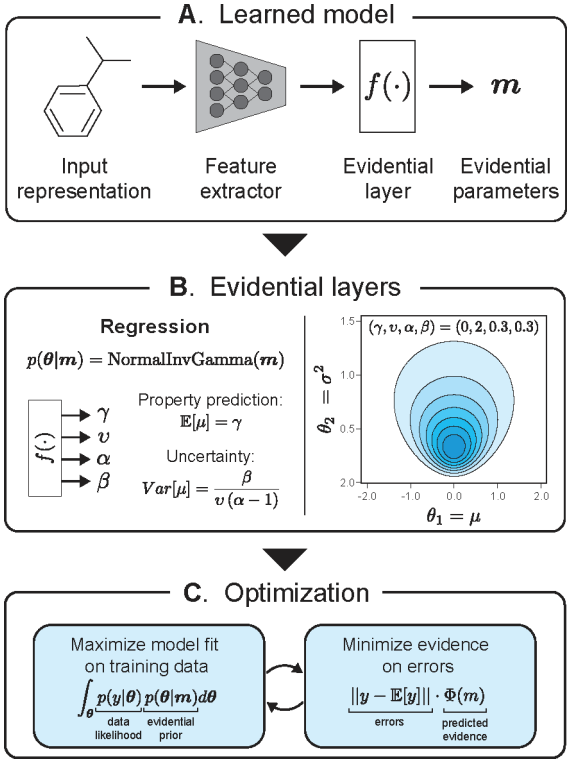


Figure 2. Building and training evidential models. (A) Evidential layers can be added to the end of exisiting molecular feature extractor neural networks. The output of the evidential layer is the parameters ($\boldsymbol{m}$) defining the molecule's evidential distribution. (B) For continuous regression learning problems, the evidential distribution $p(\boldsymbol{\theta}|\boldsymbol{m})$ can take the form of a Normal-Inverse-Gamma distribution, where $\boldsymbol{m} = \{\gamma, \upsilon, \alpha, \beta\}$, to yield both property prediction and uncertainty estimates. Color represents likelihood density (darker = greater density). (C) The model (feature extractor and evidential layer) is trained end-to-end using backpropagation with a multi-objective loss that jointly maximizes model fit and inflates uncertainty (i.e., minimizes evidence) on errors.

outputs four values per target, corresponding to the four parameters of $\boldsymbol{m}$, and is trained using a multiobjective loss that aims to jointly maximize model fit while minimizing evidence on errors (Figure 2C).

In this work, we demonstrate that this evidential learning framework can be coupled with molecular feature extraction networks to predict target properties and also estimate uncertainty (Figure 2). We accomplish this by taking the molecular representation learned by a feature extractor (e.g., a neural network operating on 2D molecular graphs) and feeding it into a dense evidential layer which maps these higher dimensional learned features to the four evidential parameters $\boldsymbol{m}$ which yield both property prediction and uncertainty estimates. All models are trained end-to-end, from molecule input to evidential property output, via backpropagation by optimizing the evidential loss function. [Full model details can be found in the Supporting Information, and code for implementation can be accessed at https://github.com/aamini/chemprop.]

RESULTS

Uncertainty Benchmarking. We first sought to demonstrate that our evidential learning algorithm could produce desirable uncertainties across both molecular and atomistic property prediction tasks. Given our emphasis on downstream tasks that require choosing the correct molecule from a larger screening library (Figure 1B), we evaluated whether predicted uncertainties are correctly ranked with respect to error; that is, predictions with the lowest uncertainty should also be expected to have the lowest error.

We integrated the evidential regression method into a directed message passing neural network (D-MPNN) architecture[50] and assessed its performance in the "lower-N" (data set of ≤10,000 molecules) regression setting on commonly used benchmarking data sets[29] of aqueous solubility (Delaney), solvation energy (Freesolv), lipophilicity (Lipo), and atomization energy (QM7) (Figure 3A). With smaller data sets, sampling approaches such as model ensembling are not prohibitively expensive in practice, so evidential regression must demonstrate more calibrated uncertainty predictions relative to standard sampling-based UQ methods to justify its adoption.

Evidential regression performed well in its ability to rank uncertainties with respect to error (Table 1, Figure S1). Specifically, the evidential method achieves the lowest error across all methods tested when considering only the top 5% most certain predictions for three of the four data sets tested (Delaney, Freesolv, QM7; Table 1). On both the Delaney and QM7 data sets, error returned by the evidential model is well below the second best performing method by the 50% confidence cutoff (Figure 3B,C). The drastic improvement over ensembles in QM7 is consistent with previous observations that single neural network models are more accurate than ensembles in the top confidence percentiles on QM7.[15] Still, in the lower-N setting, there is some variance in performance across data sets. On the lipophilicity data set, the RMSE computed at uncertainty cutoff percentiles of 0.25 and below for evidential regression is higher (worse) than the dropout-based sampling method, showing no advantage in selecting the most accurately predicted test set molecules over dropout (Table 1, Figure S1). Furthermore, the evidential method yields higher rank correlation between uncertainty and error than both ensembles and dropout on two of the four lower-N data sets tested and is at least within one standard deviation of the ensemble method for three of the four data sets tested, supporting its ability to better rank predictions (Figure S3).

To further evaluate performance in the lower-N setting, we conducted a similar analysis on three additional lower-N data sets acquired from the Therapeutics Data Commons[30] with tasks to predict hepatocyte clearance ("Clearance"), median lethal dose ("LD50"), and plasma protein binding rates ("PPBR"). We find that evidence is competitive with sampling-based methods here as well, with RMSE that is at least as low as the top performing sampling method on all three data sets tested (Table S1). Furthermore, error for the evidential method decreases steeply as a function of predicted certainty (Figure S2), and rank correlation between error and uncertainty

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367

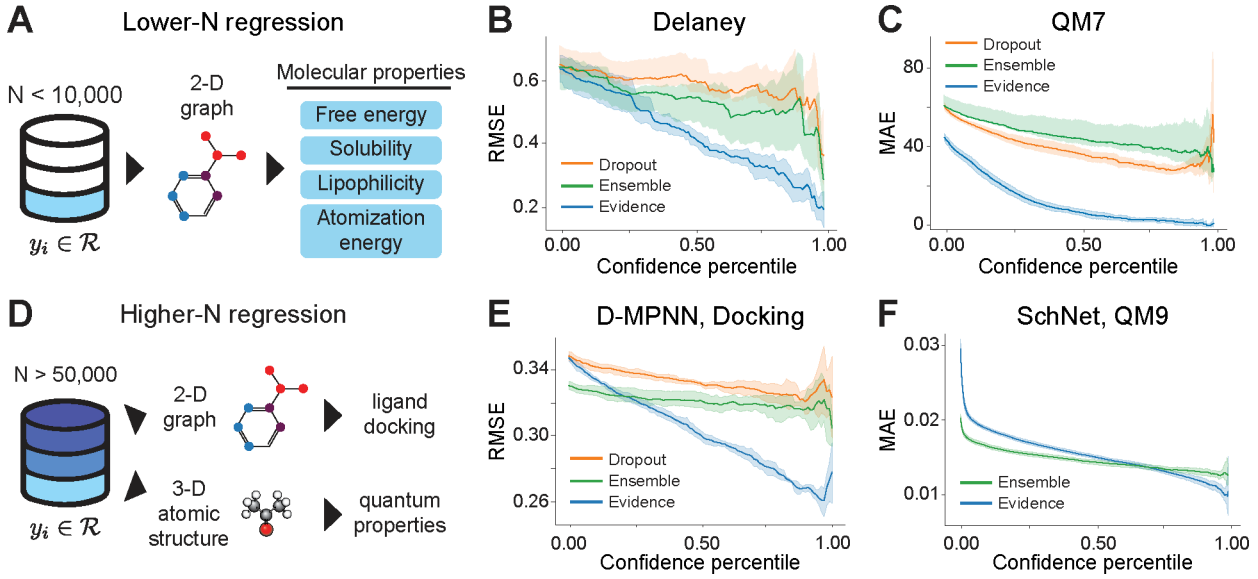


Figure 3. Benchmarking evidential uncertainty for molecular property prediction. (A) Lower-N regression tasks using 2D molecular representations for uncertainty benchmarking. (B, C) Prediction error, measured as root mean squared error (RMSE) or mean average error (MAE), at different confidence percentile cutoffs for the Delaney (B) and QM7 (C) data sets. Mean ± 95% confidence interval (c.i.), n = 10 independent trials. (D) Higher-N regression tasks using 2D or 3D molecular representations. (E, F) Prediction error at different confidence percentile cutoffs for the Docking (E) and QM9 (F) data sets for 2D direct message passing (D-MPNN; E) and 3D atomistic (SchNet; F) neural networks, respectively. Mean ± 95% c.i., n = 5 independent trials.

is highest for the evidential method on two out of the three data sets tested (Figure S3).

Given these promising results in the lower-N setting, we next evaluated the generalizability of the evidential method to larger scale data sets of ≥50,000 data points (Figure 3D). Data sets of this size often represent large-scale chemical libraries[31] or are generated via expensive physics-based simulations.[32] In these settings, the ability to quickly rank a larger library based upon confidence is advantageous for guiding downstream analyses and experimentation.[33] To this end, we compared UQ methods for 2D message passing networks evaluated on each of two "higher-N" (≥50,000 molecules) data sets: the QM9 data set containing computer-generated quantum mechanical properties for small organic molecules[29,32] and a ligand docking data set containing scores of 50,240 molecules docked against thymidylate kinase using AutoDock Vina.[34,35]

For both these data sets, evidential regression predictions have lower error than both ensemble- and dropout-based methods for all confidence percentile cutoffs greater than 50%, demonstrating steeper declines in error as a function of confidence (Figure 3E, Table 1). Compared to the ensemble-based method, evidential regression also displays higher rank correlation between uncertainty and error on both the docking data set ($\rho_{evidence} = 0.163 \pm 0.009$, $\rho_{ensemble} = 0.040 \pm 0.018$) and the QM9 data set ($\rho_{evidence} = 0.469 \pm 0.084$, $\rho_{ensemble} = 0.244 \pm 0.097$) (Figure S3). On the QM9 data set with 2D molecular representations, the evidential method exhibits steeper declines in error as a function of confidence cutoffs relative to the ensemble baseline, both for individual tasks (Figure S4) and on the aggregated uncertainty averaged across tasks (Figure S5A,B).

To demonstrate the utility of the evidential learning approach for a variety of chemistry-specific neural model architectures, we integrated the evidential regression loss function into an atomistic neural network, implemented via the SchNetPack[51] software [https://github.com/atomistic-machine-learning/schnetpack], that operates on 3D molecular conformers (Figure 3D). While ensembles are more accurate with no cutoff calculations ($MAE_{ensemble} = 2.04 \times 10^{-2}$, $MAE_{evidence} = 2.98 \times 10^{-2}$; Table 1), the evidential method still produces uncertainties that correlate well with error (Figure 3F). When considering only predictions in the 95% confidence percentile, the evidential method displays a quantitative improvement over the ensemble method ($MAE_{ensemble} = 1.29 \times 10^{-2}$, $MAE_{evidence} = 1.12 \times 10^{-2}$; Figure 3F, Table 1). Rank correlation between error and uncertainty also reflects the steeper decrease of the evidential method relative to the ensemble-based method ($\rho_{evidence} = 0.361 \pm 0.007$ vs $\rho_{ensemble} = 0.220 \pm 0.012$; Figure S3C). Taken together, these results demonstrate the promise of evidential regression in achieving well-ranked uncertainty estimates across different data set sizes and molecular representations, highlighting the modularity of this method.

Calibration and Tunability. After observing steep reductions in error with increasing confidence, we next investigated the calibration of the predicted uncertainties—a critical property for the translation of any UQ method. With a perfectly calibrated classifier, we expect to find the true target value in the 90% credible interval 90% of the time.[36] However, if a regression model is overconfident, we would find the true value in the 90% credible interval less than 90% of the time and vice versa for an underconfident model. Here, we explore the calibration properties of evidential learning for molecular property prediction.

Evidential learning methods introduce regularization terms that minimize model evidence in instances of high predictive error.[26,27] Specifically, in the evidential regression method, the training loss takes the form

$$L(x) = L_{NLL}(x) + \lambda L_{REG}(x)$$

where the log-likelihood term L_{NLL} captures model fit and λ controls the strength to which overconfident predictions are penalized by the regularization term L_{REG}. Thus, λ provides a

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367

Table 1. Model Error at Various Confidence Percentile Cutoffs[a]

	Delaney			Freesolv			Lipo			QM7 ($\times 10^2$)		
cutoff	dropout	ensemble	evidence	dropout	ensemble	evidence	dropout	ensemble	evidence	dropout	ensemble	evidence
0.0	0.68 ± 0.02	0.65 ± 0.03	0.66 ± 0.02	1.00 ± 0.06	0.94 ± 0.06	0.96 ± 0.07	0.55 ± 0.01	0.53 ± 0.02	0.55 ± 0.02	1.18 ± 0.02	1.12 ± 0.02	1.15 ± 0.03
0.5	0.62 ± 0.03	0.55 ± 0.03	0.44 ± 0.01	0.79 ± 0.07	0.45 ± 0.04	0.42 ± 0.04	0.52 ± 0.01	0.40 ± 0.01	0.50 ± 0.01	0.88 ± 0.06	0.88 ± 0.06	0.39 ± 0.03
0.75	0.59 ± 0.03	0.50 ± 0.05	0.35 ± 0.02	0.85 ± 0.12	0.41 ± 0.05	0.36 ± 0.04	0.50 ± 0.02	0.38 ± 0.02	0.51 ± 0.02	0.65 ± 0.03	0.81 ± 0.06	0.23 ± 0.04
0.90	0.55 ± 0.03	0.51 ± 0.09	0.28 ± 0.02	0.66 ± 0.20	0.40 ± 0.06	0.35 ± 0.08	0.46 ± 0.03	0.38 ± 0.02	0.53 ± 0.03	0.69 ± 0.05	0.71 ± 0.11	0.10 ± 0.04
0.95	0.53 ± 0.06	0.45 ± 0.06	0.22 ± 0.02	0.75 ± 0.30	0.27 ± 0.04	0.38 ± 0.12	0.49 ± 0.04	0.36 ± 0.03	0.50 ± 0.04	0.73 ± 0.08	0.69 ± 0.11	0.10 ± 0.04

	Enamine D-MPNN			QM9 D-MPNN			QM9 Atomistic ($\times 10^{-2}$)		
cutoff	dropout	ensemble	evidence	dropout	ensemble	evidence	dropout	ensemble	evidence
0.0	3.40 ± 0.12	4.47 ± 0.18	5.60 ± 0.20	0.35 ± 0.00	0.33 ± 0.00	0.35 ± 0.00	0.35 ± 0.00	2.04 ± 0.03	2.98 ± 0.08
0.5	3.64 ± 0.05	2.12 ± 0.02	1.55 ± 0.12	0.33 ± 0.00	0.32 ± 0.00	0.30 ± 0.00	0.30 ± 0.00	1.45 ± 0.02	1.52 ± 0.02
0.75	3.42 ± 0.04	1.94 ± 0.04	1.04 ± 0.13	0.33 ± 0.00	0.32 ± 0.00	0.28 ± 0.00	0.28 ± 0.00	1.36 ± 0.02	1.33 ± 0.02
0.90	3.30 ± 0.06	1.80 ± 0.03	0.63 ± 0.12	0.32 ± 0.00	0.32 ± 0.01	0.27 ± 0.00	0.27 ± 0.00	1.31 ± 0.03	1.18 ± 0.03
0.95	3.26 ± 0.05	1.79 ± 0.05	0.42 ± 0.01	0.33 ± 0.01	0.32 ± 0.01	0.26 ± 0.01	0.26 ± 0.01	1.29 ± 0.03	1.12 ± 0.03

[a]For a given confidence percentile cutoff, top performing methods based on prediction standard error of the mean (±s.e.m.) are boldface. A cutoff of 0.95 indicates that only the top 5% most confident predictions are considered. Full confidence plots for all data sets are shown in Figure 3 and Figures S1, S4, and S5. Mean ± s.e.m. (RMSE for all D-MPNN models, MAE for atomistic); n = 10 independent trials for lower-N data sets, n = 5 independent trials for higher-N data sets.

tunable hyperparameter capable of modulating the calibration of any model trained with evidential loss (Figure 4A). To investigate the effect of λ on uncertainty calibration, we trained separate models with different regularization strengths and computed empirical calibration curves that compare the fraction of test set points that fall within a credible interval against the fraction of test set points expected to fall within the predicted credible interval.[11] As expected, trained evidential D-MPNNs move from overconfident to underconfident regions as λ is increased, as shown on the Delaney data set (Figure 4B) and across other lower-N data sets (Figure S6).

To quantify the calibration accuracy, we calculated the area between the observed calibration curve and the parity line (perfect calibration) for each value of λ evaluated across all lower-N data sets (Figure 4C). For all lower-N data sets except QM7, there exists a value of λ at which evidential regression is more calibrated than the ensemble baseline (Figure 4C). Based on these results, we choose a default of $\lambda = 0.2$, as cutoff RMSE is robust to small changes in λ (Figure S7), and use this value for all evaluations unless stated otherwise. This regularization strength yields predictions calibrated at least as well as those of ensemble-based methods across all data sets and tasks tested (Figures S1, S5, and S8). While methods have been developed to recalibrate and augment uncertainty predictions,[36] the ability to tune λ via a hyperparameter search with the evidential regression formulation presents an additional, attractive option for chemical science practitioners to quickly calibrate uncertainty before applying general purpose recalibration techniques.

Application I: Uncertainty-Guided Learning. Having verified that evidential uncertainties were well-calibrated to errors on property prediction tasks, we next sought to use these uncertainties to guide learning toward improved sample efficiency or accelerated molecular optimization. Concretely, in this section we investigate two applications, active learning and Bayesian optimization, that utilize UQ to intelligently prioritize sample acquisition (Figure 5A).

Active Learning for Sample Efficient Training. As a first validation of the utility of evidential uncertainties for guided learning, we turned to the QM9 data set,[32] a standard data set for molecular property prediction that captures geometric, energetic, electronic, and thermodynamic properties, and asked whether uncertainty-guided sample acquisition could yield a more sample-efficient learning process. For QM9 active learning experiments, data acquisition was simulated as iterative selection from the library repeated six times after initialization with a random 15% subset of the training data. At each step, the uncertainty was evaluated across the remainder of the training data (i.e., samples that had not yet been selected). For explorative selection, the k samples with the greatest estimated uncertainties at each iteration were added to the training set, and the model was subsequently retrained using this expanded data set and then evaluated on a held out test set (Figure 5A). For all evaluations, random sample selection served as a baseline for each uncertainty quantification method considered.

We find that active selection based on evidential uncertainties yields significantly improved sample efficiency, reaching the same level of performance of the full training data set with over 60% less data (Figure 5B). Further, acquisition using evidential uncertainties results in increased data efficiency relative to dropout-based selection. For example, to achieve an RMSE of 7.0, evidence-guided models required an average of 21% of the entire training data compared to 55% for dropout-guided models (Figure 5B). We observe, consistent with prior literature,

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367

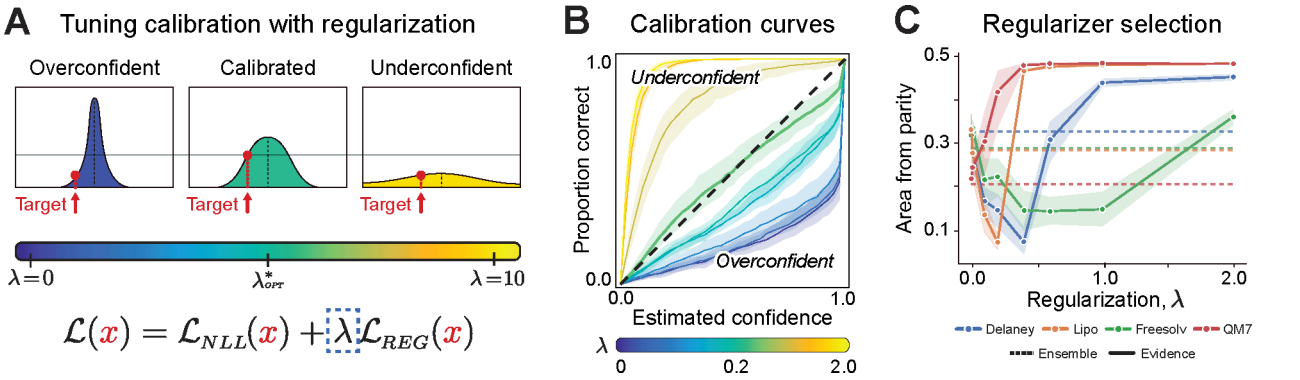


Figure 4. Tunability of the evidential uncertainty. (A) The evidential regression method can be fine-tuned with a single hyperparameter, λ, in order to achieve more calibrated predictions for a given data set. (B) Estimated confidence (cumulative probability) against the observed proportion correct for an evidential D-MPNN evaluated on the Delaney data set. The dotted line represents perfect calibration. Mean ± 95% c.i., $n = 5$ independent trials. (C) Area between the observed calibration curve and the perfect calibration line across several lower-N data sets for evidential D-MPNNs trained with varying λ. Dotted lines represent calibration of an ensemble of models. Mean ± 95% c.i., $n = 10$ independent trials.

that stochastically trained networks (e.g., trained with dropout) also suffer from a baseline performance drop relative to their deterministically trained counterparts, even when all data is considered. As expected, ensemble-based selection shows the greatest improvement over random selection, which is consistent with the advantages that training multiple independent models affords. However, this comes at a significant computational cost, given that multiple independent models must be completely retrained at each active learning step. With growing interest in using uncertainties to inform full molecular dynamics simulations[17] and to decide when to perform density functional theory simulations,[16] retraining 5 to 10 times as many models can drastically increase the expense and overhead of a simulation. In contrast, evidential learning enables resource-efficient uncertainty estimation at the cost of just a single deterministic model, while still achieving increased training efficiency relative to random acquisition at a level nearly on par with ensemble-based selection (Figure 5C). Specifically, while model ensembling achieves a better overall error, the evidential method exhibits superior improvements in sample efficiency across different stages of learning, attaining up to a 18% improvement in error relative to the random baseline (Figure 5C). At any stage in learning, evidence-guided explorative sampling yields an equal to or greater drop in error over random acquisition when compared to either dropout or ensembling.

Bayesian Optimization for Accelerated Molecular Discovery. Instead of acquiring samples solely based on uncertainty, as with purely explorative active learning, Bayesian optimization provides a framework to discover high performing compounds (e.g., those with desired molecular properties) from a large search space by incorporating both predicted property scores and uncertainties to guide sample acquisition.[34,37] In this scheme, uncertainties can be used to explore the search space more conservatively and to broadly enhance the overall diversity of acquired sample batches.[38]

To this end, we investigated the utility of evidential uncertainty for Bayesian optimization settings, where the aim is to rapidly discover compounds with target molecular properties. We turned to the ligand docking data set, previously benchmarked in Figure 3D, of 50,240 molecules docked *in silico* against thymidylate kinase.[34] Given this library of ca. 50*k* molecules, we aim to identify those with the best ligand docking scores by only observing ground truth docking scores for a small subset of the library. Data acquisition is initiated by training on a random 1% subset (ca. 500 molecules) and then simulated as the iterative selection of new samples based on an upper confidence bound (UCB) acquisition function according to a given UQ method. In these experiments, a D-MPNN is used as the surrogate model by which docking scores and uncertainties are estimated.

For all three UQ methods evaluated (dropout, ensemble, and evidence), UCB acquisition yields clear improvements over the random baseline, representative of a brute-force search, as measured by the percentage of top-500 (ca. top 1%) of scores found as a function of the number of compounds explored (Figure 5D). Specifically, the evidential method discovers over 50% of the top-500 docking molecules from the pool of 50*k* molecules after exploring fewer than 2*k* molecules (less than 4% of the search space). Similar to the active learning experiments, we also observe that the evidential method outperforms dropout sampling but does not exceed the performance of ensembling (Figure 5D, Table S2). While previous studies on this data set have shown that using greedy sampling based upon predicted docking score outperforms UCB,[34] we additionally evaluate the structural diversity of the newly acquired pool, relative to the training set, in both greedy and UCB sampled molecules after one round of acquisition (Figure 5E). Relative to its greedy baseline, the evidential UCB method results in a statistically significant increase in the average Tanimoto distance between sampled molecules and their respective 10-nearest training set neighbors, while the dropout- and ensemble-based UCB methods do not (Figure 5E). Together, these results support the use of evidential uncertainties within Bayesian optimization frameworks for accelerated virtual screening and molecular discovery.

Application II: Uncertainty-Guided Inference for Virtual Screening. Though virtual screening is a common tool in computer-aided molecular discovery, all *in silico* predictions of QSAR models must be experimentally validated, and often only a small fraction of predictions or candidates nominated by QSAR models holds true in the real world.[4,5] Therefore, there remains a need for integrated methods that can help ensure the robustness of QSAR predictions. Fast and scalable UQ methods have the potential to meet this need by guiding *in silico* discovery toward molecular candidates associated with greater predictive confidences, based on the hypothesis that high confidence candidates are better suited for downstream experimental validation. To this end, we next

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367

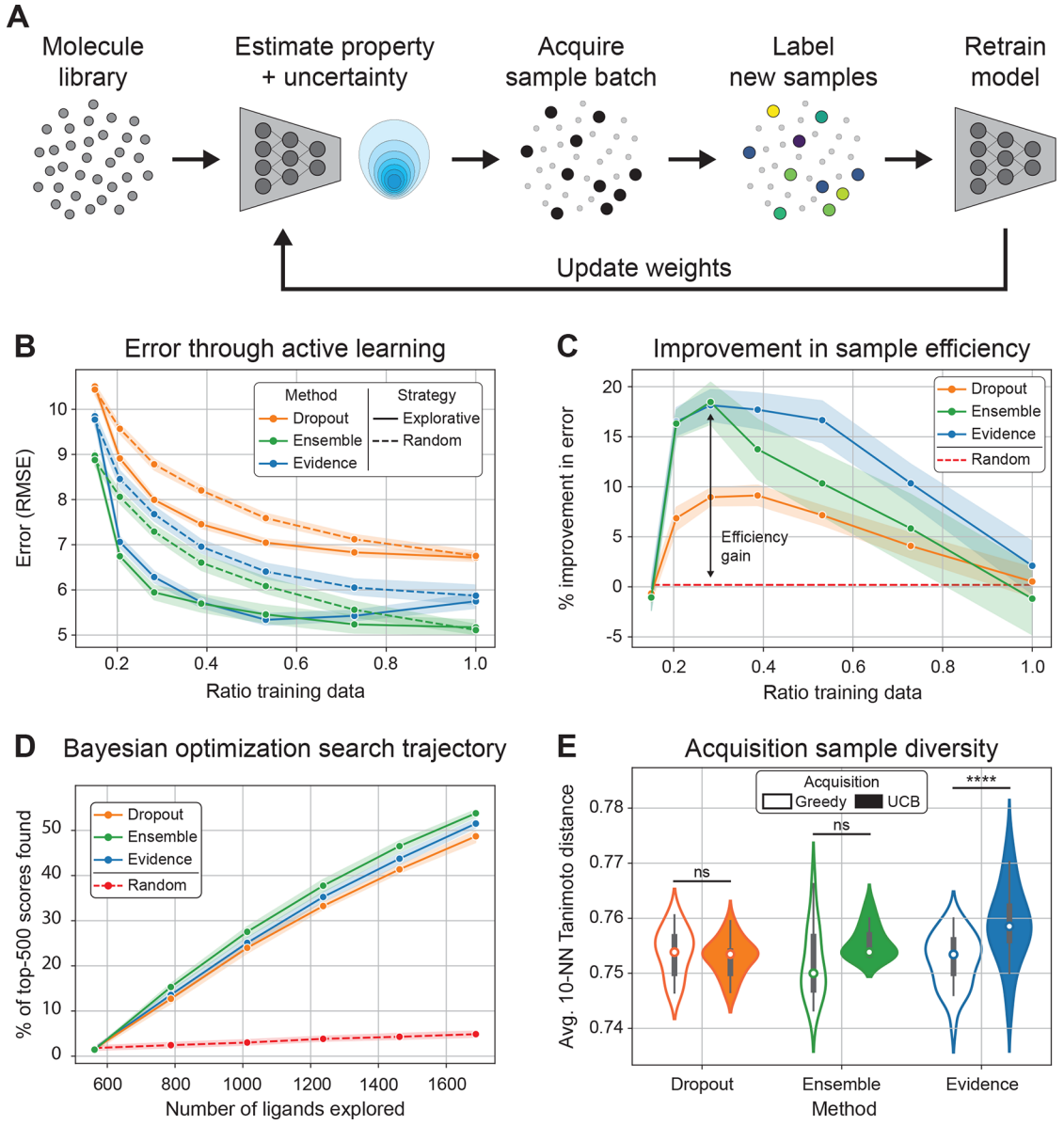


Figure 5. Evidential active learning and Bayesian optimization. (A) Experimental scheme. (B) Active learning with explorative (solid) versus random (dashed) sampling for D-MPNN evaluated on the QM9 data set. Mean ± 95% c.i., $n = 10$ independent trials. (C) Change in sample efficiency for explorative acquisition in (B), evaluated as the percent decrease in predictive error relative to a randomly selected training set. (D) Bayesian optimization performance on Enamine 50k data, measured by the percentage of top-500 scores found as a function of the number of ligands explored. Solid traces represent an upper confidence bound (UCB) acquisition strategy. Mean ± 95% c.i., $n = 10$ independent trials. (E) Average 10-nearest training set neighbors (10-NN) Tanimoto distance for batch samples after the first round of acquisition in Bayesian optimization experiments. Dots represent median; bars represent interquartile range; lines represent upper and lower adjacent values. $n = 10$ independent trials, two-tailed unpaired t test, ****$P < 0.0001$.

investigate the potential for evidential deep learning to discover high confidence drug candidates in retrospective virtual screening campaigns. Concretely, we consider a virtual screening pipeline for antibiotic discovery[39] and demonstrate how evidential uncertainties can be integrated to more accurately prioritize drug repurposing candidates for use as antibiotics by additionally filtering large screening libraries based on confidence in addition to predicted activity.

To achieve this, we develop a framework for uncertainty-guided prioritization in virtual screening (Figure 6A). A large, labeled data set of small molecules is used to train an evidential model which is in turn applied to a smaller, unlabeled discovery data set to predict both molecular properties as well as uncertainties. From these predictions, candidate molecules are subsequently ranked by their associated property values and then filtered further based on confidence thresholds ranging from the 50^{th} to 100^{th} percentiles of greatest predictive confidence (i.e., lowest uncertainties). Finally, among this filtered subset, experimental hit rates (i.e., correlation of true versus predicted activities) are determined either retrospectively, as in this work, or prospectively to assess the relative benefit of uncertainty-guided prioritization versus naive nomination (i.e., without confidence filtering).

To concretely demonstrate the utility of this approach, we considered the question of antibiotic discovery and leveraged a recent data set[39] of small molecules and their *in vitro* growth

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367

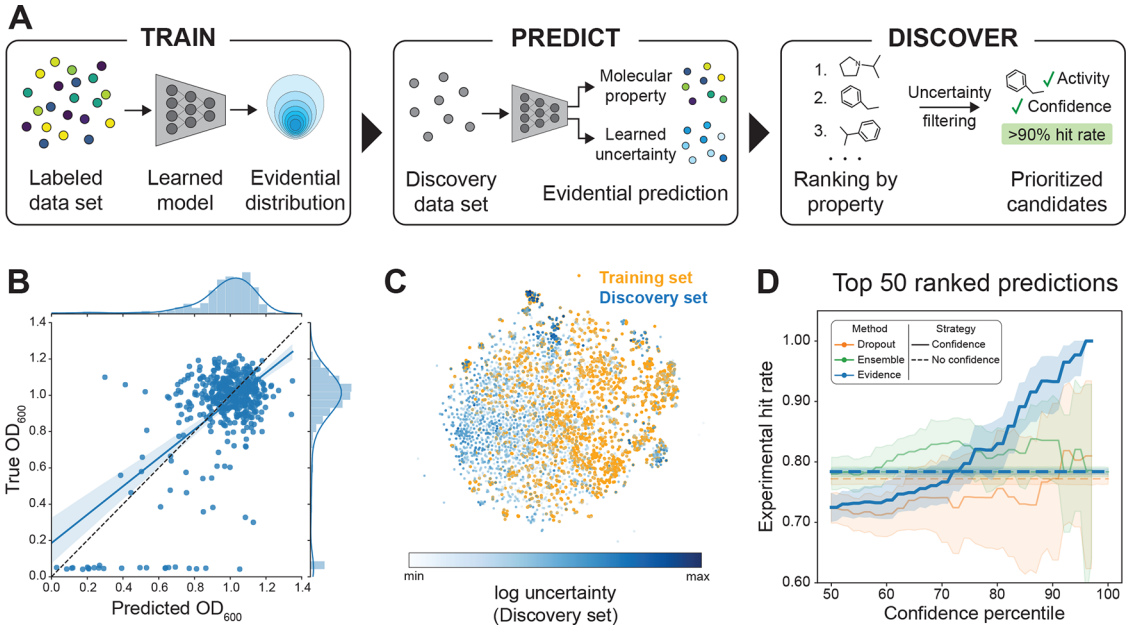


Figure 6. Uncertainty guided nomination in virtual screens. (A) Experimental framework in which trained uncertainty-aware models are deployed on a discovery data set and molecules are ranked based on predicted properties. Uncertainty filtering is used to prioritize candidates among the top ranking molecules. (B) Performance of evidential D-MPNN after training to predict *E. coli* growth inhibition. (C) t-SNE visualization of training set (orange) and discovery data set (Broad library), colored by predicted evidential uncertainties (blue). (D) Application of confidence filters to prioritize sets of antibiotic candidates with high experimental hit rates. Mean ± 95% c.i., $n = 10$ independent trials.

inhibition against *Escherichia coli* (*E. coli*; inhibition measured by OD_{600}) (Figure S9A). We frame the prediction as a regression problem and train a D-MPNN with evidential loss and outputs on this data set of 2,335 small molecules to predict their OD_{600} values. The accuracy of the resulting model is shown on a held-out validation set (Figure 6B).

Next, extending off the virtual screening pipeline presented by Stokes et al.,[39] we applied the trained evidential D-MPNN model to an independent, unlabeled discovery data set with the aim of identifying high confidence candidate antibiotics. We leverage the Broad Drug Repurposing Hub[40] as the discovery data set and generate predictions for both estimated antibiotic activities as well as learned evidential uncertainties (Figure S9B). To begin to understand how uncertainties scale and extend to this discovery data set, we visualize the structural overlap in chemical space between molecules in the discovery (Broad) data set compared to those in the training data set and annotate molecules in the discovery data set with their estimated evidential uncertainties (Figure 6C). Qualitatively, this analysis revealed select regions of chemical space associated with higher evidential uncertainties that also exhibit less overlap with the training set, consistent with the expected inflation of uncertainties for out-of-distribution or distribution shifted domains. Furthermore, comparison of predicted growth inhibition to evidential uncertainty demonstrates that predicted active molecules (lower predicted OD_{600}) trended toward higher uncertainties (Figure S9B), an observation consistent with the stark imbalance and skewness of the training set (Figure S9A).

We then utilized evidential uncertainties to prioritize high confidence candidate antibiotics from the discovery data set, with the goal of identifying molecule sets with high experimental hit rates, i.e., high likelihoods of having true growth inhibitory activity in the real world (Figure 6A). To this end, following prediction of both antibiotic activity and the associated evidential uncertainty, we rank molecules in the discovery data set according to their predicted antibiotic activities (i.e., lowest to highest predicted OD_{600}, where lower is better) and then select the top k ranking molecules based on predicted activity, as outlined in a previous virtual screen on this data set.[39] We subsequently filtered the resulting set of k molecules based on confidence estimates for varying confidence thresholds. Specifically, for a given confidence threshold p, molecules with estimated confidences below the associated p^{th} percentile are removed from the list of top k molecules, with p ranging from the 50^{th} to 100^{th} percentiles of greatest predictive confidence.

Experimental hit rates (true $OD_{600} < 0.2$) for these model-nominated compounds were then determined using the subset of candidates for which growth inhibitory activity against *E. coli* had been determined[39] (Figure S9C). This analysis revealed that augmenting network predictions with confidence-based filtering with evidential uncertainties can increase the experimental hit rate relative to that of an unfiltered set of candidates (Figure 6D). Increasing confidence percentiles enriched the candidate set for experimental hits, from a hit rate of 78% for naive filtering to over 95% after confidence filtering using our evidential method (Figure 6D). While filtering with ensemble-derived uncertainties also increased the experimental hit rate above the baseline, this difference was not as great as the relative increase provided by the evidential method. These results suggest that evidential uncertainties can be used to prioritize high confidence drug candidates in virtual screens in order to ultimately guide discovery toward greater likelihoods of experimental success.

DISCUSSION

Contributions. In this work, we establish evidential deep learning as a scalable, efficient, and easy-to-use uncertainty quantification method for molecular property prediction in the chemical and physical sciences. By integrating our algorithm into both message passing and atomistic networks operating on 2D

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367

graphs and 3D conformers, respectively, we demonstrate its modularity across different network architectures and its applicability across a range of tasks in both lower-N and higher-N settings. Through benchmarking experiments against model ensembling and dropout sampling methods, we show that our evidential algorithm exhibits strong calibration and yields uncertainties that scale with prediction errors, supporting the utility of our method for prioritizing candidate molecules from large screening libraries. Furthermore, we validate the utility of evidential deep learning for uncertainty-guided learning and compound prioritization in virtual screening. We find that evidential uncertainties can effectively guide sample acquisition to improve training efficiency and to accelerate virtual screening in active learning and Bayesian optimization settings. Finally, by leveraging an evidential message passing network to identify high confidence candidate antibiotics, we show that evidential uncertainties can be used to direct retrospective virtual screening campaigns toward compound sets with increased experimental validation rates.

Advantages of Evidential Learning in Chemical Science. The evidential deep learning framework offers several advantages relative to existing UQ approaches for neural models in the chemical sciences. In contrast to other methods such as Bayesian neural networks that require modifying architectures to output probability distributions over network weights,[41] our algorithm can be incorporated into an existing network architecture by modifying the loss function and the network's final output layer. Furthermore, our method presents key scalability and efficiency advantages over sampling-based approaches for QSAR UQ, namely traditional model ensembling[13] and dropout sampling,[14] which necessitate training and/or evaluation of multiple surrogate models in order to obtain approximations of epistemic uncertainty. While widely used, these methods can incur high computational costs which may be prohibitive in settings that are resource-constrained, that require iterative training, or that use large networks or large data sets. Our method overcomes this limitation by directly modeling a higher-order probability distribution over the likelihood function and requires only a single forward pass through a network to obtain uncertainty estimates.[26,27]

Opportunities and Applications in Molecular Property Prediction. Because of its efficiency and ease of use, evidential deep learning may be particularly relevant to uncertainty-guided virtual screening of large scale chemical libraries. Virtual screening workflows often involve exhaustive prediction of the properties and performance of compounds in large virtual libraries prior to prioritization of candidates for experimental validation.[31] In this setting, evidential deep learning may be used to efficiently obtain uncertainty estimates to understand when the predictions of QSAR models may not be trusted and furthermore to accelerate downstream sample annotation or acquisition, for example via active learning. We envision that evidential learning can be incorporated as an efficient, modular UQ method for virtual screening and compound discovery campaigns.

Our methods may also prove useful in the context of neural networks as surrogate models for quantum mechanical and molecular dynamics simulations,[42,43] where there is increasing interest in using uncertainties to actively guide simulation experiments to determine when machine learned predictions can no longer be trusted.[16,17] Furthermore, though in this work we integrated evidential deep learning into the SchNet architecture,[44] continued development and integration of evidential methods into new atomistic machine learning architectures[18] could help advance their deployment for prediction of potential energy surfaces and quantum mechanical properties, tasks that demand computational scalability and efficiency.

Scope and Future Work. While evidential deep learning provides key advantages over existing methods for UQ in neural models, there are several considerations that motivate opportunities for future work. First, the vast majority of our analyses here focus on regression problems in which networks are trained to predict a continuous molecular property. Evidential deep learning models were originally presented in the setting of multiclass classification[26] and, as such, are also applicable to these domains as well. However, since the natural form of many molecular properties is continuous (not discrete), we focus our analysis in this work on the applicability of evidential methods specifically in the regression domain. We hypothesize that the benefits of evidential UQ for classification will also be apparent in *multiclass* settings, where a single input is being classified as one discrete class from a set of options, such as in protein secondary structure or amino acid prediction, among other applications.[45] Future research to this end will be important to validate the generality of evidential UQ for classification settings.

Furthermore, in this work, we focus on several metrics to evaluate the quality of uncertainty estimates: confidence percentile cutoff errors, Spearman rank correlation coefficients between error and uncertainty, and miscalibration area. We also acknowledge that these UQ methods would be best evaluated in terms of their performance on realistic applications and accordingly demonstrate the use of evidential uncertainties for efficient Bayesian optimization, active learning, and virtual screening. Thus, future work is needed to identify and formulate other impactful applications where effective UQ methods yield improvements in downstream performance. This could help solidify the utility of uncertainty for machine learning in the chemical, biological, and physical sciences, where guarantees in not only model performance but also confidence are ultimately needed for wide-scale adoption.

Finally, in this work we use our evidential UQ method to guide learning and compound screening in the retrospective setting, through active learning experiments and evaluation on an antibiotic discovery data set. While these evaluations support the utility of evidential deep learning, and UQ more broadly, for similar analyses, they remain retrospective. Further work to explore the utility of evidential uncertainties in the prospective setting,[46] for example to identify new, high confidence drug candidates that may in turn be experimentally tested in the real world, could help facilitate the adoption of UQ approaches in discovery and engineering pipelines.

CONCLUSION

In summary, we have developed a flexible, scalable, and efficient approach to uncertainty estimation in neural networks for molecular property prediction in the chemical and physical sciences. We demonstrated that evidential deep learning provides well-calibrated uncertainties in structure−property prediction and validated its relevance to uncertainty-guided active learning and to prioritization of candidates in virtual screening. We expect that evidential deep learning, which can readily be incorporated into existing network architectures and be applied to a variety of predictive learning tasks, could help facilitate the robust and reliable deployment of uncertainty-

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367

aware neural models for molecular property prediction, discovery, and design.

METHODS

All analyses shown in this work were performed *in silico*, and no unexpected or unusually high safety hazards were encountered. All scripts were written in Python; PyTorch[47] was used for building all machine learning architectures; RDKit[48] was used for various cheminformatics calculations.[49−55] Additional methods and implementation details are available in the Supporting Information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscentsci.1c00546.

Detailed methods; table of benchmarking results on all data sets evaluated; statistical significance tests for Enamine Bayesian optimization; uncertainty benchmarking on lower-N MoleculeNet[29] data sets; uncertainty benchmarking on lower-N Therapeutic Data Commons[30] data sets; correlation of error and uncertainty; task-specific cutoff analysis; uncertainty benchmarking on higher-N data sets; effect of λ on uncertainty calibration and cutoff RMSE; task-specific calibration analysis; and overview of antibiotic discovery data sets (PDF)

AUTHOR INFORMATION

Corresponding Author

Connor W. Coley − *Department of Chemical Engineering, MIT, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-8271-8723; Email: ccoley@mit.edu

Authors

Ava P. Soleimany − *Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts 02139, United States; Graduate Program in Biophysics, Harvard University, Boston, Massachusetts 02115, United States; Microsoft Research New England, Cambridge, Massachusetts 02142, United States;* orcid.org/0000-0002-8601-6040

Alexander Amini − *Department of Electrical Engineering and Computer Science, MIT, Cambridge, Massachusetts 02139, United States*

Samuel Goldman − *Computational and Systems Biology, MIT, Cambridge, Massachusetts 02139, United States*

Daniela Rus − *Department of Electrical Engineering and Computer Science, MIT, Cambridge, Massachusetts 02139, United States*

Sangeeta N. Bhatia − *Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts 02139, United States; Department of Electrical Engineering and Computer Science, MIT, Cambridge, Massachusetts 02139, United States; Howard Hughes Medical Institute, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-1293-2097

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.1c00546

Author Contributions

○A.P.S., A.A., and S.G. contributed equally.

Notes

The authors declare no competing financial interest.

All code to reproduce experiments and results for this article may be accessed at https://github.com/aamini/chemprop.

ACKNOWLEDGMENTS

The authors acknowledge the MIT SuperCloud and Lincoln Laboratory SuperComputing Center for providing computational resources that have contributed to the results reported. The authors thank Lucas Liebenwein for development and maintenance of the DRL GPU cluster. A.P.S. acknowledges support from the NIH Molecular Biophysics Training Grant NIH/NIGMS T32 GM008313 and the National Science Foundation Graduate Research Fellowship. A.A. acknowledges support from the National Science Foundation Graduate Research Fellowship. S.N.B. is a Howard Hughes Medical Institute Investigator. S.G. and C.W.C. thank the Machine Learning for Pharmaceutical Discovery and Synthesis consortium.

REFERENCES

(1) Nigam, A.; Pollice, R.; Hurley, M. F.; Hickman, R. J.; Aldeghi, M.; Yoshikawa, N.; Chithrananda, S.; Voelz, V. A.; Aspuru-Guzik, A. Assigning Confidence to Molecular Property Prediction. 2021, arXiv preprint arXiv:2102.11439. https://arxiv.org/abs/2102.11439 (accessed 2021-07-19).

(2) Lamb, G.; Paige, B. Bayesian Graph Neural Networks for Molecular Property Prediction. 2020, arXiv preprint arXiv:2012.02089. https://arxiv.org/abs/2012.02089 (accessed 2021-07-19).

(3) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2020**, *2*, 573−584.

(4) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977−5010.

(5) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.

(6) Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? 2017, arXiv preprint arXiv:1703.04977. https://arxiv.org/abs/1703.04977 (accessed 2021-07-19).

(7) Eyke, N. S.; Green, W. H.; Jensen, K. F. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *React. Chem. Eng.* **2020**, *5*, 1963.

(8) Hie, B.; Bryson, B. D.; Berger, B. Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Systems* **2020**, *11*, 461−477.

(9) Roy, A. G.; Ren, J.; Azizi, S.; Loh, A.; Natarajan, V.; Mustafa, B.; Pawlowski, N.; Freyberg, J.; Liu, Y.; Beaver, Z., et al. Does Your Dermatology Classifier Know What It Doesn't Know? Detecting the Long-Tail of Unseen Conditions. 2021, arXiv preprint arXiv:2104.03829. https://arxiv.org/abs/2104.03829 (accessed 2021-07-19).

(10) Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chemical science* **2019**, *10*, 7913−7922.

(11) Tran, K.; Neiswanger, W.; Yoon, J.; Zhang, Q.; Xing, E.; Ulissi, Z. W. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology* **2020**, *1*, 025006.

(12) Jospin, L. V.; Buntine, W.; Boussaid, F.; Laga, H.; Bennamoun, M. Hands-on Bayesian Neural Networks−a Tutorial for Deep Learning

Users. 2020, arXiv preprint arXiv:2007.06823. https://arxiv.org/abs/2007.06823 (accessed 2021-07-19).
(13) Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems* **2017**, 6402−6413.
(14) Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning* **2016**, 1050−1059.
(15) Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 3770−3780. PMID: 32702986.
(16) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.
(17) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Computational Materials* **2020**, *6*, 20.
(18) Klicpera, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. 2020, arXiv preprint arXiv:2011.14115. https://arxiv.org/abs/2011.14115 (accessed 2021-07-19).
(19) Klicpera, J.; Groß, J.; Günnemann, S. Directional message passing for molecular graphs. arXiv preprint arXiv:2003.03123. 2020, https://arxiv.org/abs/2003.03123 (accessed 2021-07-19).
(20) Nix, D.; Weigend, A. Estimating the Mean and Variance of the Target Probability Distribution. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*; 1994; Vol. *1*, pp 55−60, DOI: 10.1109/ICNN.1994.374138
(21) Bishop, C. M. *Mixture density networks*; 1994.
(22) Gilitschenski, I.; Sahoo, R.; Schwarting, W.; Amini, A.; Karaman, S.; Rus, D. Deep Orientation Uncertainty Learning based on a Bingham Loss. *International Conference on Learning Representations*; 2019.
(23) Amini, A.; Soleimany, A. P.; Schwarting, W.; Bhatia, S. N.; Rus, D. Uncovering and mitigating algorithmic bias through learned latent structure. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*; 2019; pp 289−295, DOI: 10.1145/3306618.3314243.
(24) Gurevich, P.; Stuke, H. Gradient conjugate priors and multi-layer neural networks. *Artificial Intelligence* **2020**, *278*, 103184.
(25) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning Based Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2697.
(26) Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems* **2018**, 3179−3189.
(27) Amini, A.; Schwarting, W.; Soleimany, A.; Rus, D. Deep evidential regression. *Advances in Neural Information Processing Systems*; 2020; Vol. *33*.
(28) Goodfellow, I.; Bengio, Y.; Courville, A. 6.2.2.3 softmax units for multinoulli output distributions. *Deep learning*; 2016; pp 180−184.
(29) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(30) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics data Commons: machine learning datasets and tasks for therapeutics. 2021, arXiv preprint arXiv:2102.09548. https://arxiv.org/abs/2102.09548 (accessed 2021-07-19).
(31) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Das, K. M. P.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **2020**, *580*, 663−668.
(32) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 14002.
(33) Jastrzębski, S.; Szymczak, M.; Pocha, A.; Mordalski, S.; Tabor, J.; Bojarski, A. J.; Podlewska, S. Emulating docking results using a deep neural network: a new perspective for virtual screening. *J. Chem. Inf. Model.* **2020**, *60*, 4246−4262.
(34) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. 2020, arXiv preprint arXiv:2012.07127. https://arxiv.org/abs/2012.07127 (accessed 2021-07-19).
(35) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(36) Kuleshov, V.; Fenner, N.; Ermon, S. Accurate uncertainties for deep learning using calibrated regression. *International Conference on Machine Learning* **2018**, 2796−2804.
(37) Frazier, P. I. A tutorial on Bayesian optimization. 2018, arXiv preprint arXiv:1807.02811. https://arxiv.org/abs/1807.02811 (accessed 2021-07-19).
(38) Hernández-Lobato, J. M.; Requeima, J.; Pyzer-Knapp, E. O.; Aspuru-Guzik, A. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. *International conference on machine learning* **2017**, 1470−1479.
(39) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackerman, Z.; et al. A deep learning approach to antibiotic discovery. *Cell* **2020**, *180*, 688−702.e13.
(40) Corsello, S. M.; Bittker, J. A.; Liu, Z.; Gould, J.; McCarren, P.; Hirschman, J. E.; Johnston, S. E.; Vrcic, A.; Wong, B.; Khan, M.; et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **2017**, *23*, 405−408.
(41) Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight uncertainty in neural network. *International Conference on Machine Learning* **2015**, 1613−1622.
(42) Imbalzano, G.; Zhuang, Y.; Kapil, V.; Rossi, K.; Engel, E. A.; Grasselli, F.; Ceriotti, M. Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.* **2021**, *154*, 074102.
(43) Rufa, D. A.; Macdonald, H. E. B.; Fass, J.; Wieder, M.; Grinaway, P. B.; Roitberg, A. E.; Isayev, O.; Chodera, J. D. Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning/molecular mechanics potentials. *BioRxiv* **2020**, DOI: 10.1101/2020.07.29.227959.
(44) Schütt, K.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems* **2017**, 991−1001.
(45) Townshend, R. J.; Vögele, M.; Suriana, P.; Derry, A.; Powers, A.; Laloudakis, Y.; Balachandar, S.; Anderson, B.; Eismann, S.; Kondor, R., et al. ATOM3D: Tasks On Molecules in Three Dimensions. 2020, arXiv preprint arXiv:2012.04035. https://arxiv.org/abs/2012.04035 (accessed 2021-07-19).
(46) Kearnes, S. Pursuing a Prospective Perspective. *Trends in Chemistry* **2021**, *3*, 77.
(47) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. 2019, arXiv preprint arXiv:1912.01703. https://arxiv.org/abs/1912.01703 (accessed 2021-07-19).
(48) Landrum, G.; et al. Rdkit: Open-source cheminformatics software. *GitHub and SourceForge* **2016**, *10*, 3592822.
(49) Murphy, K. P. Conjugate Bayesian analysis of the Gaussian distribution. *def* **2007**, *1*, 16.
(50) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(51) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K.-R. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.
(52) Wenlock, M.; Tomkinson, N. *Experimental in Vitro DMPK and Physicochemical Data on a Set of Publicly Disclosed Compounds*; 2015.
(53) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative structure- activity relationship modeling of rat

acute toxicity by oral exposure. *Chem. Res. Toxicol.* **2009**, *22*, 1913−1921.
(54) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(55) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261−272.

https://doi.org/10.1021/acscentsci.1c00546
ACS Cent. Sci. 2021, 7, 1356−1367