Evidential Deep Learning for Guided Molecular Property Prediction and Discovery

Ava P. Soleimany[1,2*], **Alexander Amini**[3*], **Samuel Goldman**[4,5*],
Daniela Rus[3], **Sangeeta N. Bhatia**[1], **Connor W. Coley**[4]
[1] Health Sciences & Technology, MIT
[2] Graduate Program in Biophysics, Harvard University
[3] Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT
[4] Chemical Engineering, MIT
[5] Computational and Systems Biology (CSB), MIT
[*] Equal contribution

Abstract

While neural networks (NNs) achieve state-of-the-art accuracy for many tasks in quantitative structure-activity relationship (QSAR) modeling, they can struggle with generalization to out-of-domain examples, poor sample efficiency, and uncalibrated predictions for drug discovery. In this paper, we leverage advances in evidential deep learning to demonstrate a new approach to uncertainty quantification for molecular structure-property regression at no additional computational cost for both message passing and atomistic NNs on the QM9 benchmark dataset. 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.

1 Introduction

Because neural networks (NNs) are susceptible to failure modes in out-of-distribution regimes, it is critical to understand their predictive confidence, particularly for drug discovery and virtual screening applications where model predictions can guide time- and resource-intensive experimentation. Methods for uncertainty quantification (UQ) can help address these needs and facilitate robust application of neural models across a variety of tasks in the chemical sciences.

Existing approaches to *epistemic* (model) UQ for cheminformatics tasks include sampling based methods, such as ensembling and Monte Carlo (MC) dropout, and Bayesian neural networks [13, 9]. However, these approaches only generate approximations to the underlying uncertainty functions via stochastic sampling, yet incur significant computational costs and runtimes, hindering their application to iterative active learning strategies and their deployment in resource constrained settings.

In contrast, NNs can be trained to predict the parameters of the underlying probability distribution and obtain closed-form solutions of *aleatoric* (data) uncertainty, without sampling [12, 2, 7, 8]. Evidential deep learning [16, 1] frames learning as an evidence acquisition process to infer the parameters of an evidential distribution and simultaneously model both epistemic and aleatoric uncertainty [4].

In summary, the contributions of this work are as follows (Fig. 1):

1. Demonstration of evidential deep learning as a new approach to UQ for molecular structure-property prediction with well-calibrated uncertainties;

Machine Learning for Molecules Workshop at NeurIPS 2020. `https://ml4molecules.github.io`

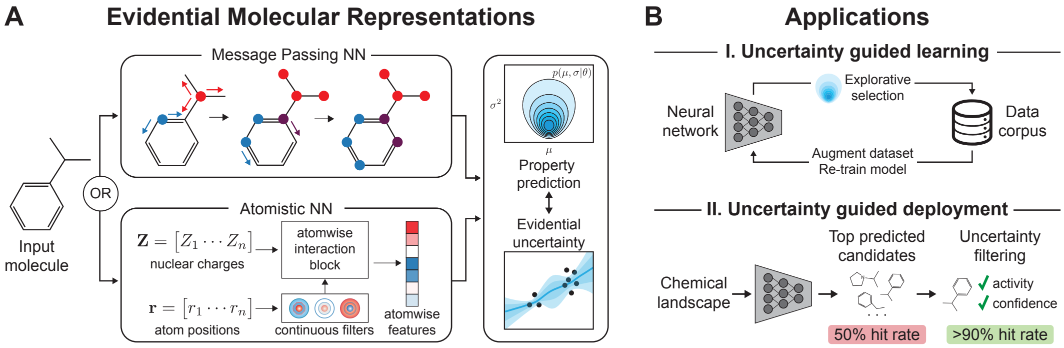

Figure 1: **Evidential uncertainty for molecular prediction and discovery.** **A**. Evidential direct message passing or atomistic neural networks learn molecular representations and predict target properties as well as the parameters of an underlying evidential distribution, capturing the evidence in support of each prediction and enabling uncertainty estimation. **B**. Uncertainties are applied during learning (I) to increase sample efficiency and during deployment (II) to discover high confidence candidates with increased empirical success rates.

2. Validation of its relevance to (a) active learning for sample efficient model training and (b) prioritization of candidates in virtual screening to improve validation rates.

2 Related Work

While a plethora of distance-based and non-parametric methods for UQ have been developed [10, 19], ensembling [11] and MC-dropout [6] are still accepted as state of the art for epistemic UQ for molecular NNs, due in part to their model-agnostic nature and ease of implementation [13, 5, 17]. However, recent analyses [9, 13] have revealed an overwhelming lack of consensus on top performing aleatoric and epistemic UQ methods across molecular property prediction datasets. Most relevant to the applications considered in this work, the atomistic network ANI for atomic energies was improved by repeated acquisition of new training data via ensemble-based query by committee, incurring the cost of retraining multiple models at each acquisition step [17].

3 Methods and Datasets

Evidential deep learning for regression. Evidential models [1, 16] train the network to directly output the parameters of the underlying probability distribution. For continuous (regression) targets, $\mathbf{x}$, these *evidential* distributions can be parameterized with a Normal Inverse-Gamma (`NIG`) over the lower order likelihood parameters: $p(\mu, \sigma^2|\gamma, \lambda, \alpha, \beta)$. The network, which outputs the four `NIG` parameters ($\mathbf{m} = \{\gamma, \lambda, \alpha, \beta\}$), is trained using a multi-objective loss which jointly aims to maximize model fit, $\log\, p(\mathbf{x}|\mathbf{m})$, and minimize evidence on errors. Full training details are in the Appendix.

Network architectures. To show its broad applicability in molecular modeling, we integrate evidential regression into a state-of-the-art D-MPNN model, Chemprop [22], and the end-to-end atomistic NN, SchNet [15], to show performance on 2D graphs and 3D conformers, respectively. The D-MPNN is implemented with default Chemprop parameters (i.e., hidden dimension of 300, 3 layers, and no dropout) [22]. SchNet is implemented with parameters of available pretrained *schnetpack* models (i.e., 128 features, 50 Gaussians, a cosine cutoff of 10 and 6 interaction layers) [14]. Final layers of both models infer a single evidential distribution for each task, with each parameterized by four outputs (e.g., predicting 12 tasks uses 48 outputs).

Datasets. For benchmarking and active learning D-MPNN experiments, we predict all 12 output tasks of the QM9 dataset containing computer-generated quantum mechanical properties [21]. The single task of total formation energy, U_0 [14], is used for SchNet. For virtual screening experiments, models are trained on small molecules and their *in vitro* growth inhibitory activity against *E. coli* [18]. Models were evaluated on the Drug Repurposing Hub library [3], and predictions were compared to empirically determined activities for a subset of these molecules [18].

4 Results

4.1 Uncertainty calibration and benchmarking

We first sought to demonstrate that our evidential learning algorithm could yield well-calibrated uncertainties on molecular and atomistic property prediction, such that the lowest uncertainty samples have the lowest error. We compared to Monte Carlo dropout and ensembling (both with $n = 5$ samples) as baselines for UQ and utilized both D-MPNN and SchNet architectures.

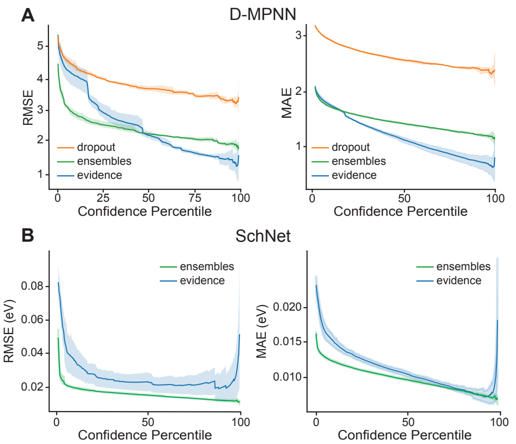

Figure 2: **Calibration of evidential uncertainties**. Average test set RMSE and MAE when considering only a subset of the most confident predictions for D-MPNN (**A**) and SchNet (**B**). Mean $\pm$ s.d. over $n = 5$ runs. Dropout is used during training only with $d = 0.1$. Cutoffs are computed every 30 datapoints, exclusively.

We trained networks on QM9, and evaluated error and uncertainties on held-out test sets using a random $(0.8, 0.1, 0.1)$ train-validation-test split. To compare performance, we ranked the test set by predicted confidence and computed mean average error (MAE) and root mean squared error (RMSE) at different confidence cutoffs (Fig. 2). Because the D-MPNN models report $n_{\text{tasks}} = 12$ different confidence values for each molecule, cutoff scores were computed for each output task separately before averaging across tasks to produce a single performance metric. We report individual task curves in Appendix Fig. S1 and Fig. S2.

In the 2D setting (D-MPNN), evidential uncertainty outperformed ensemble and dropout based methods in terms of the steepness of MAE decrease with increasing confidence (Fig. 2A). Despite showing decreased performance on the RMSE metric, evidential uncertainty demonstrated a steeper decline in error as a function of confidence and had lower RMSE than both dropout and ensembles at high confidence percentiles (Fig. 2A). In the 3D setting (SchNet), evidential learning did not match ensemble model performance and displayed uncalibrated confidence for the 40% *highest* confidence predictions for the RMSE metric (Fig. 2B). While RMSE is less calibrated than the ensembling approach, MAE still declines as a function of confidence, yielding uncertainties without increased compute cost or model architecture changes (e.g. dropout) (Fig. 2B).

4.2 Active learning using evidential uncertainties

One application of UQ in a downstream learning task is to guide the training process through active learning, where data is iteratively added to the training dataset according to an *acquisition function* quantifying its utility; here, we use the estimated uncertainty as the acquisition function. We evaluate the efficacy of explorative active learning for each of the dropout, ensembling, and evidential learning UQ methods described above. Data can be acquired either greedily or sampled stochastically.

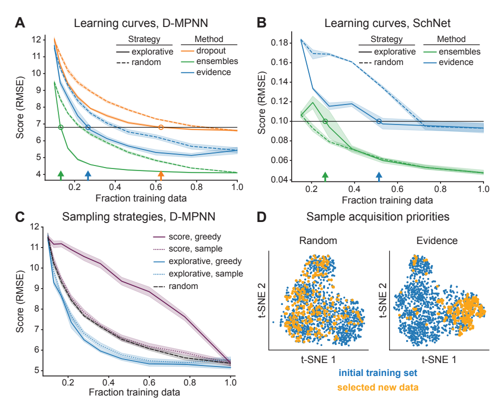

Figure 3: **Evidential active learning on QM9**. Performance of explorative versus uniform (random) sampling for D-MPNN (**A**) and SchNet (**B**). (**C**) Effect of different acquisition strategies (explorative, score, random) for D-MPNN. (**D**) t-SNE visualization of actively and randomly selected training samples.

Active learning trials were initialized with a random 10% or 15% subset of the training data. At each step, the uncertainty was evaluated across the remainder of the training data, and used to iteratively add new samples for the next round of training. Model error was evaluated on a held out test set. For all evaluations, random sample selection served as a baseline for each uncertainty quantification method considered.

For both the D-MPNN and SchNet, we found that active selection based on evidential uncertainties improved sample efficiency by 36% and 53% over random acquisition, respectively (Fig. 3A, B). Further, for the D-MPNN, acquisition using evidential uncertainties resulted in increased data efficiency relative to dropout-based selection; as an example, to achieve an RMSE of 6.8, evidence-guided models required an average of 26% of the entire training data compared to 62% for dropout-guided models (Fig. 3A). For the D-MPNN, uncertainties derived from model ensembling resulted in the greatest improvement over random selection; however, ensembling requires training multiple independent models at each active learning step and thus carries a significant computational expense. In contrast, evidential learning enables resource efficient uncertainty estimation at single-model cost and still achieves increased training efficiency compared to random acquisition.

We next considered whether data acquisition would be improved by access to an oracle that measures absolute error on held out training data as the acquisition function score. Score-based strategies did not improve D-MPNN training relative to random selection (Fig. 3C). Furthermore, this difference was consistent for both stochastic sampling and greedy selection of the acquisition function. These results highlight that principled uncertainty estimates can be more informative than predictive error or random selection in identifying data that add new knowledge to the model (Fig. 3D).

4.3 Discovery of high confidence drug candidates via retrospective virtual screening

Lastly, we investigated the potential for evidential deep learning to discover high confidence drug candidates in a retrospective virtual screening campaign. We trained a D-MPNN with the evidential loss on a recent antibiotic discovery dataset of 2,335 small molecules and their *in vitro* growth inhibition against *E. coli* (measured as OD_{600}) [18]. Training details are available in the Appendix. The resulting model achieved robust performance on a held-out subset of this dataset (Fig. 4A).

We applied the trained model to the Broad Drug Repurposing Hub [3] and visualized the structural overlap between these test molecules and the training set as well as their estimated evidential uncertainties (Fig. 4B). Comparing predicted antibacterial activity to evidential uncertainty demonstrated that predicted active molecules (lower predicted OD_{600}) trended towards higher uncertainties (Fig. 4C), as expected, due to the stark imbalance and skewness of the training set (Fig. 4A).

We utilized evidential uncertainties to prioritize high confidence candidate antibiotics, with the goal of nominating compound sets with high experimental hit rates. In line with the approach outlined in [18], the top k Drug Repurposing Hub molecules were selected based on their predicted activity. Molecules with confidence values below the p^{th} percentile within this set were removed using varying thresholds, p. Experimental hit rates (true $OD_{600} < 0.2$) for these model-nominated compounds were estimated using the subset of candidates for which empirically determined antibiotic activity was reported [18] (Fig. 4D). This analysis revealed that augmenting NN predictions with confidence-based filtering could increase the validation rate relative to that of the unfiltered set of $k = 50$ candidates.

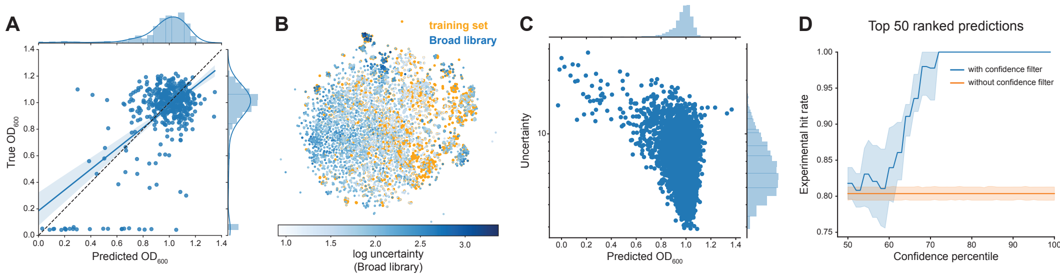

Figure 4: **Uncertainty guided nomination of candidate antibiotics. A.** Performance of evidential D-MPNN after training to predict *E. coli* growth inhibition. **B.** t-SNE visualization of training set (orange) and the Broad library, colored by predicted evidential uncertainties (blue). **C.** Predicted growth inhibition of *E. coli* against estimated uncertainty for compounds in Broad library. **D.** Application of confidence filters to prioritize sets of antibiotic candidates with high experimental hit rates. Mean $\pm$ s.d., $n = 5$.

5 Conclusion

We demonstrate how recently developed evidential deep learning methods can be used for computationally inexpensive UQ in cheminformatics. By benchmarking on two separate architectures for small molecule 2D graphs and 3D conformers, we showcase the modularity of evidential UQ. These uncertainties prove useful in both active learning of quantum mechanical surrogate models and the discovery of novel antibiotic compounds, supporting the generalizability and promise of evidential deep learning for molecular machine learning.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33, 2020.

[2] Christopher M Bishop. Mixture density networks. 1994.

[3] Steven M Corsello, Joshua A Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E Hirschman, Stephen E Johnston, Anita Vrcic, Bang Wong, Mariya Khan, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nature medicine*, 23(4):405–408, 2017.

[4] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[5] Natalie S Eyke, William H Green, and Klavs F Jensen. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry & Engineering*, 2020.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[7] Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, and Daniela Rus. Deep orientation uncertainty learning based on a Bingham loss. In *International Conference on Learning Representations*, 2019.

[8] Pavel Gurevich and Hannes Stuke. Gradient conjugate priors and multi-layer neural networks. *Artificial Intelligence*, 278:103184, 2020.

[9] Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020. PMID: 32702986.

[10] Jon Paul Janet, Chenru Duan, Tzuhsiung Yang, Aditya Nandy, and Heather J Kulik. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chemical science*, 10(34):7913–7922, 2019.

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[12] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, June 1994.

[13] Gabriele Scalia, Colin A. Grambow, Barbara Pernici, Yi-Pei Li, and William H. Green. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning Based Molecular Property Prediction. *Journal of Chemical Information and Modeling*, 2020. ISBN: 1549-9596 Publisher: ACS Publications.

[14] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, and K.-R. Müller. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *Journal of Chemical Theory and Computation*, 15(1):448–455, January 2019. Publisher: American Chemical Society.

[15] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in neural information processing systems*, pages 991–1001, 2017.

[16] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.

[17] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of chemical physics*, 148(24):241733, 2018. ISBN: 0021-9606 Publisher: AIP Publishing LLC.

[18] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackerman, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[19] Kevin Tran, Willie Neiswanger, Junwoong Yoon, Qingyang Zhang, Eric Xing, and Zachary W Ulissi. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 2020.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[21] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018. Publisher: Royal Society of Chemistry.

[22] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

S1 Supplementary Methods

S1.1 Evidential learning

We consider learning problems with continuous targets, where the observed targets, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, are drawn i.i.d. from a Gaussian distribution with *unknown mean and variance* (μ, σ^2), which we seek to probabilistically estimate. We model this by placing a conjugate prior distribution over these two likelihood parameters. Specifically, we place a Gaussian prior on our unknown mean, $p(\mu)$, and an Inverse-Gamma prior on our unknown variance, $p(\sigma^2)$. Our aim is to learn the joint posterior distribution, referred to as the evidential distribution, through observation of our targets $p(\mu, \sigma^2 | \boldsymbol{m}; \boldsymbol{x})$:

$$p(\underbrace{\mu, \sigma^2}_{\boldsymbol{\theta}} | \underbrace{\gamma, \upsilon, \alpha, \beta}_{\boldsymbol{m}}) = \frac{\beta^\alpha \sqrt{\upsilon}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{2\beta + \upsilon(\gamma-\mu)^2}{2\sigma^2}\right\}. \tag{S1}$$

Practically, this amounts to training a neural network to infer the four model parameters $\boldsymbol{m} = \{\gamma, \upsilon, \alpha, \beta\}$ defining the evidential distribution for any given input. For a given realization of $\boldsymbol{m}$, we have a closed form equation for the distribution's density function (Eq. S1 to compute the first and second order moments (i.e., the prediction and uncertainty) directly:

$$\underbrace{\mathbb{E}[\mu] = \gamma,}_{\text{prediction}} \qquad \underbrace{\mathbb{E}[\sigma^2] = \tfrac{\beta}{\alpha-1},}_{\text{aleatoric}} \qquad \underbrace{\mathrm{Var}[\mu] = \tfrac{\beta}{\upsilon(\alpha-1)}}_{\text{epistemic}}. \tag{S2}$$

The loss function for evidential regression networks is a multi-objective function that simultaneously aims to maximize model evidence ($\mathcal{L}_{\text{NLL}}$) and minimize incorrect evidence on errors ($\mathcal{L}_{\text{REG}}$). The model evidence, also commonly referred to as the marginal likelihood, $p(\boldsymbol{x}, \boldsymbol{m})$, can be computed through a double integral marginalizing over both likelihood parameters, (μ, σ^2). In the case of the `NIG` distribution, this marginalization reduces to:

$$p(\boldsymbol{x}|\boldsymbol{m}) = \mathrm{St}\left(\boldsymbol{x}; \gamma, \frac{\beta(1+\upsilon)}{\upsilon\,\alpha}, 2\alpha\right). \tag{S3}$$

where $\mathrm{St}\left(x; \mu_{\text{St}}, \sigma^2_{\text{St}}, \upsilon_{St}\right)$ is the Student-t distribution evaluated at x with location μ_{St}, scale σ^2_{St}, and υ_{St} degrees of freedom.

The total loss, $\mathcal{L}(\boldsymbol{x})$, is computed as:

$$\mathcal{L}(\boldsymbol{x}) = \mathcal{L}_{\text{NLL}}(\boldsymbol{x}) + \lambda\, \mathcal{L}_{\text{REG}}(\boldsymbol{x}) \tag{S4}$$

$$= \log p(\boldsymbol{x}|\boldsymbol{m}) + \lambda \cdot ||\boldsymbol{x} - \gamma||_p \cdot \Phi \tag{S5}$$

where λ is a regularization coefficient, $||\cdot||_p$ is a p-norm, and $\Phi = 2\upsilon + \alpha$ is the cumulative evidence of the predicted distribution. For additional details on the evidential loss please refer to [**?**].

S1.2 Experimental details

S1.2.1 Training hyperparameters

2D D-MPNN models were trained using the Adam optimizer and Noam scheduler as detailed in [20] and implemented with D-MPNN's in [22]. We use the default Chemprop learning rate parameters, specifically a batch size of 50, 2.0 warmup epochs, an initial learning rate of $1e-4$, max learning rate of $1e-3$, and final learning rate of $1e-4$ after decay.

For benchmarking the 3D SchNet models, we use this same Noam scheduler, increasing the batch size to 100 for consistency with [14]. These benchmarking parameters are summarized in Table S1.

For active learning experiments with 3D SchNet models, we found instability early in the sampling procedure and find empirically better performance with a maximum learning rate, $\eta = 2e-4$. In all other active learning experiments with the D-MPNN, we maintain the training hyperparameters used in benchmarking.

Table S1

Hyperparameter	**Value**
Batch size	(50, 100)
Optimizer	Adam
Initial η	$1e{-}4$
Max η	($1e{-}3$, $2e{-}4$)
Final η	$1e{-}4$
Train, val, test split	80-10-10
Epochs	100
Warmup epochs	2

Table S2: Hyperparameter selection for D-MPNN and SchNet model training. Where differences exist between the two models, both hyperparameters are provided in a tuple.

S1.2.2 Active learning sampling strategies

In this work we evaluated three different forms of acquisition function. The primary baseline is the random (uniform) sampling selection, wherein new data points are selected from the data corpus at random with a uniform prior. Next, we evaluate acquisition functions computed according to the estimated uncertainty of making a prediction with that data point. Points with larger uncertainty will have a larger acquisition score and vice versa. Finally, for score-based acquisition we assume access to an oracle labeler to inform the score or error of the data point. Points with larger error will have larger acquisition likelihood. The score baseline is a non-realistic baseline, but provides a valuable comparison as it demonstrates the concrete benefit of leveraging uncertainty, especially where we do not have ground truth labels as in the active learning domain.

Given an acquisition function we also provide comparison with two different sampling strategies: (1) greedy and (2) stochastic. In a greedy strategy, the data point with the top acquisition score will be sampled deterministically. Alternatively, we can use the acquisition function to define a categorical distribution over each data point weighted accordingly. A new data point is then selected by stochastically drawing a random choice from this categorical distribution.

S1.3 Retrospective virtual screening

While the original work [18] binarized the antibiotic dataset and trained a classification model for predicting activity, here we reconsider this as a regression task and learn on the raw continuous targets to predict each compound's growth inhibition of *E. coli*, measured as OD_{600} (lower, more growth inhibitory).

Given the limited quantity of data available, we augmented representations learned by the D-MPNN with 200 molecular features computed in RDKit as in the original analysis [18]. We train a D-MPNN model to predict evidential uncertainties over OD_{600} using a 80-10-10 split size and the same learning rate parameters used in benchmarking.

After training models to predict evidential uncertainties on the original $2,335$ molecule dataset, we apply models to the Broad Drug Repurposing Hub [3] and select the top-k compounds with the highest predicted inhibition. In the original analysis, the top $k = 99$ predicted molecules were empirically tested for hits ($OD_{600} < 0.2$) to compute ground truth inhibition values for evaluation [18]. We select the top $k = 50$ to maximize overlap with the originally selected 99 molecules. Given the top predictions we apply an additional round of filtering using the uncertainties of each molecule. For a set percentile cutoff threshold, p, we remove any of the k predictions which fall below the bottom p-percentile of uncertainty. In other words, we keep only the predicted molecules which exhibit the highest relative confidence. We compute the empirical hit rate of the remaining molecules based on overlap with the original experimentation [18]. We plot the empirical hit rate as a function of the confidence cutoff (p) across our top k predicted molecules.

S2 Supplementary Figures

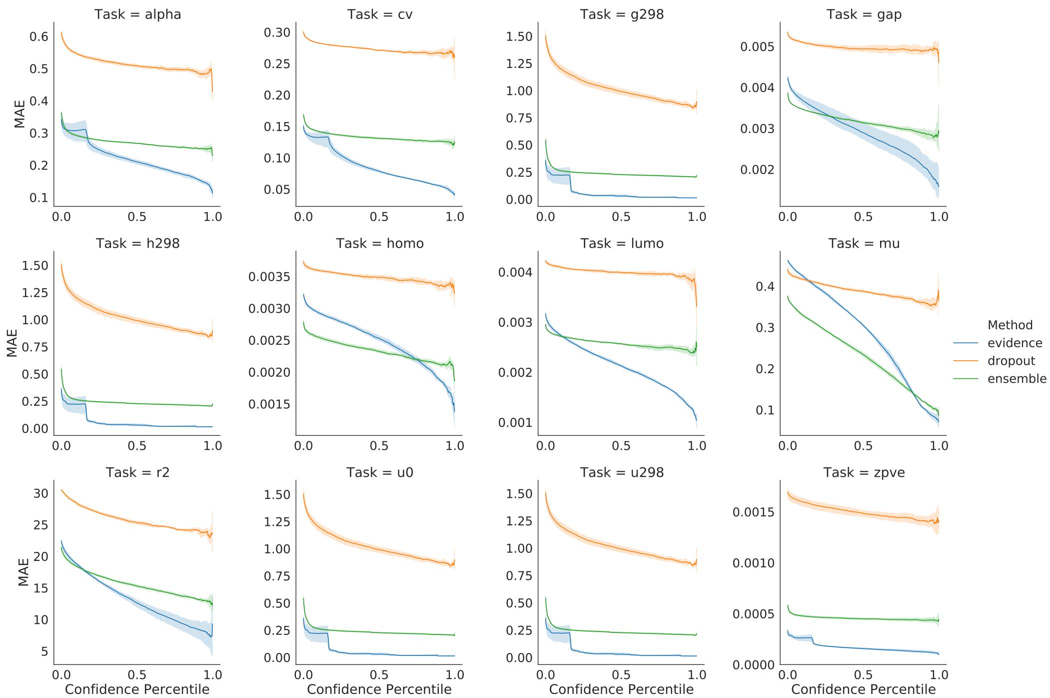

Figure S1: **Task-specific D-MPNN uncertainty benchmarks**. D-MPNN average test set MAE at different uncertainty cutoffs, separated by task. Mean $\pm$ s.d. over $n = 5$.

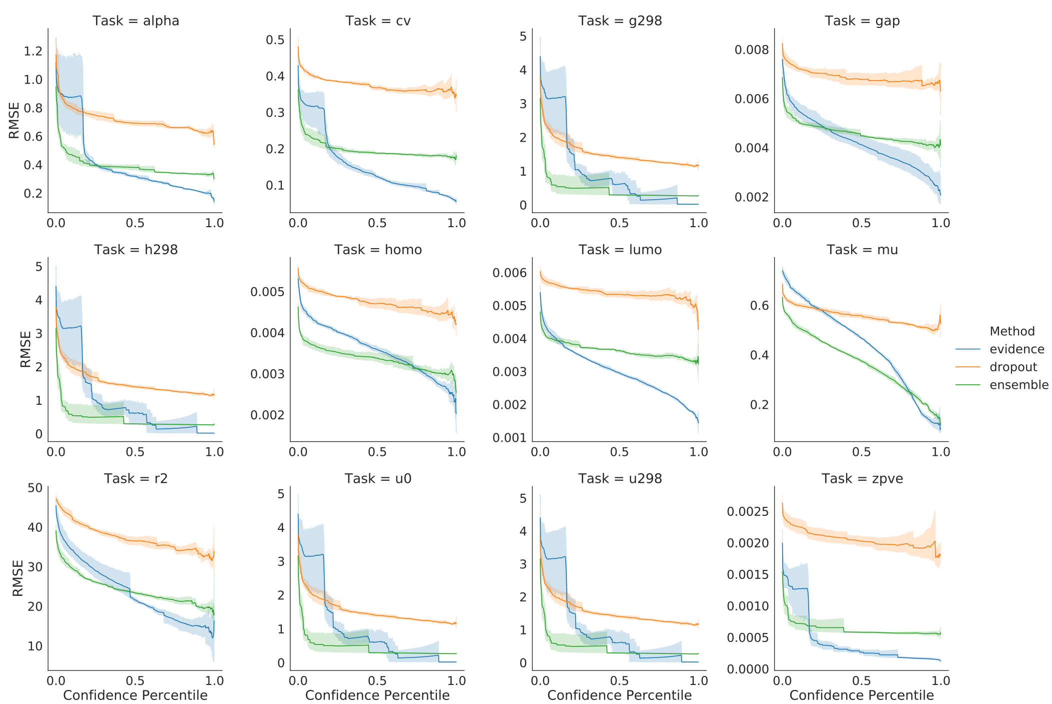

Figure S2: **Task-specific D-MPNN uncertainty benchmarks**. D-MPNN average test set RMSE at different uncertainty cutoffs, separated by task. Mean $\pm$ s.d. over $n = 5$.