Reliable Graph Neural Networks for Drug Discovery Under Distributional Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

The concern of overconfident mispredictions under distributional shift demands extensive reliability research on Graph Neural Networks used in critical tasks in drug discovery. Here we first introduce **CardioTox**, a real-world benchmark on drug cardiotoxicity to facilitate such efforts. Our exploratory study shows overconfident mispredictions are often distant from training data. That leads us to develop distance-aware GNNs: **GNN-SNGP**. Through evaluation on CardioTox and three established benchmarks, we demonstrate GNN-SNGP's effectiveness in increasing distance-awareness, reducing overconfident mispredictions and making better calibrated predictions without sacrificing accuracy performance. Our ablation study further reveals the embeddings learned by GNN-SNGP improves distance-preservation over its base architecture and is one major factor for improvements.

1 Introduction

In recent years, Graph Neural Networks (GNNs) have illustrated remarkable performance in scientific applications such as physics [Battaglia et al., 2016, Sanchez-Gonzalez et al., 2018], biomedical science [Fout et al., 2017], and computational chemistry [Duvenaud et al., 2015, Kearnes et al., 2016]. One important example is the early-phase drug discovery, where GNNs have shown promise in critical tasks such as *hit-finding* and *liability screening* (i.e., predicting the *binding affinity* and the *toxicity* of candidate drug molecules, respectively) [McCloskey et al., 2020, Siramshetty et al., 2020].

However, a key reliability concern that hinders the adoption of GNN in real practice is the *overconfident mispredictions*. For example, in liability screening, an Overconfident False Negative (OFN hereafter) prediction leads the GNN model to mark a toxic molecule as safe, causing severe consequences by leaking it to the next stage of drug development. Such concern is further exacerbated by a second key characteristic of the drug discovery tasks: *data distributional shift*; drug discovery tasks often explicitly evaluate novel molecules by moving into areas that are not previously represented in training data. As a result, the testing molecules are characteristically distinct from the training data, and can carry novel toxic signatures that were previously unseen by the model. A model with just high in-distribution accuracy is not sufficient under those situations; quantifying model reliability and designing techniques to improve robustness against overconfidence under distributional shift become especially relevant.

Currently, there lacks a drug discovery benchmark that targets realistic concerns about model reliability under distributional shifts. Thus, we introduce **CardioTox**, a data benchmark based on a real-world drug discovery problem and is compiled from 9K+ drug-like molecules from ChEMBL, NCATS and FDA validation databases [Siramshetty et al., 2020]. To evaluate model reliability, we generate additional molecular annotations and propose novel metrics to measure the models against the real-world standards around the responsible application of GNN. This is our first contribution.

Using CardioTox, our second contribution is an exploratory study on the root causes behind the overconfident mispredictions under distributional shift, and principled modeling approaches to miti-

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

gate it. In particular, we observe that many overconfidently mispredicted molecules are structurally distinct from training data (i.e., they are "far" from training data based on molecule-fingerprint graph distance [Rogers and Hahn, 2010], see Section 3). This failure mode suggests that improving the *distance-awareness* of a GNN model would be an effective solution: a test molecule that is far away from the decision boundary (e.g., a toxic but novel molecule, due to its novelty, may possess few prior toxic signatures defining the decision boundary) should still get low confidence if it's distant from training data. To this end, the recently proposed Spectral-normalized Neural Gaussian Processes (SNGP) [Liu et al., 2020] demonstrates a concrete approach to achieve this goal. Specifically, it imposes a distance-preserving regularization (i.e., spectral normalization) to the feature extractor, and replaces the dense output layer with a distance-aware classifier (i.e., random-feature Gaussian process). SNGP has shown promising robustness improvements in vision and language problems. To our best knowledge, this is the first study to bring the distance-aware design principle to GNN models to to improve reliability performance on the molecule graph.

To summarize, our contributions are the following (Appendix F summarizes the related work):

- **Data**: We introduce **CardioTox**, a real-world drug discovery dataset with multiple test sets (IID and distribution-shifted ones), to the robustness community to facilitate reliability research of graph models. The distributional shift challenge reflected in CardioTox is realistically faced by the field: test domain often comes from a data source that's different from the training source and has considerable amount of novel molecule graph structures (see also Figure S3 in Appendix E). We further design distance-based data splits for CardioTox to quantitatively measure model's distance-awareness.

- **Model**: We develop an end-to-end trainable **GNN-SNGP** architecture as well as its ablated version GNN-GP. Empirically, this method outperforms its base architecture in not only accuracy (e.g., AUROC) but also robustness (e.g., ECE) especially in reducing overconfident mispredictions. Measured by CardioTox's distance-based data splits, GNN-SNGP shows higher distance-awareness under data shifts than the baselines, which explains its ability in overconfident misprediction reduction. Our implementation together with CardioTox dataset will be open sourced via Github (link hidden to preserve anonymity).

- **Ablation Study**: We carry out extensive ablation studies which confirm that the robustness improvements come from both distance-aware classifier (GP-layer) and the distance-preserving latent representations. We also investigate the generalizability of this approach by testing on other graph modeling domain: molHIV, BBBP and BACE and obtained consistent results.

2 Methods

GNN baseline We use a vanilla Message Passing Neural Network (MPNN) [Gilmer et al., 2017] as the GNN baseline in this study. Specifically, the message function is modeled by a dense layer $M(h_v, h_w, e_{vw}) = W_1[h_v||h_w||e_{vw}]$, where W_1 is the weight matrix. After getting aggregated message m_v for each node, the hidden node feature h_v is updated via a Gated Recurrent Unit $U(h_v, m_v) = \mathsf{GRU}(h_v, m_v)$. Finally, we read out graph level representation by $R = \sum_{v \in V} \sigma(W_2[h_v||h_v^0]) \odot (W_3 h_v)$, where W_2, W_3 are two weight matrices and σ is the sigmoid function. The graph level representation is fed into a final dense layer to generate logits.

GNN-GP: improving distance-awareness of classifier Following Liu et al. [2020] who proposed Gaussian process layer (GP-layer hereafter) in vision and language models, we introduce it to the graph domain and developed GNN-GP model to increase distance-awareness. Figure S1 (Appendix B) shows the high-level architectural changes.

As Figure 1 shows in detail, Gaussian processes are made end-to-end trainable with GNN by approximating GP kernel function via random Fourier feature generation [Rahimi and Recht, 2007]. Specifically, ω_i, b_i are randomly sampled at initialization and kept fixed during training. The distribution that ω_i, b_i are sampled from depends on what Gaussian processes kernel we'd like to approximate. Take the Gaussian kernel as an example, we have $\omega_i \sim N(0, 1); b_i \sim U(0, 2\pi)$. During inference time, each sample would get logit predictions as well logit variances, both of which are utilized to compute predictive probabilities via men-field approximation [Lu et al., 2020].

GNN-SNGP: incorporating distance-preserving feature extraction Due to feature collapse in feature extraction [Liu et al., 2020, van Amersfoort et al., 2021], neural representation may not faithfully preserve distance in the input manifold. Liu et al. [2020] propose to preserve input distance

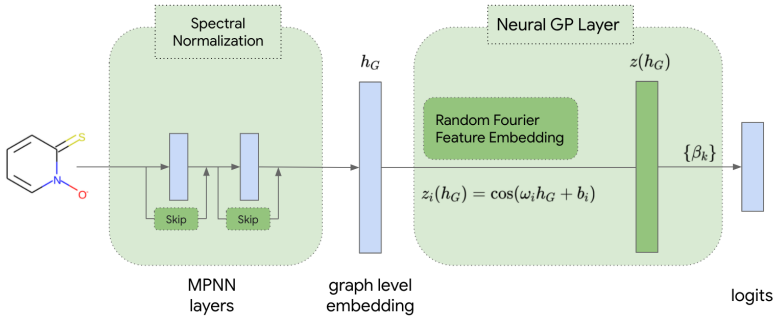


Figure 1: Improving GNN architecture via distance-preserving feature extractor (skip connection and spectral normalization) and distance-aware classifier (neural Gaussian Processes layer).

in feature extraction by applying Spectral Normalization (SN) [Gouk et al., 2021, Miyato et al., 2018] to the residual networks. As a result, combining SN and GP would increase the model's overall distance-awareness, helping OOD detection and other robustness metrics.

Since our vanilla MPNN model (GNN baseline) does not have residual connections, we create GNN-SNGP through two following changes (Figure 1). First, we have message function modeled by a dense layer with SN: $M(h_v, h_w, e_{vw}) = W_1^{SN}[h_v||h_w||e_{vw}]$ where W_1^{SN} is the dense layer weight matrix whose spectral norm is regularized. Second, we add residual connection to the message passing layer through the node update function: $U(h_v, m_v) = \mathsf{GRU}(h_v, m_v) + h_v$.

3 CardioTox: Drug cardiotoxicity under distributional shift

Dataset information In early drug discovery stages, two types of models are widely used: hit-finding model and anti-target (i.e., liability) model. The former aims to score molecule based on its potential of making a drug (e.g., whether a molecule could bind tightly to the disease-causing protein target), while the latter aims to filter out those molecules with potentially high liability (e.g., whether a molecule could be toxic). One such liability is cardiotoxicity, which occurs when a molecule inhibits hERG, a protein target that is related to heart-rhythm control [Smith et al., 1996, Vandenberg et al., 2012]. Due to previous failing incidents, FDA has required new drugs to pass drug cardiotoxicity examination [Center for Drug Evaluation and Research, 2005].

Thus in this work, we introduce CardioTox, a benchmark based on a real-world drug discovery problem and is compiled from 9K+ drug-like molecules from ChEMBL and NCATS databases [Siramshetty et al., 2020]. To evaluate GNN model reliability, we add graph structural information (e.g., node features) and set up three test sets that reflect distributional shifts: Test-IID is a set that's sampled from the same distribution as the Train set. Test-OOD1 and Test-OOD2 are molecule sets coming from NCATS and FDA respectively: 84% of Test-OOD1 and 82% of Test-OOD2 are novel molecules that are distant from Train set (Figure S3). We further generate additional molecular annotations: *close* sample or *far* sample based on Tanimoto fingerprint distance [Bajusz et al., 2015] to the train set (Appendix E) . This setup allows us to quantitatively assess models' distance-awareness.

Evaluation metrics We evaluate popular metrics for uncertainty such as Expected Calibration Error (ECE), Brier score (Brier), Negative Log Likelihood (NLL) (cf. [Ovadia et al., 2019] for an overview). To assess the models' abilities in capturing graph distance and mitigating over-confidence, we also introduce two new metrics: *Percentage of Over-confident False Negatives* (OFNs %): $\frac{\sum_i \mathbb{1}(\hat{y}_i < 0.1 \text{ and } y_i^{\text{true}} = 1)}{\sum_i \mathbb{1}(\hat{y}_i < 0.1)}$ where $\hat{y}_i, y_i^{\text{true}}$ are predictive probability and label for the positive class (i.e., toxic class in CardioTox) for the i-th test sample, and *Distance-Awareness AUC* (DA-AUC) that evaluates if the model assigns higher confidence to test inputs in *close set* than those in *far set* (find details Appendix C). Note that OFNs % can be "cheated" by making all predictions under-confident, therefore an ideal uncertainty model should perform well *both* in calibration metrics and in OFNs %.

Accuracy and robustness performance Table 1 shows that GNN-GP outperforms GNN baseline in both AUROC and robustness metrics for the CardioTox task. This is the case for both the in-distribution (Test-IID) and the shifted test sets (Test-OOD1 and Test-OOD2). GNN-SNGP shows additional gains in robustness. If resource allows, Deep Ensemble [Lakshminarayanan et al., 2017] of

GNN-SNGP further boosts performance, eliminating all OFNs in Test-OOD2. It is worth noting that our GNN-SNGP ensemble has outperformed previous state-of-the-art neural models [Siramshetty et al., 2020] in AUROC performance.

Table 1: Accuracy (AUROC), robustness (ECE, Brier, NLL) and overconfidence (OFNs) performance for Drug cardiotoxicity benchmark. Results are averaged over 10 seeds.

Test-IID	AUROC (↑)	ECE (↓)	Brier (↓)	NLL (↓)	OFNs% (↓)	DA-AUC (↑)
GNN baseline	0.919±0.003	0.037±0.003	0.194±0.007	0.352±0.018	4.05±0.26	0.500±0.004
GNN-GP	**0.937±0.001**	0.036±0.002	**0.176±0.008**	**0.298±0.015**	**3.51±0.17**	**0.523±0.004**
GNN-SNGP	0.932±0.001	**0.028±0.001**	**0.179±0.005**	**0.295±0.005**	**3.56±0.21**	0.517±0.004
SNGP Ensemble	**0.942±0.000**	**0.013±0.001**	**0.162±0.000**	**0.264±0.001**	**2.54±0.07**	**0.525±0.002**
Test-OOD1	**AUROC (↑)**	**ECE (↓)**	**Brier (↓)**	**NLL (↓)**	**OFNs% (↓)**	**DA-AUC (↑)**
GNN baseline	0.786±0.004	0.102±0.012	0.343±0.071	0.578±0.125	1.68±0.08	0.604±0.006
GNN-GP	0.823±0.003	0.090±0.010	0.327±0.061	0.546±0.107	1.41±0.12	**0.632±0.002**
GNN-SNGP	**0.836±0.003**	**0.074±0.008**	**0.316±0.047**	**0.503±0.072**	**1.31±0.09**	**0.635±0.007**
SNGP Ensemble	**0.850±0.002**	**0.039±0.002**	**0.277±0.005**	**0.428±0.006**	**1.22±0.08**	**0.643±0.003**
Test-OOD2	**AUROC (↑)**	**ECE (↓)**	**Brier (↓)**	**NLL (↓)**	**OFNs% (↓)**	**DA-AUC (↑)**
GNN baseline	0.831±0.007	0.082±0.012	0.284±0.060	0.492±0.113	1.73±0.20	0.630±0.008
GNN-GP	0.873±0.006	0.074±0.007	0.261±0.047	0.442±0.086	**0.98±0.19**	0.656±0.011
GNN-SNGP	**0.885±0.007**	**0.044±0.006**	**0.238±0.040**	**0.389±0.068**	**1.02±0.11**	**0.678±0.008**
SNGP Ensemble	**0.896±0.002**	**0.021±0.002**	**0.210±0.003**	**0.333±0.005**	**1.06±0.11**	**0.682±0.003**

Why does distance-awareness help reduce overconfident mispredictions? We observe a significant portion of OFNs are distant from the train set (i.e., 60% of them have Tanimoto distance > 0.65 [Bajusz et al., 2015], also see Figure S2b). This could happen when a GNN model lacks of distance-awareness: a toxic but novel molecule can be far away from the wrong side of model's decision boundary, due to lacking known toxic signatures that is present in train set. Without being aware of the distance to train set, the GNN baseline model tends to base its prediction on distance to the decision boundary and give high confidence for such cases.

To this end, GNN-GP leverages GP's distance-awareness and is able to naturally incorporate distance into predictive uncertainty. As shown in Table 1, the SN version GNN-SNGP achieves highest distance-awareness (DA-AUC) in the two data-source-shifted test sets, correlating well with OFNs reduction. Figure S2b shows a decreasing trend (from GNN baseline to GNN-GP to GNN-SNGP) of the percentage of distant samples among OFNs. Overall, using GP is able to improve uncertainty estimate for over 80% of the baseline OFNs while also improving calibration (Figure S4).

Additional ablations to assess relative contributions of SN and GP In order to understand relative performance contributions of the two modeling components (i.e., distance-preserving feature extractor vs distance-aware classifier), we carry out an extensive ablation study in Appendix I. We take the latent representations learned by GNN baseline, GNN-GP and GNN-SNGP (increasing distance-preservation) and feed them to classifiers with increasing distance-awareness: Dense layer, GP-layer, exact Gaussian processes classifier (GPC hereafter). Table S2 suggests there's synergy between distance-preservation of neural representation and distance-awareness of the classifier. With the least distance-aware classifier (i.e., Dense layer), AUROC drops when increasing distance-preservation in neural representation. With the most distance-aware classifier (i.e., exact GPC), increasing neural representation's distance-preservation benefits accuracy, robustness as well as overconfidence reduction. We find a GNN model achieves best performances when both are present in the architecture (GNN-SNGP embeddings with GPC). Interesting, naively increasing distance-preservation along is not sufficient in guaranteeing good generalization; we observe that the pre-defined representation (i.e., the molecule fingerprint FP) gives relatively low AUROC under data shifts (e.g., on Test-OOD2) despite being perfectly distance-preserving. This is related to the trade-off in representation learning between dimension reduction and information preservation. Appendix I discusses in further detail.

Additional results on existing benchmarks Consistent with the results obtained in CardioTox, GNN-GP also outperforms GNN baseline on three established graph classification benchmarks [Wu et al., 2018]: molHIV, BBBP and BACE (see Appendix H). We can make a few observations on the results in Table S1. First, GNN-GP consistently achieves higher AUROC, improves calibration as well as reduces overconfident mispredictions across all the benchmarks than the baseline. Second, with limited accuracy performance drop, the spectral normalized version GNN-SNGP can further improve model robustness compared with GNN-GP. Our conclusion is included in Appendix A.

References

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.*, 7:20, May 2015.

Peter W Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction Networks for Learning about Objects, Relations and Physics. In *NeurIPS*, 2016.

Center for Drug Evaluation and Research. S7B Nonclinical Evaluation of the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation) by Human Pharmaceuticals, 2005.

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *NeurIPS*, 2015.

Boyuan Feng, Yuke Wang, Zheng Wang, and Yufei Ding. Uncertainty-aware attention graph neural network for defending adversarial attacks. *arXiv preprint arXiv:2009.10235*, 2020.

Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein interface prediction using graph convolutional networks. *NeurIPS*, 2017.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(Database issue):D1100–7, January 2012.

Simon Geisler, Daniel Zügner, and Stephan Günnemann. Reliable graph neural networks via robust aggregation. *arXiv preprint arXiv:2010.15651*, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural Message Passing for Quantum Chemistry. *ICML*, 2017.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. *NeurIPS*, 2017.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W Coley. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.*, 60 (8):3770–3780, August 2020.

Doyeong Hwang, Grace Lee, Hanseok Jo, Seyoul Yoon, and Seongok Ryu. A benchmark study on reliable molecular supervised learning via Bayesian learning. *arXiv preprint arXiv:2006.07021*, 2020.

Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-RevNet: Deep Invertible Networks. *ICLR*, 2018.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.*, 30(8):595–608, August 2016.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 2017.

Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *NeurIPS*, 2020.

Zhiyun Lu, Eugene Ie, and Fei Sha. Uncertainty estimation with infinitesimal jackknife, its distribution and mean-field approximation. *arXiv preprint arXiv:2006.07584*, 2020.

Kevin McCloskey, Eric A Sigel, Steven Kearnes, Ling Xue, Xia Tian, Dennis Moccia, Diana Gikunju, Sana Bazzaz, Betty Chan, Matthew A Clark, John W Cuozzo, Marie-Aude Guié, John P Guilinger, Christelle Huguet, Christopher D Hupp, Anthony D Keefe, Christopher J Mulhern, Ying Zhang, and Patrick Riley. Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit Finding. *J. Med. Chem.*, June 2020.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *NeurIPS*, 2019.

Ali Rahimi and Ben Recht. Random features for large-scale kernel machines. *NeurIPS*, 2007.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5): 742–754, May 2010.

Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. In *ICML*, 2018.

Vishal B Siramshetty, Dac-Trung Nguyen, Natalia J Martinez, Noel T Southall, Anton Simeonov, and Alexey V Zakharov. Critical Assessment of Artificial Intelligence Methods for Prediction of hERG Channel Inhibition in the "Big Data" Era. *J. Chem. Inf. Model.*, December 2020.

P L Smith, T Baukrowitz, and G Yellen. The inward rectification mechanism of the HERG cardiac potassium channel. *Nature*, 379(6568):833–836, February 1996.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409*, 2021.

Jamie I Vandenberg, Matthew D Perry, Mark J Perrin, Stefan A Mann, Ying Ke, and Adam P Hill. hERG k(+) channels: structure, function, and clinical significance. *Physiol. Rev.*, 92(3):1393–1478, July 2012.

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.

A Conclusion

In this study, we introduce GNN-GP and GNN-SNGP together with a new benchmark CardioTox from drug discovery setting. Through evaluation on four datasets, we demonstrate their effectiveness in reducing overconfident mispredictions and making better calibrated GNN models without sacrificing accuracy performance. The improved robustness appears to come from the boost in distance-awareness. We further discover that the embedding space induced by SN and GP addition improves distance-preservation over its base architecture and is one major factor to bring improvements in accuracy, general robustness and overconfidence performance.

B GNN-SNGP architecture

We present high-level architectural changes based on GNN baseline model in Figure S1: distance-preserving feature extractor by adding skip connection and spectral normalization regularization, distance aware classifier using neural GP layer.

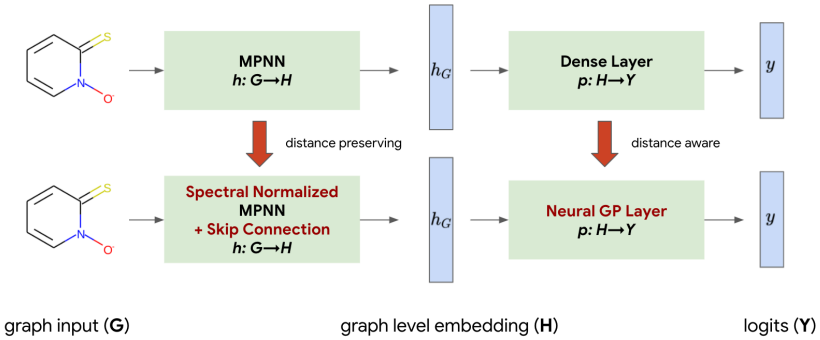


Figure S1: Improving GNN architecture via distance-preserving feature extractor (skip connection and spectral normalization) and distance-aware classifier (neural Gaussian Processes layer).

C Evaluation Metrics

In this study, we examine for each accuracy performance, general robustness performance, overconfidence performance as well as distance-awareness performance.

Accuracy performance AUROC for the classification benchmark.

Robustness performance Expected Calibration Error (abbr. ECE), Brier Score (abbr. Brier), and Negative Log Likelihood (abbr. NLL).

Overconfidence performance: OFNs% and OFPs% Occurance of overconfident false negatives and false positives. Depending on actual applications, we may care more about Overconfident False Negatives (OFNs) in the drug cardiotoxicity task whereas in the molHIV task we care about general overconfidence, therefore both OFNs and OFPs (Overconfident False Positives). In this study, an overconfident misprediction is defined as any test sample whose predictive confidence (computed via Maximum Softmax Probability, i.e., $\max(p_k)$ is higher than 90% yet its prediction is wrong, namely:

$$\max_k(p_k) > 90\% \quad \text{and} \quad \arg\max_k(p_k) \neq y^{\text{true}} \tag{1}$$

where $k \in \{0, 1\}$ is class logit index, and $y^{\text{true}} \in \{0, 1\}$ is the ground truth class.

In this study we measure OFNs% and OFPs%, defined as percentages as follows:

$$\text{OFNs}\% = \frac{\sum_i \mathbb{1}(\hat{y}_i < 0.1 \text{ and } y_i^{\text{true}} = 1)}{\sum_i \mathbb{1}(\hat{y}_i < 0.1)} \tag{2}$$

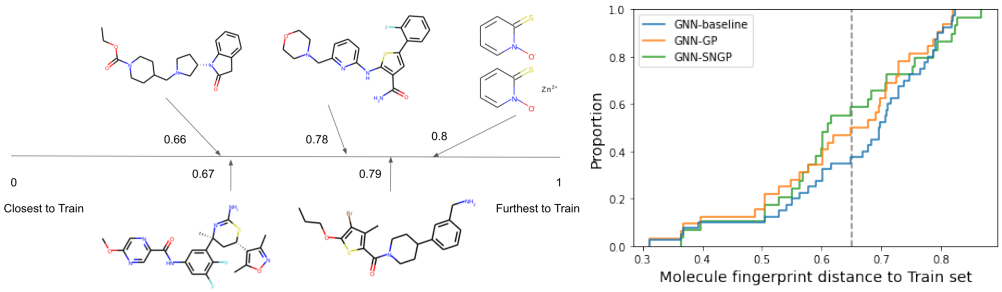
$$\text{OFPs}\% = \frac{\sum_i \mathbb{1}(\hat{y}_i > 0.9 \text{ and } y_i^{\text{true}} = 0)}{\sum_i \mathbb{1}(\hat{y}_i > 0.9)} \tag{3}$$

where $\hat{y}_i$ is i-th sample's predictive probability for positive class, i.e., $p^i_{k=1}$.

Distance-awareness performance: DA-AUC We design a classification task for a given test set: any molecule belonging to the *close set* (short distance to Train set, defined in E) gets ground-truth label 0, any molecule belonging to the *far set* (long distance to Train set, defined in E) gets ground truth label 1. We use predictive uncertainty (computed via 1-max(p_k)) to classify if a test sample is in the *close set* or the *far set*. DA-AUC is the AUROC measurement for this task.

D Distance analysis for OFNs in CardioTox

Here we present our analysis on OFNs incurred in GNN baseline model. One of the outstanding observations is that a good portion of OFNs molecules are distant from the train set. Figure S2a shows a few such example molecules. More quantitatively, Figure S2b suggests over 60% of OFNs have Tanimoto distance > 0.65, which is often regarded as a condition of having novel molecule structure. As we introduce models with more distance-awareness (e.g., GNN-GP, GNN-SNGP), distant molecules becomes less dominant in OFNs.



(a) (b)

Figure S2: Graph distance analysis for the overconfident false negatives (OFNs) in CardioTox benchmark. (a) shows five overconfident false negatives with their distance (in 0-1 scale) to the Train set. (b) Cumulative distributions over fingerprint distance from the OFNs (incurred in corresponding models) to the Train set.

E Drug cardiotoxicity data split

We re-organized the original hERG (cardiotoxicity anti-target) dataset [Siramshetty et al., 2020] into **CardioTox** to facilitate graph model reliability research. The original dataset comes with a Train set, a compilation from ChEMBL [Gaulton et al., 2012] and NCATS (National Center for Advancing Translational Sciences), a prospective validation set from NCATS Siramshetty et al. [2020] and a test set from FDA (see Figure S3).

As a first step, we added graph structural information such as node features, edge features and adjacency matrix for each molecule record and created three test sets: Test-IID (randomly sampled 20% from the Train distribution), Test-OOD1 (the prospective validation set from NCATS) and Test-OOD2 (from FDA). The shift is mainly from input distribution (i.e., graph structures): majority of Test-IID are similar molecules to Train set, while 84% of Test-OOD1 and 82% of Test-OOD2 have novel structures.

As a second step, we further split each test set into *close set* and *far set*. Using Tanimoto graph distance [Bajusz et al., 2015] defined by molecule fingerprint representation [Rogers and Hahn, 2010], we group test samples with distance (average distance to top 8 nearest training samples) < 0.7 into *close set* and remaining samples go to *far set*. Figure S3 shows the detailed splitting flow and final counts.

F Related work

On GNN robustness research, Geisler et al. [2020] proposed a novel GNN aggregation function, Soft Medoid, to improve robustness against adversarial attacks from edge perturbation. Feng et al. [2020] constructed a two-branched GNN architecture where the first branch computes model uncertainty and data uncertainty, the second utilizes those uncertainties to adjust attention during aggregation.

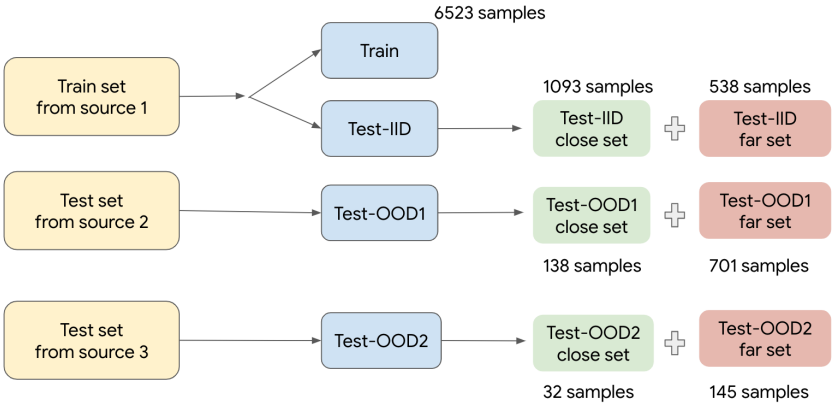


Figure S3: CardioTox benchmark designed for investigation on accuracy and robustness performance under various levels of distributional shift. Majority of Test-IID are similar molecules to Train set, while 84% of Test-OOD1 and 82% of Test-OOD2 have novel structures.

Those studies are organized around defending adversarial attacks, while our work aims to mitigate overconfidence issue by introducing targeted techniques.

In the area of molecule graph learning, Hwang et al. [2020] have recently applied Deep Ensemble, Monte Carlo dropout, Stochastic Gradient Langevin Dynamics (SGLD), Stochastic Weight Averaging (SWA), and Stochastic Weight Averaging Gaussian (SWAG) to GNN models and those techniques can be categorized as different approaches to create ensemble models. For example, SWA picks ensemble members along the training procedure. Our work focuses on a single GNN model and introduces architectural changes to improve its distance-awareness to reduce overconfident mispredictions. Another related work is from Hirschfeld et al. [2020] where a GNN model is first trained and its latent representation gets fed to a downstream Gaussian Processes. In contrast to that, we develop an end-to-end trainable GNN-SNGP without computational constraint of kernel computation and storage that come with exact Gaussian processes, making GNN-SNGP better suited to processing large scale datasets such as high throughput screening data in drug discovery [McCloskey et al., 2020].

On the distance-preservation/awareness side, compared with SNGP [Liu et al., 2020] techniques used in this work, two-sided gradient penalty [Gulrajani et al., 2017, Van Amersfoort et al., 2020] constrains gradient deviation by placing penalty on the loss and can be sensitive to tune. Deep Invertible Networks [Jacobsen et al., 2018] achieves high level of distance-preservation via invertiable layers but are often difficult to train and expensive in memory.

G Uncertainty improvements for OFNs

We have carefully looked into the OFNs generated by the GNN baseline, and assess their uncertainty improvements by monitoring uncertainty increase ratio (UIR):

$$\text{UIR} = \frac{U_{\text{GNN-GP}}}{U_{\text{GNN baseline}}} \tag{4}$$

where $U_{\text{GNN-GP}}, U_{\text{GNN baseline}}$ are the uncertainty estimates by the GNN-GP model and GNN baseline model. UIR > 1 indicates that the GNN-GP becomes less overconfident than the GNN baseline. Figure S4 lists UIRs for the 40 top overconfident false negatives from the GNN baseline. We observe over 80% of them have improved uncertainty estimates (UIR>1). As annotated in Figure S4, many highly improved OFNs are the ones distant from the Train set, while the ones getting worse uncertainty estimates often are close to the Train set. This is expected in a model with high distance-awareness.

H Results on existing benchmarks

In addition to **CardioTox**, we have applied our GNN-GP and GNN-SNGP models to three established molecule benchmarks: molHIV, BBBP and BACE. Specifically, molHIV is a dataset of HIV antiviral

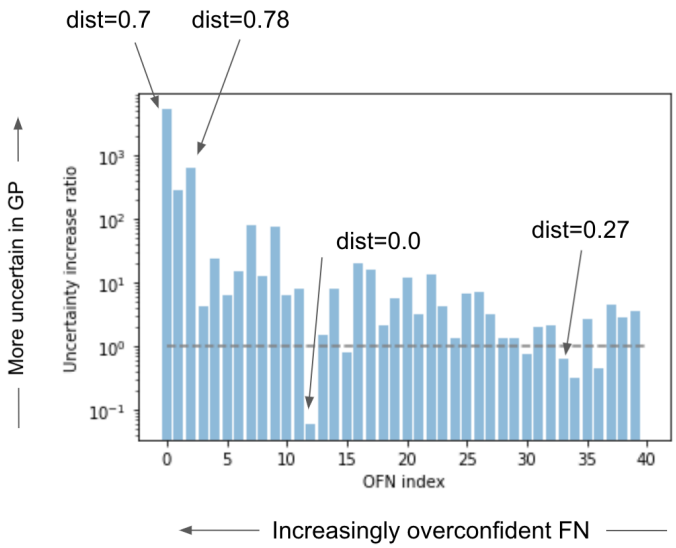


Figure S4: Uncertainty increase ratio for the top 40 overconfident false negatives. Two distant (distance = 0.7 and 0.78) and two close (distance = 0.0 and 0.27) OFNs are annotated.

activity (each molecule has an active or inactive label), BBBP a dataset of Brain-Blood Barrier Penetration (each molecule has a label indicating whether it can penetrate through brain cell membrane to enter central nervous system) and BACE a dataset of binding affinity against human beta-secretase 1 (each molecule has a label indicating whether it binds to human beta-secretase 1). Table S1 shows the results.

Table S1: Accuracy (AUROC) and robustness (ECE, Brier, NLL) performance for molHIV, BBBP, BACE benchmark. Results are averaged over 10 seeds.

molHIV	AUROC (↑)	ECE (↓)	Brier (↓)	NLL (↓)	OFNs% (↓)	OFPs% (↓)
GNN baseline	0.738±0.004	0.019±0.001	0.058±0.002	0.143±0.009	2.38±0.05	1.58±0.40
GNN-GP	**0.756±0.004**	0.013±0.001	**0.054±0.003**	0.125±0.006	**2.14±0.05**	0.16±0.15
GNN-SNGP	0.745±0.003	**0.008±0.001**	**0.055±0.002**	**0.122±0.003**	**2.16±0.05**	**0.00±0.00**
BBBP	**AUROC (↑)**	**ECE (↓)**	**Brier (↓)**	**NLL (↓)**	**OFNs% (↓)**	**OFPs% (↓)**
GNN baseline	0.773±0.002	0.202±0.007	0.495±0.029	0.911±0.059	29.50±1.32	11.43±0.88
GNN-GP	**0.792±0.005**	0.136±0.012	**0.427±0.035**	0.679±0.066	**27.23±1.44**	6.05±0.93
GNN-SNGP	0.776±0.007	**0.119±0.010**	**0.426±0.022**	**0.651±0.032**	**26.18±4.29**	**4.33±0.78**
BACE	**AUROC (↑)**	**ECE (↓)**	**Brier (↓)**	**NLL (↓)**	**OFNs% (↓)**	**OFPs% (↓)**
GNN baseline	0.789±0.004	0.181±0.011	0.460±0.064	0.784±0.101	27.90±1.71	5.77±0.61
GNN-GP	**0.807±0.004**	0.145±0.010	**0.426±0.034**	0.674±0.074	24.15±1.79	2.24±0.46
GNN-SNGP	0.803±0.003	**0.105±0.013**	**0.435±0.051**	**0.642±0.073**	**21.03±2.08**	**0.30±0.19**

I Ablation study on performance contribution

Distance-preservation in representation learning and distance-awareness in classifier both impact the performances we care about in this study. In order to understand relative contributions of them to the performance improvement, we further carried out an extensive ablation study. We take the latent representations learned by GNN baseline, GNN-GP and GNN-SNGP (increasing distance-preservation) and feed them to classifiers with increasing levels of distance-awareness: Dense layer (using logistic regression), GP-layer, exact Gaussian processes classifier (GPC hereafter). We also added experiments using pre-defined representation (i.e., molecule fingerprint denoted as FPs) which should provide 100% of distance preservation. Table S2 shows the results.

We find a GNN model achieves best performances when both distance-preservation (via SN) and distance-awareness (via GP) are present in the architecture (i.e., GNN-SNGP embeddings with GPC). However, one interesting note is that naively increasing distance-preservation does not guarantee

generalization, either in-domain or OOD. For example, we observe that the pre-defined representation (i.e., molecule fingerprint denoted as FPs) offers perfect distance-preservation, but tends to give relatively low AUROC (compared with neural representation) under data shifts such as Test-OOD2 even with exact GPC as its classifier. We hypothesis that this is related to a general trade-off in representation learning between dimension reduction and information preservation. The dimension reduction aspect tries to discard information noisy and/or less relevant to prediction task so that given limited data, models could escape the curse of dimensionality and achieve good performance under finite data. However, this could also lead to the feature collapse phenomenon that harms model robustness especially under distributional shifts. On the other hand, the information preservation aspect recovers robustness by mitigating feature collapse, but keeping around noisy/irrelevant features may demand more data for convergence and thus negatively impact accuracy given limited data. This work tries to find a good balance between these two aspects by designing a distance-preserving representation that can be optimized toward the task at hand.

Table S2: Ablation study results for Accuracy (AUROC), robustness (ECE, Brier, NLL) and over-confidence (OFNs) performance for CardioTox benchmark. Note that since Dense (implemented by logistic regression) and GPC use deterministic optimization instead of stochastic gradient descent, the error bars are zero.

Test-IID	AUROC (↑)	ECE (↓)	Brier (↓)	NLL (↓)	OFNs% (↓)
FPs + GPC	**0.921±0.000**	0.024±0.000	**0.201±0.000**	**0.332±0.000**	4.74±0.00
FPs + GP-layer	0.881±0.001	**0.020±0.003**	0.250±0.012	0.396±0.015	4.97±0.46
FPs + Dense	0.895±0.000	0.065±0.000	0.221±0.000	0.436±0.000	6.43±0.00
GNN-baseline embed + GPC	0.923±0.000	0.035±0.000	0.181±0.000	0.335±0.000	4.29±0.00
GNN-baseline embed + GP-layer	0.920±0.001	0.037±0.002	0.189±0.003	0.328±0.005	4.36±0.00
GNN-baseline embed + Dense	0.919±0.000	0.037±0.000	0.194±0.000	0.352±0.000	4.05±0.00
GNN-GP embed + GPC	0.934±0.000	0.005±0.000	**0.169±0.000**	0.282±0.000	2.98±0.00
GNN-GP embed + GP-layer	0.933±0.001	0.033±0.002	0.177±0.003	0.298±0.007	3.77±0.00
GNN-GP embed + Dense	0.904±0.000	0.042±0.000	0.217±0.000	0.377±0.000	4.37±0.00
GNN-SNGP embed + GPC	**0.935±0.000**	**0.002±0.000**	0.172±0.000	**0.280±0.000**	**2.92±0.00**
GNN-SNGP embed + GP-layer	0.932±0.001	0.028±0.001	0.179±0.005	0.295±0.005	3.56±0.21
GNN-SNGP embed + Dense	0.894±0.000	0.021±0.000	0.223±0.000	0.375±0.000	4.63±0.00
Test-OOD1	**AUROC (↑)**	**ECE (↓)**	**Brier (↓)**	**NLL (↓)**	**OFNs% (↓)**
FPs + GPC	**0.834±0.000**	0.097±0.000	**0.133±0.000**	**0.248±0.000**	1.21±0.00
FPs + GP-layer	0.764±0.009	**0.076±0.010**	0.192±0.099	0.325±0.124	1.87±0.39
FPs + Dense	0.761±0.000	0.112±0.000	0.326±0.000	0.611±0.000	2.67±0.00
GNN-baseline embed + GPC	0.794±0.000	0.069±0.000	0.290±0.000	0.487±0.000	1.83±0.00
GNN-baseline embed + GP-layer	0.796±0.001	0.094±0.006	0.334±0.034	0.540±0.052	1.47±0.05
GNN-baseline embed + Dense	0.786±0.000	0.102±0.000	0.343±0.000	0.578±0.000	1.68±0.00
GNN-GP embed + GPC	0.805±0.000	0.050±0.000	0.279±0.000	0.463±0.000	1.48±0.00
GNN-GP embed + GP-layer	0.810±0.001	0.072±0.004	0.295±0.023	0.472±0.039	1.55±0.03
GNN-GP embed + Dense	0.773±0.000	0.084±0.000	0.315±0.000	0.523±0.000	2.15±0.00
GNN-SNGP embed + GPC	**0.841±0.000**	**0.034±0.000**	**0.270±0.000**	**0.429±0.000**	1.17±0.00
GNN-SNGP embed + GP-layer	0.836±0.003	0.074±0.008	0.316±0.047	0.503±0.072	1.31±0.09
GNN-SNGP embed + Dense	0.769±0.000	0.073±0.000	0.324±0.000	0.506±0.000	**1.16±0.00**
Test-OOD2	**AUROC (↑)**	**ECE (↓)**	**Brier (↓)**	**NLL (↓)**	**OFNs% (↓)**
FPs + GPC	**0.784±0.000**	0.074±0.000	**0.167±0.000**	**0.287±0.000**	**0.00±0.00**
FPs + GP-layer	0.716±0.016	**0.051±0.004**	0.219±0.080	0.357±0.101	4.10±0.55
FPs + Dense	0.604±0.000	0.176±0.000	0.397±0.000	0.783±0.000	5.71±0.00
GNN-baseline embed + GPC	0.842±0.000	0.081±0.000	0.265±0.000	0.451±0.000	1.89±0.00
GNN-baseline embed + GP-layer	0.838±0.004	0.081±0.006	0.266±0.028	0.426±0.042	1.95±0.14
GNN-baseline embed + Dense	0.831±0.000	0.082±0.000	0.284±0.000	0.492±0.000	1.73±0.00
GNN-GP embed + GPC	0.853±0.000	0.024±0.000	0.253±0.000	0.434±0.000	**0.00±0.00**
GNN-GP embed + GP-layer	0.850±0.003	0.064±0.004	0.257±0.018	0.417±0.032	1.42±0.15
GNN-GP embed + Dense	0.825±0.000	0.087±0.000	0.284±0.000	0.512±0.000	2.04±0.00
GNN-SNGP embed + GPC	**0.891±0.000**	**0.021±0.000**	**0.208±0.000**	**0.347±0.000**	**0.00±0.00**
GNN-SNGP embed + GP-layer	0.885±0.007	0.044±0.006	0.238±0.040	0.389±0.068	1.02±0.11
GNN-SNGP embed + Dense	0.824±0.000	0.050±0.000	0.253±0.000	0.414±0.000	1.10±0.00